UNIVERSITY OF SOUTHAMPTON

FACULTY OF LAW, ARTS & SOCIAL SCIENCES

School of Management

Domain Knowledge Integration in Data Mining for Churn and Customer Lifetime Value Modelling: New Approaches and Applications

by

Elen de Oliveira Lima

Thesis for the degree of Doctor of Philosophy

January 2009

UNIVERSITY OF SOUTHAMPTON

FACULTY LAW, ARTS & SOCIAL SCIENCES

SCHOOL OF MANAGEMENT

Doctor of Philosophy

DOMAIN KNOWLEDGE INTEGRATION IN DATA MINING FOR CHURN AND CUSTOMER LIFETIME VALUE MODELLING: NEW APPROACHES AND APPLICATIONS

By Elen de Oliveira Lima

ABSTRACT

The evaluation of the relationship with the customer and related benefits has become a key point for a company's competitive advantage. Consequently, interest in key concepts, such as customer lifetime value and churn has increased over the years. However, the complexity of building, interpreting and applying customer lifetime value and churn models, creates obstacles for their implementation by companies. A proposed qualitative study demonstrates how companies implement and evaluate the importance of these key concepts, including the use of data mining and domain knowledge, emphasising and justifying the need of more interpretable and acceptable models. Supporting the idea of generating acceptable models, one of the main contributions of this research is to show how domain knowledge can be integrated as part of the data mining process when predicting churn and customer lifetime value. This is done through, firstly, the evaluation of signs in regression models and secondly, the analysis of rules' monotonicity in decision tables. Decision tables are used for contrasting extracted knowledge, in this case from a decision tree model. An algorithm is presented, which allows verification of whether the knowledge contained in a decision table is in accordance with domain knowledge. In the case of churn, both approaches are applied to two telecom data sets, in order to empirically demonstrate how domain knowledge can facilitate the interpretability of results. In the case of customer lifetime value, both approaches are applied to a catalogue company data set, also demonstrating the interpretability of results provided by the domain knowledge application. Finally, a backtesting framework is proposed for churn evaluation, enabling the validation and monitoring process for the generated churn models.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DECLARATION OF AUTHORSHIP

I, Elen de Oliveira Lima, declare that this thesis entitled "Domain Knowledge Integration in Data Mining for Churn and Customer Lifetime Value Modelling: New Approaches and Applications" and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been accepted for publication as: Domain Knowledge Integration in Data Mining using Decision Tables: Case Studies in Churn Prediction, submitted to the Journal of Operational Research Society, forthcoming.

Signed: …….

Date:…….

ACKNOWLEDGEMENTS

I would like to give special thanks to my supervisors, Dr Bart Baesens and Dr Christophe Mues for providing valuable intellectual support throughout the writing of the thesis. I also would like to thank Dr Edward Malthouse for providing me with a very interesting data set, which enabled me to progress with my research into customer lifetime value and also the Center for Customer Relationship Management for making available two very interesting churn data sets. I am also very grateful to the University of Southampton for providing the financial support for my studies. I also give especial thanks to Professor Sally Brailsford and Professor Paul Harper for believing in me and for giving me valuable support in the beginning of my PhD and throughout the research.

To the support from my Brazilian friends, Daniela Ross, Mariella Scofield, Elaine Zanazi, Maira Rodrigues, Danielle Moraes, Luciana Hazin, Solange Correa, Denise Cysneiros, Denise Silva, Alinne Veiga, Katia Maciel and Isabel Carvalho, both here and overseas, all became a part of my life and will always be in my heart. To Noor Maya and to my dear friend Melinda Taylor, who, following the same path as me, shared the same tensions and excitements in different moments of our journey, which included long talks and support that motivated us to continue. Their friendship is one of the main prizes I acquired doing this research.

To Charlie, whose love, companionship, care, support and understanding, made it possible for me to go on and was always there, in all moments, good and bad, making me see and believe that everything was going to be fine. His presence in my life made everything make sense. To his family, Mary, David, Sue, Andy, Charlie, Kayleigh, Graham, Ruth, Will and Emily, who became my family and supported me all the way in this journey.

To my parents, Edgar and Valdelice, whose love and admiration made me feel proud of this final product, which is of such important value. And finally to my dearest sister, Edneide, my fortress, my confident, whose wise words kept me on the correct path and made me feel strong.

Thank you all. Without you, this journey was not going to be the fruitful path that was laid before me.

Chapter 1. Introduction

Over the years, the relationship a company develops with the customer has become a key point for competitive advantage and one of the main elements for the survival in the marketplace. The implementation of what is called Customer Relationship Management (CRM) proved successful in many cases, but failed in many others. Concepts such as retention, satisfaction, loyalty and acquisition became parts of the vocabulary in many organisations and their understanding and measurement became a synonym of success.

In this research, two more recent concepts from the CRM umbrella are considered, customer lifetime value and churn. Their analysis has grown quickly over the years, but the idea of creating models that are more acceptable and understandable is still in need. As a result, domain knowledge will be used as part of the data mining process in order to adapt the models and to extract useful information for strategy development.

1.1. Customer Lifetime Value and Churn Context

When evaluating the relationship a customer has with a company, two concepts that may be used as part of an analytical approach are customer lifetime value and churn. Customer lifetime value (CLV) corresponds to the estimated present value of the customer's future cash flows, and involves elements such as cost, revenue, discount rate and time horizon (Berger and Nasr, 1998; Jain and Singh, 2002; Malthouse and Blattberg, 2005). Another important aspect is retention rate or, conversely, probability of churn. As emphasised by Kim *et al.* (2006) and Hwang *et al.* (2004), it is necessary to consider the churn probability in the lifetime evaluation of a customer, instead of only looking at past profit history.

However, there is little available information that would clearly indicate how companies are implementing churn and CLV predictions. There are many formulas and approaches to CLV prediction, but most research is done in an academic environment, most often in partnership with companies. It is necessary to understand how companies view this type of analysis, and also, how this type of prediction could be more understandable for decision making and strategy definition.

Hence, the focus of this research is the use of data mining to investigate and define CLV and churn prediction models. The analysis of these elements is growing in importance to companies, creating the need to make the data mining models more understandable and interpretable, which is achieved through the use of domain knowledge (Anand *et al.*, 1995; Kopanas *et al.* 2002; Martens *et al.*, 2006). Domain knowledge or common sense is related to information about a specific domain or data, which can be applied in different parts of the data analysis. Decision makers need to trust the models and use the information extracted from them to define strategies.

As a result, rather than simply applying data mining and choosing the best predictive model, the approach proposed is to use domain knowledge integrated with data mining analysis, in order to better understand and select the best predictors that could, firstly, influence the probability of a customer churning or not churning, and secondly, evaluate how the predictors for CLV would impact on its value.

1.2. Research Questions and Contributions

After emphasising the need for more integration and understanding in churn and CLV prediction models, this section sets forth the main research questions and contributions of this work.

1.2.1. To Better Integrate and Understand the Concepts of Churn and CLV

In terms of CLV, some of the main points are related to: How to operationalise CLV in practice? How to choose a suitable discount rate? How to define the best time horizon for the calculation, finite or infinite?

In terms of churn, some aspects are: how to use the concept? How to integrate it into CLV? How to identify churners effectively?

To be able to achieve this, the research will start by examining the concepts of churn and CLV, how they have been investigated in previous research, thereby emphasising their importance (see Chapter 2). It will draw important conclusions about the theoretical aspects of CLV and churn.

Then, a benchmarking study will be done in Chapter 4, exploring some key data mining techniques and their performance when evaluating churn. The data mining techniques will be briefly explained, but the focus will be on the evaluation of results. Chapter 8 will also present a benchmarking study, which will be related to CLV evaluation.

1.2.2. To Obtain a Fresh View of CLV and Churn in Practice

The research question in this case is: From a qualitative perspective, how is the CRM approach (e.g., CLV, churn, and retention) assessed by companies?

To support the research, a qualitative survey will be executed to evaluate if and how companies implement their customer analysis, investigating how key concepts (e.g., churn, CLV, domain knowledge, data mining, decision tables) are perceived by companies. This will enable the possibility to acquire a fresh view on how the concepts of CLV and churn are being approached in practice by companies, with the results described in Chapter 3. Research suggests that such a study has not yet been conducted in the literature.

1.2.3. Integration of Domain Knowledge into Data Mining

The main research question in this case is: From a data mining perspective, how to make data mining models understandable and compliant with domain knowledge?

One of the main difficulties when developing CLV and churn models is to obtain models that not only give the results in the time frame expected, but also present results that can be trusted and used for decision making. The purpose is then to integrate domain knowledge into data mining, checking whether the knowledge contained in the models is in accordance with domain knowledge.

In this research, this is done by exploring the concept of domain knowledge in Chapter 5, also describing how it can be integrated with the data mining analysis. Subsequently, the empirical analysis will demonstrate how domain knowledge can be integrated as part of the data mining process when predicting churn, through, firstly, the evaluation of signs in the logistic regression in Chapter 6 and secondly, the analysis of rules' monotonicity in decision tables in Chapter 7. Both approaches will be applied to two

telecom data sets, in order to empirically demonstrate how domain knowledge can facilitate the interpretability of results, when predicting churn.

Finally, it will be discussed how to deal with wrong signs in a logistic regression model, through the use of a method developed to assist the enforcement of correct signs in combination with the logistic regression evaluation. Secondly, decision tables will be used to facilitate the analysis and interpretation of rules generated by the decision tree model. An algorithm that allows checking whether the knowledge contained in a decision table is in accordance with the domain knowledge available is then presented.

An empirical evaluation of CLV will also be performed in Chapter 8, using a DMEF (Direct Marketing Educational Foundation) data set originated from a catalogue company, in order to evaluate the CLV results after the use of domain knowledge. Firstly, a benchmarking analysis will be done, comparing different data mining techniques, in order to evaluate the models generated. Secondly, domain knowledge will be applied to selected models, with the purpose to evaluate if its use in this context is relevant and presents valid results.

In conclusion, the idea is not only to help companies discover which customers are more valuable or will churn. The purpose is to help them identify, through the use of domain knowledge, the main elements in their data that can contribute positively or negatively to the relationship with the customer. Based on that, it is possible to define strategies that would benefit both companies and customers. The findings and results from these chapters are major contributions of this research.

1.2.4. A Framework for Monitoring and Backtesting Churn Models

The final research question is: How to backtest churn models to guarantee that an established model is still applicable in a future context?

This research question will be addressed in Chapter 9, which will adapt a methodology of backtesting, creating a framework for churn model monitoring. First, the monitoring approaches will be discussed. Then, the methodology for backtesting will be adapted for churn models and it will be applied to a out-of-time data set from one of the churn models previously investigated in Chapter 6 and Chapter 7, comparing the results with

the initial training model and initial scoring data set. Research suggests that such a study proposing a backtesting framework for churn models has not yet been conducted; therefore, this is also a major contribution of this research.

Chapter 2. An Overview of Customer Lifetime Value and Churn

2.1. Introduction

The value of a customer changes during their established relationship with an organisation (Kumar *et al.*, 2004). Customers vary extensively in a range of attributes, including product preferences, price sensitivity, cost-to-serve, retention rates and responses to marketing strategies. As a result of these and other factors, customers differ widely in the value they represent to an organisation (McDougall *et al.*, 1997).

Meltzer (2002) states that many organisations do not know which of their customers they should focus on. Customers are not created equal, yet the services provided by many organisations seem to make precisely this assumption. It is necessary to know how much it costs to serve existing customers and how much profit they bring to the organisation. To enable organisations to proceed with the development of strategies, concepts of customer lifetime value (CLV) and churn are of key interest. This research focuses on associating them with elements that are related to the concept of understanding and evaluating the customer. Below a diagram is proposed (Figure 1), illustrating the connections between these key elements.

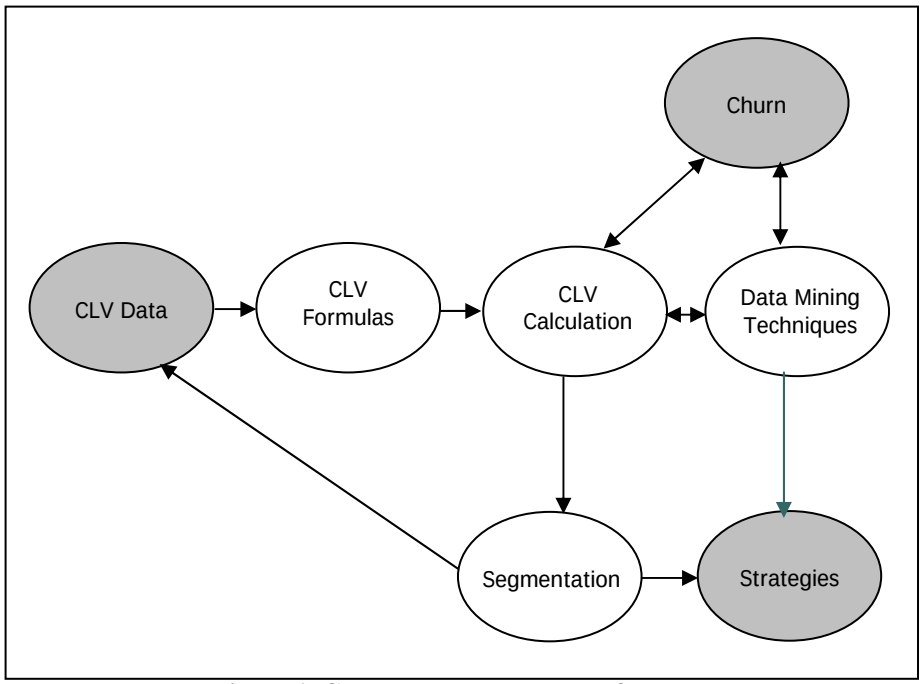

Figure 1. CLV and churn research framework

There is a virtual connection between strategies and the main elements of churn and CLV. It is a cycle, where the feedback from the strategies together with the measurements of CLV and churn allow for the evaluation of the strategy used, recalculating CLV, and making the whole process happen. This occurs in order to define the best strategies to maintain or increase the lifetime value of customers. Segmentation could be used in a final stage, as strategies can be defined for an individual, groups or the whole population.

2.2. The concept of Customer Lifetime Value

Mason (2003) states that some customers are more valuable than others. It is normally believed that long-lifetime customers are more profitable to an organisation (Reichheld, 1996; Jain and Singh, 2002). The difficult part is determining how to distinguish the more profitable from the less profitable.

Choosing the most profitable customers and accurately targeting and nurturing them, while virtually deselecting the least profitable customers, is one way of vastly improving profits (Meltzer, 2002).

Stahl *et al.* (2003) emphasise that since not all customers are financially attractive to the organisation, it is crucial that their profitability be determined and that resources be allocated according to the customer's lifetime value.

The consistency of profits per customer is brought into question by Berger and Nasr (1998), suggesting that a key advantage in retaining customers is that profits created by them tend to increase faster over time. They explore Reichheld's (1996) arguments to justify acceleration in customers' profits: first, revenues from customers usually grow over time; second, existing customers are more efficient to serve, resulting in cost savings; third, satisfied customers act as referrals who recommend the business to others; fourth, in some industries, old customers pay higher prices than new ones.

Berger and Nasr (1998) continue emphasising the importance of retaining customers, suggesting that any change in retention rate is expected to have a greater effect on CLV. However, an organisation cannot assume that high-profit customers in the past will be

profitable in the future, nor can they assume that historically low-profit will be low-profit customers in the future (Malthouse and Blattberg, 2005). A study by Reinartz and Kumar (2000), carried out with the customers of a catalogue dealer, resulted in the finding that long-term customers are not necessarily profitable customers. The dynamics of costs and revenues seem to depend on the nature of the customer relationship.

Ryals and Knox (2005) state that the difficulty in measuring customer profitability is related to the fact that accounting and reporting systems usually reflect product profitability rather than customer profitability. In other cases, profitability may not be a good predictor of the future, generating results that can be harmful to longer-term value creation, when applied to relationship strategies (Ryals, 2002; Ryals and Knox, 2005). Ryals (2002) also argues that profitability may not be reliable as a guide to the future, as changes in a customer's life circumstances or preferences can modify purchasing behaviour in different periods. Because of this, it is necessary to consider whether single period customer profitability can and should be used as a guide for future actions.

The challenge for any organisation is to implement a distinctive combination of targeted strategies focussing on customers, in order to maximise the profits earned by the organisation for every customer lifetime. Reichheld *et al.* (2000) support this argument, stating that profits are important because they allow the company to improve value, providing incentives to facilitate employees, customers and investors to remain loyal to the company.

CLV represents the present value of the expected benefits less the costs of initialising, maintaining and developing the customer relationship (Malthouse and Blattberg, 2005). Different academics explore the CLV concept (Dwyer, 1997; McDougall *et al.*, 1997; Berger and Nasr, 1998; Jain and Singh, 2002; Mason, 2003; Stahl *et al.*, 2003; Hwang *et al.*, 2004; Kumar *et al.*, 2004; Pfeifer and Bang, 2005; Ryals and Knox, 2005), which can be summarised as the sum of accumulated cash flows of a customer over their entire lifetime with the organisation.

CLV has become central to relationship marketing (Berger and Nasr, 1998; Ryals, 2003; Malthouse and Blattberg, 2005). The marketing focus moved from being product focussed to being customer focussed (Rust *et al.*, 2000).

Many researchers have been exploring the concept of CLV in the last two decades, increasingly so in recent years. The literature review table presented below (Table 1) shows the focus of some of these academics when exploring the concept of CLV, demonstrating their depth of research and which main elements they explore together.

These connections are going to be explored in more detail throughout the next sections. However, it is useful to be able to visualise beforehand how the research in this subject has been approached.

Reference	Data	Techniques	Variable Construction	Observations	Sector
Dwer (1997)	Specific values were used to formulate examples.	CLV formulas.	CLV: gross margin, costs, time horizon, discount rate.	Retention and migration model for CLV.	Examples: Consumer magazine and catalogue distribution.
McDougall *et al.* (1997)	No data used.	CLV formula.	Customer value: acquisition cost, revenue stream, cost stream, length of relationship. Apply discount rate to get the net present value (NPV).	Theoretical approach. Example of churn calculation to help define CLV.	Not specified.
Berger and Nasr (1998)	Specific values were used to formulate examples.	Mathematical formulas.	Based on the financial components of CLV: gross contribution, retention rate, period, discount rate, costs.	Presentation of mathematical models for determination of CLV.	Fictitious examples: insurance company, healthy club, car dealership, credit card company.
Rust *et al.* (2000)	No data used.	None.	None.	A theoretical approach to CLV.	Not specified.
Jain and Singh (2002)	No data. Only explanation of formulas.	Different CLV formulas. Average CLV. Segmented CLV.	Based on the financial components of CLV: cash flow, revenue, cost, period.	Presentation of models for determination of CLV.	Not specified.
Meltzer (2002)	No data used.	None.	Can consider churn in the CLV calculation.	A theoretical approach to profitability and CLV.	Bank and telecommunication company examples.
Ryals (2002)	No data used.	Recommends the use of data mining tools	CLV elements: duration of	A theoretical approach to CLV.	General examples: airlines, office supplies (staples).

		in general.	relationship, revenues, costs, discount rate, probability of retention (risk).		
Berger *et al.* (2003)	Cohort 1: 6,094; Cohort 2: 7,166; Cohort 3: 7,758.	Customer migration model for CLV evaluation.	CLV: based on average gross margin, cost per passenger, discount rate and period.	Practical calculation of CLV.	Cruise-ship company.
Gupta and Lehmann (2003)	Data obtained from companies annual and financial reports.	CLV formulas.	CLV: margin, retention rate, discount rate and period.	CLV with infinite time horizon (period).	E*Trade, Amazon, Ameritrade, Capital One, Ebay, CDNow, AT&T.
Mason (2003)	Use of internal customer data to evaluate the lifetime of customers: 7,953.	Use of purchase history – IT analyst created reports.	CLV: net cash flow, discount rate, time horizon.	Some explanations and details about how the calculations were done.	Catalogue Company (Tuscan Lifestyles).
Ryals (2003)	No Data. Specific values were used to formulate examples.	A CLV definition and calculation.	CLV elements: duration of relationship, revenues, costs, discount rate, probability of retention (risk).	Introduction of risk in the calculation of CLV.	Not specified.
Stahl *et al.* (2003)	No Data. Specific values were used to formulate examples.	CLV approach and calculation.	CLV: function of the potential revenue stream, the costs associated, the duration of the relationship, and risk.	CLV components: Base potential, growth potential, networking potential and learning potential.	Industrial manufacturer.
Hwang *et al.* (2004)	Dataset with 200 data fields and 16,384 records. Final sample: 101 data fields, 2,000 records.	Decision tree, artificial neural network, logistic regression.	Consider churn in the CLV calculation.	Misclassification rate and lift chart were used to compare the data mining techniques.	Wireless telecommunication industry in Korea.
Kumar *et al.* (2004)	No data used.	Use of net present value (general and individual).	Average CLV. Individual CLV.	Applications examples. Concept of active and	General application.

				inactive customers. CLV applied to develop strategies.	
Rust *et al.* (2004)	100 surveys completed; 17 coefficients.	Markov switching matrix.	CLV: frequency, quantity, brand patterns and contribution margin.	Customer equity as the sum of the CLVs of current and potential customers.	Airline industry.
Venkatesan and Kumar (2004)	Cohort1: 1,316; Cohort2: 873.	Markov Chain Monte Carlo.	CLV: purchase frequency, contribution margin and marketing costs.	Maximise CLV for resource allocation strategy.	B2B manufacturer.
Malthouse and Blattberg (2005)	Four samples: Service Company: 150,000 Not-for-profit organisation: 191,779 Business-to-business company: 100,000 Catalogue company: 106,284	Regression and neural network applied to the data.	CLV: net contribution, discount rate and time horizon.	Use of Historical information to predict CLV.	Service company, not-for-profit organisation, business-to-business company, catalogue company.
Pfeifer and Bang (2005)	No Data. Numerical example to illustrate the different ways of calculating an average CLV.	Different average CLV formulas.	CLV: cash flow (revenue less variable cost, not including acquisition and fixed costs), discount rate, customer lifetime (units of time).	Different ways of calculating an average CLV, examine which one of them is the best.	Not specified.
Ryals and Knox (2005)	From hundred of customers, 12 key accounts were chosen to be evaluated.	Individual calculations of CLV and EV for the 12 key accounts.	To calculate CLV: the anticipated lifetime of the customer relationship in months or years; the profit in each future period (revenue less cost), and a discount rate.	Concept of Economic Value (EV) of a customer: Forecast CLTV + future customer risk.	Zurich Insurance company.
Glady *et al.* (2006)	Sample of 10,000 customers.	Logistic regression, decision trees	CLV: cash flows, period, products,	Churn estimation based on CLV.	Belgian retail financial company.

		and neural networks for churn.	discount rate.		
Kim *et al.* (2006)	Dataset with 200 data fields and 16,384 records. Final sample: 101 data fields, 2,000 records.	Decision tree, neural network and logistic regression.	Consider churn in the CLV calculation.	Misclassification rate and lift chart were used to compare the data mining techniques.	Wireless telecommunication company.
Kumar (2006)	Numerical examples to illustrate the calculation of CLV.	Average and individual CLV.	CLV: future contribution margin and cost, period, discount rate.	Strategic discussion to maximise CLV.	Not specified.
Kumar *et al.* (2006)	303,431 customers	Linear and logistic regression.	CLV: purchase frequency, contribution margin and marketing costs.	An empirical application for the calculation of individual CLV.	Retail setting.
Ryals and Knox (2007)	From eighteen major customers, twelve key accounts were chosen to carry out risk evaluations.	Individual calculations of CLV for ten key accounts.	To calculate CLV: revenues adjusted by the risk, period, cost and a discount rate (WACC).	Comparison of CLV using, different discount rates and future customer risk.	B2B Insurer.

Table 1. CLV approach in the literature review

2.2.1. Practical Application of Customer Lifetime Value

After defining CLV in the previous section, it is possible to use the information presented in Table 1 to explore some of the practical applications for CLV.

Mason (2003, p.56) states that "although a widely held maxim in direct marketing is that past behaviour is a good predictor of future behaviour, for a brand new customer there is little past behaviour on which to build a forecast." She uses the analysis of lifetime value to evaluate if a new customer's early behaviour can be an indicator of future behaviour.

Stahl *et al.* (2003) state that an accurate measurement of CLV requires, a) an exact allocation of costs to customer relationships according to the resources employed; b) an estimation of all monetary and non-monetary benefits created by the customer

relationship; c) a consideration of cost and revenue changes over the estimated duration of a customer relationship; d) the discounting to the present of future cash flows generated over the estimated duration of a customer relationship; and e) an estimation of the relationship risks. They also argue that changes in the surrounding environments of companies (e.g., technological, political, economic, judicial or social) may affect sales directly or indirectly through their potential impact on industry and competitive conditions. As a result, risk must be included in the measurement of CLV.

For organisations that consider customers as manageable assets (Jain and Singh, 2002; Ryals, 2002; Ryals, 2003; Stahl *et al.*, 2003; Kumar *et al.*, 2004), CLV can become the measure that leads investments, such as infrastructure and marketing activities. Understanding how valuable the customers are and which of them are the most profitable is essential information in the process of customer retention.

Effectively, as argued by Ryals (2002), companies need to predict the future purchasing behaviour of key customers to arrive at their CLV. For Ryals and Knox (2005), this value represents the retained customer worth to the organisation, based on predicted future transactions and costs. They also argue that customer relationships should be measured and managed for value rather than profit. For them, CLV calculations provide a better guide for customer strategy than current period profit. An adjustment to CLV is needed to counterbalance the profits earned and the risk of investing in the customer relationship.

Based on this, Ryals and Knox (2005) go on to explore the concept of economic value of a customer, which is composed by forecast CLV plus the future customer risk. To demonstrate their analysis, they measured CLV and economic value for twelve clients of Zurich Insurance Company. They came to the conclusion that new accounts tend to be more profitable than old ones and that retention of an unprofitable customer that has a high profile sometimes is necessary for reference and referral purposes. This refers to the idea of social network effects and can be supported by Stahl *et al.* (2003), when they state that referrals might lead to additional sales and lower acquisition costs as new customers are attracted through word-of-mouth advertising and that referrals from a customer with a positive reputation may strengthen the loyalty of other customers.

To examine how accurately CLV can be predicted, Malthouse and Blattberg (2005) investigate four types of companies in their case studies: a service company, a not-for-profit organisation, a business-to-business company, and a catalogue company. They used regression and neural network models for each of the data sets, in order to estimate the future CLV for individuals or households using the available information, especially past purchase behaviour. They came to the conclusion that most companies will not be able to accurately forecast the future behaviour of customers. This is due to the fact that historical value is not a very precise predictor of future behaviour (Stahl *et al.*, 2003; Malthouse and Blattberg, 2005).

Malthouse and Blattberg (2005) also argue that in situations where the future cannot be predicted accurately, historically valuable customers may be the wrong customers in which to invest marketing resources. As there will always be a certain level of unpredictability in a customer's purchases, an organisation will be relying on chance purchases when they use historical information to allocate marketing investments.

Supporting this argument, Ryals (2002) states that current and historic profitability data is not necessarily reliable to be used for future value prediction. McDougall *et al.* (1997) emphasise this point, arguing that in an ideal world, customer value could always be obtained from existing historical information. However, in the real world this is not always possible, for a number of reasons. Firstly, the necessary historical data may not be available; secondly, current customers may not be representative of future customers; and lastly, historical data may not accurately reflect future market behaviour.

Nevertheless, even with some contradiction, historical data still is the most used and accurate type of data used for prediction. That is why it is necessary to fully understand the data in use, and also, to make sure of its reliability and consistency.

Berger *et al.* (2003) focus on the practical application of CLV, evaluating how CLV can impact on marketing strategy. They have applied customer migration models on their evaluation. They calculated an average CLV based on the data of five consecutive years, considering only the customers that started on a specific route on the first year of evaluation. They executed the analysis on three different cohorts (1993, 1994, and 1995). The CLV calculation was based on actual past data, in order to predict an

average value for each customer. It was used to support investments in advertisement, based on the estimation of the customer's worth.

Drye *et al.* (2001) argue that identifying the customers most likely to stay the longest may help to focus recruitment campaigns towards those customers who are most likely to maintain their interest, support and membership for the longest, and as a result have a significantly higher lifetime value. They also say that there will be a variety of factors that influence whether or not customers will continue their purchases, emphasising that some of these factors will be known and others will not. As a result, they claim that statistical analysis enables a range of factors to be included and also allows testing of whether different strategies should be used for different types of customers. In their study, they use survival analysis to analyse events that occur many times for the same individual, taking into consideration the time between events and censored observations, in order to investigate the chances of the customer defecting or repeating their purchase. They argue that survival analysis can better identify the most regular customers, and those more likely or unlikely to repeat purchase.

When considering the key components of the process of measuring and managing CLV, Kumar *et al.* (2004) focus on customer acquisition and customer retention. Stahl *et al.* (2003) support this, stating that the CLV measurement is essential for developing and maintaining long-term profitable customer relationships, as it is fundamental for customer acquisition and retention decisions. Kumar *et al.* (2004) also say that the gross contribution and the marketing costs of every customer are also important to compute CLV. They argue that the measurement of average CLV can be used to evaluate competitor organisations, as it is possible to have access to published sources like annual reports and financial statements, which contain various elements of the average CLV formula. Another application would be to support merger and acquisition decisions, in order to help identify the market worth of the organisation.

Rust *et al.* (2000) emphasise that there is little literature regarding studies that can predict CLV on the individual level. This might be due to the fact that individual customer behaviour is unpredictable and affected by a number of factors which often are not observed. Kumar *et al.* (2004) also emphasise that the average CLV does not provide the precise lifetime value of each individual customer. They say that the variation around the average CLV is the key to tailoring differential treatment for

individual customers. Each customer might have different patterns of active and inactive periods. Calculating the probability of a customer being active together with the customer's previous gross contribution, helps to identify the expected gross contribution for that customer. They then explore some applications for the individual CLV, for example:

- To compute the profitable lifetime duration and decide when to discontinue or scale down marketing efforts for a customer, by computing the net present value of the expected gross contribution from the customer over a period. If the net present value drops below the planned marketing cost, it indicates that the lifetime duration of the customer is not profitable anymore, being necessary to discontinue or scale down the level of marketing efforts directed to the customer.

- To prioritise and select customers on the basis of CLV, as it predicts revenue and profits, anticipating and modelling future customer behaviour.

- To understand the variables that explain the differences in the profitable lifetime duration between customers to identify areas for managerial intervention (identify variables that impact customer profitability). Customers who demonstrate a moderate but stable time interval between successive purchases are likely to exhibit longer profitable durations compared to other customers. As a result, an organisation can proactively define strategies such as cross-selling, loyalty and rewards programmes, based on their understanding of the effects of these variables on CLV.

What is important is that the individual or average CLV will be used in accordance with the organisation needs. For example, if the organisation wants to evaluate a large amount of customers to analyse the overall situation of the company, the average CLV is the best choice. However, if the company has only a few large clients or only wants to investigate a few large clients the individual CLV should be applied.

Supporting this argument, Jain and Singh (2002) argue that the use of CLV models will depend on the type of organisation. They state that companies that have a few and identifiable customers might benefit from models that measure CLV on the individual

basis whereas firms having large number of customer with small sales, they might benefit from models that help segment the customer base on the basis of the average CLV.

Following the customer evaluation, Ryals and Knox (2005) argue that some customers cost more to serve than others. This is usually influenced by the customer behaviour. For example, new customers may require a more customised service and products, but only consume the organisation's products on a limited basis or as a trial. Customers that buy on a regular basis and that have routine and predictable purchasing habits are reasonably easy to serve (Ryals, 2002; Ryals and Knox, 2005).

Other factors also have an effect on the lifetime value of a customer. For example, the influence of churn on the lifetime value presented a gap in knowledge. Only recently this relationship started to be explored. Kim *et al.* (2006) and Hwang *et al.* (2004) consider the churn probability and past profit history on the lifetime evaluation, using neural networks, decision tree and logistic regression to predict churn, using the result from the best performing model in the calculation of CLV.

Glady *et al.* (2006) adopt an approach where CLV is used to estimate churn. Churn is calculated based on the probability that the customer value decreases below a pre-established threshold, which is calculated based on CLV. Rosset *et al.* (2003) also emphasise the importance of CLV in churn analysis. For them, the CLV information is complementary to churn analysis, as it demonstrates how much is being lost due to churn, working as a recommendation of how much effort should be concentrated on that customer or segment under evaluation. They use the churn concept in their CLV calculation, incorporating the customer's churn probability as a survival function that indicates if the customer will be active at the time of evaluation.

These cases demonstrate how the concepts of churn and CLV can be interconnected and influence each other. The concept of churn will be further explored in section 2.4.

2.2.2. Customer Lifetime Value Models and Formulas

Dwyer (1997) made the concept of CLV more widely known. He introduced the evaluation of CLV taking into consideration a retention or migration model, according

to how the relationship with the customer is established. If a customer is considered to be lost-for-good, it is better modelled as customer churn, where it means that if a customer leaves and then returns, they are considered as a new customer.

In the always-a-share approach, the customer never leaves the company: they can reduce their purchase behaviour, but they will always be considered as the same customer; in this case, a customer migration model would be recommended. The choice will depend on the type of analysis that is being done, as well as the type of industry under evaluation.

Building on Dwyer (1997)'s work, Berger and Nasr (1998) discuss a series of mathematical models for determination of CLV, based on some general cases of customer behaviour, such as customer retention situations and a customer migration model. They conclude that determining customer value can aid decision making, as it is possible to evaluate the impact of different strategies on the value of CLV.

Jain and Singh (2002) discuss the focus given to CLV in different studies, for example, the development of models to calculate CLV for the individual customer. These applied models are more relevant to practitioners who wish to use CLV as a basis for making strategic or tactical decisions. Some cases are expressed below:

- Blattberg and Deighton (1996) propose a model for managerial decision making to maximise CLV, in order to find the best balance between spending on customer acquisition and retention. They suggest that future and present customers are segmented into homogeneous groups, based on their behaviour and attitude, in order to estimate spending in acquisition and retention.

- Pfeifer and Carraway (2000) propose that Markov Chain Models are appropriate for modelling customer relationships, as they can be used to model both customer retention and migration circumstances. Customer migration is taken into consideration when a customer is inactive for a period and as they return, they are treated as a retained customer.

Jain and Singh (2002) also discuss models for customer base analysis, involving the use of methods to analyse information about existing customers and the value of their future

transactions. These models take into account the past purchase behaviour of all customers, in order to obtain probabilities of purchase in the next time period. For example:

- The Pareto/NBD (Schmittlein *et al.*, 1987) model is applicable in situations where the time when the customer becomes inactive is unknown and the customer shifts from inactive to active at any time and how many times they want. To calculate CLV, it is essential to establish the number of active customers and the number of future active customers in each time period.

- Reinartz and Kumar (2000) extended the Pareto/NBD model into a dichotomous "alive/dead" measure and used it to test Reichheld's (1996) statements, regarding the profitability of long-lifetime customers, concluding that, a) the relationship between customer-lifetime and profitability is positive, but weak; b) they found no support for the argument that profits from long life customers increase over time; c) the notion that customers with long tenure are associated with lower promotional costs is rejected; and d) long lifetime customers do not pay higher prices. These findings contradict the results from Reichheld (1996). This fact could be directly linked to the type of sector under evaluation, which provided Reinartz and Kumar (2000) with different results. This emphasises the fact that each industry should be analysed separately and that different points of reference can be extracted from these analyses.

In a different view, Ryals (2003) draws a parallel between brands and customer relationships, involving the notion of portfolio. She explores the application of modern portfolio theory to customers or customer segments and shows that modern portfolio theory and the capital asset pricing model can be suitably applied to a customer portfolio. To do this, it is necessary to be able to identify customer returns and risk. She argues that customer returns are generally measured as customer profitability or CLV. As single-period measurement of customer profitability may not be a reliable guide to the true value of a customer, CLV is then considered a better measure of customer return (Ryals, 2003, Ryals and Knox, 2005). To assess customer risk, the discount rate used to calculate CLV can be adjusted for the customer risk.

Ryals and Knox (2005) state lifetime value analysis and risk-adjustment process can represent the way to develop strategies to customer retention. Customer risk and a customer economic value are significant when developing customer strategies. The appropriate strategy will depend on whether and to what extent the risk of churn can be reduced and the customer economic value increased. Although they argue that not enough research and literature on customer risk indicate that the concept is not properly appreciated, it is acquiring strength over time.

As suggested by Jain and Singh (2002), research is needed into estimation methods that provide stable, consistent and unbiased estimates of CLV. In general, the basic idea of CLV, which is explored by a number of researchers (Meltzer, 2002; Ryals, 2002; Ryals, 2003; Stahl *et al.*, 2003), is represented by the equation below:

$$\text{CLV} = \sum_{i=1}^{n} \frac{[(R_i * r_i) - C_i]}{(1+d)^i}$$

where:

R_i = Revenue, meaning the gross contribution from a customer or segment of customers at time period *i*;

C_i = Costs involved with acquiring, servicing and maintaining the customer or segment in time period *i*. In this case and in most cases, it does not include acquisition costs.

r_i = Retention rate, which represents the individual probability or proportion of customers expected to continue buying the company's goods or services in the subsequent time period *i*. However, not all academics use this risk element in their calculations.

n = Period, duration of relationship, time horizon.

d = The discount rate used in determining the present value of future cash flows. This will be explored in more detail in section 2.3.

Gupta and Lehmann (2003) demonstrate how to use publicly available data, such as financial information, in order to estimate CLV. They explore the general formulation of CLV, including a margin, discount rate and retention rate as key factors for the CLV calculation, but also adding the idea of an infinite time horizon (Gupta *et al.*, 2004; Fader *et al.*, 2005). Based on their analysis and assuming that the profit margin *m* and the retention rate *r* are constant over time (to simplify the explanation), the CLV formula would be as follows:

$$\text{CLV} = \sum_{i=1}^{\infty} \frac{m * r^i}{(1+d)^i} = m\left(\frac{r}{1+d-r}\right)$$

For Fader *et al.* (2005), because they are using the Pareto/NBD approach to calculate the discounted expected transactions (one of the CLV components), it would make more sense to switch the discounted expected transactions' estimation from a discrete-time to a continuous-time formulation, computing the CLV over an infinite time horizon. Gupta *et al.* (2004) consider that the value of the customer base for a company is composed by the sum of the CLV from current and future customers and they also developed their CLV model assuming an infinite time horizon. This decision was based on the facts that: firstly, they do not need to specify how long the customers will stay with the company; second, the retention rate accounts for the fact that the chances of a customer staying with the company will decrease over time; thirdly, a finite time horizon overestimates CLV; fourthly, retention and discount rate will guarantee that distant future values will contribute less to CLV; and lastly, the use of infinite time horizon simplifies the CLV formulation. Conversely, Berger *et al.* (2003) assume that a time horizon greater than five years would have a limited effect on CLV, resulting in an underestimated customer value.

What can be concluded is that the CLV formula is not something fixed. It can and in certain ways should be adapted to the reality of the organisation. The complexity depends on the variables involved in the CLV calculation, depending on the type of organisation. The main idea is to keep the CLV calculation simple, but in a way that it is going to give the expected and adequate results, according to the company's data.

2.2.3. Data Mining and Customer Lifetime Value

Wynn and Crawford (2001) state that companies believe that customers are what sustain any business and that they have a lifetime value. They also argue that organisations need to use specific data mining techniques to estimate CLV based on the information collected. This is essential to develop and enhance a continuous relationship strategy, with special focus on building relationships with customers who present attractive lifetime value. Table 2 shows examples of the use (or not) of data mining when calculating CLV.

Method	Source	Data Characteristics
Regression and neural network	Malthouse and Blattberg (2005)	Service Company: 150,000; Not-for-profit Organisation: 191,779; B2B company: 100,000; Catalogue company: 106,284.
Markov chain Monte Carlo	Venkatesan and Kumar (2004)	B2B manufacturer: 1,316 (cohort 1) and 873 (cohort 2); 4 years and 3 years, respectively.
Markov switching matrix	Rust *et al.* (2004)	CLV: Airline industry - 100 surveys completed, 17 coefficients. Then, generalised to around 40 millions U.S. customers.
Linear and logistic regression	Kumar *et al.* (2006)	Retailer: 303,431 customers; 4 years of data; 3 years CLV prediction.
Pareto /NBD	Reinartz and Kumar (2003)	US catalogue retailer: 11,992 households (3 cohorts); 3 years. High-tech B2B: 4128 customers; 8 years.
Pareto/NBD	Fader *et al.* (2005); Fader *et al.* (2007)	Internet company: 23,570 customers - 78 weeks. Catalogue C.: 7,953 customers – 5 years.
Specific values were used to formulate	Stahl *et al.* (2003); Berger and Nasr (1998); Ryals (2003); Pfeifer and	Industrial manufacturer; insurance company, health club, car dealership, credit card company; No specification of

examples	Bang (2005); Kumar (2006)	customer amount; No specification of time horizon.
Did not specify any technique used	Ryals and Knox (2005)	Zurich Insurance Company: From hundreds of customers, 12 key accounts were chosen to be evaluated. Whole past lifetime until present moment under evaluation.

Table 2. Data mining and CLV

Ryals (2002) argues that data mining that investigates undiscovered patterns in customer transactions is essential to CLV analysis. Nevertheless, few examples can be found in the literature regarding the application of data mining techniques when calculating CLV. Many academics used only the theoretical approach to CLV (McDougall *et al.,* 1997; Jain and Singh, 2002; Kumar *et al.,* 2004). Others used specific values to formulate examples applying the CLV models and formulas (Berger and Nasr, 1998; Ryals, 2003; Stahl *et al.,* 2003; Pfeifer and Bang, 2005; Kumar, 2006).

In other research, CLV was calculated for a specific organisation, but the data mining techniques used were omitted, not specified, or their use was dismissed in the analysis. This is the case of Ryals and Knox (2005), where they evaluated the economic value of a customer, analysing the data of twelve clients of Zurich Insurance Company.

Nevertheless, some researchers specify the importance and use of data mining in their data evaluation (Rust *et al.*, 2004; Venkatesan and Kumar, 2004; Malthouse and Blattberg, 2005). As explored in subsection 2.1.1, Malthouse and Blattberg (2005) used regression and neural network models for each of the case studies investigated, with the purpose to estimate future CLV for individuals and households.

When measuring CLV, Venkatesan and Kumar (2004) provide a framework for its calculation involving customer selection and marketing resource allocation. They used Markov chain Monte Carlo to estimate the purchase frequency of each customer, which was then used for the calculation of individual CLV. In their case, the main elements of CLV are purchase frequency, contribution margin and marketing costs, although they do emphasise that the CLV components will vary depending on the industry being analysed.

The concept of customer equity as the sum of the CLVs of current and potential customers of a company is explored by Rust *et al.* (2004). They use CLV as part of a marketing strategy evaluation, integrating different aspects in their analysis. In the case of CLV, they used a Markov switching matrix to model the customer retention, defection and possible return, assuming that a customer can buy different brands, at different times, and can always return to purchasing a brand they have stopped buying in the past. The idea is to calculate the CLV of a customer for each one of the companies they buy from. In their example, the computed individual CLV from the survey respondents (100 customers) was averaged and generalised to project the customer equity for the company under investigation, American Airlines.

Reinartz and Kumar (2000) use the CLV formula to calculate a finite individual CLV. They use the customer lifetime resulting from the adapted Pareto/NBD to group customers into segments of low to high lifetime, in order to analyse which segments are more valuable. Customer lifetime is the period between the customer's first purchase and the date associated with the probability of the customer being alive, specified by the Pareto/NBD model. This probability was adapted to a dichotomous alive/dead measure based on a threshold of 0.5. Reinartz and Kumar (2003) then go further extending their earlier research, identifying the period beyond which the customer stops being profitable. They calculate the net present value of the customer using the probability of the customer being alive as an integrated part of the net present value, taking into consideration all the profits generated by the customer. Based on this, they decide if it is worth to continue investing in this relationship. This is done by comparing the cost of contacting the customer with the net present value of the expected contribution for that specific moment in time. Doing this, they can evaluate each period of their evaluation time, identifying the moment where the customer is considered to have ceased the relationship with the company. They emphasise that this model permits the estimation of the net present value for future periods, which would imply the calculation of the CLV for this customer in a specific future period in time.

Fader *et al.* (2005) also proposed the use of an alternative model to the Pareto/NBD, in order to simplify the prediction of a customer's future purchase based on past purchase behaviour. They adapted the Pareto/NBD in order to estimate the discounted expected transactions, which is a key element in the CLV calculation.

Similarly to Reinartz and Kumar (2000), Reinartz and Kumar (2003) and Fader *et al.* (2005), Fader *et al.* (2007) also apply an adapted Pareto/NBD model when estimating CLV for the Tuscan Lifestyles case previously explored by Mason (2003). In this case, there is no availability on detailed transaction, so they use aggregated data in their analysis. This means that the Pareto/NBD model could not be used in its original form. Again, they adapted the Pareto/NBD in order to estimate the discounted expected transactions. However, because of specific peculiarities of the data set under evaluation, the assumption made by Fader *et al.* (2005) could not be used for this analysis. In this case, the Pareto/NBD was further adapted taking these factors into consideration, and then used in the CLV calculation.

Kumar *et al.* (2006) evaluate CLV at the individual level, with the purpose to maximise profitability. They use an empirical analysis of retail data to support the use of individual CLV to this type of industry. They adapted the CLV formula from Venkatesan and Kumar (2004) and used linear regression to model the contribution margin for each customer, which was part of the CLV calculation. They then used logistic regression to identify the variables that would influence the grouping of customers into segments, based on the value of their CLV.

In all these cases, it shows that the calculation of CLV can present different elements, depending on the industry and the type of data available, which demonstrates the flexibility necessary in the CLV calculation. The use of data mining facilitates and makes the prediction of elements more accurate, providing more reliability to the results.

In conclusion, the importance of data mining for calculating CLV is unquestionable. It works as an essential step in acquiring the main elements to be used in the calculation. The amount of data available and the rapid change of the market, both require powerful tools that can extract essential and quick information that can support the marketing strategies of the organisation. The overall research literature not only suggests a lack in comparing different data mining techniques, but also the use and interpretability of the models are seldom questioned. This research will address these issues in Chapter 8, through a benchmarking comparison of different data mining techniques and the use of domain knowledge to make the models more understandable and interpretable.

2.3. Discount Rate Evaluation for Customer Lifetime Value

The discount rate is usually considered to be the same as the cost of capital and it is one of the main elements in the CLV estimation. As Jacobs *et al.* (2001) state, it converts future cumulative gross margin into the present value. It discounts future cash flows into their present values, transforming a larger future value into a smaller present value.

Based on financial evaluations, some key methods can be used to define the discount rate: weighted average cost of capital, capital asset pricing model, risk adjusted discount rate, and separate cash flow discounting method. These methods will be discussed below.

2.3.1. Weighted Average Cost of Capital

Taking into consideration the discount rate as being the cost of capital of a company, the weighted average cost of capital (WACC) (Awerbuch and Deehan, 1995; Babusiaux and Pierru, 2001; McClure, 2003) is the measurement commonly used to assess the cost of capital of a firm, consequently being usually adopted in the literature of CLV as the discount rate used in its calculation.

To obtain the WACC, each category of capital is proportionately weighted. All capital sources (e.g., common stock, preferred stock, bonds and any other long-term debt) are included in a WACC calculation. WACC is calculated by multiplying the cost of each capital component by its proportional weight, as exemplified below (Slack, 1999):

$$WACC = \frac{E}{V}\mathrm{Re} + \frac{D}{V}Rd(1 - Tc)$$

where:

- Re = cost of equity
- Rd = cost of debt
- E = market value of the firm's equity
- D = market value of the firm's debt
- V = E + D

- E/V = percentage of financing that is equity
- D/V = percentage of financing that is debt
- Tc = corporate tax rate

This would be the basic equation, which should be adapted in case the capital of the company is not homogenously composed of equity and debt. The cost of debt is obtained through the use of the current or appropriate market rate that the company is paying on its debt. The cost of equity includes business and financial risk in its calculation (Slack, 1999) and it can be obtained through the capital asset pricing model equation, which will be described later in this section.

Ryals (2002) argues that when calculating the profit of a company, its cost of debt interest is deducted, but the cost of equity, which is often more significant, is usually not taken into account. It is necessary to include the cost of equity in the calculation of WACC, because any investment will only create shareholder value if the return on capital exceeds its cost of capital.

Using WACC as the discount factor would be adequate in situations where (Slack, 1999):

1. The total risk of the project should be the same as the total risk currently faced by the company;

2. The project should be financed so that the long-term capital structure of the company remains unchanged;

3. The project should be small (or marginal) in relation to the overall value of the company.

If the company uses WACC for all projects, and the capital asset pricing model is not used to define the cost of equity, there will be a tendency to accept or reject projects incorrectly. This would generate inconsistencies that could be damaging to the financial state of the company.

Babusiaux and Pierru (2001) explore how WACC can be adjustable, emphasising that it can be used to analyse the profitability of a project when its revenue is subject to a different tax rate from the one used to calculate the discount rate.

The WACC can be used as the correct discount rate for projects that have the same risk as the company's existing business. However, if the project has different risk than the company, WACC would not provide an appropriate discount factor, unless the cost of equity is obtained through capital asset pricing model (Slack, 1999). Nevertheless, for this to be applicable, the other two rules above have to remain true. This is supported by the fact that if the financial structure of the project changes, the overall cost of capital will change; in this case, WACC will not reflect reality.

2.3.2. Capital Asset Pricing Model

The capital asset pricing model (CAPM) is used to calculate an appropriate rate of return, working as a correct discount rate, taking into account the market risk (Awerbuch and Deehan, 1995; Emhjellen and Alaouze, 2002). This risk is identified in its formula as the beta (β) value, as shown below:

$$\text{Rate of Return} = \text{Discount Rate} = Rf + \beta(Rm - Rf)$$

where:

- Rf = risk free rate of interest;
- β = beta, the sensitivity to the market risk;
- Rm = the expected return of the market to compensate the extra risk of investment.

CAPM can be used as the discount rate, but it is also used as a main element in the definition of other discount rates. In the case of WACC for example, it is used when calculating the cost of equity. In this case, risk is taken into consideration when estimating the discount rate. Other applications of the CAPM are shown in the next subsections.

2.3.3. Separate Cash Flow Discounting Method

Emhjellen and Alaouze (2002) argue that using WACC as a discount rate when a project has two different cash flows streams (revenue and costs, with different risk) results in an incorrect net present value, making it necessary to use a separate cash flow discounting method.

They argue (p. 456) that "The correct project NPV (net present value) is obtained when the expected after tax revenue of a project is discounted using the required rate of return of this expected cashflow and the expected after tax cost cashflow is discounted using the required rate of return of this cashflow." They use value weighted CAPM betas to estimate the rate of return for the revenue cash flow, assuming that the rate of return of the cost cash flow is known.

This means that the cash flows have to be independently discounted and then be used to obtain the net present value. As a result, it can provide better estimates of project's net present values than WACC when considering separate cash flows.

2.3.4. Risk Adjusted Discount Rate

Awerbuch and Deehan (1995) emphasise that the WACC method ignores financial risk, leading to unreliable results. They argue that WACC can be used when evaluating the net cash flows of homogeneous projects, whose risk characteristics are similar to those of the firm as a whole.

To compare projects with different risk levels, they argue that a market based risk adjusted discount rate must be estimated using the CAPM or a similar approach. This can provide a proper discount rate for a specific level of risk. WACC assumes the same discount rate for all the projects. Using risk adjusted discount rate method each project will have a specific discount rate, according to its own risk. The riskier the project, the higher the discount rate is.

However, Ryals (2003) argue that adjusting the discount rate to reflect risk may not bring the benefits expected. Firstly, the impact on CLV is small, showing only a reduced variation in its value, as CLV is usually calculated for a small period of time.

Secondly, as the risk is calculated based on the average customer portfolio, if there are any changes in the customer portfolio and this portfolio is composed of only a few key customers, this would result in the wrong discount rate risk; if the portfolio contains many customers, these changes may not be a problem.

This argument does not eliminate the possibility of using risk adjusted discount rate. An evaluation is required of the results and of the current situation when the possibility of using the risk adjusted discount rate is considered.

2.3.5. Discount Rate Analysis

Taking into account all these various ways of defining a discount rate, it is safe to assume that all these strategies have their usefulness and applicability. What should be taken into consideration is the reality of the company that is being evaluated, and what kind of risk and scenario is in place. More than one discount rate strategy can be applied at the same time. The key is to choose the most adequate one, which will give the more realistic and direct results.

In the case of CLV, the WACC still is the most commonly used discount rate, due to its easy accessibility by companies. However, one aspect that could be taken into consideration is the application of different discount rates at the same time, in order to compare the final results, having a good sensitivity analysis in operation. Based on that, the best result can be chosen and used to define strategies.

Supporting this argument, Ryals and Knox (2007) recommend the use of WACC, as it moves the CLV concept closer to the shareholder value analysis. To support their recommendation, they evaluated the resulting CLV based on a regular discount rate of 12% and on the WACC of 4.8% defined for the company.

They have calculated the CLV for ten major clients of a B2B insurer. It shows that there was not a major impact on the CLV using the two different discount rates. With an average CLV of four years, the difference was about 13.8%. They argue that the hardly significant differences in CLV values based on different discount rates are due to the fact that the CLV was predicted for a short period of time. The longer the predicted CLV, the bigger would be the impact of the discount rate on the present value.

This evaluation supports the fact that the choice of discount rate will depend on the situation under analysis. If in doubt of which discount rate to use, it is worth to use different ones and compare their results. This is applicable especially if it does not represent a big increase in monetary expenses that are not justifiable.

2.4. The Concept of Churn

Within the concept of customer analysis, the churn element is prominent not only for its role in the CLV evaluation, but also as a concept in its own right. As described by Kim *et al.* (2006), churn can be defined as the number or percentage of regular customers who decide to end a relationship with a service provider. This view of churn being related to a customer leaving a company is widely held within the churn research field.

Churn is closely related to the retention concept, representing the opposite effect: churn = 1 – retention. While the focus of the retention investigation is to find out why customers stay, churn focus on the reasons why customers leave. Churn investigation has grown over the years and some of the key studies published in the area of churn are displayed in Table 3.

Source	Method	Data Characteristics
Masand *et al.* (1999)	Simple regression, nearest neighbour, decision trees and neural networks	GTE wireless: 20 largest markets (over 3 million customers).
Wei and Chiu (2002)	Decision trees	Telecom: 21 million customers
Nath and Behara (2003)	Naïve bayes and association rules	Telecom: 171 variables. Three data files: 100,000; 51,306; 100,462.
Hwang *et al.* (2004)	Neural networks, decision tree and Logistic regression	Telecom - Final sample: 2,000 records, 101 variables.
Van den Poel and Larivière (2004)	Survival analysis and proportional hazard model	Financial company: 47,157 customers, 23 variables.
Larivière and Van den	Survival analysis and	Belgian financial provider:

Poel (2004)	proportional hazard model	519,046 customers.
Kim *et al.* (2005)	SVM and Back-propagation NN	Credit card: 4,650 customers, 8 variables.
Yu *et al.* (2005)	Churn-Strategy Alignment Model	Telecom: 100,000 customers, 171 variables.
Zhao *et al.* (2005)	One-class SVM, NN, DT and naïve bayes	Wireless: 2,958 customers, 171 variables.
Buckinx and Van den Poel (2005)	Logistic regression, ARD neural network and random forests	Fast moving consumer goods retailer: 32,371 customers, 61 predictors.
Ahn *et al.* (2006)	Logistic Regression	Telecom: 5,789 customers, 14 variables.
Coussement and Van den Poel (2008a)	Support vector machines, Logistic regression and random forests	Newspaper subscription: 90,000 customers, 82 variables.
Glady *et al.* (2006)	Logistic regression, decision trees and neural networks, AdaCost and cost-sensitive decision tree	Financial: 10,000 customers.
Kim *et al.* (2006)	Neural networks, decision tree and Logistic regression	Telecom - 16,384 records, 101 variables.
Neslin *et al.* (2006)	Logistic regression, decision trees, neural networks, discriminant analysis, cluster analysis and bayes	Telecom: 171 variables. Three data files: 100,000; 51,306; 100,462.
Burez and Van den Poel (2007)	Logistic regression, Markov chains and random forests	Pay-TV company: all active customers in Feb 28th 2002, and 29 variables.

Table 3. Churn literature review

All these studies use data mining or statistical methods, in order to predict the probabilities of churn. Some researchers developed their own tools, using traditional data mining techniques as the basis for their models (Masand *et al.*, 1999; Wei and Chiu, 2002). Masand *et al.* (1999) used CHAMP, software for churn prediction,

evaluating a customer's propensity to churn within the next sixty days in a wireless company. They used decision trees and genetic algorithm techniques to rank and group variables. Subsequently, they built simple regression, nearest neighbour, decision trees and neural networks models to predict churn, finding neural networks to have the best performance. The neural network model is then learnt by CHAMP, in order to better predict churn for different data sets. Wei and Chiu (2002) propose a churn prediction technique that uses call pattern changes and contractual data to identify potential churners at a contract level. They argue that in the mobile industry, customers can hold several mobile contracts at a time and it is necessary to take this into consideration when identifying churn. They then used the decision tree approach as the basis for the development of their technique.

As discussed below, other researchers use the traditional data mining techniques, comparing their results in order to determine the best technique for churn prediction, based on their data. This would be the case, for example, of Hwang *et al.* (2004) and Kim *et al.* (2006) when comparing the results of decision trees, neural networks and logistic regression; Buckinx and Van den Poel (2005) also compare logistic regression, neural network and random forests when predicting churn.

Hwang *et al.* (2004) use a three dimension approach, viz. current value, potential value and customer loyalty, in order to consider customer defection (churn) in the calculation of CLV. The importance of taking churn into consideration in the CLV model is because it affects future profit generation to the organisation. They then suggest a model that focusses not only on past profit contribution, but also on potential profit generation and future financial contribution of a customer. This model was applied to a wireless communication company in Korea, in order to use the customer value to calculate the lifetime value of customers and to define segmentation strategies. Decision trees, artificial neural networks and logistic regression were used to predict churn, and the result from the best performing model (in their case, logistic regression) was used in the calculation of CLV.

Following the same pattern of analysis, Kim *et al.* (2006) propose a CLV model and customer segmentation strategy considering churn and cross-selling opportunities. Their model was applied to a wireless telecommunication company and used decision trees, neural networks and logistic regression to predict the churn element to be used in the

calculation of CLV. They wanted to know what improvements in the marketing strategy would be necessary to make customers more valuable in each segment. One of their results shows that it is more important to retain customers than to develop new customers.

Buckinx and Van den Poel (2005) use logistic regression, automatic relevance determination (ARD) neural networks and random forests to build a model for predicting partial defection (churn). It is referred to as partial churn because the analysis is done in a non-contractual company, in this case a retailer environment, where customers will not stop all purchases at once; they will tend to start reducing the purchases, which can be considered as a partial churn. With time, this partial churn could become a total churn, so it is necessary to identify the partial churn to avoid total loss of the customers, especially focussing on the more loyal and valuable customer. All techniques presented similar performance, but random forests consistently performed marginally better. They came to the conclusion that it is possible to identify future partial churners, and it is possible to identify the best predictors, which could facilitate the definition of marketing strategies.

As cited in subsection 2.1.1., Glady *et al.* (2006) use a pre-established threshold on CLV to predict if a customer is going to churn or not. A churner is defined as someone with a decreasing CLV. Logistic regression, decision trees, neural networks, AdaCost and cost-sensitive decision tree are then used to compare the accuracy of classifying churn, after obtaining its value from the CLV calculation**.**

When comparing the performance of logistic regression, markov chains and random forests to predict churn, Burez and Van den Poel (2007) identified that random forests was the best model. It was then used to select a proportion of customers (top 30%) that obtained the highest churn probability. They then implemented a churn prevention programme on these customers to evaluate the reduction in churn provided by the programme. Based on this, it is possible to measure the profit contribution based on their churn prevention programme currently in operation. CLV and churn are used as main components in the profit contribution formula, which was obtained from Neslin *et al.* (2006).

Evaluating churn models for a tournament, Neslin *et al.* (2006) analysed the results from different data mining techniques used to calculate churn, which were applied together with the CLV calculation to evaluate gain in profit. They demonstrate that logistic regression and decision trees were the most used techniques for churn estimation, and conclude that they performed well and are good techniques to be used when analysing churn.

Naïve bayes and association rules are also applied for churn analysis, as demonstrated by Nath and Behara (2003). They discuss the use of naïve bayes and association rules to calculate churn, using the first for the analysis of a telecom data set. The idea is to identify the customers that are more likely to churn and define strategies to keep them, in order to control churn rates.

As emphasised by Larivière and Van den Poel (2004), companies not only need to understand churn, but they need to be able to convert the probability of churn into strong relationships with the customers. As a result, they used survival analysis to define the timing of churn, and more specifically, they estimated a proportional hazard model to test the impact of cross-selling at the moment of churn likelihood, evaluating if this could prevent churn. Van den Poel and Larivière (2004) also use survival analysis and the proportional hazard model to evaluate the churn behaviour, emphasising that the appropriate use of the results from the churn analysis may positively influence increasing retention rates.

The most commonly used data mining techniques for churn prediction are logistic regression (e.g., Hwang *et al.*, 2004; Buckinx and Van den Poel, 2005; Coussement and Van den Poel, 2008a; Glady *et al.*, 2006; Neslin *et al.*, 2006), decision trees (e.g., Coussement and Van den Poel, 2008a; Glady *et al.*, 2006; Neslin *et al.*, 2006; Burez and Van den Poel, 2007) and neural networks (e.g., Kim *et al.*, 2005; Zhao *et al.*, 2005; Glady *et al.*, 2006; Kim *et al.*, 2006). The use of support vector machines for churn prediction has also been receiving particular attention in recent research (Kim *et al.*, 2005; Zhao *et al.*, 2005; Coussement and Van den Poel, 2008a). In order to investigate how effectively support vector machines would detect churn, Kim *et al.* (2005) compared support vector machines and back-propagation neural network for predicting churn on a data set from a credit card company. Zhao *et al.* (2005) performed a similar analysis, comparing one-class support vector machines, neural networks, decision trees

and naïve bayes, also with the intention of empirically evaluating the predictive accuracy of support vector machines. In both studies, support vector machines performed better. The purpose again was to compare the accuracy of different data mining techniques when predicting churn.

Coussement and Van den Poel (2008a) also evaluate the predictive accuracy of support vector machines, benchmarking the result against logistic regression and random forests. However, in their study they used a sufficiently large data set to explore the churn prediction. Their research suggests that support vector machines presents a generalised good performance, but random forests still present the overall better performance. They also emphasise the importance of identifying the best churn predictors, to enable the adaptation of marketing strategies that can be used to reduce churn.

Some studies focus on identifying the reasons for churn (Nath and Behara, 2003; Hadden *et al.*, 2007; Zhao *et al.*, 2005). For example, churn could be caused by elements such as dissatisfaction with the product or service, and a better or similar offer from a competitor. These are unnatural defections that should be avoided. Nevertheless, there will always be a point of natural defection in the lifetime of a customer, such as moving to another country, which cannot be controlled (Van den Poel and Larivière, 2004). Yu *et al.* (2005) discuss a churn analysis that can be useful to managers, helping them define strategies. They developed a method, Churn-Strategy Alignment Model – CSAM, with the purpose to integrate key aspects of competitive strategy models and use it for churn evaluation. The idea is to focus on why customers churn, not only on the dichotomy churn or not churn.

When investigating the churn determinants and effects for a mobile phone company, Ahn *et al.* (2006) define various hypotheses that identify key elements that have an impact on customer churn. For example, they argue that call drop and failure rates influence positively the probability a customer will churn. They also suggest that the number of complaints has a positive influence on churn. They tested these and other hypotheses using a telecom data set. They emphasise that the idea is to identify the reasons and elements that influence customers' churn and by doing so, retain customers through churn reduction.

It can be concluded that in practice predictive power is often not the only criterion for model selection; ideally models should not only be accurate but also intuitive. Masand *et al.* (1999) emphasise that churn models should be complemented with intuitive explanations, such that they can be successfully used by marketing analysts and decision makers.

Understanding and distinguishing the reasons for churn is essential in the analysis of customer behaviour. This research will address this issue by interpreting and evaluating the data mining results, showing how to effectively incorporate domain knowledge into the churn modelling process.

2.5. Conclusions

CLV models assess the long-term value of customers based on their entire lifetime. They provide the means to understand upon which customer to focus. Nevertheless, most companies will not be able to accurately forecast the future behaviour of customers. This is due to the fact that historical value may not be a very precise predictor of future behaviour. However, this is the most common and relevant type of data that is available for analysis. What needs to be emphasised is that misclassification rates will always exist, but they can be minimised.

Valid data that is well understood represents a valuable source of information. Attention to the variables and knowledge of the organisation under investigation help to put this valuable information into the best use. One of the most important aspects is to be aware of not overcomplicating the CLV calculation. This should make it possible to obtain reliable results, avoiding the risk of being misled or using parameters that are not adequate.

The same aspects should also be taken into consideration when evaluating churn. In understanding and managing CLV and churn, a company can more efficiently allocate resources to its customers, becoming better able to concentrate on developing long-term customer relationships and well-defined strategies.

A key aspect for any company is to retain their customers, and the evaluation of churn and CLV can provide the elements to increase the level of retention. This research has

no intention of invalidating previous methods of investigating CLV and churn. The purpose is to use data mining to develop models for their prediction that could be more understandable and easier to interpret. This is done in order to facilitate their acceptability in the decision making process when defining strategies to deal with the customers. The main point is not only to be able to identify the customers under investigation, but to recognise which factors can influence the desirable (or undesirable) behaviour from the customers.

Chapter 3. Qualitative Investigation of the Customer Analysis

3.1. Introduction

From the literature on churn and CLV, it is possible to visualise how the researchers implement their analysis, but no information is provided on the companies' perceptions. Important questions arise regarding how companies perceive and implement their key customer strategies and what kind of analysis is performed.

This chapter explores the use of qualitative surveys, where companies were asked questions regarding their customer analysis process. The purpose was to investigate how key concepts (e.g., churn, CLV, domain knowledge, data mining, decision tables) are perceived by these companies.

First, there will be an examination of how the qualitative investigation was set. Then, the results from the pilot study will be discussed, followed by the evaluation of the final survey. Finally, conclusions will be drawn based upon the key findings, facilitating the analysis and understanding of the elements under investigation.

3.2. Survey Design for Qualitative Evaluation

When proceeding with qualitative analysis (Silverman, 2005; Mason, 2002), this research will focus on obtaining information from customer strategies adopted by companies. The qualitative method allows for the collection of key strategic information, which can then be used to compare different company processes.

During the design of the qualitative analysis, a method of getting involved with the interviewee and add flexibility was required. With that in mind, the approach adopted was semi-structured interviews with a guiding questionnaire (Mason, 2002). This approach permitted not only that questions could be asked in different order, but also added the possibility of including new points as the discussion progressed.

Another option would have been structured questionnaires, which are usually used to minimise bias. This is achieved by efforts to standardise the questions and the way they

are asked. This is based on the assumption that bias can be controlled or removed and follows a 'stimulus-response' model; once the stimulus is standardised, any variation in responses will be a true measure and not the result of the method used. Such an approach is appropriate for research which requires absolute responses to questions: 'yes', 'Elizabeth', 'Wednesday', for example, the type of data which Mason (2002, p.65) refers to as "…the kind of broad surveys of surface patterns."

The questions of bias interpretation could be a concern in some types of qualitative research. However in this case, this does not apply. This is due to the fact that the research in question is a focussed strategy comparison, not a behavioural examination.

This part of the research involved an enquiry into the opinions, explanations and descriptions which people made and it was necessary to use a qualitative approach which would enable the understanding of those perspectives, in order to attempt to see through the eyes of those being studied. Semi-structured interviews were the most appropriate method as they provided a framework upon which the respondents could be guided towards particular areas of interest but express themselves in their own terms and use of language, while conveying the depth of data which was required.

Based on the qualitative research characteristics, a smaller number of cases may be sufficient for analysis, as the focus is more on the details of the results (Silverman, 2005; Yardley, 2000). Hence, the focus was not on the amount of companies investigated, but on the information they provided and its comparison.

This research focus is not on similar patterns of behaviour, but of the information as a whole. Furthermore, the quantity of the data was manageable for manual analysis. As a result, there was no need to use qualitative analysis software. Also, the focal point was not on subjective information, but on clear indication of customer strategies' implementation and understanding. This can be better obtained through a complete look at the information provided by each company. Based on the information extracted, there can be comparisons to other companies' results. This approach was used in the pilot and final survey.

For the research, the interview questionnaire was designed to take approximately one hour, involving aspects of the CRM concept (e.g., churn, CLV, retention and satisfaction) and data mining.

After obtaining the results from the first three interviews and also as the research progressed, new elements were added to the survey, including aspects of domain knowledge and decision tables. This would add more value to the contribution of the findings, generating concise and specific results. Consequently, the questionnaire interview would take around one hour, but this time would vary, depending on the interviewee.

The main questions addressed in the interviews are related to the following aspects:

- Concept definitions: How do the companies define customer, CLV and churn?
- Customer metrics: Which of the customer measures (e.g., acquisition, loyalty, CLV and churn) are implemented and what are their levels of importance?
- Data manipulation and analysis: What kind of data is used for analysis? Which data mining techniques are used and how their performance is measured?
- Use of domain knowledge: How is domain knowledge used?

The findings from the pilot and final survey will be shown in sections 3.3 and 3.4, respectively.

3.3. Survey Results

For the survey analysis, seven companies were interviewed:

- company “A” is an airline company;
- company “B” is an airline company;
- company “C” is a telecom company;

- company “D” is from the automotive industry, especially car insurance and breakdown services;

- company “E” is an office retailer;

- company “F” is from the aerospace industry;

- company “G” is a medical insurance company.

In each company, the interviewee had key involvement in the customer relations process. For the purpose of confidentiality, any information that could identify the companies was eliminated. The key findings are discussed below.

3.3.1. Concept Definitions

Most of the companies interviewed in this research held similar definitions of the customer concept, where a customer is anyone that uses their services. Companies “B” and “D” have a broader approach, identifying the customer as anybody that makes contact, which can be through a transaction or even prospective customers in their database.

With regard to CLV definition, companies “C”, “D” and “G” define it as how much the customer is going to generate for them throughout their lifetime, which is consistent with the literature (Dwyer, 1997; Jain and Singh, 2002; Malthouse and Blattberg, 2005). This is realised through the calculation of the churn rate, life expectancy, average customer usage over that time and length of relationship with the company. Conversely, when asked about their understanding of CLV, the remaining companies argue that it is difficult to put into facts and figures, because their customers do not use their service all the time. Company “B” is trying to build up a profile of behaviour and propensity of service use, in order to enable cross-selling so they can increase the CLV. Although CLV is viewed as important, it is currently evaluated only by two of these companies.

When considering churn, most companies share the same concept: a customer churns when they stop paying. For company “D”, a customer churns if they do not renew their

service. They spend a great deal of effort when dealing with active churn, because the customer contacts them. They then use the CLV results to define a method to regain this customer, whilst trying to avoid passive churn when it is identified. On the other hand, companies "A" and "B" have distinctive knowledge and approaches. Company "A" is unfamiliar with the term churn, but they conclude that for them it is like market share. They can measure the defection on a monthly base for their business customers, but they do not have mechanisms in place to monitor and contact their individual customers. Similarly, company "B" is not aware of many measures of churn, but it is something they want to evaluate in the future.

In terms of churn probabilities, only companies "C" and "D" were able to provide some information. Company "C" considers that less than 10% of their customers churn every month, but they could not guarantee that this information was totally accurate. Company "D" evaluates churn in terms of groups and some of these groups will have a 3% or 4% churn rate, whereas other groups, for example customers for less than 12 months and less than 25 years old, will have a churn rate of 45 or 50%.

3.3.2. Customer Metrics

Acquisition, retention, loyalty and satisfaction are still the main customer metrics under evaluation by companies. In terms of importance, retention and loyalty are considered to have equal value by company "A", both being key priorities and being measured through their loyalty program. They believe churn will naturally reduce as a consequence of success in the previous metrics (retention, loyalty, satisfaction and acquisition). Finally, they find CLV very difficult to measure; they believe it is more difficult to quantify in comparison to the other elements, and as a result of this there is no immediate intention to measure it.

For company "B", satisfaction is their first priority, as it leads to loyalty, with satisfied customers increasing the likelihood of purchases and positive referrals. Consequently, retention will increase and churn will be smaller, leading to an increase in CLV. Acquisition is the only element analysed independently. For example, in markets where the company is not well known or where the company is just starting to operate, they believe acquisition should be their number one priority. Overall, they consider all metrics important with the sequence depending on the market under evaluation.

Company “C” is developing a strategy to detect churn, with their existing system detecting churn only after the customers have already gone. For them, churn analysis would be used to detect whether the customer is going to leave the company, which corresponds to the existing literature definition (Zhao *et al.*, 2005; Kim *et al.*, 2006). However, they do not consider churn as being easily identified. It is difficult to measure churn in their market, because the customer can stop using the service and then restart after a period of time, therefore being reactivated in their system. At the same time, a customer can have already churned and still be active in the database.

Whilst companies “D”, “E”, “F” and “G” were less forthcoming on their interaction with customers, they did provide some useful insights towards the importance and application of domain knowledge. In addition to this, companies “D” and “G” are the only ones implementing the CLV calculation. For both companies a finite time horizon is chosen. They argue that this time horizon is defined based particularly on company policies, as most of their financial evaluations assume a finite time frame. Also for both companies, CLV is calculated at the individual level and the churn probability is used as an element within the CLV calculation. Company “D” implements the calculations using different values for time horizon and discount rate (with the WACC as their main point of analysis). The purpose is to see how their segments would behave, and as long as the methodology creates the same relationship between groups of customers, the absolute CLV amount is not important.

3.3.3. Data Manipulation and Analysis

In terms of predictions, the companies use all available records and variables in their database, the idea being to use the same data to perform any calculations needed. This process of data preparation is still under way for companies “A”, “B” and “C”, who are experiencing difficulty in key variable identification and evaluation. Companies “D”, “E” and “G” consider that the same data source is used for all their customer analysis, but key variables are important for specific calculations. For example, company “G” selects a sample of around 50,000 customers for their calculations from their data set. They use the same data for calculating different elements, including CLV and churn, but develop different methods for each one of the aspects measured.

An assessment of initial data mining use shows that companies “A” and “E” rely mostly on Microsoft Excel for their analysis, whereas companies “B”, “C”, “D” and “G” rely on data mining use for their customer analysis. Company “F” only has a few big clients, and because of that, they consider that there is no need to create any models for customer evaluation.

Company “B” uses decision trees, more specifically classification trees, in order to group customers into specific segments. Neural networks are the main data mining technique used by company “C”. From their model, which is still under development, they extract the algorithm and apply it to the newest available data, scoring the model against the new variables, trying to predict how the new customers are going to behave. Both companies “D” and “G” use regression for their customer analysis, which includes, for example, the churn and the lifetime value prediction. Company “G” also uses decision trees if it is necessary to implement something simpler. Neural networks were tried, but found to be less useful in their context.

The concept of decision tables is of relevant importance for this research and it will be described and further investigated in Chapter 7. However, it is not known or used by any of the companies.

In terms of performance metrics for the data mining models, lift chart is used by companies “C”, “D” and “G”. Whilst companies “C” and “G” use lift chart to compare the performance of the original data and of the control group (analysing how well the model is being estimated), company “D” uses lift chart and ROC curves to choose the model that has a better performance (e.g., for churn). The lift chart and the ROC curve, together with other performance measurements that will be used in this research, will be further explained in Chapter 4.

3.3.4. Use of Domain Knowledge

In a general sense, domain knowledge is considered as common sense, being applied in different parts of the data analysis. This view is shared by most of the companies. Company “D” uses common sense in their data preparation, for example, when choosing the type of discount rate and the amount of years for their time horizon; however, there is not a strong use of domain knowledge in their analysis, except when

the final results seem contradictory, which leads to further investigation. On a similar approach, company "G" usually relies on statistics. Only in the final results, if the model presents a variable that did not make sense in the analysis, they would eliminate it to avoid questioning the validity of the model.

The domain knowledge aspect will be further discussed in section 5.4, Chapter 5.

3.4. Conclusions

When evaluating different companies, with different sizes and sectors, it would be expected to see great variation in their attitude towards customers. In general for this research, this was not the case.

Based on the results presented over this chapter, the key points found in this survey are:

- The customer is generally someone that uses the company's service, but some companies are also taking into consideration future prospects.

- CLV is viewed as important, but not evaluated by the majority of companies interviewed. Only two companies calculate it, and in both cases it required dedication and convincing of the methodology.

- Churn is also important, viewed as something to detect if the customer is going to leave the company or not. Most of the companies emphasise the difficulty of predicting it. Three companies are developing or using strategies to detect it and in two of the cases, the churn prediction is used as an element in the CLV calculation. This demonstrates that the most advanced companies in customer evaluation visualise and understand the importance of integrating these two elements.

- In general, acquisition, retention, loyalty and satisfaction are still main factors, but the idea that all the elements are interconnected is growing. With this is mind, they can develop a more concise and stable customer analysis.

- In terms of data selection, their data usually comes from one source and their idea is to always group them together. The variables and models used would depend on the measurement in place. Some companies are more advanced in this aspect, but in general the idea is the same.

- Domain knowledge is a key element in the data analysis and it is approached in different ways. The idea of using common sense leads to a more detailed evaluation, using the knowledge acquired to guide and influence the analysis. This knowledge would be applied especially in the final results, to ensure the validity of the models generated.

- Decision tables, which will be an important element in this research, are not used by any of the companies enquired.

- In terms of data mining, some companies are not using any of the techniques, still manipulating their data through the use of excel. These are the companies that have a less advanced measurement in place, focussing more on specific statistic calculations. For the other companies that are developing or implementing a deeper customer evaluation, which includes churn and CLV predictions, the use of data mining is irrefutable. The techniques used are decision trees, neural networks, logistic and multiple regressions, which are in tune with the common practice, and each company has their own preference.

- In terms of performance measurements, some companies are not aware of any of them yet. They are still in a position of thinking or starting to implement the techniques, and the performance measurements are something they still need to take into consideration. Again, only the most advanced companies in customer evaluation are using performance measurements, including the use of lift charts and one example ROC curve measurement.

From this, it can be concluded that there is always the chance the customer is going to leave, and the companies are aware of that. What they are trying to do is to define methods that can help them identify which customers to focus on, which ones have bigger propensity to leave and which ones are going to stay. However, this is not an easy task and not all the companies are prepared for this type of analysis.

One important aspect is the use of domain knowledge in their analysis (which will be further investigated in Chapter 5), and to do that, they need to understand their data, prepare it well and make sure that the results presented are coherent and useful. For example, companies tend to focus on retention to make sure that the customer is not defecting. They have a percentage of churn, but they do not analyse why the customer defected.

In conclusion, the higher levels of the company usually need convincing about the reliability of the models results. To do this, it is necessary to understand the results presented and trust them. The performance of the models would be a good argument for convincement, but the focus on the main variables of the models would support the implementation of key strategies. This would create a bigger impact on the company's positioning in the marketplace, especially where the definition of such models is still under development.

Chapter 4. Benchmarking Predictive Methods for Churn Prediction

4.1. Introduction

A key aspect of data analysis is the use of techniques and methods to facilitate the manipulation of data and extraction of valid information. Because of the amount of data currently available and the necessity of obtaining more accurate and concise results, the use of data mining became a natural part of data analysis.

Data mining refers to the process of extracting information and knowledge from large amounts of data (Hand *et al.*, 2001; Giudici, 2003; Nath and Behara, 2003; Witten and Frank, 2005). To achieve this, data mining techniques explore the repositories of data and extract the information needed.

In this chapter, the use of four data mining techniques is explored, evaluating their results on three different data sets. Firstly, the key concepts of the four data mining techniques are described. Secondly, the data set descriptions are presented. Lastly, the data analysis results are evaluated. The purpose of this approach is to set a benchmarking of how the techniques perform for the prediction of churn, defining the two techniques that will be further evaluated in subsequent chapters.

4.2. Logistic Regression

Linear regression is a type of regression analysis in which data are modelled by a least squares function which is a linear combination of the model parameters and depends on one or more independent variables (Witten and Frank, 2005; Moore and McCabe, 2006). Linear regression expresses the target variable as a linear combination based on the predictive variables that are relevant for the target variable. For m variables and i = 1,…, n individuals, the linear regression formula can be represented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m$$

where:

- y is the target variable for each individual i.
- β_0 is the constant parameter, which represents the intercept.
- β_j represents the weight given to the specific variable x_j associated to it (j = 1, … m).
- x_1 , ..., x_m represents the predictive variables for each individual i, from which y is to be predicted.

However, linear regression would not be the best choice in the churn context, as the target variable is not continuous and the distribution is not normally distributed. As a result, logistic regression would then be chosen.

Logistic regression (So, 1995; Allison, 2001; Giudici, 2003; Witten and Frank, 2005; Moore and McCabe, 2006) is a predictive modelling technique, where the dependent variable is discrete or categorical, for example, churn (1) or not churn (0). The logistic regression models the relationship between p (the probability of the target event occurring - e.g., churn = 1) and x_1 , ..., x_m (the predictive variable (s) for the model):

$$p(y = 1 \mid x_1, \ldots, x_m) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m)}}$$

where y is the binary class label {0,1}, x is the input data, and the parameters β_0 (intercept) and β_1 to β_m are typically estimated using the maximum likelihood procedure (Allison, 2001). This resulting formula tries to estimate the probability $p(y = 1 \mid x_1 , \ldots, x_m)$.

For example, if the target variable in the logistic model is “churn”, after analysing the data set and obtaining the logistic regression equation, if the value of p is bigger than some selected cut-off value (e.g. 0.5), it means the event is satisfied (in this case, customer i is going to churn). This evaluation can be done for all customers in the data set, making logistic regression a useful and powerful tool for prediction. As an example, evaluating a churn data set where the target variable is “churn”, “Inter” represents if the customer has international plan, “Vmail” if the customer has voice mail plan, and

"CustServ" means if the customer contacted the customer service centre, the probability of "*p*" would be represented by:

$$p(churn = 1 \mid Inter, Vmail, CustServ) = \frac{1}{1 + e^{-(0.12 + 1.13*Inter - 1.1*Vmail + 0.56*CustServ)}}.$$

The table below (Table 4) shows the results for this specific situation:

Inter	Vmail	CustServ	*p*	Churn
Yes	Yes	5	0.95	Yes
No	Yes	5	0.86	Yes
No	No	5	0.94	Yes
No	No	1	0.66	Yes
No	Yes	0	0.27	No

Table 4. Logistic regression example

Only in the last situation where p = 0.27, assuming a cut-off of 0.5, would the model predict that the customer did not churn.

As mentioned before, the maximum likelihood procedure is the one usually adopted to estimate the parameters of the logistic regression model (Allison, 2001; Witten and Frank, 2005). It presents consistent results, and as the sample size grows, the probability of the estimate being near to its true value also grows, making it a good predictor for large samples. Using the same notation for the variables in the logistic formula, the log-likelihood equation is represented as:

$$LL = \sum_{i=1}^{n} y_i \log(p(y = 1 \mid x_1, ..., x_m)) + (1 - y_i) \log(1 - p(y = 1 \mid x_1, ..., x_m))$$

In this case, *n* represents the number of data points of the training set under evaluation and the variable y_i will assume value zero or one. The log-likelihood should then be maximised. To obtain this, a widely used method is the Newton-Raphson algorithm (Allison, 2001; Hastie *et al.*, 2001). For a given data set and probability model, maximum likelihood estimation produces estimates for the model regression

coefficients that would generate the observed data at hand more likely than any other parameter values.

To test the significance of the coefficients obtained through the maximum likelihood procedure, test statistics are available, more specifically the Wald chi-square statistic (Allison, 2001; Moore and McCabe, 2006). This is obtained by dividing each coefficient by its standard error and squaring the result. The standard error is a measure of the accuracy (or estimate of the error) of the coefficient estimate due to the randomness caused by using a sample instead of the entire population. In this case, the bigger the Wald chi-square statistic, the more predictive the variable is to the model. Figure 2 shows an example of the maximum likelihood coefficient estimation obtained from SAS software, which also shows the Wald chi-square values.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.8834	0.0733	145.3110	<.0001
plan_chosen	1	1	0.4001	0.0631	40.2073	<.0001
plan_chosen	2	1	-2.7032	0.1568	297.2585	<.0001
plan_chosen	3	1	1.0117	0.0621	265.7206	<.0001
new_mobile	N	1	-0.2042	0.0762	7.1899	0.0073
new_mobile	U	1	0.4862	0.0464	109.8079	<.0001
Prizm_G	Group1	1	-0.5118	0.0381	180.4875	<.0001
Prizm_G	Group2	1	0.0527	0.0338	2.4300	0.1190
total minutes		1	-0.00229	0.000156	214.2370	<.0001

Figure 2. Example of likelihood estimation in SAS

In this case, it demonstrates that most variables are highly significant to the target variables (in this example, churn).

An important point to be emphasised is that predictive power makes logistic regression a commonly selected technique in prior studies of churn (Hwang *et al.*, 2004; Buckinx and Van den Poel, 2005; Ahn *et al.*, 2006; Coussement and Van den Poel, 2008a; Glady *et al.*, 2006; Kim *et al.*, 2006; Neslin *et al.*, 2006; Burez and Van den Poel, 2007), which encourage the further investigation of its results.

These studies were described in more detail in Chapter 2, but it is important to underline the predictive power of this technique and its capacity to generate understandable and concise results.

4.3. Decision Trees

The idea when building a decision tree (Giudici, 2003; Harper and Winslett, 2006; Witten and Frank, 2005) is to group similar records in each node, making the classification as pure as possible. It works with a usually discrete target variable, using a set of predictive variables to determine the classification.

Figure 3 shows a decision tree example, where the target variable is churn (churn = 1, not churn = 0), and the predictive inputs are "IntlCalls" (international calls), "Vmail" (voicemail), and "CustServ" (customer service calls). In this case, for example, it considers that if the customer does not make international calls but makes service calls, this customer will churn.

.

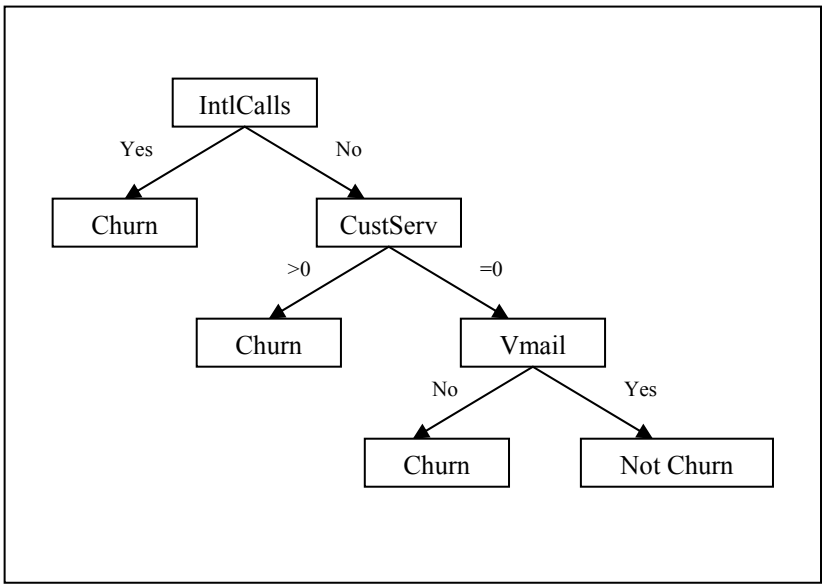

Figure 3. Decision tree example

This is only a small example to demonstrate the tree structure, with a top-down classification, starting from the root and working one's way down according to the outcomes from the internal nodes, until a leaf node has been reached and the corresponding class is assigned.

There are many algorithms and programs for computing empirical decision trees. Three of the main algorithms are CART (Breiman *et al.*, 1984; Larose, 2005; Harper and Winslett, 2006), C4.5 (Quinlan, 1993; Baesens *et al.*, 2003b; Larose, 2005), and CHAID (Kass, 1980; Neville, 1999).

The CART (Classification and Regression Trees) algorithm was suggested by Breiman *et al.* (1984). It produces only binary trees, with exactly two branches for each decision node. It can be used for predicting the values of a continuous or categorical variable. When the target variable is categorical, the technique is referred to as classification trees; if the target variable is continuous, the method is referred to as regression trees.

For regression trees, CART may identify a finite list of possible splits based on the number of different values that the variable actually takes in the data set (Larose, 2005). For classification trees, the idea is to stop when the final nodes are sufficiently pure, containing only one type of information (e.g., only churn, or only not churn), finding the optimal split for each node.

The C4.5 algorithm was generated by Quinlan (1993). Like the CART algorithm, it recursively goes to each decision node, selecting the optimal split using an entropy-based criterion until no further splits are possible. A difference lies in the fact that the tree generated is not restricted to binary splits, having the possibility of one branch for each value of a categorical attribute. Another important aspect is that the C4.5 algorithm does not accept continuous variables as target variables.

The CHAID (Chi-Squared Automatic Interaction Detection) algorithm was introduced by Kass (1980). It recursively partitions the data with a nominal target variable using different splits on the input variables (Neville, 1999). The splitting criterion is based on p-values from the chi-square distribution. The p-values are adjusted to accommodate multiple testing. After a split is adopted for an input, its p-value is adjusted, and the input with the best adjusted p-value is selected as the splitting variable. If the adjusted p-value is smaller than a specified threshold, then the node is split. The tree construction ends when all the adjusted p-values of the splitting variables in the non split nodes are above the specified threshold.

Decision tree algorithms use criteria such as entropy and gain to select the best splitting decision for the nodes. If p_1 and p_0 are the proportion of records of class 1 (e.g., churn) and 0 in the sample S, the entropy of S is:

$$\text{Entropy}(S) = -p_1 \log_2(p_1) - p_0 \log_2(p_0)$$

where p_0 = 1 - p_1, and the maximum entropy occurs when $p_1 = p_0 = 0.5$ and minimal when $p_1 = 0$ or $p_0 = 0$. The Gain (S, x_i) is then calculated, and based on it, the node split will be decided:

$$\text{Gain}(S, x_i) = \text{Entropy}(S) - \sum_{v \in values(x_i)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

where S_v characterises a sub-sample of S where the attribute x_i has one specific value. The Gain aims to measure the weighted reduction in entropy, based on the splitting on the attribute x_i. The smaller the entropy the better the splitting rule applied.

However, this method favours splits on attributes that have many distinct values when deciding about the node splits. If a data set uses a distinctive ID attribute for each customer, the technique would select it as the best splitting decision. To correct this, the C4.5 algorithm applies a normalisation and uses the gain ratio criterion:

$$\text{Gainratio}(S, x_i) = \frac{\text{Gain}(S, x_i)}{\text{SplitInformation}(S, x_i)}$$

where the SplitInformation is defined as:

$$\text{SplitInfomation}(S, x_i) = -\sum_{v \in values(x_i)} \frac{|S_v|}{|S|} \log_2 \frac{|S_v|}{|S|}$$

In this case, the SplitInformation is the entropy of S relative to the values of x_i. It measures how the attribute x_i splits the data and it discourages the selection of attributes with many uniformly distributed values. The C4.5 algorithm computes the Gain of all attributes and then selects the attribute with the highest Gainratio, taking into consideration that its Gain is at least as large as the average Gain over all attributes examined.

Another splitting method that could be used is based on the gini measure, which is normally used by the CART algorithm. It is interpreted as the probability that any two records selected at random are different. It can be represented as:

$$\text{Gini}(S, x_i) = 1 - {p_0}^2 - {p_1}^2 = 2p_0p_1$$

In this case, a pure node would have a gini index of zero (0). As the number of classes becomes less evenly distributed, the gini index would approach zero.

The chi-square test can also be used for making the splitting decision. This method measures the difference between what is expected to happen and what actually happens, with one degree of freedom:

$$\chi^2 = \sum_{i=1}^{m} \frac{(Expected_i - Actual_i)^2}{Expected_i} \approx \chi^2(1)$$

where m is the number of categories of a variable under evaluation. If the chi-square value is large, then the p-value associated with the chi-square is small, and it represents a significant difference. The best split will be the one with the smallest p-value. By default, the p-values are adjusted to take into account multiple testing. For CHAID, the splitting decision is based on this. Partitioning stops when no split meets the level of significance stipulated by the model.

These methods can be approximated in SAS Enterprise Miner. In this research then, an approximation of the CHAID algorithm was employed, through the use of the chi-square test for splitting decision. The tree is then constructed by means of recursive partitioning, until the leaf nodes are pure (contain only instances of a single class) or until they cannot be improved. As this would result in a complex tree, strategies can be used to solve the problem.

In terms of stopping decision, the idea is to avoid very small and very big trees. Very small trees will not describe the data well whereas very big trees will present too many leaves with little data, which may not be reliable for predictions. If the tree continues splitting, it will end up with one training example per leaf and the tree is overfitting. The strategy of early stopping could be used to avoid the tree overfitting.

Breiman *et al.* (1984) suggest the use of pruning instead of stopping rules. The argument is to let the tree grow until all the final nodes are pure or nearly pure, and then pruning would be used to reduce the tree to a size that would represent the lowest misclassification rate. This pruning process is also suggested by Quinlan (1993), where the pruning process is executed retrospectively after the full tree is obtained.

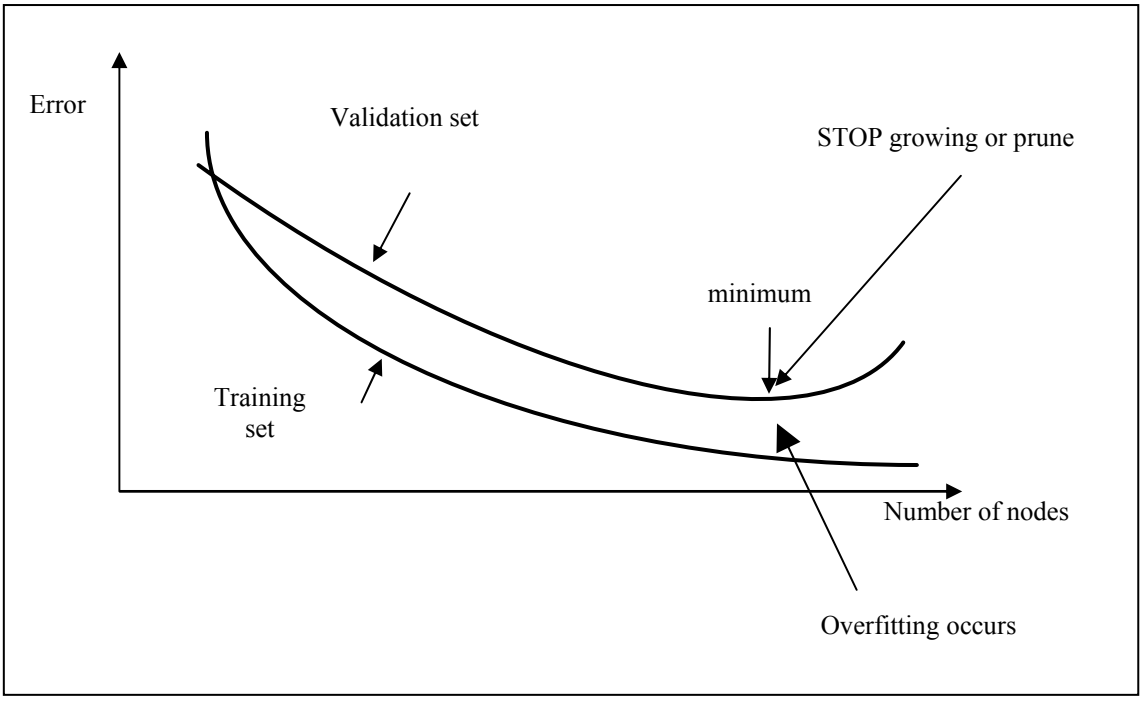

Figure 4. Using a validation set for building decision trees

In general, the idea is to generate the best tree that would not overfit. A tree overfits when the good fit of the training set is not replicated when applied to a different sample (Neville, 1999). As demonstrated in Figure 4, the idea is to use a training sample to grow the tree and use a separate validation sample to decide on optimal size of the tree. This can be done by growing the tree, monitor the error rate on the validation set and stop growing when the latter starts to increase; or to grow a full tree, and prune retrospectively using the validation set. In this research, the second option will be adopted. Note that this error can be represented by the misclassification rate or a cost based measure. For the classification trees, the misclassification rate will be the one used. For regression trees, the mean squared error (MSE) will be the one used.

In conclusion, decision trees produce models that are easy to understand, accept different data types, are useful for preliminary variable selection, and are capable of

handling missing values. They have good interpretability, being able to visualise and identify the decision rules resulting from its definition.

These reasons, together with its predictive power, make it an often used technique for churn prediction (Masand *et al.*, 1999; Wei and Chiu, 2002; Hwang *et al.*, 2004; Buckinx and Van den Poel, 2005; Zhao *et al.*, 2005; Coussement and Van den Poel, 2008a; Glady *et al.*, 2006; Kim *et al.*, 2006; Neslin *et al.*, 2006; Burez and Van den Poel, 2007). These churn applications were explored in detail in Chapter 2.

4.4. K-Nearest Neighbours

K-nearest neighbours (KNN) (Henley and Hand, 1997; Baesens *et al.*, 2003b; Witten and Frank, 2005) classify a data instance by considering only the k most similar data instances in the training set. The KNN algorithm takes a data set and a target observation, where each observation in the data set is composed of a set of variables and the target observation has one value for each variable.

The KNN are commonly determined by the Euclidean distance between an observation and the target observation, which can be defined as:

$$d(x_i, x_j) = \| x_i - x_j \| = \sqrt{\sum_{e=1}^{m} (x_{ie} - x_{je})^2}$$

where x_i and x_j are the input vectors of data instance *i* and *j*, respectively, which are part of the data set under investigation, taking into consideration all *m* variables. The Euclidean metric considers that a neighbour is regarded as nearest if it has the smallest distance in the input space. The distance between an observation and the target observation is then calculated. The k observations that have the smallest distances to the target observation are the KNN to that observation. The class label is then assigned in accordance with the class of the majority of the KNN.

Table 5 shows the voting approach of the neighbours for a binary target variable when different values of k are specified. In this example, five observations (10, 15, 50, 108, and 333) are the closest observations to the target observation. Their ranking is shown

in Table 6. The KNN are the first k observations that have the closest distances to the target observation. If the value of k is set to three, then the target values of the first three nearest neighbours (108, 15, and 10) are used. The target values for these three neighbours are C, NC, and C. Therefore, the posterior probability for the target observation to have the target value C is 2/3 (67%). This means that based on the target values of the KNN, each of the KNN votes on the value for a target observation. The votes then lead to the posterior probabilities for the binary or nominal target variable.

K	Nearest neighbours ID	Target Value of nearest neighbours	Probability of target Observation
1	108	C	p(C) = 100% p(NC) = 0%
2	108,15	C, NC	p(C) = 50% p(NC) = 50%
3	108,15, 10	C, NC, C	p(C) = 67% p(NC) = 33%
4	108,15, 10, 333	C, NC, C, NC	p(C) = 50% p(NC) = 50%
5	108,15, 10, 333, 50	C, NC, C, NC, C	p(C) = 60% p(NC) = 40%

Table 5. Voting approach with different values of k

Neighbour ID	Target: churn (C)/ not Churn (NC)	Ranking (1 = closest; 5 = farthest)
10	C	3
15	NC	2
50	C	5
108	C	1
333	NC	4

Table 6. Ranking power of neighbours

Figure 5 illustrates the situation from Table 6, demonstrating the 5-nearest neighbour classification rule, where starting from the target observation X, the classifier grows a spherical region until the five nearest training points are enclosed. The classifier will then assign the label “C” to X since three of the five nearest neighbours have this label.

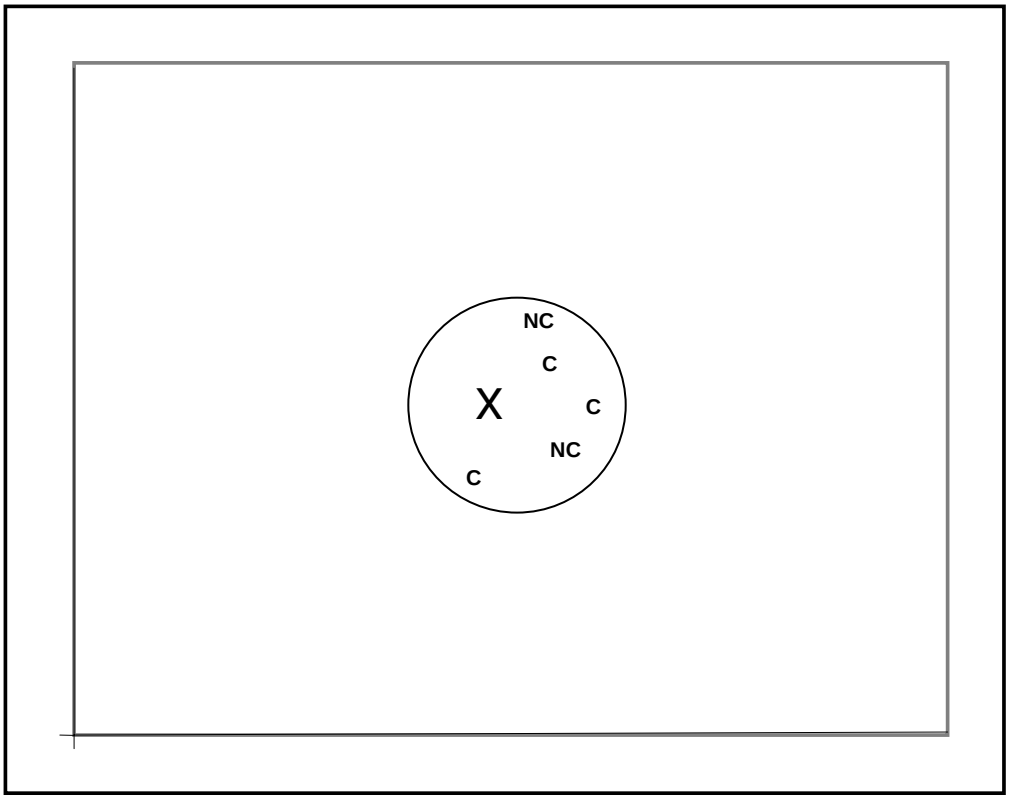

Figure 5. The 5-nearest neighbour classifier

More advanced distance measures have been proposed in the literature, but for the purpose of this research the Euclidean metric will be used. As a result, these other metrics will not be explored but they can be further investigated in Henley and Hand (1997).

In conclusion, KNN classifiers are simple and often work well, but as discussed by Hadden *et al.* (2007), they have received little attention in the customer research community. Masand *et al.* (1999) used KNN when testing different techniques for predicting churn but it was not the one that performed better. Using KNN in this research enables comparison of its performance with other data mining techniques when predicting churn. Good results would emphasise the possibility of its use in the churn context; otherwise, it will corroborate the fact that KNN do not perform well in customer analysis.

4.5. Neural Networks

Neural networks (NNs) are mathematical representations designed to model the functioning of the human brain. They are nonlinear models that recognise patterns.

Many types of NNs have been suggested in the literature for both supervised and unsupervised learning [Bishop, 1995; Ripley, 1996).

One of these methods is the multilayer perceptron (MLP), which is the most popular NN for classification (Desai *et al.*, 1996; Crone *et al.*, 2006; Witten and Frank, 2005).

A MLP is generally composed of at least three layers: one input layer, one (or more) hidden layer(s) and an output layer, each consisting of several neurons (Desai *et al.*, 1996; Baesens *et al.*, 2003a; Baesens *et al.*, 2003b; Crone *et al.*, 2006).

Figure 6 illustrates an MLP example with one hidden layer and one output neuron, where each neuron processes its inputs and generates one output value which is transmitted to the neurons in the subsequent layer.

The output of the hidden neuron *i* is then computed by processing the weighted inputs and its bias term $b_i^{(1)}$:

$$h_i = f^{(1)}(b_i^{(1)} + \sum_{j=1}^{n} w_{ij} x_j)$$

where w_{ij} represents the weight connecting input *j* to hidden unit *i*. Taking this into consideration, the result of the output layer is calculated as:

$$z_i = f^{(2)}(b_i^{(2)} + \sum_{j=1}^{n_h} v_j h_j)$$

with n_h being the number of hidden neurons and v_j represents the weight connecting hidden unit *j* to the output neuron. The bias inputs play a similar role as the intercept term in a classical linear regression model.

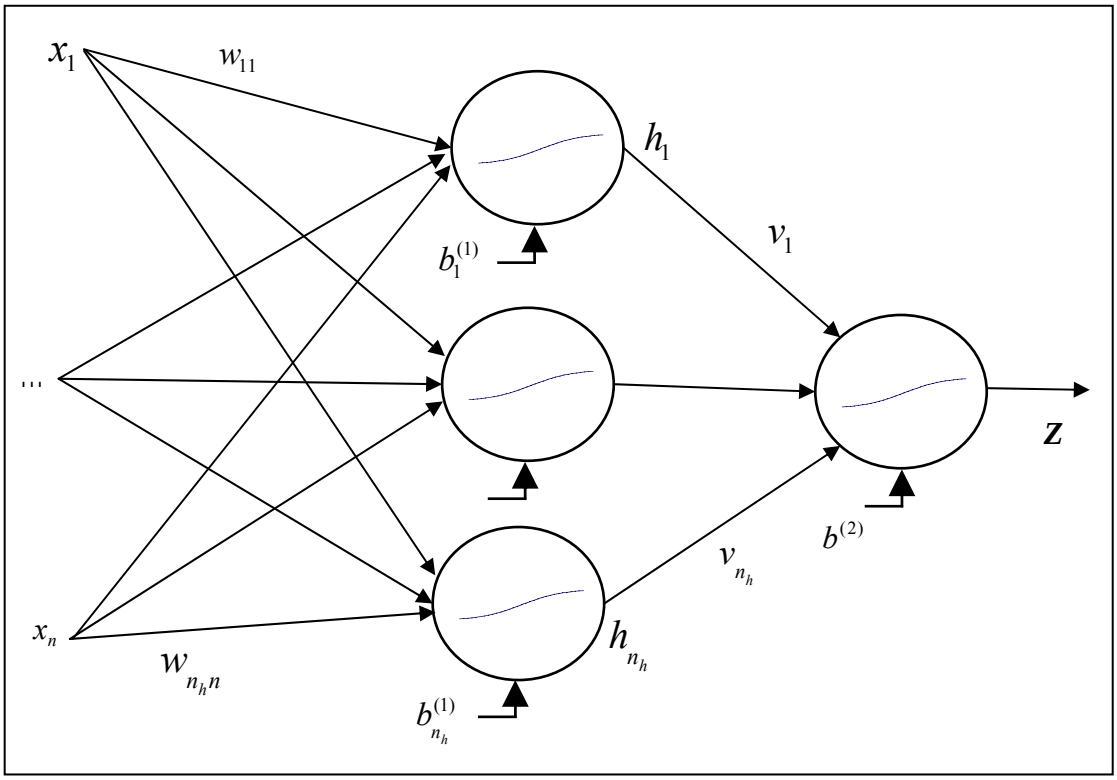

Figure 6. Example of MLP with one hidden layer

The transfer functions $f^{(1)}$ and $f^{(2)}$ allow the NN to model nonlinear relationships in the data. Some transfer functions that are commonly used are the sigmoid function:

$$f(x) = \frac{1}{1+\exp(-x)},$$

the hyperbolic tangent function:

$$f(x) = \frac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)}$$

and the linear transfer function:

$$f(x) = x.$$

For a binary classification problem, it is convenient to use the logistic transfer function in the output layer ($f^{(2)}$), since its output is limited to a value within the range [0; 1]. This allows the output y of the MLP to be interpreted as a conditional probability of the form $p(y = 1|x)$ (Bishop, 1995). In that way, the NN naturally produces a score per data

instance, which allows the data instances to be ranked accordingly, for example, for churn classification purposes.

Note that multiple hidden layers might be used, but theoretical works have shown that NN with one hidden layer are universal approximators capable of approximating any continuous function to any desired degree of accuracy on a compact interval (universal approximation property) (Bishop, 1995).

The weights *w* and *v* are essential parameters of the NN and need to be estimated during the training process. Many algorithms have been suggested to perform this estimation (Bishop, 1995). For example, given a training set of *n* data points,

$$D = \{(x_i, y_i)\}_{i=1}^{n},$$

with input data x_i that belongs to the domain under investigation and corresponding binary class labels y_i (0 or 1), the weights of the network are first randomly initialised and then iteratively adjusted so as to minimise the objective function *G*. Since the target output is categorical, an appropriate objective function is the cross-entropy function:

$$\mathrm{G} = -\sum_{i=1}^{n} \{y_i \log(z_i) + (1 - y_i) \log(1 - z_i)\}$$

With z_i being the predicted network output for observation *i*. It is possible to notice that this error function reaches its minimum when $z_i = y_i$ for all $i = 1,\ldots, n$. The logistic regression classifier is essentially a simplified NN with only one neuron. It has a sigmoid activation function and is trained to minimise the cross-entropy error.

The purpose of using NNs is to produce a model which performs well on new data that has not been evaluated before. To achieve this, it is necessary to prevent the NN from fitting the noise in the training data, which can be done by monitoring the error on a separate validation set during training of the NN. When the error measure on the validation set starts to increase, training is stopped (early stopping). However, this causes a loss of data which cannot be used for estimating the weights. Also, this method may not be appropriate for small data sets.

An alternative way is to add a penalty term to the objective function (Bishop, 1995):

$$f(w) = G + \alpha E_W$$

with $E_W = \frac{1}{2}\sum_i w_i^2$

In this case, *w* is the weight vector representing all weights *w* and *v*. This method for improving generalisation constrains the size of the network weights and is referred to as regularisation. When the weights are kept small, the network response will be smooth, decreasing the tendency of the network to fit the noise in the training data (Baesens *et al.*, 2002).

In conclusion, NN can yield significant predictive accuracy when compared to other techniques and that is why it is becoming widely used in the prediction of churn (Masand *et al.*, 1999; Hwang *et al.*, 2004; Buckinx and Van den Poel, 2005; Kim *et al.*, 2005; Zhao *et al.*, 2005; Glady *et al.*, 2006; Kim *et al.*, 2006; Neslin *et al*., 2006). However, as emphasised by Baesens *et al.* (2003a) and Wei and Chiu (2002), the resulting model lacks interpretability, preventing it from being used as an effective management tool for decision making. If the purpose is only to generate an accurate model, NNs work very well, but if the results need to be evaluated, it is necessary to adapt its final results.

4.6. Performance Metrics

4.6.1. Classification Accuracy, Sensitivity and Specificity

To measure the performance of the classification models with the binary target variable, the prediction matrix in Table 7 can be used. Based on this table, it is possible to extract the information to calculate the classification accuracy (CA), sensitivity and specificity.

Predicted	Actual value	
	+	-
+	True Positive (TP)	False Positive (FP)
-	False Negative (FN)	True Negative (TN)

Table 7. Confusion matrix for binary classification

The CA measures how many observations were correctly classified (Witten and Frank, 2005). It is represented as:

$$CA = \frac{TP + TN}{TP + FP + TN + FN}$$

The sensitivity indicates how many events (churn) turned out to be predicted correctly (Van Erkel and Pattynama, 1998; Braga and Oliveira, 2003; Witten and Frank, 2005):

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

The specificity indicates how many non-events (not churn) turned out to be predicted correctly (Van Erkel and Pattynama, 1998; Braga and Oliveira, 2003; Witten and Frank, 2005), which is defined as:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

It is important to notice that CA, sensitivity and specificity vary together as the threshold (cut-off) on a classifier is varied between its extremes, 0 and 1. To demonstrate the calculations of these performance measures and the ones that will be discussed in the next subsections, the data set in Table 8 was used. For this example, there are 40% churners and 60% non-churners. In the way this data was prepared, the intention was to identify which customers were non-churners, but the process to identify churners would follow the same proceedings. This is only a fictional generated data set, in order to demonstrate the definitions of the performance measures.

			Output by classifier	Actual status	KS Curve		Lift			ROC			
Customer	**Age**	**Income**	***Score***	***Actual Value***	***p(s|NC)***	***p(s|C)***	**rank**	***Score***	***Actual Value***	***Sens.***	***Spec.***	***1-specificity***	***ROC Increment***
C1	18	1,000	0.06	Churn	0	0.1	1	0.96	not churn	1	0.1	*0.9*	0.1000
C2	25	2,300	0.11	Churn	0	0.2	2	0.92	not churn	1	0.2	*0.8*	0.1000
C3	65	3,000	0.2	not churn	0.067	0.2	3	0.87	not churn	0.933	0.2	*0.8*	0.0000
C4	73	1,500	0.25	Churn	0.067	0.3	4	0.83	churn	0.933	0.3	*0.7*	0.0933
C5	32	2,500	0.33	Churn	0.067	0.4	5	0.82	not churn	0.933	0.4	*0.6*	0.0933
C6	54	4,000	0.35	not churn	0.133	0.4	6	0.81	churn	0.867	0.4	*0.6*	0.0000
C7	22	1,500	0.4	Churn	0.133	0.5	7	0.78	not churn	0.867	0.5	*0.5*	0.0867
C8	46	5,000	0.41	not churn	0.2	0.5	8	0.73	not churn	0.8	0.5	*0.5*	0.0000
C9	57	8,000	0.48	not churn	0.2	0.5	9	0.72	not churn	0.733	0.5	*0.5*	0.0000
C10	41	3,000	0.51	Churn	0.267	0.6	10	0.7	churn	0.733	0.6	*0.4*	0.0733
C11	21	1,000	0.56	not churn	0.333	0.6	11	0.67	not churn	0.667	0.6	*0.4*	0.0000
C12	58	3,200	0.59	not churn	0.4	0.6	12	0.64	not churn	0.6	0.6	*0.4*	0.0000
C13	55	3,100	0.63	Churn	0.4	0.7	13	0.63	churn	0.6	0.7	*0.3*	0.0600
C14	54	3,900	0.64	not churn	0.467	0.7	14	0.59	not churn	0.533	0.7	*0.3*	0.0000
C15	48	6,000	0.67	not churn	0.467	0.7	15	0.56	not churn	0.467	0.7	*0.3*	0.0000
C16	29	5,800	0.7	Churn	0.533	0.8	16	0.51	churn	0.467	0.8	*0.2*	0.0467
C17	52	5,500	0.72	not churn	0.6	0.8	17	0.48	not churn	0.4	0.8	*0.2*	0.0000
C18	52	4,200	0.73	not churn	0.667	0.8	18	0.41	not churn	0.333	0.8	*0.2*	0.0000
C19	33	3,300	0.78	not churn	0.733	0.8	19	0.4	churn	0.267	0.8	*0.2*	0.0000
C20	31	2,800	0.81	Churn	0.733	0.9	20	0.35	not churn	0.267	0.9	*0.1*	0.0267
C21	44	5,100	0.82	not churn	0.8	0.9	21	0.33	churn	0.2	0.9	*0.1*	0.0000
C22	30	2,200	0.83	Churn	0.8	1	22	0.25	churn	0.2	1	*0*	0.0200
C23	69	4,000	0.87	not churn	0.867	1	23	0.2	not churn	0.133	1	*0*	0.0000
C24	60	7,000	0.92	not churn	0.933	1	24	0.11	churn	0.067	1	*0*	0.0000
C25	71	3,500	0.96	not churn	1	1	25	0.06	churn	0	1	*0*	0

Table 8. Fictional data set for performance measures demonstration

From Table 8, assuming a cut-off of 0.51 based on the predicted value on column 4, the value for the sensitivity is 73.33% and specificity is 60%, which would be obtained as well with the application of the formulas. In this case:

$$\text{Sensitivity} = \frac{11}{11+4} = 0.7333\ (*\ 100) = 73.33\%$$

$$\text{Specificity} = \frac{6}{6+4} = 0.60\ (*\ 100) = 60.00\%$$

$$\text{CA} = \frac{11+6}{11+4+6+4} = 0.68\ (*\ 100) = 68.00\%$$

This example is only for illustration purposes. To obtain these performance measurements in this research, the Kolmogorov-Smirnov statistic will be used as the base for cut-off, which will be further explained in subsection 4.6.2.

4.6.2. The Kolmogorov-Smirnov Statistic

To analyse the performance of the models in this research, CA, sensitivity and specificity will be measured on training and test sets, assuming the Kolmogorov-Smirnov statistic as the base for cut-off (Thomas *et al.*, 2002; Wilkie, 2004; Thomas, 2007).

The Kolmogorov-Smirnov (KS) statistic is defined as the maximum distance between the cumulative score distributions for the churners, $C_C(s)$, and non-churners, $C_{NC}(s)$, as follows:

$$KS = \max_s | C_C(s) - C_{NC}(s)|$$

This is illustrated in Figure 7. The classification threshold where the KS statistic is observed will then be used for the analysis. Using the data set from Table 8, it is possible to more clearly illustrate how the KS curve is defined. This is done in Figure 8.

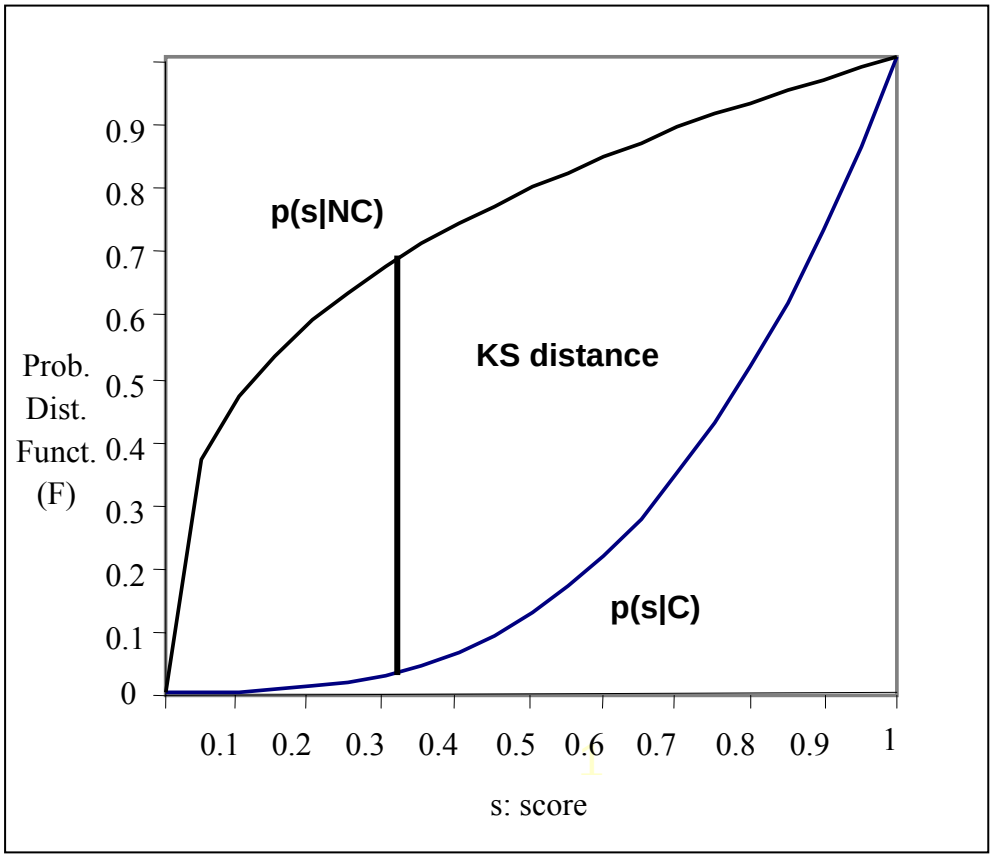

Figure 7. KS curve example for churners and non-churners

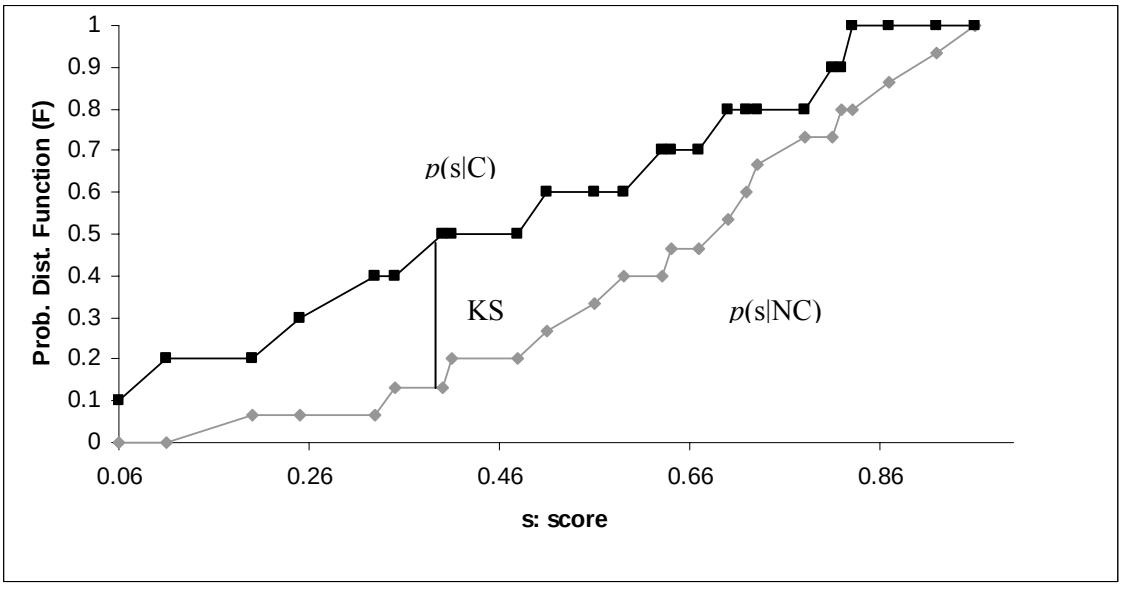

Figure 8. KS curve for data set in Table 8

The probability distribution functions of non-churners are obtained by dividing the amount of non-churners classified below a specific cut-off (represented by the scores) by the total amount of non-churners in the data set. For example, for the score of 0.35,

$p(s|NC) = 2/15 = 0.133$. The same process is used to obtain the probability distribution functions of churners.

The motivation for choosing the score s that yields this maximum distance cut-off is that it provides a good balance between sensitivity and specificity. This is emphasised by the relationship between the KS estimation and the elements of the ROC, as its value can be obtained through the sensitivity and specificity. The KS plots the functions of churners and non-churners against the score and the ROC curve plots the functions against each other, which corresponds to the sensitivity and 1-specificity, respectively. The KS can then be expressed as:

$$KS = \max_s | C_C(s) - C_{NC}(s)| = \max_s |\text{sensitivity} + \text{specificity} - 1|$$

Based on the ROC curve graph, the KS distance corresponds to the maximum vertical distance between the ROC curve and the diagonal (see Figure 9). Also, the graph in Figure 8 could be done using the sensitivity and 1-specificity in the place of the probability distribution functions; the graph would change the positions on the vertical axis, but the KS metric would be the same. This can be verified using the data set in Table 8.

4.6.3. The Area under the Receiver Operating Characteristic Curve

A Receiver Operating Characteristic Curve (ROC curve) is a graphical illustration of the sensitivity (true events) on the vertical axis versus "1-specificity" (false events) on the horizontal axis for all values of the classification threshold (cut-offs) (Van Erkel and Pattynama, 1998; Braga and Oliveira, 2003; Witten and Frank, 2005).

As an example, examining the probability of churn (churn = 1), the vertical axis would represent the percentage of true churners predicted as churners, and the horizontal axis would represent the percentage of non-churners predicted as churners. This is illustrated in Figure 9.

When using the ROC curve as a measure of performance, usually the area under the curve is measured (AUC), which would represent the behaviour of a classifier independently of the cut-off chosen or misclassification cost. For the model to be a good

model, the ROC curve needs to be above the baseline (linear line). This baseline (which represents a value of 50%) represents that any model that presents a performance below its values would perform worse than a random selection on the data set.

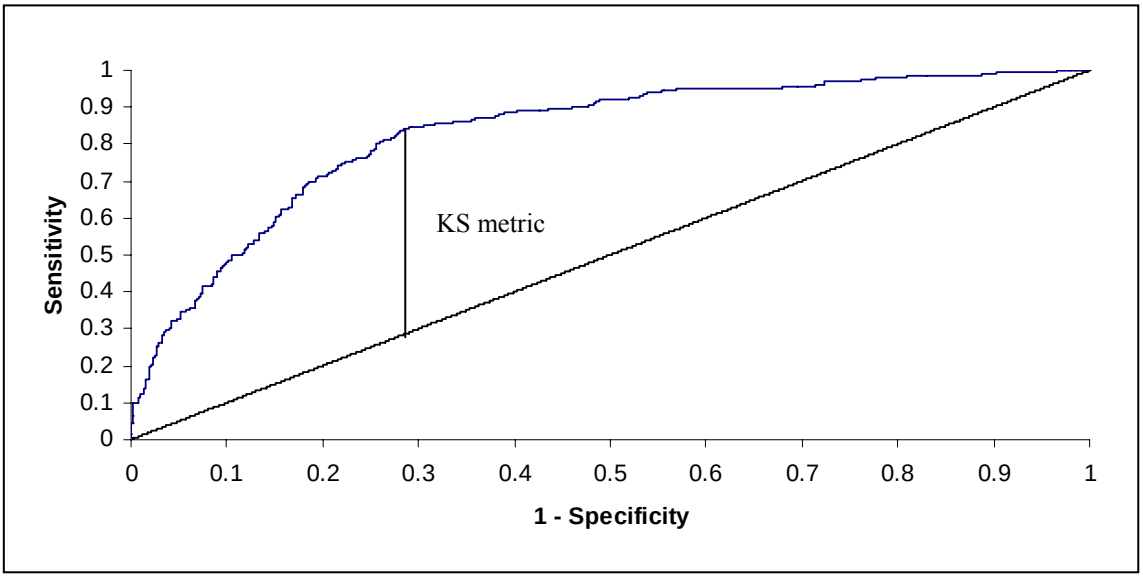

Figure 9. Sample ROC curve

The area under the ROC curve (AUC) was also used as a performance measure in this research, as it measures the ranking power regardless of the cut-off established (Hanley and McNeil, 1982; Hanley and McNeil, 1983; Van Erkel and Pattynama, 1998). It provides an estimate of the probability that a randomly chosen positive instance is correctly rated higher than a randomly selected negative instance.

Using the data set from Table 8, it is possible to plot the ROC curve, using the values of sensitivity and 1-specificity. This is illustrated in Figure 10.

Note that this graph does not present a smooth shape, as there are only a few points defined. Again, this was done only for illustration purposes. In this graph, as the classifier focussed on predicting the probability of not churning, the vertical axis would represent the percentage of non-churners predicted as non-churners, and the horizontal axis would represent the percentage of churners predicted as non-churners.

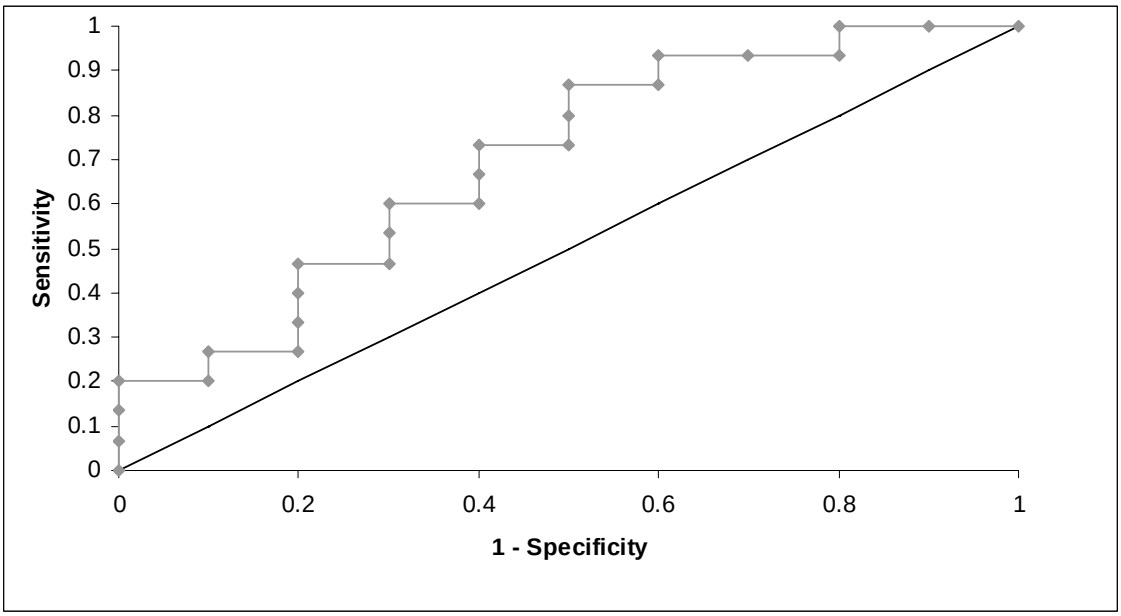

Figure 10. Sample ROC curve for data set in Table 8

In this research, the AUC is obtained by combining the ROC increments, which adds all the points of the ROC curve (represented by the sensitivity versus 1-specificity for each one of the records examined). It uses the trapezoidal rule, by adding the areas of inscribed trapezoids (Braga and Oliveira, 2003), even though it has been noted by Hanley and McNeil (1982) that this method systematically underestimates the true AUC.

As an example, by adding the ROC increments in Table 8, it is possible to obtain the value for the AUC, which in this case is 70%, representing a reasonably good performance for the model.

For this research, the ROC increments are obtained for logistic regression based on the output probabilities of the target variable, in this case, churn = 1. For decision trees, the confidence of the rules is used, i.e. the percentage of churners in the leaf node. These confidences are then applied to training and test sets to obtain the individual probability of each customer.

For the KNN model, the probability estimate is obtained as the number of churning neighbours divided by k, for each one of the customers. For the NN model, logistic transfer functions are used in the output layer, and these outputs are interpreted as the class probabilities.

The probabilities extracted from the four models are used not only to calculate the AUC, but also to define the other performance measurements (CA, sensitivity and specificity).

When comparing the models' performance, the DeLong *et al.* (1988) test is used to compare the AUCs. They suggested a nonparametric approach where the covariance matrix is estimated using the theory on generalised U-statistics. After some mathematical calculation, they arrived at the following test statistic:

$$(\hat{\theta}-\theta)c^T[cSc^T]^{-1}c(\hat{\theta}-\theta)^T,$$

which has a chi-square distribution with degrees of freedom equal to the rank of cSc^T with $\hat{\theta}$ the vector of the AUC estimates, S the estimated covariance matrix and c a vector of coefficients such that $c\theta^T$ represents the desired contrast. For more details, please refer to DeLong *et al.* (1988).

4.6.4. The Lift Chart

As explained in Chapter 3, lift chart arranges customers into deciles based on their predicted probability of response, and then plots the actual percentage of respondents. It makes it possible to visualise the effectiveness of the model under investigation.

For example, the lift chart in Figure 11 shows that in the top 10%, 85% of customers are considered churners. The baseline rate (in this case, 15%) is used for comparison purposes. The baseline is an estimate of the percentage of churners that would be expected if it was taken from a random sample.

Using the data set in Table 8, the lift chart in Figure 12 shows that in the top 20%, 80% of customers are classified correctly as non-churners. The baseline rate (in this case, 60%) is used for comparison purposes and it indicates the total amount of non-churners present in the data set.

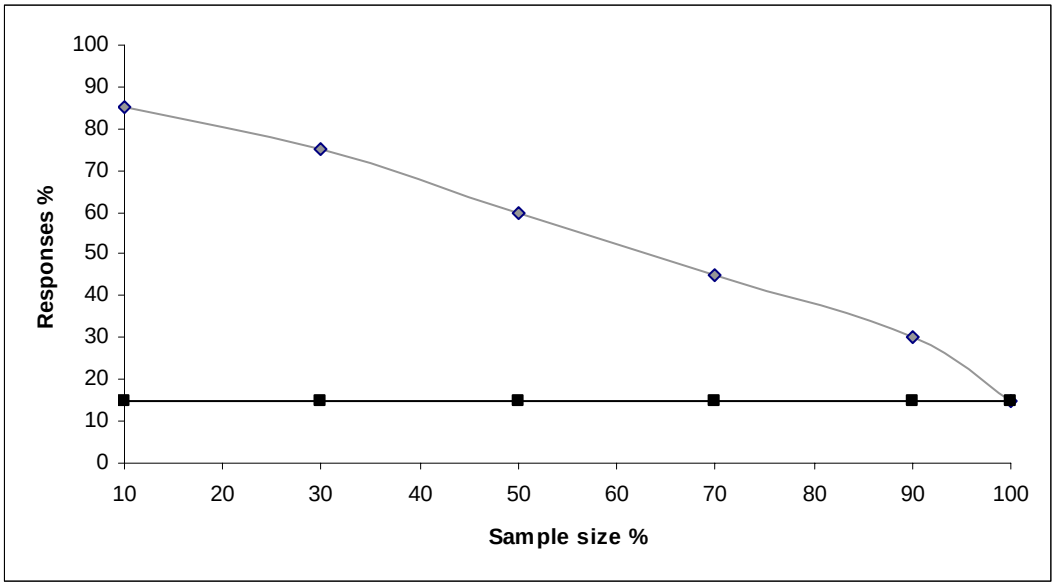

Figure 11. Lift chart example

To calculate the success proportions for the lift chart, the sample is sorted in descending order of predictive probability of non-churners, then from the top of the list, separated into samples. For each sample size, which in the case of Figure 12 were 20%, 40%, 60%, 80% and 100%, the lift is obtained by counting the number of positive instances (in this case non-churners) in each sample and dividing by the sample size. For example, for the top 20% customers, four out of five were non-churners, giving the 80% success rate. In this case, the customers were not arranged into deciles as it would create 2.5 customers to each decile.

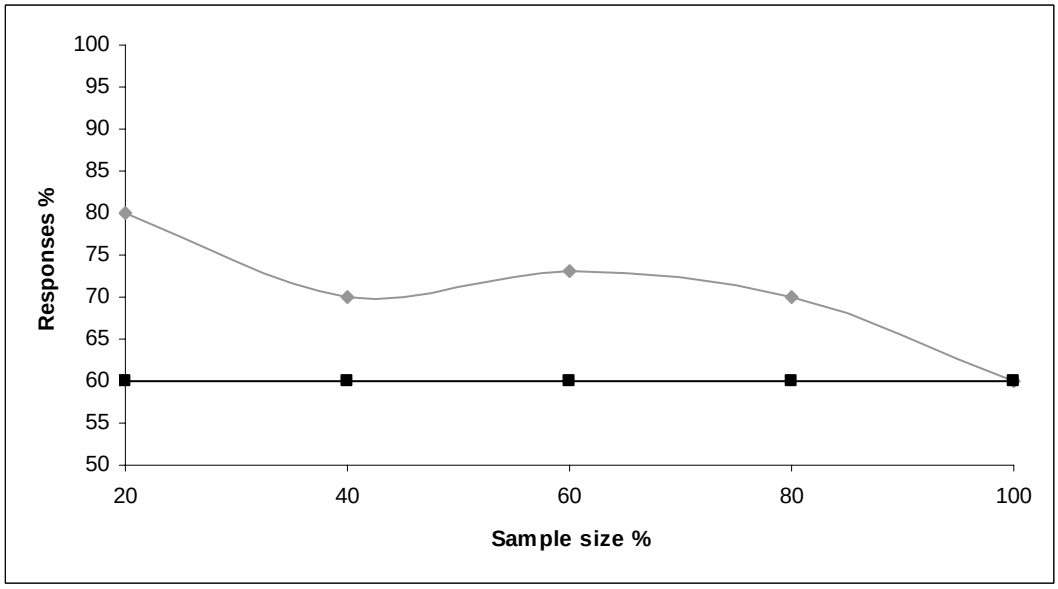

Figure 12. Lift chart for data set in Table 8

The lift chart can be represented in a cumulative or non-cumulative way. For the cumulative chart for example, which is the case in Figure 12, the vertical axis is the cumulative number of responders in each respective 20% of the sample (or decile for the default lift chart).

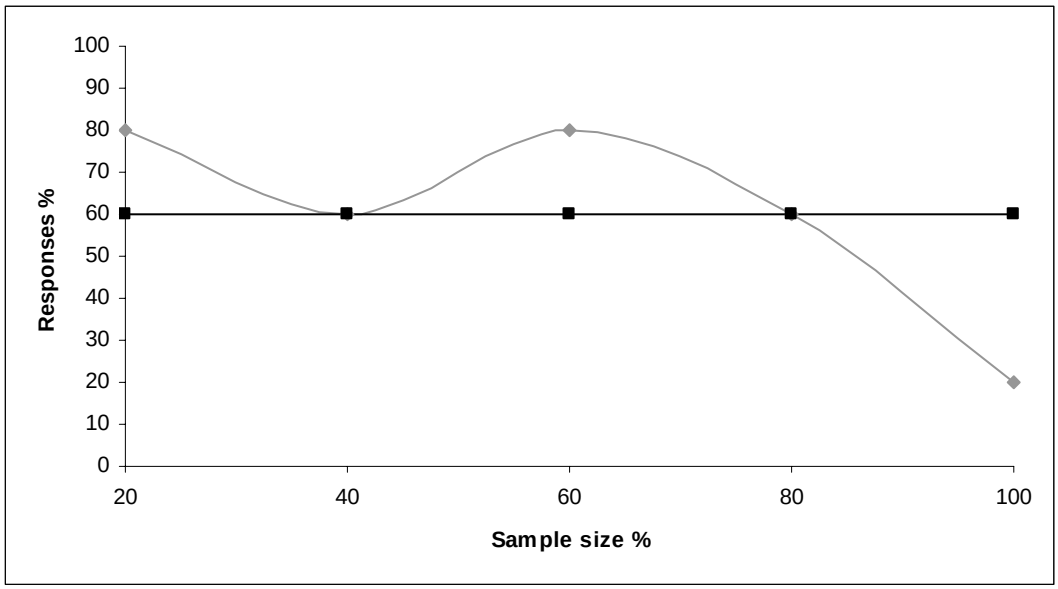

Figure 13. Non-cumulative lift chart from Figure 12

For the non-cumulative chart, the vertical axis is the number of responders in each respective decile (on in this case 20%), making it possible to verify the effectiveness of the model for each level of the score. The non-cumulative lift chart is demonstrated in Figure 13. This chart reveals that the predictive power of the model drops after the top 20% of scores, but recovers at 60% of the scores. By the eighth decile the model performs the same as a random selection.

In conclusion, Table 9 shows some performance metrics used in the churn literature, also specifying the data mining techniques applied. It shows that the lift chart is a common chosen technique, as well as AUC and CA (also called PCC – percentage of correctly classified). This research will report CA, sensitivity, specificity and AUC. The KS statistic will be used for setting the cut-off for CA, sensitivity and specificity, and the DeLong *et al.* test will be used to compare the AUC performance for the models.

Source	Method	Performance Metrics
Masand *et al.* (1999)	Simple regression, nearest neighbour, decision trees and neural networks	Lift chart.
Hwang *et al.* (2004)	Neural networks, decision tree and Logistic regression	Misclassification rate and lift chart.
Kim *et al.* (2005)	SVM and Back-propagation NN	Classification accuracy.
Zhao *et al.* (2005)	One-class SVM, NN, DT and naïve bayes	Lift curve and classification accuracy.
Buckinx and Van den Poel (2005)	Logistic regression, ARD neural network and random forests	PCC and AUC.
Kim *et al.* (2006)	Neural networks, decision tree and Logistic regression	Misclassification rate and lift chart.
Glady *et al.* (2006)	Logistic regression, decision trees and neural networks, AdaCost and cost-sensitive decision tree	PCC, AUC and two new measures defined by them: AUPROC and cumulative profit percentage.
Coussement and Van den Poel (2008a)	Support vector machines, Logistic regression, and random forests	PCC, AUC (using DeLong *et al.* test) and top-decile lift.
Burez and Van den Poel (2007)	Logistic regression, Markov chains and random forests	PCC, AUC and lift chart.

Table 9. Performance metrics in the churn literature

4.7. Data Sets Description

Three data sets were selected to compare the performance of various churn prediction approaches. The characteristics of all data sets are presented in Table 10.

Data Set	# Variables	# Obs	Training	Test	Churn Rate (%)
Telecom1	21	5,000	3,350	1,650	14.1
Telecom2	22	15,000	10,000	5,000	1.8
Telecom3	8	12,499	8,332	4,167	39.32

Table 10. Characteristics of churn data sets

The data set Telecom1 is a churn data set which is publicly available, obtained from the KDD library (http://www.datalab.uci.edu/data/mldb-sgi/data/ accessed 03 July 2006). The data set contains 5,000 observations and twenty-one variables, divided in discrete and continuous variables, with a target variable, churn. For a description of the full data set, please refer to Larose (2005).

The data set Telecom2 is a telecom data set used in the churn tournament 2003, organised by Duke University. The data set and descriptive documentation were obtained from the Centre for Customer Relationship Management at Duke University (http://www.fuqua.duke.edu/centers/ccrm/datasets/download.html accessed 10 May 2007). For this analysis, two distinctive samples of 10,000 and 5,000 customers were selected from the calibration and score data set, respectively, with a total of 171 predictor variables that were reduced to twenty-one predictive variables plus the target variable, churn.

The data set Telecom3 is a telecom data set supplied by a wireless service provider, also obtained from the Centre for Customer Relationship Management at Duke University. For this analysis, the last monthly position of the subscriber was selected, indicating if they churned or not. Based on this, a sample of 12,499 was extracted, which was divided into training and test set, and presented initially twelve variables, including the variable churn and a unique identification field. After the pre-processing stage, a final amount of eight variables were kept.

Regarding Telecom1 and Telecom3, the same proportion of churners and non-churners were applied to both training and test set, as the data was presented as a whole and no churn rate was specified. In this case, it was defined based on the proportion of churners in the data sets (14.1% for Telecom1 and 39.32% for Telecom3).

For Telecom2, the training set was stratified with the same proportion of true and false. The proportion of churners was oversampled in order to give the predictive model a better capability of detecting discriminating patterns. This oversampling did not make copies of existing data, but real distinct churn records were selected from the population. The test set was not oversampled to provide a more realistic test set, according to a monthly churn rate of 1.8%. Also, in order to obtain a reasonable amount of variables suitable for modelling and analysis, variable selection was performed on the training set, and the resulting variables were also kept in the test set.

To use these three data sets in the calculations it was necessary to first understand and investigate the data available. The pre-processing (Rud, 2001; Larose, 2005) stage was essential to eliminate some mistakes and inconsistent data, as well as to guarantee the reliability of the data. Analysis of correlations (Moore and McCabe, 2006), missing values (Witten and Frank, 2005; Larose, 2005) and outliers (Moore and McCabe, 2006; Witten and Frank, 2005; Larose, 2005) were also essential to explore the data. Any variables presenting too many unknown or inadequate values were eliminated from the analysis.

A more detailed description about Telecom1 and Telecom2 evaluation and pre-processing will be presented in Chapter 6, where it is necessary to understand the whole process to proceed with the analysis. The same type of selection and understanding was applied to Telecom3.

4.8. Data Analysis and Evaluation

To select the best model generated from the KNN algorithm, the data sets (Telecom1, Telecom2 and Telecom3) were split into training, validation and test sets and they were evaluated for five different values of k (1, 5, 10, 50 and 100) (see Table 11). The validation sets were obtained by dividing the original training sets in half (see training sets values in Table 10), obtaining training and validation sets with the same proportion of churners as the original training sets. The training sets were used to build the models and the validation sets were used for model selection. The models that had the highest performance on the validation sets were used as the optimal k to report the test sets performance. The test sets were then used to evaluate the performance of the selected models.

Data Set	Metric	K=1	K=5	K=10	K=50	K=100
Telecom1	CA	80.26	82.30	69.46	84.52	73.60
	Sensitivity	32.20	39.41	59.32	47.46	64.41
	Specificity	88.19	89.38	71.14	90.64	75.12
	AUC	60.20	65.94	69.97	**73.44**	**73.69**
Telecom2	CA	53.90	55.32	56.84	59.96	60.12
	Sensitivity	54.04	53.96	46.8	54.76	57.00
	Specificity	53.76	56.68	66.88	64.36	63.24
	AUC	53.90	57.14	58.49	**61.93**	**61.96**
Telecom3	CA	72.08	74.20	76.64	76.86	74.82
	Sensitivity	47.25	58.42	52.62	54.33	58.42
	Specificity	88.17	84.41	92.21	91.46	85.44
	AUC	67.71	72.45	74.82	75.83	**76.81**

Table 11. Performance on validation set - selection for KNN model

As a result, in order to select the best value of k, the Delong *et al.* test was used. For Telecom1, the resulting comparison is shown in Figure 14. Based on that, it can be seen that there was significant difference between the AUC with k=1, k=5 and k=10, and between them and k=50 and k=100, assuming a 95% confidence level. The test also showed that there was no significant difference between k=50 and k=100. As these present the highest values for AUC, it would be at the researchers' discretion to choose which one to use. In this case, k=50 was chosen.

A similar analysis was done for Telecom2 and Telecom3. For Telecom2, the comparison demonstrated that there was significant difference between the AUC with k=1, k=5, k=10 and k=50, and between them and k=100, assuming a 95% confidence level. It also showed that there was no significant difference between k=50 and k=100. As these present the highest values for AUC, it would also be at the researchers' discretion to choose which one to use. As with Telecom1, k=50 was chosen. A summary result can be seen in Figure 15.

With regard to Telecom3, the resulting analysis can be observed in Figure 16. All the models presented significant differences between themselves. In this case, the model that had the biggest AUC was chosen, k=100.

AUC and DeLong Comparison at 95% Confidence Intervals

	AUC
K=1	0.6020
K=5	0.6594
K=10	0.6999
K=50	0.7344
K=100	0.7369

Contrast Coefficients

	bp1	bp2	bp3	bp4	bp5
Row1	1	-1	0	0	0
Row2	1	0	-1	0	0
Row3	1	0	0	-1	0
Row4	1	0	0	0	-1
Row5	0	1	-1	0	0
Row6	0	1	0	-1	0
Row7	0	1	0	0	-1
Row8	0	0	1	-1	0
Row9	0	0	1	0	-1
Row10	0	0	0	1	-1

Tests and 95% Confidence Intervals for Contrast Rows

	Estimate	Std Error	Chi-square	P-value
Row1	-0.0575	0.0161	12.7602	0.0004
Row2	-0.0979	0.0173	31.9580	<.0001
Row3	-0.1324	0.0185	51.2008	<.0001
Row4	-0.1349	0.0191	49.9885	<.0001
Row5	-0.0404	0.0116	12.2115	0.0005
Row6	-0.0749	0.0161	21.5409	<.0001
Row7	-0.0774	0.0170	20.8577	<.0001
Row8	-0.0345	0.0158	4.7663	0.0290
Row9	-0.0370	0.0177	4.3900	0.0362
Row10	-0.0025	0.0093	0.0721	0.7883

Overall P-value <.0001

Figure 14. DeLong evaluation of validation set performance for KNN model - Telecom1

AUC and DeLong Comparison at 95% Confidence Intervals

	AUC
K=1	0.5390
K=5	0.5714
K=10	0.5849
K=50	0.6193
K=100	0.6196

Tests and 95% Confidence Intervals for Contrast Rows

	Estimate	Std Error	Chi-square	P-value
Row1	-0.0324	0.0074	19.3690	<.0001
Row2	-0.0459	0.0081	32.0503	<.0001
Row3	-0.0803	0.0089	81.3523	<.0001
Row4	-0.0806	0.0090	79.9729	<.0001
Row5	-0.0135	0.0050	7.2459	0.0071
Row6	-0.0479	0.0075	40.7356	<.0001
Row7	-0.0482	0.0078	37.8681	<.0001
Row8	-0.0344	0.0063	29.7358	<.0001
Row9	-0.0347	0.0068	26.0317	<.0001
Row10	-0.0003	0.0031	0.0069	0.9340

Overall P-value <.0001

Figure 15. DeLong evaluation of validation set performance for KNN model - Telecom2

AUC and DeLong Comparison at 95% Confidence Intervals

	AUC
K=1	0.6771
K=5	0.7245
K=10	0.7483
K=50	0.7583
K=100	0.7681

Tests and 95% Confidence Intervals for Contrast Rows

	Estimate	Std Error	Chi-square	P-value
Row1	-0.0474	0.0040	138.4179	<.0001
Row2	-0.0711	0.0050	206.2211	<.0001
Row3	-0.0812	0.0063	163.8043	<.0001
Row4	-0.0910	0.0061	222.0347	<.0001
Row5	-0.0238	0.0030	63.1083	<.0001
Row6	-0.0338	0.0054	39.6680	<.0001
Row7	-0.0436	0.0052	71.5475	<.0001
Row8	-0.0100	0.0049	4.2277	0.0398
Row9	-0.0199	0.0048	17.0670	<.0001
Row10	-0.0099	0.0046	4.5515	0.0329

Overall P-value <.0001

Figure 16. DeLong evaluation of validation set performance for KNN model - Telecom3

A similar analysis was done in order to define the best selection for the NN model. The model was estimated for each one of the data sets using a different number of hidden neurons. The lower the number of hidden neurons the more linear the model is. Using a similar approach to the one adopted for the KNN model, the NN model was estimated with 2, 4, 6 and 8 hidden neurons (Table 12).

Data Set	Metric	2 Hidden Neurons (HN)	4HN	6HN	8HN
Telecom1	CA	81.76	81.04	84.94	90.22
	Sensitivity	74.58	77.54	78.39	77.97
	Specificity	82.95	81.62	86.02	92.24
	AUC	83.88	84.32	87.66	**89.31**
Telecom2	CA	62.52	61.90	61.62	61.62
	Sensitivity	66.64	77.40	63.32	68.36
	Specificity	58.40	46.40	59.92	55.20
	AUC	**65.86**	65.96	66.33	65.44
Telecom3	CA	77.34	58.14	75.66	77.24
	Sensitivity	71.73	95.05	48.17	53.85
	Specificity	80.97	34.22	93.47	92.41
	AUC	**84.87**	70.61	77.55	81.06

Table 12. Performance on validation set - selection for NN model

The resulting models where then compared using the DeLong *et al.* test to evaluate the performance on the validation set. The same approach adopted for the KNN model was used for the NN model selection, based on the model significance in comparison with the other resulting NN models. It was then concluded that for Telecom1, the best model would be the one with 8 hidden neurons; for Telecom2 and Telecom3, the best models were the ones with 2 hidden neurons. This indicated that Telecom2 and Telecom3 are more linear models than Telecom1. After scoring the results of the three data sets, the performance measures are shown in Table 13. It was then possible to evaluate their performance against each other.

Data Set	Metric	LR (%)	DTree (%)	KNN (%)	NN (%)
Telecom1 Training	CA	76.51	89.44	83.62	87.94
	Sensitivity	74.36	71.61	50.85	81.36
	Specificity	76.87	92.38	89.03	89.03
	AUC	81.17	84.21	76.67	90.60
Telecom1 Test	CA	73.65	90.64	82.66	86.01
	Sensitivity	83.83	72.34	50.21	82.55
	Specificity	71.98	93.64	88.05	86.58
	AUC	82.91	84.94	71.96	90.79
Telecom2 Training	CA	62.43	60.58	61.68	63.78
	Sensitivity	68.11	68.20	70.04	64.20
	Specificity	56.77	52.96	53.32	63.36
	AUC	66.28	64.20	66.45	67.24
Telecom2 Test	CA	51.23	52.30	72.96	65.98
	Sensitivity	77.78	74.44	51.11	63.33
	Specificity	50.75	51.89	73.36	66.03
	AUC	67.64	64.25	63.52	68.32
Telecom3 Training	CA	67.52	66.59	69.13	69.97
	Sensitivity	59.86	83.82	49.33	69.96
	Specificity	72.49	55.42	81.96	69.98
	AUC	73.47	76.16	72.60	78.25
Telecom3 Test	CA	62.47	70.07	64.17	68.03
	Sensitivity	76.01	57.08	63.31	71.37
	Specificity	53.70	78.50	64.73	65.88
	AUC	70.62	71.30	68.81	76.86

Table 13. Performance measures for the data mining techniques

Using the Delong *et al.* test, it was possible to identify if there was any difference between the AUCs and if any of the models would perform better. For Telecom1, the DeLong evaluation for the test set can be seen in Figure 17. Some significant conclusions can be drawn from that.

Figure 17 shows that there were significant differences between the KNN model and all the other models, and also between the NN model and all the other models. On the other hand, there was no significant difference between the decision tree and logistic regression models. Based on these results, the model that offered the best performance for Telecom1 was the one generated by the NN algorithm; however, the logistic regression and decision tree model also performed relatively well.

AUC and DeLong Comparison at 95% Confidence Intervals

	AUC
LR	0.8291
DTree	0.8494
KNN	0.7196
NN	0.9079

Contrast Coefficients

	LR	DTree	KNN	NN
Row1	1	-1	0	0
Row2	1	0	-1	0
Row3	1	0	0	-1
Row4	0	1	-1	0
Row5	0	1	0	-1
Row6	0	0	1	-1

Tests and 95% Confidence Intervals for Contrast Rows

	Estimate	Std Error	Chi-square	P-value
Row1	-0.0203	0.0178	1.2935	0.2554
Row2	0.1095	0.0227	23.2583	<.0001
Row3	-0.0788	0.0112	49.5384	<.0001
Row4	0.1298	0.0262	24.5771	<.0001
Row5	-0.0585	0.0157	13.8214	0.0002
Row6	-0.1883	0.0205	84.4424	<.0001

Overall P-value <.0001

Figure 17. DeLong evaluation of data mining techniques for Telecom1

For Telecom2, the DeLong evaluation for the test set can be seen in Figure 18. It demonstrated that with an overall p-value of 0.1936, there was no significant difference between the performances of any of the models. In this case, any of the models could be chosen for analysis. As NN and logistic regression presented the highest values for the AUC, they could be chosen for the analysis. What would influence the choice in this case, would be the usability of the final results. If the overall prediction is the main

element expected from the model, NN would be chosen. If facility of interpretability is the key element, logistic regression would be a better choice.

AUC and DeLong Comparison at 95% Confidence Intervals

	AUC
LR	0.6764
DTree	0.6425
KNN	0.6352
NN	0.6832

Contrast Coefficients

	LR	DTree	KNN	NN
Row1	1	-1	0	0
Row2	1	0	-1	0
Row3	1	0	0	-1
Row4	0	1	-1	0
Row5	0	1	0	-1
Row6	0	0	1	-1

Tests and 95% Confidence Intervals for Contrast Rows

	Estimate	Std Error	Chi-square	P-value
Row1	0.0339	0.0184	3.4031	0.0651
Row2	0.0412	0.0250	2.7140	0.0995
Row3	-0.0068	0.0108	0.3899	0.5323
Row4	0.0073	0.0245	0.0892	0.7652
Row5	-0.0407	0.0211	3.7359	0.0533
Row6	-0.0480	0.0282	2.9021	0.0885

Overall P-value = 0.1936

Figure 18. DeLong evaluation of data mining techniques for Telecom2

With regard to Telecom3, the DeLong evaluation for the test set is illustrated in Figure 19. Based on this, the model that offers the best performance is NN. All the other models differ significantly from the NN results, but they assume a very similar pattern of results. The logistic regression model is not significantly different from the decision tree and KNN models. However, there is a significant difference between the decision tree and KNN models. In this case, it would be interesting to also evaluate the results from the best model of these three, especially if the interpretability of the variables and results would be something to be taken into consideration. In this case, the decision tree would give the best performance, even if it is not significantly different from the logistic regression model.

The best AUCs were then highlighted in Table 14, appearing in bold and underlined. From this table, it can be conclude that NN achieved very good performances compared to the other techniques under examination. However, logistic regression also achieved reasonably good performance which indicates that the data sets used are only weakly

nonlinear. Decision trees also presented a relatively good performance, with easy to interpret results. The KNN models in general presented the worst performance, but their performances were only marginally below some of the other techniques.

AUC and DeLong Comparison at 95% Confidence Intervals

	AUC
LR	0.7063
DTree	0.7130
KNN	0.6881
NN	0.7686

Contrast Coefficients

	LR	DTree	KNN	NN
Row1	1	-1	0	0
Row2	1	0	-1	0
Row3	1	0	0	-1
Row4	0	1	-1	0
Row5	0	1	0	-1
Row6	0	0	1	-1

Tests and 95% Confidence Intervals for Contrast Rows

	Estimate	Std Error	Chi-square	P-value
Row1	-0.0067	0.0090	0.5624	0.4533
Row2	0.0182	0.0097	3.5289	0.0603
Row3	-0.0623	0.0074	70.6629	<.0001
Row4	0.0249	0.0066	14.2150	0.0002
Row5	-0.0556	0.0044	157.1465	<.0001
Row6	-0.0805	0.0061	172.6242	<.0001

Overall P-value <.0001

Figure 19. DeLong evaluation of data mining techniques for Telecom3

Data Set	Metric	LR (%)	DTree (%)	KNN (%)	NN (%)
Telecom1	CA	73.65	90.64	82.66	83.73
	Sensitivity	83.83	72.34	50.21	86.38
	Specificity	71.98	93.64	88.05	83.30
	AUC	***82.91***	***84.94***	71.96	**89.43**
Telecom2	CA	51.23	52.30	72.96	65.98
	Sensitivity	77.78	74.44	51.11	63.33
	Specificity	50.75	51.89	73.36	66.03
	AUC	**67.64**	64.25	63.52	**68.32**
Telecom3	CA	62.47	70.07	64.17	68.03
	Sensitivity	76.01	57.08	63.31	71.37
	Specificity	53.70	78.50	64.73	65.88
	AUC	***70.62***	***71.30***	68.81	**76.86**

Table 14. Indication of best performing technique on the test set - based on AUC

In summary, the choice of technique will depend of the data used and the kind of application it will have. It is recommended to compare the results of different techniques and then, choose the most applicable. Table 11 and Table 12 demonstrated the results of the validation sets for the KNN and NN model selection, respectively, whereas Table 14 demonstrated the performance metrics on the test sets for the four data mining techniques, in order to evaluate which of the final models has the best performance.

4.9. Improving Churn Models from the Data Perspective

One important aspect is that, independent of the data mining technique used, a good way to augment the performance of a churn model is by improving data quality. Data quality is related to terms such as accuracy, completeness, consistency and relevance when selecting and preparing the data for analysis (Rajagopalan and Isken, 2001; Bert-Équille, 2007). Different methods can then be used to assure the quality of the data under investigation.

With consideration to the essential need of maintaining data consistency, this research applies the treatment of missing values and outliers as ways of improving data quality. Missing values could be assigned using the mean or median for the continuous attributes and the most frequent category for the categorical attributes. In this research, two of the data sets did not present any missing values. For Telecom2, as there were too many variables for analysis, the variables already presenting an excessive amount of missing values were eliminated from the study.

Checking for outliers is recommended as they can affect the classification and predictive precision of the models. They could be treated, for example, by truncating these values and replacing then by the 99^{th} percentile. This approach was not used for these data sets, but it will be adopted in Chapter 8. For these data sets, outliers were checked for the continuous variables against the target variable “churn”, using box plots for analysis. In this case, the outliers encountered were not treated. For example, the variable “CustServ_Calls” presented values between 0 and 9, and the values from 4 to 9 classified as outliers were not considered a problem. In this research, it is appropriate to know about their existence and be able to identify and treat them, where necessary.

Another option would be, if applicable to the research under investigation, the improvement of data definition. The creation of ratio variables and derived variables represent examples of data definition. Derived variables can be useful in contexts where time spans of analysis need to be adapted, for example the transformation of a variable represented in months instead of weeks.

Ratio variables could be used to represent the information provided by two other variables, simplifying the analysis or reducing any problem caused by the existence of these original variables in the analysis, for example the existence of multicollinearity. An example of ratio variable would be the creation of a variable "charge per minute", which could be originated from the division of "total charge amount" and "total minutes usage". In this case, the last two variables could add bias to the analysis if used at the same time, and the ratio variable would properly represent the meaning of the two original variables. However, the use of ratios has to be properly controlled, as it can also generate inadequate values. These aspects are further explained in Chapter 6.

In this research, the data definition approach was not used because the idea was to explore the individual effects of variables, choosing the ones that would better influence the results.

Domain knowledge, which will be used in the data mining evaluation, is also applied in the pre-processing stage, but in a more intuitive manner. It is used to identify inconsistencies and interpret these values, and the appropriate actions then taken (e.g., removal or substitution). If the data used in the analysis is badly collected or incorrectly pre-processed, it increases the chances of incorrect results. As a result, data quality is essential in the data mining analysis and the reliability of the data is of distinctive importance.

Another important aspect is the use of data enrichment, for example, in the form of network based data. Web pages connected by hyperlinks, research papers connected by citations and social networks (such as LinkedIn and Facebook) are examples of network based data. They are important because distinct factors can influence the customer behaviour in different ways. In terms of social networks, a customer opinion can be transferred to many other friends at the same time, and this could influence their behaviour. The use of social networks is growing fast and many people could use the

information that circulates in their groups as a base for making their decisions. For example, if one customer complains about a specific service provider and disseminates their dissatisfaction in the network, this attitude could influence other customers to be aware of the problem and leave the company as well. It would also influence the decision of other customers that were planning on starting a relationship with that service provider.

The method to investigate this would be by using network based learning, in order to estimate the class for an instance based on the class of the neighbours. To achieve this, it is necessary to have non-relational classifiers, a relational model and collective inference. The non-relational classifiers are information specific for that customer, estimated with traditional methods and used in the relational learning and collective inference. The relational model uses the relations or links in the network. The collective inference determines how the unknown values are estimated together, influencing each other.

Macskassy and Provost (2003) suggest a relational neighbour classifier that predicts based on class labels of related neighbours, with no need to use learning or inherent attributes. They assume that some class labels are known and that linked instances have the propensity to belong to the same class. They provide a good overview related to relational learning models, emphasising that their model performed well in comparison to other existing methods, assuming the characteristics above. For example, they used an internet movie database to predict movie box-office receipts (whether a movie would be a blockbuster), using relations such as “has the same actor”, “has the same director” and “movies from the same company”.

In the example above, they came to the conclusion that it is possible to increase the performance of the model by considering more than one relation when they are available. However, they emphasise that the relational structure of the data makes it difficult to separate into test and training sets without losing some of the relational information. In this analysis, they have used the whole data for prediction, which could indicate that their results might be over-optimistic.

A network based example in the churn context could involve the relational data linking one customer to another. The more calls customer X makes to customer Y, the higher

the weight linking these two customers, in contrast with a smaller weight with customer Z to whom customer X made less calls. This relational informational could then be used in the mining process. Based on the data sets available for this research, this relational analysis was not performed. If network data was available, this analysis could bring important factors that would influence the probability of a customer churning and the model's performance.

4.10. Conclusions

Logistic regression, decision trees, k-nearest neighbours and neural networks are classification techniques that can be used to predict the probability of the categorical dependent variable, in this case, churn. Also, decision trees and logistic regression are well-known techniques for churn and general prediction problems; neural networks usually present good performance and their use is growing; finally k-nearest neighbours are another technique that is easy to implement.

For further investigation of churn in this research, the data mining techniques chosen were logistic regression and decision trees. They were chosen based on their predictive characteristics and facility of interpretation and understanding, especially related to the type of variables investigated. Neural networks could also be used, especially as they presented the best results, but the complexity added would make them difficult to interpret. K-nearest neighbours presented the worst performance measurement and can require large computing power, since for classifying an instance its distance to all the other instances in the training set has to be calculated. Additionally, when many irrelevant attributes are present, the classification performance may become worse when observations have distant values for these attributes. As a result, this technique is excluded from further investigation.

In conclusion, logistic regression and decision trees are commonly chosen techniques for churn evaluation. As a result, this motivated their further investigation, in order to make their results more compliant with the knowledge in the company. This would then facilitate their acceptability in the decision making process.

Chapter 5. Importance of Domain Knowledge

5.1. Introduction

One of the main points of this research is to explore how data mining techniques can assist and help customer predictions, making the final results more understandable. The idea is to make the resulting models compliant with domain knowledge, in a way that would facilitate their interpretation and usability by the company.

With this in mind, it is necessary to understand and explore how domain knowledge is applied in data evaluation and analysis, and how it will influence the final results.

Subsequently, a methodology is defined to incorporate domain knowledge into the data mining analysis, using an intuitive and practical approach that facilitates the interpretability of the models.

5.2. The Domain Knowledge Concept

Domain knowledge relates to information about a specific domain or data that is collected from previous systems or documentation, or elicited from a domain expert (Anand *et al.*, 1995). It can come not only from consolidated information that is transferred to the data analysis by business experts, but can also be generated from common sense, a knowledge that is based on intuition and experience, without having pre-defined elements.

It can be found and used in many general tasks within data analysis, even when it is not properly defined. Data preparation, final analysis and problem definition are some of the key examples of domain knowledge being applied. It can also be used by inserting rules in the data analysis in order to guide and define the way to proceed.

Figure 20 illustrates different ways of incorporating domain knowledge. This figure does not exhaust all the possibilities of domain knowledge. The purpose is to illustrate some of the situations where its use would be appropriate.

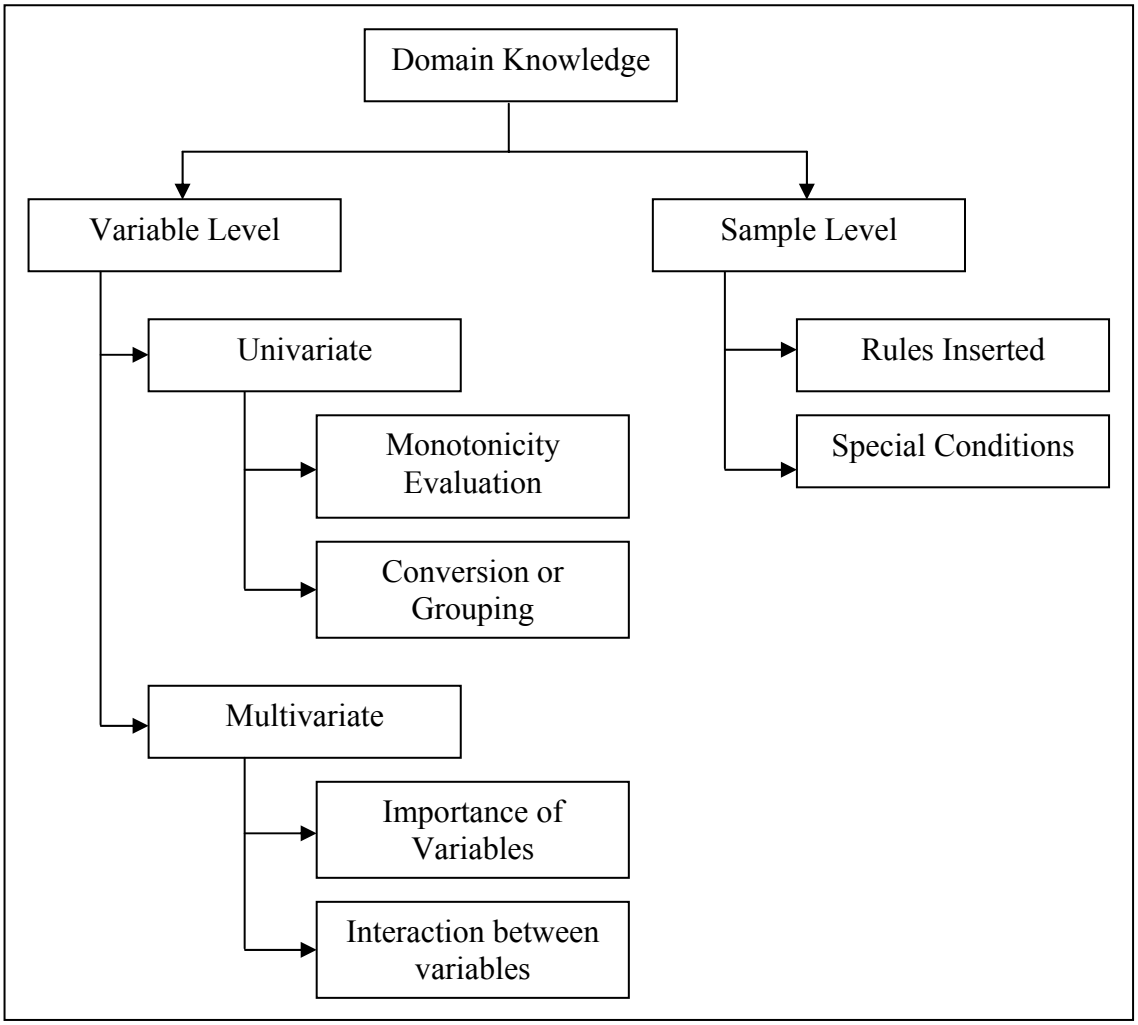

Figure 20. Possibilities for domain knowledge incorporation

Within this figure, domain knowledge at the variable level is related to the values associated with a variable, and how this analysis is carried out. In the case of univariate analysis, this refers to the effects and limitations imposed to one single variable. Two examples are shown below:

- The monotonicity evaluation assumes that a variable will present a unique influence on the target, increasing or decreasing the target value, based on the expected sign of that variable. The sign evaluation indicates a pattern of influence. This means, for example, that if the value of "charged amount to customer" increases, if there is no variation on the other variables (ceteris paribus), it should indicate an increase in the probability of churn. This monotonicity aspect will be further discussed in section 5.3.

- Conversion or grouping relates to categorical variables. Their values can be converted into numbers to proceed with a more intuitive sign evaluation, or they can be grouped into smaller sets to facilitate interpretability of results. For example, a variable "State" which contains fifty-one categorical values could be grouped into a few numeric values or a few categorical values, based on the probability of customer churn. This is related to the use of weight of evidence and coarse classification respectively, further discussed in Chapter 6.

In the multivariate analysis the effects are analysed taking into account several variables simultaneously:

- Depending on the scenario evaluated, different variables can have different levels of importance. In a churn evaluation for example, the amount of minutes used by a customer might be more important than the customer's age, and this should be taken into consideration. This degree of importance should be considered when defining which variables should be evaluated based on the monotonicity constraint.

  The bigger the impact the variable has on the analysis, the more effort should be spent on analysing its influence on the resulting model. For a telecom analysis for example, the amount of service calls a customer makes could result in the addition of a constraint indicating that the higher the number of service calls, the greater the probability of a customer churning. On the other hand, the variable related to the American State the customer lives may not necessarily require a constraint, as the model result could provide this information.

- With regards to the interaction between variables, the type of influence one variable can have on another must be evaluated; for example, an increase in amount charged by a telecom company can influence the minutes' usage of a customer. This analysis is important, for example, to evaluate the level of correlation between two or more variables. If two variables are highly correlated, multicollinearity can be introduced in the model, which could in turn affect the monotonicity evaluation. The multicollinearity problem is further discussed in Chapter 6.

When discussing domain knowledge at the sample level, two examples can be mentioned:

- Rules may be applied to limit the data selected for usage or to enforce a specific condition on the data. For example, a rule stating that all values in the variable "age" that are above a specified threshold will be reduced to the value of this threshold. In the case this threshold is established at seventy-five years old for the purpose of the analysis (it could have been any value based on the needs of the analysis), any value above this would be reduced to seventy-five. This example rule could be used to eliminate outliers. Rules like this one can be used in different parts of the data analysis, and in many cases are added as an intrinsic part of a knowledge algorithm.

- Special conditions have different purposes than the insertion of rules and should be used in particular situations. Domain knowledge is used to pre-define the selection of data for analysis, having in mind that the condition needs to be satisfied in the data selection and evaluation. For example, a rule could be added that only a customer with an international call plan will be evaluated. In this case, this rule is limiting the search space previously selected to the data mining process, eliminating from the analysis any customer that does not have the specific plan. These special conditions can be added and removed at any time, being used to obtain different results in the analysis of a data set.

In all the cases at variable and sample levels, the domain knowledge needs to be directly related to the area where it is being applied. The rules will be specific for the context under evaluation and non-transferable. They should be properly specified and their effects investigated, as they will affect the general direction of the analysis.

When evaluating the results from the data analysis, domain knowledge is often used with the purpose of helping to understand the results and to extract the information required. Most of the time, these results are not challenged; they are accepted as they are presented. However, as companies need to understand what the results mean and how better to use the outcomes, these results could be questioned and evaluated, with domain knowledge being essential in this process.

5.3. Domain Knowledge in the Literature

Domain knowledge can be used for discovery of meaningful information (Djoko *et al.*, 1997), which can then be employed as a guide in the discovery process. It can be useful in different ways, for example, to make patterns visible, to filter unimportant features and to discover accurate knowledge.

Some key applications of domain knowledge in the data mining process and data analysis are included in Table 15.

Source	Sector	Area of application
Matheus *et al.* (1993)	Evaluation of KDD systems	Use of domain knowledge for knowledge discovery.
Anand *et al.* (1995)	Rules examples of personnel, housing and insurance databases	Use of domain knowledge to reduce the data before the data mining analysis.
Ben-David (1995)	Five data sets: financial, employee related and academic	Use of domain knowledge for monotonicity constraint evaluation.
Bejar *et al.* (1997)	UCI Repository data sets (e.g., Soya bean)	Use of domain knowledge for knowledge discovery.
Djoko *et al.* (1997)	Substructure discovery	Use of domain knowledge for knowledge discovery.
Sill (1998)	Corporate bonding rates	Use of domain knowledge for monotonicity constraint evaluation.
Barzilay and Brailovsky (1999)	Texture recognition	Use of domain knowledge for feature selection.
McClean *et al.* (2000)	Only small example. No domain specified	Use of domain knowledge to reduce missing values and outliers.
Almeida and Torgo (2001)	Financial time series prediction	Use of domain knowledge for feature construction.
Alonso *et al.* (2002)	Medical diagnosis domain	Use of domain knowledge in different phases of the data analysis

		process (e.g., building phase and validation of results).
Kopanas *et al.* (2002)	Decision support system for a telecom company	Use of domain knowledge in different phases of the data analysis process (e.g., data pre-processing and problem definition).
Feelders and Pardoel (2003)	Bankruptcy and Windsor housing data sets + four data sets from UCI repository	Use of domain knowledge for monotonicity constraint evaluation.
Velikova and Daniels (2004)	Price prediction model	Use of domain knowledge for monotonicity constraint evaluation.
Altendorf *et al.* (2005)	Five data sets from UCI repository	Use of domain knowledge for monotonicity constraint evaluation.
Martens *et al.* (2006)	Five data sets from UCI repository	Use of domain knowledge for monotonicity constraint evaluation.
Prinzie and Van den Poel (2006)	CRM cross-sell application	Comparison of random and expect feature selection.
Velikova *et al.* (2006)	House pricing model	Use of domain knowledge for monotonicity constraint evaluation.
Van Gestel *et al.* (2007)	Credit risk evaluation	Use of domain knowledge for monotonicity constraint evaluation (signs of coefficients).
Yu *et al.* (2007)	Stock exchange domain	Use of domain knowledge for feature selection.
Sinha and Zhao (2008)	Credit risk analysis (loan)	Use of domain knowledge in variable creation.

Table 15. Literature review on domain knowledge

The approach proposed by Anand *et al.* (1995) combines the use of the knowledge by the business expert with a discovery process that can help uncover patterns in the data that are useful for the analysis. They consider that the customer provides information in two forms, domain knowledge and bias information, which facilitate the incorporation of domain knowledge into the data mining effort. They go on to apply domain

knowledge to reduce the search space before the data mining analysis, hereby making patterns more intuitive. This is an accepted approach, but it will not be used in this research. For this research instead, the data will be pre-processed using domain knowledge, but the patterns will be evaluated only after the data mining process is executed.

Alonso *et al.* (2002) use expert knowledge entered into a system not only in the model building phase, but also to clean and pre-process the data for the data mining analysis. Their system allows experts to validate the data mining results and, if required, support the preparation of new data mining tasks, to make the solution as accurate as possible.

Their approach is more in agreement with the aims of this research; however, an expert system will not be developed. The domain knowledge will be applied directly to the data mining results, which will be manipulated (if necessary) in accordance with the domain knowledge employed.

Kopanas *et al.* (2002) tries to identify how domain knowledge is used during different phases of the data analysis process, for example, for data pre-processing and problem definition, where the role of the domain expert is essential. They also discuss the use of domain knowledge examining practices and rules of a specific company and emphasising the need for both types of knowledge (business specific and general domain knowledge) when performing the analysis.

This approach is more in accordance with the one that will be used in this research, where domain knowledge will be applied to data pre-processing and the interpretation of the features selected by the data mining techniques. Although the feature selection method used (based on their frequent incidence in the decision tree model) is different from the approach adopted in this research, the use of domain knowledge is evident in their evaluation and creation.

Almeida and Torgo (2001) use domain knowledge represented as technical analysis indicators to execute feature construction during the pre-processing stages. After generating the features based on the domain knowledge representation language, the features are ranked and filtered and only the highest ranking features are kept for the analysis; data mining is then used for prediction.

This research will also use statistical evaluation to select the best features for prediction, but it does not manipulate variable creation in general, especially in the churn analysis. If it is necessary to manipulate a variable in the pre-processing stage, it will be manipulated to avoid problems in the calculation; however, the objective is to keep the variable's values and purpose correct, not to create new variables.

McClean *et al.* (2000) use domain knowledge in the pre-processing stage, with the purpose to reduce missing values and outliers. The domain knowledge is added to the system as integrity constraints or other type of knowledge. A logic programming algorithm re-engineers the data, generating new values for missing fields or outliers, based on this domain knowledge.

In this research, domain knowledge will also be applied in the pre-processing stage, but in a more intuitive manner. For example, as general values for groups are not given if the records present missing values, then the records could be eliminated; in the case of outliers, domain knowledge would be used to identify and interpret these values, and the appropriate actions would be taken (removal or substitution).

Barzilay and Brailovsky (1999) examine the use of feature selection when using support vector machines. They came to the conclusion with their experiments that the classifiers present better results when the number of input features is relatively small. They then use available domain knowledge to perform feature selection, which shows an improvement of the results.

Prinzie and Van den Poel (2006) are also in favour of feature reduction, as unnecessary features only complicate the analysis and can introduce multicollinearity in the case of a multinomial logistic regression model. Feature selection can improve accuracy and comprehensibility of a model, but they argue that feature and model selection needs to be addressed at the same time, in order to guarantee the best results. They go on to compare the use of random feature selection and expert feature selection, applied for redundancy removal and performance enhancement, which are used simultaneously with the model selection. In their research, it is shown that the random selection approach presents higher performance, illustrating that the domain knowledge application, for their case, would not be the better option.

In this current research however, the initial feature selection will be independent of the model, as the purpose is to evaluate different models using the same features. Also, the individual evaluation of features will happen after the models are extracted, in order to apply the domain knowledge to the final features and not to the feature selection. The feature selection will be based on statistical significance and also on the model applied after the initial feature selection.

Yu *et al.* (2007) argue that irrelevant features do not enhance performance to a model, but can introduce noise, which can be damaging to the model performance. They propose an approach to select an optimal set of features for a specific algorithm, containing the major relevant features, excluding redundant effects. To achieve this, domain knowledge is applied in the form of constraints, which guide the tuning process of the kernel for feature selection. This kernel, which is an approximation to the real mutual information between features, is responsible for selecting the features that will be used for prediction and the domain knowledge constraints are responsible for the accuracy of the measurements. Based upon these constraints, the features kept are the ones that have the highest correlation with the target variable, eliminating redundant features whose correlation with other features is higher than their correlation with the target variable.

More recently, Sinha and Zhao (2008) investigate if the performance of data mining classifiers would improve with the incorporation of domain knowledge. They use a small loan data set to compare the performance of different data mining techniques (naïve bayes, logistic regression, decision tree, decision table, neural networks, k-nearest neighbour and support vector machines) with and without the presence of domain knowledge. This knowledge was represented by rules from an expert inserted into the data mining technique through the creation of a credit rating attribute. Their analysis demonstrate that the overall performance of the classifiers improve with the incorporation of domain knowledge, the exception being for decision trees, which showed that there was no significant difference.

For this research, domain knowledge will be used for pre-processing of the variables, but the feature selection will be based on statistic measurements, without imposing

values to the model. The idea is to use domain knowledge to evaluate the resulting models, and not to impose the initial model.

In many cases, domain knowledge is used in the knowledge discovery process, not in the manipulation of results from the data mining analysis. This is the case of Djoko *et al.* (1997), Bejar *et al.* (1997) and Matheus *et al.* (1993).

Djoko *et al.* (1997) add domain knowledge to a knowledge discovery system, in order for it to assist the knowledge discovery process, making the discoveries more meaningful to the system user. Matheus *et al.* (1993) use domain knowledge to influence the search when discovering patterns in databases. These discoveries would then be stored to be used in future pattern discoveries.

Bejar *et al.* (1997) define the domain theory as constraints, used as a guide in the inductive process in an unsupervised learning algorithm, which increases the stability and quality of the classification. For them, in the same way that some attributes can be ignored in the predictions, they could also be specifically used to increase predictiveness.

As emphasised on beforehand, rules will not be forced into the data analysis to influence the results in this research; they will be evaluated after the models prediction, and domain knowledge will be used with this purpose.

One of the main aspects related to the domain knowledge integration is the monotonicity constraint (Ben-David, 1995; Sill, 1998; Feelders and Pardoel, 2003; Velikova and Daniels, 2004; Altendorf *et al.*, 2005; Velikova *et al.*, 2006; Martens *et al.*, 2006; Van Gestel *et al.*, 2007). A monotonicity constraint specifies that a variable has a uniquely determined impact (either increasing or decreasing) on the target.

To incorporate this monotonicity requirement, different techniques have been suggested. This is, for example, the case for decision trees (Ben-David, 1995; Feelders and Pardoel, 2003; Velikova and Daniels, 2004), Bayesian network learning algorithms (Altendorf *et al.,* 2005), AntMiner+ (Martens *et al.*, 2006), neural networks (Velikova *et al.*, 2006) and logistic regression (Van Gestel *et al.*, 2007).

Ben-David (1995) and Velikova and Daniels (2004) incorporate monotonicity constraints directly into the decision tree model, manipulating the resulting tree in order to obtain monotone trees. Feelders and Pardoel (2003) also test the monotonicity directly on the leaves of the classification tree, creating a large overfitted tree and pruning it towards monotone subtrees. For them, the monotone trees have comparable performance as to the original trees. Feelders and Pardoel (2003) and Velikova and Daniels (2004) also state that the monotone trees are considerably smaller, making them more interpretable.

Sill (1998) developed a monotonic network, by taking maximum and minimum operations on groups of hyperplanes. This is done in order to enforce monotonicity constraints by constraining the signs of the hyperplane weights. This means that if the desired monotonicity is supposed to increase, all the weights connected to that specific input are constrained to be positive; if the intention is a decreasing monotonicity, the weights are constrained to be negative. The performance of the monotonic networks is compared against a linear model and a standard neural network model, and the monotonic networks perform better.

Altendorf *et al.* (2005) discuss how knowledge about qualitative monotonicities can be formalised and incorporated into Bayesian network learning algorithms. They have constructed the structure of the Bayesian network based on previous domain knowledge, inserting monotonicity annotations on each of the network links. These annotations represent a positive or negative influence that a variable X has on a variable Y. They conclude that the monotonicity analysis can improve the performance of the Bayesian algorithm, especially on very small data sets.

Velikova *et al.* (2006) suggest the use of partially monotone neural network models, in order to generate models that are more in accordance with the knowledge from the decision makers. The idea is that the output variable depends monotonically on some of the input variables but not all. They use the monotonic networks defined by Sill (1998) as the approach for their partial monotonic model. They compare their results with standard neural networks and partially monotone linear models and conclude that their model provides a better fit to the data.

Van Gestel *et al.* (2007) use domain knowledge for monotonicity constraint incorporation through the evaluation of signs of coefficients in the logistic regression model. Variables that present coefficients with an opposite sign to what is expected by the financial analysts are not selected for analysis. They use this variable selection as part of their analysis, considering it a key step for successful model implementation.

Furthemore, Martens *et al.* (2006) came up with the terminology of acceptable classification models for implementation, based on the idea of providing models that are accurate, comprehensible and justifiable, through the use of a knowledge fusion process that incorporates domain knowledge from the experts into the data mining analysis. They are able to incorporate domain knowledge into the AntMiner+ algorithm. The rules generated by the AntMiner+ algorithm are validated using decision tables. The analysis of counter-intuitive rules in the decision table enables problem visualisation. They then are able to incorporate domain knowledge into the AntMiner+ algorithm by changing its heuristic values. Based upon their definition on monotonic constraints, they evaluate which of these variables have a constraint and should be manipulated by the domain expert. They conclude that, in terms of generalisation behaviour, integrating knowledge into the model does not tend to influence the final performance that much. This justifies the interpretation and intervention of the domain expert when evaluating data mining models.

In this research, the monotonicity constraint will be achieved by adapting the results obtained from the logistic regression, linear regression and decision tree models, based on common sense and knowledge of the domain under investigation. The monotonicity evaluation could be applied to all or some variables. It will depend upon the variables selected by the model and on the existing knowledge of these variables.

All these previous research papers present some key elements that need to be considered in the application of domain knowledge. The next section presents the purpose and methodological evaluation of domain knowledge in this research's experiments and its integration with data mining in the domain of customer evaluation.

5.4. Practical Application of Domain Knowledge

When evaluating the practical use of domain knowledge, the company interviews (also discussed in Chapter 3) provided some interesting insights. All companies agree the use of domain knowledge is common sense and that it can be applied in different stages of the data analysis. What they call common sense is based on their experience, on how similar customers behaved in the past and this knowledge is used in their day-to-day analysis.

For some companies, domain knowledge is used throughout their analysis, for example, in data preparation, data analysis and evaluation of results. This approach is in accordance with the literature (e.g. Alonso *et al.*, 2002; Kopanas *et al.*, 2002), in which domain knowledge is an essential part of the whole data evaluation process. Other companies consider that it can be used in a retrospective manner: if something goes wrong, they can go back and re-evaluate the data. Another approach is to rely on the statistical investigation, but when the final results seem contradictory, this leads to further investigation. For example, if the model presents a variable that did not make sense in the analysis, they would eliminate it to avoid questioning the validity of the model. This approach presents similarities with Van Gestel *et al.* (2007). As described in the previous section, they use domain knowledge for monotonicity constraint incorporation, in order to ensure that the variables present in the model have their coefficients with the expected sign. However, this analysis of signs performed by the companies in this research is not formal; it is based on their intuition and needs of the business, but it is not fixed (which would require it to be followed in every data analysis).

In terms of domain experts, two distinctive beliefs are identified. For some companies, only a few key people that work directly with the data analysis are considered domain experts. On the other extreme, some companies consider that everybody that works in customer service (which could represent a significant amount of people) is a domain expert in the area associated with the customer. The knowledge acquired by them is used in the data evaluation and should be a continuous source of support and information.

For some of these companies, their knowledge or common sense can be extensively used when offering deals to retain the customers. Based on their evaluation of customer value (taking into consideration the customer analysis model), it is possible to define how much promotion can be allocated to that customer, in a way that it would not be damaging to the profit generated by the customer.

An important aspect in all cases is the trust in the generated models. If the methodology of generating the models is stable, the practical use of the models becomes the key factor. By stable it is meant that the company can trust the model to help the decision making process. Some companies trust their models, but they adjust and adapt them based on, for example, customer feedback, history, or even due to the fact that they know the source for their predictions and the type of assumptions made for their calculations.

What companies should do is to prepare their data, making sure that it is properly pre-processed, guaranteeing data quality. A domain expert is very important in this stage. Then the data mining procedures would be employed and domain knowledge could be used before, during or after this stage, as described in section 5.3 (Ben-David, 1995; Sill, 1998; Feelders and Pardoel, 2003; Velikova and Daniels, 2004; Altendorf *et al.*, 2005; Velikova *et al.*, 2006; Martens *et al.*, 2006; Van Gestel *et al.*, 2007).

The model needs to be adjusted by the domain knowledge to become interpretable, but the performance metrics also represent a strong part of the data analysis. If the models generated do not present a reasonable performance, they would not be useful for implementation. Also, different models could be generated, and the best model would then be chosen for deployment. Guaranteeing that the model is interpretable but robust at the same time is an important aspect of the analysis. The trade-off between interpretability and predictive power should not be too high. The methodology presented in the next section represents a good approach to incorporate domain knowledge into the data mining analysis.

Another point to be taken into consideration is how to weigh the importance from knowledge mined from data versus knowledge elicited from the expert. These two elements have to work together, in order to provide the best interpretability and the best results for the model. If the knowledge extracted from the data is significantly different

from the knowledge provided by the expert, the data analysis should be checked to ensure that the data was properly prepared. Also, the expert knowledge needs to be validated, based on other domain experts or consolidated knowledge in the area. It is important to emphasise that new knowledge may come out of the data analysis, which is in accordance with the mining process to discover new insights from the data.

Finally, after the model is generated and validated, it could then be deployed for strategy definition. As an example in a churn prediction context, some of the most predictive variables could be used for the definition of strategies to target the general population. This would have a broader effect on the customer base of the company. In addition to this, as budgets for customer strategies in reality are generally restricted, if the model presents a good predictive power, the customers highly classified as churners will be the ones target for marketing campaigns. These points emphasise how important it is to have a reliable and interpretable model.

5.5. Methodology for Empirical Evaluation using Domain Knowledge

This research has the purpose to explore how data mining techniques can facilitate and help churn and CLV prediction, making the resulting models compliant with domain knowledge, in such a way that would facilitate their interpretation and usability by the company.

However, some reasons why data mining can be different from domain knowledge are, (a) poor data, which was not properly treated or is not from a reliable source; (b) multicollinearity still in the model; and c) spurious correlations not identified. Spurious correlations are defined as a misleading correlation between two variables, caused by a third variable. This will be further explained in Chapter 6.

To observe and correct any such discrepancies, hence making the data mining analysis compliant with domain knowledge, this research explores some key analysis aspects. It means that it is necessary to, (a) perform correct and proper pre-processing, and if necessary, redefine variables; (b) evaluate the signs of the regression coefficients; and (c) analyse counter-intuitive rules in decision tables.

To proceed with the evaluation, using data mining and domain knowledge, the analysis will be done in accordance with Figure 21, which describes the methodology applied.

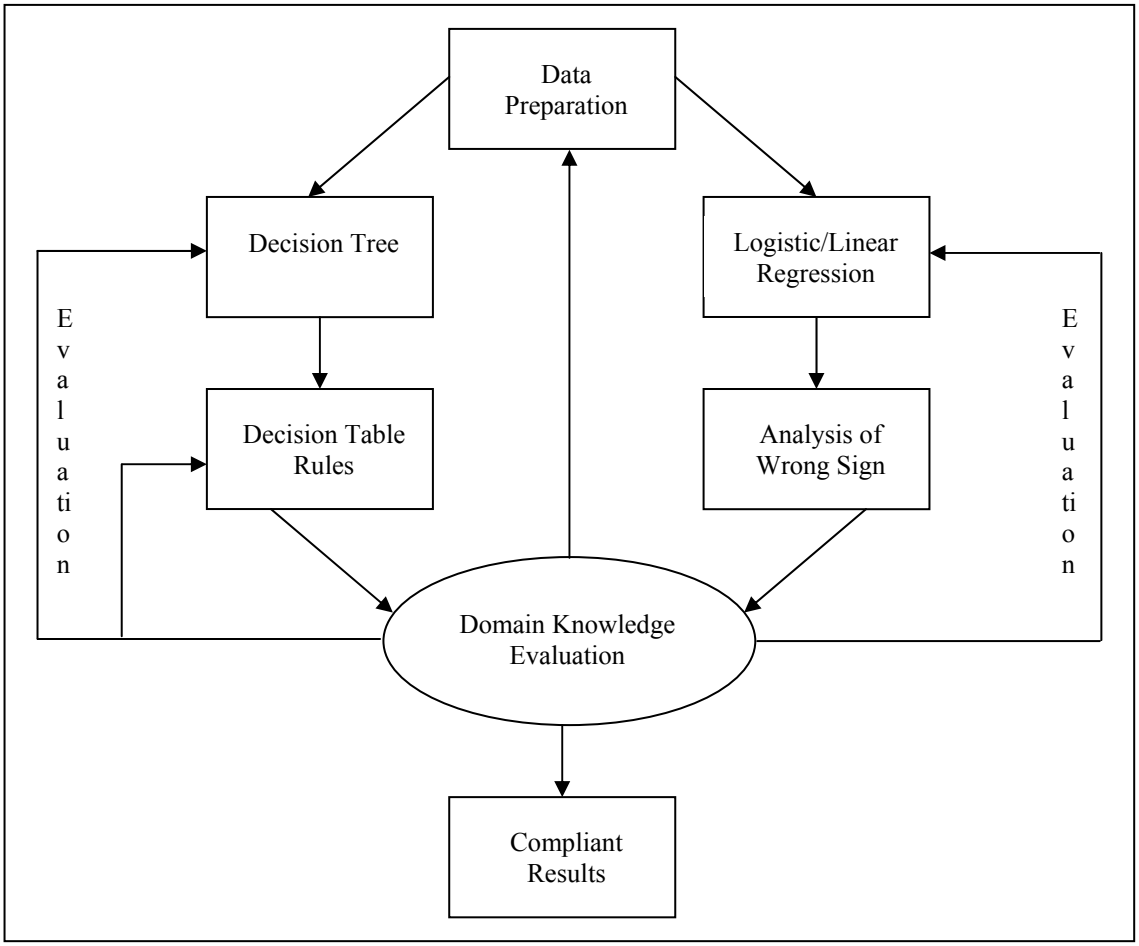

Figure 21. Research methodology

In this figure, domain knowledge is applied in the pre-processing stage, making sure that the variables are in accordance with what is expected of them, based upon a proper cleaning and preparation of variables. After the data is prepared, the linear regression, logistic regression and decision tree models are trained.

With regard to the logistic and linear regression models, domain knowledge will be expressed in terms of the expected sign of the variable coefficients. Next, the analysis of coefficient signs will be undertaken. Using this procedure, inputs will be added to the model based on their significance, and only inputs with the correct expected sign will be kept. The logistic regression approach will be further discussed in Chapter 6 in relation to churn analysis. The linear regression approach will be further discussed in Chapter 8 in relation to CLV analysis.

For the decision tree model, decision tables will be used to facilitate the domain knowledge evaluation. They will be used to check for expected effects of variables one-by-one, evaluating their monotonicity by moving them down in the condition order. This will be further discussed in Chapter 7 and Chapter 8.

As stated in section 5.3, the monotonicity analysis will depend on the variables selected by the model and on the existing knowledge about these variables. It could be applied to all or some variables, but the purpose is to generate simpler and more interpretable models than those originally generated. If the constraint adaptation for an input would not significantly influence the result and add too much complexity to the model, its change will not be supported. The idea is to generate models that are acceptable by decision makers, so more complex models are not adequate.

As stated by Martens *et al.* (2006), the concept of hard and soft constraints should be taken into consideration. The hard constraints are mandatory and should not be violated. These are the ones that need to be changed in accordance with the domain knowledge. The soft constraints are preferred but not mandatory, and the values could be manipulated or not. As a result, the domain expert should choose which variables should incorporate a constraint, based on the knowledge available and the model under evaluation.

It is necessary to emphasise that checking expected signs is not a new approach, however this procedure is relevant for churn evaluation and no evidence was found of its use in this context. Also, other approaches exist for including monotonicity constraints in rule-based classification models, as explained in section 5.3, for example by enforcing the constraints during algorithm learning, as in the referred AntMiner+ technique (Martens *et al.*, 2006).

5.6. Conclusions

The purpose of this research is to use data mining and apply common business sense to data pre-processing and modelling, based on the domain knowledge in relation to the company's operating sector. Its objective is to make sure the results will be coherent and useful to the development of strategies.

Domain knowledge is applied in a variety of situations and its use goes from the pre-processing stage, to the data analysis and evaluation of results. It can assume many forms such as data preparation, insertion of rules, limitation of search space, and monotonicity constraints. For this research, the monotonicity approach will be used to evaluate the results from logistic regression, linear regression and decision trees. The methodology proposed will facilitate the interpretability of results provided from the data mining models.

Using this type of approach will make companies more aware of the findings, being able to better understand the results. As a consequence, the models will be more acceptable and it will be possible to make more decisions based on the information extracted.

Chapter 6. Evaluation of Signs to Support Domain Knowledge Integration

6.1. Introduction

This chapter explores the evaluation of signs in a logistic regression model, in order to integrate domain knowledge into data mining. In particular, the identification and correction of wrong signs is discussed when selecting the most predictive variables in defining the probability of a customer churning or not.

To proceed with this assessment, the concept of wrong sign is explored, evaluating the causes and how it is perceived in the literature. This is followed by an explanation on how domain knowledge is integrated with the wrong sign evaluation as presented in pseudo-code.

Finally, an empirical evaluation is undertaken using two churn data sets from the telecom sector. Finally, conclusions are drawn from the analyses and results.

6.2. Wrong Signs in Logistic Regression

Wrong signs occur when the estimated signs of variables are opposite to what is expected from them. Kennedy (2002), Mullet (1976), and Verstraeten (2005) present some of the situations that can cause wrong signs and suggest some solutions to correct the problem.

One of the reasons for wrong signs may be bad data quality. If the data used in the analysis is badly collected or incorrectly pre-processed, it increases the chances of incorrect results, leading to greater chances of wrong signs.

Remaining multicollinearity in the model would be considered a strong possibility of causing wrong sign (Mullet, 1976; Verstraeten, 2005). Multicollinearity is caused by a strong correlation between predictor variables. Collinearity increases the standard deviation, such that the coefficient might accidentally switch to the wrong sign.

As explored by Mullet (1976), taking the residual variance from the analysis of variance:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}$$

where n is the sample size, y_i represents the actual value for the target and $\hat{y}_i$ is the predicted value, the residual variance measures the dispersion between y_i and $\hat{y}_i$. The formula below is considered for the variance analysis:

$$s^2(\beta_i) = \frac{\sigma^2}{n s_i^2 (1 - R_i^2)}$$

with sample variance of a variable X_i (S_i^2), R_i^2 the multiple correlation between X_i and the other X's in the model, and assuming a sample n with σ^2 (model residual variance). From this formula, as R_i^2 increases approaching unity, the variance of the estimate β_i will also increase. As the collinearity increases, the likelihood of wrong signs increases, due to the large estimated variances.

In the formula above, the sample variance is represent as:

$$S_i^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}$$

where n is the sample size and $\overline{X}$ is the mean, and the variance measures the dispersion between X_i and the mean.

The multicollinearity problem is supported by Kennedy (2002), who suggests the creation of a ratio between the related variables, rather than using the two original values. In this research, this would not apply because the idea is to explore the individual effects of variables, choosing the ones that would better influence the results.

To solve the problem of multicollinearity, the correlations and associations between independents variables will be evaluated and only the variables that have a better relationship with the target variable will be kept.

This is supported by Mullet (1976), as he suggests the inclusion of all important variables in the model, whilst being aware of the effects this can cause to the model, emphasising the need to evaluate correlated variables. So, it is necessary to evaluate them, and consider the effects as variables are added or removed from the model.

Another problem would be inappropriate variable or data definition. This would happen, for example, when using ratios as variables. This means that if the sign of one of the values change, it may also change the meaning and interpretability of the results. Consider the case of the ratio debt/earnings in a credit scoring evaluation, as shown in Table 16.

Debt	Earning	Ratio Value	Expected effect	Customer
+	+	+	-	Good
+	-	-	+	Good

Table 16. Change of signs based on variable value

Based on Table 16, the first row indicates that the higher the value of the ratio, the smaller the probability of that customer being good. This would give a negative sign to the expected effect of the ratio when measuring the probability of a customer being good. Nonetheless, if the earnings become negative (loss), it would change the meaning of the ratio. For example in Table 17:

Ratio Value	Effect	Result
+20 (increase)	-	P(good) decreases
+10 (increase)	-	P(good) decreases
-10 (decrease)	+	P(good) decreases
-20 (decrease)	+	P(good) decreases

Table 17. Example of debt/earnings analysis

Because of this, the use of this type of variable must be well evaluated before considering its inclusion.

Kennedy (2002) discusses selection bias as another problem that can cause wrong signs. In this case, the bias is generated by the fact that the observations in the data are not randomly selected and there are factors influencing its presence. This situation must be avoided, and the understanding, correction or elimination of such inconsistency is necessary, unless it is part of a specific type of analysis.

Spurious correlations could also be another problem. This is related to a misleading correlation between two variables, caused by a third variable. The ability to identify when a correlation is not coherent is necessary. This can be done by evaluating the models and recognising patterns that are contradicting the data. An example of spurious correlations as described by Hand *et al.* (2001) was the finding that over the past thirty years when the winner of the Super Bowl championship in American football is from a particular league, a leading stock market index historically goes up in the following months.

In this research, the evaluation of spurious correlations is done by understanding the meaning of each variable, therefore facilitating the analysis of correlation. In case too many variables are present, instead of evaluating the individual correlations between the variables, the correlation was evaluated between the independent variables and the target variable, and only the most significant variables were kept in the analysis.

Amongst other factors, Kennedy (2002) also emphasises the problem related to outliers. The presence of outliers could influence the results, and the variable's sign could change based on it. In this case, it is necessary to evaluate all variables and analyse if any discrepancies in values were supposed to be present or not. If necessary, these records should be removed from the analysis, to avoid unnecessary added bias.

Another problem would be the expected signs of categorical variables. One way of solving this problem would be, for example, by applying weight of evidence (WoE) coding, based on the following formula:

$$WoE = \ln \frac{p(churn)}{p(notchurn)}$$

where *p(churn)* is obtained based on the amount of churners for that specific value of the categorical variable, divided by the total amount of churners in the data set, and *p(notchurn)* is obtained based on the amount of non-churners for that specific value of the categorical variable, divided by the total amount of non-churners in the data set.

The amount of churners and non-churners in the data set will indicate how the monotonic analysis will be represented: the higher the weight of evidence, the higher/lower the probability of churn. For example, in the case of one of the data sets evaluated (Telecom1), evaluating the categorical variable "State", the higher the weight of evidence, the higher the probability of churning in comparison with the probability of not churning. This means that, proportionally, there are more churners in that group than non-churners:

- If *p(churn)* > *p(notchurn)* then WoE > 0
- If *p(churn)* < *p(notchurn)* then WoE < 0

It also shows that the bigger the variation between *p(churn)* and *p(notchurn)*, the bigger the WoE, with negative or positive value. For example,

- for *p(churn)* = 0.094 and *p(notchurn)* = 0.046, WoE = 0.71465
- for *p(churn)* = 0.32 and *p(notchurn)* = 0.23, WoE = 0.33024
- for *p(churn)* = 0.25 and *p(notchurn)* = 0.23, WoE = 0.08338
- for *p(churn)* = 0.23 and *p(notchurn)* = 0.29, WoE = -0.2318
- for *p(churn)* = 0.10 and *p(notchurn)* = 0.20, WoE = -0.69315

In this research, weight of evidence was not used. The idea of this research is to evaluate the influence of the variables in the prediction. As a result, the categorical variables that had only a few values were kept the same. This is the case, for example, of the variable "hnd_webcap" in Telecom2, which presents three categorical values and these are kept unchanged for the data evaluation.

For the categorical variables that presented too many values, in order to reduce complexity of evaluation, coarse classification was used based on the odds of churn to create groups of variables under a few labels. For example, in the case of the categorical variable "State", the coarse classification was made based on grouping values with

similar odds. The odds were obtained based on the amount of churners divided by the amount of non-churners for that specific value of the categorical variable. The idea was to keep the significance of the variables in order to facilitate their interpretability when developing strategies.

In conclusion, wrong signs should be avoided and treated in order to facilitate the acceptance by managers and users, even if this represents an exchange between interpretability and predictive power, as long as this trade-off is not too big. The idea is to make the models more friendly and easy to interpret, without loosing too much in the predictive power.

6.3. Integration of Wrong Sign Evaluation with Domain Knowledge

When executing a feature selection with logistic regression, three techniques are of relevance: forward, backward and stepwise selection. With forward selection, when the logistic procedure computes the chi-square statistic for each variable not in the model, if it is significant at some significance level, the corresponding variable is added to the model and it is never removed from it. In the backward selection, the analysis starts with all the variables in the model, and then the least significant variable is removed, so long as it is not significant at the chosen significance level; the method of removal continues until all remaining variables are statistically significant. Forward selection has the drawback that each addition of a new variable may turn one or more of the already included variables into non significant. Backward selection has a similar problem, as sometimes variables are removed, but they would be significant when added to the final model. The solution could then be to have a compromise between forward and backward selection methods, which can be achieve by stepwise selection. With this method, the variables are entered into the model also based on the chi-square statistic, but each forward selection step may be followed by one or more backward elimination steps, assuring that only significant variables would be present in the model.

Verstraeten (2005) proposes a model to enforce signs in the logistic regression model, FSR (Forward Selection with Sign Restriction). His model is similar to the forward selection technique, differing in the fact that the features are only included when the signs of all parameters in the model correspond to their univariate counterparts. He also

shows that the fewer variables are present in the model, the more the signs will be in accordance with their univariate signs.

When comparing the results of FSR with other variable selection methods (e.g., forward and stepwise selection) Verstraeten (2005) concludes that in most of the cases, the other feature selection techniques were not significantly better than FSR. As a result, he has obtained a more robust and easier to interpret model.

In this current research, the technique used was stepwise selection, which in general presents a good or better performance than the other common selection techniques. To enforce the correct sign in combination with stepwise selection, the domain knowledge was applied based on the expected sign for the variable, also taking into consideration how significant this variable was to the model (based on the p-value).

To proceed, the signs' analysis is made following the method described in Figure 22, in the form of pseudo-code. Using this procedure, inputs were added to the model based on their significance, and only inputs with the correct expected sign were kept.

The p-value represents how significant the variable is to that model (Rud, 2001; Moore and McCabe, 2006). The smaller the p-value, the more significant this variable is to the model. The p-value is derived from the Wald chi-square measure, which is directly related to the predictivity of the variable. It is this one which is used by the logistic regression model in SAS (Allison, 2001), and works similar to a normal chi-square statistic (Rud, 2001; Daniel, 2005; Moore and McCabe, 2006).

It shows that the bigger the Wald chi-square value, the more significant the variable is to the model, resulting in a smaller p-value. So, to proceed with the evaluation and facilitate visualisation, the variables selected by the stepwise procedure are ordered ascending by their Wald chi-square value, and their signs evaluated. The Wald chi-square value is used to identify the main predictors, in case the purpose was to limit the number of predictor variables. For this analysis, it is assumed that a variable with a p-value bigger than 1% is not significant to the model, so it should be eliminated even if the sign is correct.

(1) Define domain constraint D_i over A_i
($i = 1, ..., n$, where n is the number of attributes)
(2) Generate model based on stepwise regression
(3) For each attribute A_i in the model
If the coefficient sign for A_i violates the domain constraint D_i then
Begin
Eliminate A_i from the model: $A = A - \{ A_i \}$
Run stepwise regression using A
Go back to (3)
End;

Figure 22. Algorithm for selection of correct signs for logistic regression

Based on this pseudo-code, the procedure runs a full stepwise selection, and then evaluates the signs of the variables: if compliant with the domain knowledge, keep the variable; if not, eliminate the variable and run the model again until obtaining a compliant and interpretable model. This method was applied in the section below, evaluating two Telecom data sets.

6.4. Empirical Evaluation

Using the framework presented in Figure 21 (Chapter 5), two data sets were evaluated for churn prediction. The characteristics of both data sets are presented in Table 18. The empirical results will be shown in the subsections below.

Data Set	# Variables	# Obs	Training	Test	Churn Rate (%)
Telecom1	21	5,000	3,350	1,650	14.1
Telecom2	22	15,000	10,000	5,000	1.8

Table 18. Characteristics of selected churn data sets

6.4.1. Data Set Telecom1 – Description

As briefly discussed in Chapter 4, the data set used in this section is a churn data set publicly available, obtained from the KDD library. It contains 5,000 observations and 21 variables divided in discrete and continuous, with a target variable, churn.

One of the ways to make the data mining analysis compliant with domain knowledge is to correct and properly pre-process the data, understanding and redefining variables. As explained on page 113, the variable "State" contained fifty-one values, which were

grouped into fewer categories, to facilitate the model estimation and reduce complexity. Therefore, it was coarse classified into three groups, based on the odds of the target variable (churn).

It may also be necessary to eliminate from the data set all observations that contain unknown or inadequate values. Checking for outliers and missing values is recommended as they can affect the classification and predictive precision of the models. In the case of this data set, no missing values were found and the evaluation of outliers showed coherent values.

Correlation was measured between the variables, to analyse the type of relationship that existed not just between themselves, but between them and the target variable. Pearson chi-square statistic, the Cramer's V statistic, t-statistic (*t* test), and Pearson correlation coefficients were used to analyse the correlations (Rud, 2001; Chernick and Friis, 2003; SAS Help and Documentation; Sheskin, 2004). If the analysis showed a strong relationship between the variables, it would mean that these variables should not be evaluated together in the study, in order to avoid bias or misinterpretation of results. If this strong correlation is between the independent variables and the target variable, the variable should be kept, as it indicates a strong predictive power for the independent variable. This process would work as a variable selection method in case too many variables existed for analysis. As only twenty independent variables were present, this analysis was used as a tool to help in the evaluation of signs and rules resulting from the analyses. This feature selection is in line, as far as possible, with the domain knowledge applied.

To test the correlation between the discrete variables against the discrete target variable, the Pearson chi-square statistic was used, evaluating the p-value (<.01) and the Cramer's V statistic. In the chi-square analysis, a contingency table is constructed contrasting the categorical variable with the binary target as illustrated in Table 19.

	Value 1	**Value 2**	**…**	**Value *m***	**Total**
Churn	n_{11}	n_{12}		n_{1m}	$n_{1.}$
Not Churn	n_{21}	n_{22}		n_{2m}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$		$n_{.m}$	n

Table 19. Contingency table for the chi-square analysis

Note that n_{ki} represents the number of observations in cell [k, i]. When assuming that the categorical variable is unrelated to the churn status, the number of observations in cell [k, i] should be equal to $\frac{n_{k.}n_{.i}}{n}$.

The more these numbers differ from the observed frequencies n_{ki}, the bigger the dependence between the categorical variable and the churn probability, and hence the better its predictive quality. The chi-square test then measures the dissimilarity between the reported and expected numbers assuming independence and is represented as:

$$\chi^2 = \sum_{k=1}^{2}\sum_{i=1}^{m} \frac{(n_{ki} - \frac{n_{k.}n_{.i}}{n})^2}{\frac{n_{k.}n_{.i}}{n}}$$

with m representing the numbers of variable classes and m-1 degrees of freedom. Higher values of the test statistic indicate that the independence assumption is less likely and hence the variable has good predictive power. The p-values can be used to rank order the variables, with low p-values indicating that the variable is predictive.

The Cramer's V statistic is closely related to the chi-square test and is defined below:

$$\text{Cramer's V} = \sqrt{\frac{\chi^2}{n}}$$

Note that more generally, the denominator equals n*min(k-1; m-1), with k the number of rows and m the number of columns of the contingency table. Since k equals two in the churn prediction scenario, the denominator equals n. Cramer's V is between 0 and 1, and the higher the value (closer to 1), the more predictive the variable is.

The t-statistic (t test) was used to obtain the correlation between the continuous variables against the discrete target variable, selecting the variables with the most significant p-values (<.01). Lastly, to analyse the correlation between the continuous variables, the Pearson correlation coefficients were used (in accordance with Table 20).

	Vmail_ Message	Day_ Mins	Day_ Charge	Eve_ Mins	Eve_ Charge	Intl_ Mins	Intl_ Calls	Intl_ Charge	CustServ_ Calls
Vmail_ Message	1.0000	0.00538 0.7036	0.00538 0.7039	0.01949 0.1682	0.01950 0.1681	0.00246 0.8618	0.00012 0.9930	0.00251 0.8594	-0.00709 0.6164
Day_ Mins	0.00538 0.7036	1.00000	**1.00000** **<.0001**	-0.01075 0.4473	-0.01076 0.4468	-0.01949 0.1683	-0.00130 0.9266	-0.01941 0.1699	0.00273 0.8468
Day_ Charge	0.00538 0.7039	**1.00000** **<.0001**	1.00000	-0.01075 0.4474	-0.0176 0.4470	-0.01949 0.1682	-0.00131 0.9264	-0.01942 0.1698	0.00273 0.8472
Eve_ Mins	0.01949 0.1682	-0.01075 0.4473	-0.01075 0.4474	1.00000	**1.00000** **<.0001**	0.00014 0.9923	0.00839 0.5532	0.00016 0.9910	-0.01382 0.3284
Eve_ Charge	0.01950 0.1681	-0.01076 0.4468	-0.01076 0.4470	**1.00000** **<.0001**	1.00000	0.00013 0.9926	0.00839 0.5530	0.00015 0.9913	-0.01384 0.3280
Intl_ Mins	0.00246 0.8618	-0.01949 0.1683	-0.01949 0.1682	0.00014 0.9923	0.00013 0.9926	1.00000	0.01679 0.2352	**0.99999** **<.0001**	-0.01212 0.3915
Intl_ Calls	0.00012 0.9930	-0.00130 0.9266	-0.00131 0.9264	0.00839 0.5532	0.00839 0.5530	0.01679 0.2352	1.00000	0.01690 0.2322	-0.01915 0.1758
Intl_ Charge	0.00251 0.8594	-0.01941 0.1699	-0.01942 0.1698	0.00016 0.9910	0.00015 0.9913	**0.99999** **<.0001**	0.01690 0.2322	1.00000	-0.01218 0.3892
CustServ_ Calls	-0.00709 0.6164	0.00273 0.8468	0.00273 0.8472	-0.01382 0.3284	-0.01384 0.3280	-0.01212 0.3915	-0.01915 0.1758	-0.01218 0.3892	1.00000

Table 20. Correlation matrix for the continuous variables

The continuous variables that were highly correlated are: Days_Mins & Days_Charge; Eve_Mins & Eve_Charge; Night_Mins & Night_Charge; and Intl_Mins & Intl_Charge. As an example of the evaluation, the correlation matrix between some of the continuous variables is shown in Table 20.

Note that the table only shows how the analysis was done and does not present all the continuous variables. In the case of logistic regression with stepwise selection, the algorithm will eliminate the variables that are not relevant to the model. The analysis of correct signs of the logistic regression will show if the variables kept are the correct ones. A description of the main variables is shown in Table 21.

Variable	Description
State	Categorical, used to identify the fifty States on the EUA and the District of Columbia.
Intl_Plan	Discrete, indicates if the customer has international plan.
Vmail_Plan	Discrete, indicates if the customer has voice mail plan.
Vmail_Message	Continuous, number of voice mail messages sent.
Day_Mins	Continuous, indicates the amount of minutes the customer used the service during the day.

Day_Calls	Continuous, indicates the number of calls the customer made during the day.
Day_Charge	Continuous, indicates how much the customer paid for the day calls
Eve_Mins	Continuous, indicates the amount of minutes the customer used the service during the evening.
Eve_Calls	Continuous, indicates the number of calls the customer made during the evening.
Eve_Charge	Continuous, indicates how much the customer paid for the evening calls.
Night_Mins	Continuous, indicates the amount of minutes the customer used the service during the night.
Night_Calls	Continuous, indicates the number of calls the customer made during the night.
Night_Charge	Continuous, indicates how much the customer paid for the night calls.
Intl_Mins	Continuous, indicates the amount of minutes the customer used in international calls.
Intl_Calls	Continuous, indicates the number of international calls the customer made.
Intl_Charge	Continuous, indicates how much the customer paid for the international calls.
CustServ_Calls	Continuous, indicates how many calls the customer made to customer service, during the period in examination.

Table 21. Description of main variables for Telecom1

6.4.2. Data Set Telecom2 – Description

As presented in Chapter 4, this data set is a telecom data set used in the churn tournament 2003, organised by Duke University.

As described in the data set documentation, the customers selected were with the company for at least six months and were sampled during July, September and November of 2001, and January of 2002. For each customer, predictor variables were calculated based on the previous four months. Churn was calculated based on whether the customer left the company during the period 31-60 days after the customer was

originally sampled. The one month treatment lag between sampling and observed churn was for the practical concern that in any application, a few weeks would be needed to score the customer and implement any proactive actions.

For this analysis, two distinct samples of 10,000 and 5,000 customers were selected from the calibration and score data sets, respectively, with a total of 171 predictor variables. These samples were used for training and test sets, respectively. In order to reduce the data to a reasonable amount of variables suitable for modelling and also to facilitate their understanding and analysis, a variable selection was executed on the training set. Any variable presenting excessive unknown or inadequate values were eliminated from the analysis.

Correlation was measured between the variables and the target variable. As for Telecom1, to test the correlation between the discrete variables against the discrete target variable, the Pearson chi-square statistic was used, evaluating the p-value and the Cramer's V statistic; the t-statistic was used to evaluate the correlations for the continuous variables against the discrete target variable. As too many continuous variables presented a strong relationship with the target variable, only the variables with a p-value smaller than 0.0001 (<.0001) were taken into consideration, in order to select only the most relevant variables.

Even after this, there was the need for another pre-selection. In this case a stepwise logistic regression was executed for two groups of continuous variables, each one with thirty-one variables; then, the resulting variables of both models were evaluated again with the stepwise logistic regression model. In the end, this resulted in eleven continuous variables that were evaluated together with the ten categorical variables previously selected.

From these ten categorical variables, four of them (area, crclscod, ethnic, and hnd_price) had too many values. Using the same principle used for Telecom1, coarse classification was applied on these four variables, based on the odds of the target variable (churn). The grouping of these categorical variables into fewer categories, leads to the building of a more robust predictive model, since fewer parameters need to be estimated. A description of the main variables is shown in Table 22.

Variable	Description
AVG3MOU	Average monthly minutes of use over the previous three months
AVG3REV	Average monthly revenue over the previous three months
AVG6MOU	Average monthly minutes of use over the previous six months
AVGMOU	Average monthly minutes of use over the life of the customer
DROP_VCE_MEAN	Mean number of dropped (failed) voice calls
EQPDAYS	Number of days (age) of current equipment
MOU_MEAN	Mean number of monthly minutes of use
REV_MEAN	Mean monthly revenue (charge amount)
TOTMRC_MEAN	The base cost of the calling plan regardless of actual minutes used. Charge per minute
TOTREV	Total charge amount
PHONES	Number of handsets issued
AREA	Geographic area
ASL_FLAG	Account spending limit
CRCLSCOD	Credit class code
DUALBAND	Dualband
ETHNIC	Ethnicity roll-up code
HND_PRICE	Current handset price
HND_WEBCAP	Handset web capability
KID0_2	Child 0 - 2 years of age in household
KID3_5	Child 3 - 5 years of age in household
REFURB_NEW	Handset: refurbished or new

Table 22. Description of main variables for Telecom2

6.4.3. Empirical Analysis

The data sets were evaluated using the SAS 9.1 Solution and Enterprise Miner, these being part of a statistical package that makes it possible to evaluate a great amount of data, facilitating its interpretability and manipulation.

For Telecom1, to proceed with the calculations, the data set was split into a training set (67% = 3350 observations) and a test set (33% = 1650 observations), stratified with the same proportion of churners in both subsets. The proportion of churners is approximately 14.1% in each subset.

For Telecom2, to proceed with the calculations, the training set (10000 observations) was stratified with the same proportion of churners in the training set, but the test set (5000 observations) was not oversampled to provide a more realistic test set, according to a monthly churn rate of 1.8%.

After scoring the results of both data sets, the performance measures are shown in Table 23. As described in Chapter 4, when analysing the performance of the model, the measures used were the AUC, and the CA, sensitivity and specificity on training and test sets, assuming the KS statistic as the basis for cut-off.

Data Set	Metric	Training Set (%)	Test Set (%)
Telecom1	CA	**76.51**	73.65
	Sensitivity	74.36	**83.83**
	Specificity	**76.87**	71.98
	AUC	81.17	**82.91**
Telecom2	CA	**62.43**	51.23
	Sensitivity	68.11	**77.78**
	Specificity	**56.77**	50.75
	AUC	66.28	**67.64**

Table 23. Original logistic regression performance measurements

With traditional churn evaluation, if the results obtained with the logistic regression model were better than other models, this model would automatically be selected for any analysis. However, in this case, the results from the logistic model were re-evaluated, taking the domain knowledge into consideration, as this is one of the main purposes of this research.

6.4.4. Telecom1 – Domain Knowledge Evaluation

For the logistic regression model, the maximum likelihood coefficient estimates are shown in Table 24. Domain knowledge was then used to evaluate the coefficients' signs, analysing which kind of influence the variable has on churning.

Parameter		Expected	Estimate	Wald chi-square	P-value
Intercept		+/-	-7.6182	206.48	<.0001
Intl_Plan	No	+/-	-1.0588	201.97	<.0001
CustServ_Calls		+	0.5121	163.26	<.0001
Day_Mins		-	**0.0129**	138.11	<.0001
Vmail_Plan	No	+/-	0.4805	42.34	<.0001
Eve_Mins		-	**0.00747**	41.00	<.0001
State	Group1	+/-	-0.6261	29.94	<.0001
Intl_Calls		-	-0.0982	14.68	0.0001
Intl_Charge		+	0.2944	14.62	0.0001
Night_Charge		+	0.0879	12.13	0.0005

Table 24. Analysis of maximum likelihood coefficient estimates – Telecom1

This domain knowledge is expressed in the "Expected" Column, showing the expected sign for each variable. For example, the variable "Intl_Calls" suggests that it has a negative effect on the probability of churning: if the number of calls increases, the chance of churn decreases. Regarding the variable "CustServ_Calls", different meanings of interpretation can be made. In this research, because no clear information is presented on the data set description and because of its strong correlation with the target variable, it may indicate that this variable is more strongly related to the number of complaints. That is the assumption made in this research, which is in accordance with the findings from Company "C". For this company, the more the customers call their service number, the more it is related to complaints and these customers tend to churn more.

However, if complaint behaviour results in a favourable attitude towards the company when service failure recovery is satisfactory, the effect of the variable would be the opposite (Keaveney, 1995; Maxham, 2001; Coussement and Van den Poel, 2008b). This is related to a service recovery paradox, indicating how efficient complaint handling strategies are important. As a result, Coussement and Van den Poel (2008b) argue that complaining does not necessarily mean that the customer will leave the company. In their research, the more complaints a customer has sent to the company, the more likely they stay with the company.

On the other hand, Tax *et al.* (1998) emphasise that most companies underestimate the impact of efficient complaint handling. If the customer complaint management programs are not effective, customer complaints may lead to customer churn (Solnick and Hemenway, 1992; Keaveney, 1995). This is supported by Ahn *et al.* (2006), who found evidence that the number of complaints is positively associated with the churn probability. This is the approach adopted for this data set under investigation. As a result, the variable "CustServ_Calls" suggests that it influences churn positively: if the number of service calls increases, the chance of churn also increases.

Most of the variables' signs were according to what is expected, the exception being for the variables "Day_Mins" and "Eve_Mins", for which one would expect that an increase in the number of minutes usage would represent a decrease in the probability of churn rather than an increase. Hence, these variables were eliminated one by one, and the stepwise logistic regression procedure was rerun, also comparing the performance measurements. These values are highlighted (bold) in Table 24.

After evaluating the signs, the logistic regression results (see Table 25) presented coherent variables that are more in line with the domain knowledge. The performance measurements are presented in section 6.5, in Table 28.

Parameter		**Estimate**	**Wald chi-square**	**P-value**
Intercept		-7.8545	218.66	<.0001
Intl_Plan	No	-1.0472	201.30	<.0001
CustServ_Calls		0.5152	167.33	<.0001
Day_Charge		0.0759	139.06	<.0001
Vmail_Plan	No	0.4861	43.69	<.0001
Eve_Charge		0.0861	39.83	<.0001
State – Group1	No	0.4409	26.54	<.0001
Intl_Charge		0.3122	16.68	<.0001
Intl_Calls		-0.0932	13.39	0.0003
Night_Charge		0.0880	12.34	0.0004

Table 25. Final analysis of maximum likelihood coefficient estimates – Telecom1

6.4.5. Telecom2 – Domain Knowledge Evaluation

As modelled for Telecom1, the maximum likelihood coefficient estimates for Telecom2 are shown in Table 26.

Again with this data set, after running the logistic regression, some variables' signs were not in accordance with domain knowledge. For example, the variables "avgmou", "avg3mou" and "avg6mou", which represent average monthly usage of a customer over the lifetime, three and six months, respectively; it would be expected that an increase in the number of minutes usage would represent a decrease in the probability of churn.

Furthermore, the variables "totmrc_Mean" (charge base cost) and "totrev" (charge amount) would be expected to have positive influence on the probability of churn, as an increase in the amount charged would indicate an increase in the probability of churn. These values are highlighted (bold) in Table 26.

Parameter		Expected	Estimate	Wald chi-square	P-value
Intercept		+/-	-0.3205	5.9356	0.0148
eqpdays		+	0.000921	128.7771	<.0001
Mou_Mean		-	-0.00316	74.4891	<.0001
Rev_Mean		+	0.0184	53.3883	<.0001
hnd_price	G1	+/-	-0.3020	31.8872	<.0001
totmrc_Mean		+	**-0.00618**	25.0421	<.0001
Totrev		+	**-0.00016**	24.8197	<.0001
refurb_new	N	-	-0.1521	24.3498	<.0001
Area	G1	+/-	-0.1522	21.0873	<.0001
Avgmou		-	**0.000680**	20.7470	<.0001
avg3rev		-	-0.0106	20.3961	<.0001
avg3mou		-	**0.00163**	17.1940	<.0001
avg6mou		-	**0.000874**	16.7800	<.0001
Crclscod	G1	+/-	-0.3442	15.7693	<.0001
Crclscod	G2	+/-	-0.2791	10.5278	0.0012
Phones		+	0.0552	15.6979	<.0001
Ethnic	G1	+/-	-0.2069	6.6685	0.0098

Table 26. Analysis of maximum likelihood coefficient estimates – Telecom2

Hence, these variables were eliminated one by one in the logistic regression procedure, to evaluate if the new resulting variables were in accordance with what was expected. Any new variables added to the model were also evaluated, and if their values were not significant, they would be eliminated from the analysis.

Parameter		Estimate	Wald chi-square	P-value
Intercept		-0.6106	28.1442	<.0001
eqpdays		0.00112	277.5695	<.0001
mou_Mean		-0.00058	59.4963	<.0001
refurb_new	N	-0.1830	37.9813	<.0001
hnd_price	G1	-0.2176	30.8158	<.0001
rev_Mean		0.00355	25.6603	<.0001
phones		0.0566	19.2629	<.0001
crclscod	G1	-0.3677	18.9638	<.0001
crclscod	G2	-0.2986	12.9284	0.0003
Area	G1	-0.1017	16.6614	<.0001
drop_vce_Mean		0.0100	8.1108	0.0044

Table 27. Final analysis of maximum likelihood coefficient estimates – Telecom2

After evaluating the signs and eliminating insignificant variables, the final table of maximum likelihood coefficient estimates presented coherent variables that are more in accordance with what is expected in the domain knowledge analysis (see Table 27 above). The performance measurements are shown in section 6.5, in Table 28.

6.5. Final Results

From the final results in Table 28, it can be concluded that logistic regression is a suitable choice of classifier for integrating domain knowledge into the model, as, in the analysis of both data sets, the model's performance stays relatively stable with the introduction of domain constraints when taking into consideration the AUC measure, with only fairly small variations in the performance values.

On both training sets (Telecom1 and Telecom2), the DeLong *et al.* (1988) test showed a significant difference between the originals and amended models, demonstrating that

the original models showed a better performance than the amended models (see Figure 23 and Figure 24).

Data Set	Metric	Training Set (%)		Test Set (%)	
		Original	Amended	Original	Amended
Telecom1	CA	76.51	70.19	73.65	70.35
	Sensitivity	74.36	77.12	83.83	82.13
	Specificity	76.87	69.09	71.98	68.41
	AUC	**81.17**	78.09	**82.91**	**80.82**
Telecom2	CA	62.43	61.28	51.23	64.01
	Sensitivity	68.11	64.58	77.78	64.44
	Specificity	56.77	57.99	50.75	64.00
	AUC	**66.28**	64.56	**67.64**	**65.15**

Table 28. Original and amended logistic regression performance measures

However, on both test sets, variations in performance values were still fairly small; for Telecom2, the differences in AUC between the original and amended models were statistically insignificant at the 95% level according to the test proposed by DeLong *et al.* (1988) (see Figure 26), whereas for Telecom1 the difference was significant at the 95% but not the 99% level (see Figure 25). In other words, the variation in AUC values on the test set was comparatively small, justifying the benefits of including domain knowledge in the model estimation.

AUC and DeLong Comparison at 95% Confidence Intervals

	AUC
Original - LR1	0.8117
Amended - LR2	0.7809

Contrast Coefficients

	LR1	LR2
Row1	1	-1

Tests and 95% Confidence Intervals for Contrast Rows

	Estimate	Std Error	Chi-square	P-value
Row1	0.0308	0.0060	26.2170	<.0001

Overall P-value <.0001

Figure 23. DeLong evaluation of training set for logistic regression – Telecom1

```
AUC and DeLong Comparison at 95% Confidence Intervals
                                  AUC
             Original - LR1       0.6628
             Amended - LR2        0.6456

                Contrast Coefficients
                           LR1       LR2
              Row1          1        -1

Tests and 95% Confidence Intervals for Contrast Rows
          Estimate Std Error Chi-square  P-value
   Row1  0.0172    0.0027     41.6591    <.0001

              Overall P-value <.0001
```

Figure 24. DeLong evaluation of training set for logistic regression – Telecom2

```
AUC and DeLong Comparison at 95% Confidence Intervals
                                  AUC
             Original - LR1       0.8291
             Amended - LR2        0.8082

                Contrast Coefficients
                           LR1       LR2
              Row1          1        -1

Tests and 95% Confidence Intervals for Contrast Rows
          Estimate Std Error Chi-square  P-value
   Row1  0.0208    0.0090      5.3194    0.0211

              Overall P-value 0.0211
```

Figure 25. DeLong evaluation of test set for logistic regression – Telecom1

```
AUC and DeLong Comparison at 95% Confidence Intervals
                                  AUC
             Original - LR1       0.6764
             Amended - LR2        0.6515

                Contrast Coefficients
                           LR1       LR2
              Row1          1        -1

Tests and 95% Confidence Intervals for Contrast Rows
          Estimate Std Error Chi-square  P-value
   Row1  0.0249    0.0141     3.0994     0.0783

              Overall P-value 0.0783
```

Figure 26. DeLong evaluation of test set for logistic regression – Telecom2

The increases and decreases in performance noted in the logistic regression model are due to the way the cut-off was set, using the KS statistic, where the cut-off varies from one model to the next, in order to have a balance between sensitivity and specificity. If

the same cut-off is chosen, for example 0.5, such a big variation in performance will not happen. For example, for Telecom1, the CA is 86.73% and 86.55% before and after domain knowledge analysis, respectively, but the imbalance between sensitivity and specificity increases. These results are presented in Table 29. Similar observations apply to Telecom2, as demonstrated in Table 30.

In both tables (see Table 29 and Table 30), the performance metrics are presented for the test set, following the cut-off schemes below:

- Based on the KS statistic (already explained in Chapter 4);
- Assuming a cut-off of 0.5;
- Assuming a cut-off based on the sample proportions. For example, Telecom1 presented 14.1% churners in the test set, so the cut-off was specified in such a way that almost exactly 14.1% of the observations are predicted as churners. This was done by scoring the test set customers, and then considering the 14.1% highest scores as churners. The proportion for Telecom2 is 1.8%.

	Metric (%)	KS Statistic as cut-off base	0.5 cut-off	Churn proportion as base for cut-off
Before Domain Knowledge	CA	73.65	86.67	84.51
	Sensitivity	83.83	11.49	45.11
	Specificity	71.98	99.02	90.99
After Domain Knowledge	CA	70.35	86.61	83.43
	Sensitivity	82.13	14.47	41.28
	Specificity	68.41	98.46	90.36

Table 29. Logistic regression performance measures for different cut-offs on the test set – Telecom1

Data Set	Metric (%)	KS Statistic as cut-off base	0.5 cut-off	Churn proportion as base for cut-off
Before Domain Knowledge	CA	51.23	57.99	96.75
	Sensitivity	77.78	65.56	10.00
	Specificity	50.75	57.85	98.35
After Domain Knowledge	CA	64.01	59.01	96.55
	Sensitivity	64.44	66.67	4.44
	Specificity	64.00	58.87	98.24

Table 30. Logistic regression performance measures for different cut-offs on the test set – Telecom2

From these tables, it is possible to conclude that if the same cut-off is chosen before and after domain knowledge evaluation, then there is no big variation in performance between them. However, it is also possible to notice that the balance between sensitivity and specificity ceases to exist for Telecom1 when the KS statistic is not used as the cut-off. Telecom2 still presents some balance between sensitivity and specificity at the 0.5 cut-off, but not when the churn proportion is used as the base for cut-off.

6.6. Conclusions

In this chapter, it was investigated how to make data mining models that are developed for churn prediction more understandable and compliant with domain knowledge. More specifically, it was shown how the analysis of coefficient signs in logistic regression can be used to check whether the knowledge contained in the data mining models is in accordance with domain knowledge, and how to correct any discrepancies found.

The idea is not only to help companies discover which customers are more valuable or will churn, but to help them identify the main elements in their data that can contribute positively or negatively to the relationship with the customer, and through that, define strategies that would benefit both company and customer alike.

In conclusion, logistic regression is a suitable choice of classifier for integrating domain knowledge into the model, as in the analysis of both data sets, the model's performance stayed relatively stable with the introduction of domain constraints. The shown procedures help to ensure that the variables presented in both models are presenting the expected relationship with the target variable, with only limited loss of predictive power.

Chapter 7. Decision Tables for Domain Knowledge Integration into Data Mining Models

7.1. Introduction

This chapter explores another main contribution of this research, the use of decision tables to integrate domain knowledge into data mining. More specifically, decision tables are used to evaluate the rules generated by a decision tree model when evaluating the probability of churn.

To proceed with this, first an explanation of decision tables is given, followed by an exploration of its use in the literature, and an introduction to the software used for its analysis in this research. Then, an explanation of how domain knowledge will be evaluated is presented in the form of pseudo-code, followed by an empirical study using the same data sets as used in Chapter 6. Conclusions are then drawn on the methods applied.

7.2. Visualisation of Rules using Decision Tables

Decision tables (DTs) are a tabular representation used to describe and analyse decision situations, for example churn evaluation, where the state of a number of conditions jointly determines the execution of a set of actions (Vanthienen and Wets, 1994). In this context, the conditions correspond to the antecedents of the rules whereas the actions correspond to the outcome from the rules, in this case, churn or not churn. A DT consists of four quadrants, separated by thicker lines horizontally and vertically, in accordance with Figure 27. The horizontal line divides the table into a condition and an action part. The vertical line separates subjects from entries.

condition subjects	condition entries
action subjects	action entries

Figure 27. DT quadrants

The condition subjects are the criteria that are relevant to the decision making process. They represent the attributes of the rules about which information is needed to classify a

given customer as churner or non-churner. The action subjects describe the classes of the classification problem (churn, not churn), which are the possible outcomes of the decision making process. Each condition entry describes a relevant subset of values for a given condition subject (attribute), or contains a dash symbol ('-') if its value is irrelevant within the context of that column. Subsequently, every action entry holds a value assigned to the corresponding action subject, with an 'x' entry indicating which value applies to a particular combination of conditions. Therefore, every column in the entry part of the DT comprises a classification rule, indicating what action or actions apply to a certain combination of condition states.

If each column only contains simple states, the table is called an expanded DT. If the table contains contracted or irrelevant entries, it is called a contracted DT. Table contraction can be achieved by combining columns that lead to the same action configuration. The number of columns in the contracted table can then be further minimised by changing the order of the conditions, which provides a more compact and comprehensible representation of the extracted knowledge. This can be seen in Figure 28, which shows the three situations described above.

1. International Calls	Y				N			
2. Usage of Voice-Mail	Y		N		Y		N	
3. Usage of Video-Calls	Y	N	Y	N	Y	N	Y	N
1. Churn	-	x	x	x	-	x	-	x
2. Not Churn	x	-	-	-	x	-	x	-

(a) Expanded DT

1. International Calls	Y			N	
2. Usage of Voice-Mail	Y		N	-	
3. Usage of Video-Calls	Y	N	-	Y	N
1. Churn	-	x	x	-	x
2. Not Churn	x	-	-	x	-

(b) Contracted DT

1. Usage of Video-Calls	Y		N
2. International Calls	Y	N	-
3. Usage of Voice-Mail	Y	N	-	-
1. Churn	-	x	-	x
2. Not Churn	x	-	x	-

(c) Minimised DT

Figure 28. Minimising the number of columns of a DT (based on Vanthienen and Wets, 1994)

This research is intentionally restricted to single-hit tables, in which columns have to be mutually exclusive, because of their advantages with respect to verification and validation (Vanthienen *et al.*, 1998a). So, this research will require that the condition entry part of a DT satisfies the following two criteria, a) Completeness: all possible combinations of condition values are included; b) Exclusivity: no combination is covered by more than one column.

This type of DT that can be easily checked for potential anomalies, such as inconsistencies (a particular case being assigned to more than one class) or incompleteness (no class assigned). An example that demonstrates both inconsistency and incompleteness is demonstrated below. Take the following example rule set:

- Rule 1 (R1): If Average Usage < 25 and International Plan = Y and Service Calls ≥ 3, Then Churn
- Rule 2 (R2): If Average Usage < 25 and International Plan = N, Then Churn
- Rule 3 (R3): If Average Usage ≥ 25 and International Plan = Y, Then Not Churn
- Rule 4 (R4): If Average Usage < 25 and Service Calls < 3, Then Not Churn

The resulting DT is shown in Table 31.

1. Average Usage	< 25				≥ 25			
2. International Plan	Y		N		Y		N	
3. Service Calls	< 3	≥ 3	< 3	≥ 3	< 3	≥ 3	< 3	≥ 3
1. Churn	-	x	x	x	-	-	-	-
2. Not Churn	x	-	x	-	x	x	-	-
Contributing rule (s)	R4	R1	R2 R4	R2	R3	R3		

Table 31. DT containing anomalies – example for DT verification

When these rules are transferred to the DT, it is possible to identify the following anomalies:

- Conflicting action entries: rules R2 and R4 contradict each other for Average Usage < 25, International Plan = N and Service Calls < 3
- Missing entries: no class specified for Average Usage ≥ 25 and International Plan = N

In these cases, it was easy to identify the conflicting and missing rules, making DT a useful tool for verification. Another point to be noted is that R3 is represented in two action entries, which is not considered an anomaly, but demonstrates that the DT can be contracted and become simpler. Consider also that R4 is adapted and another two rules are created as follows:

- Rule 4 (R4): If Average Usage < 25 and International Plan = Y and Service Calls < 3, Then Not Churn
- Rule 5 (R5): If Average Usage ≥ 25 and International Plan = N and Service Calls < 3, Then Not Churn
- Rule 6 (R6): If Average Usage ≥ 25 and International Plan = N and Service Calls ≥ 3, Then Churn

Applying these three rules together with R1, R2 and R3, and also contracting the final table, the resulting DT is presented in Table 32. From this table, it is possible to verify that all anomalies were eliminated.

1. Average Usage	< 25			≥ 25		
2. International Plan	Y		N	Y	N	
3. Service Calls	< 3	≥ 3	-	-	< 3	≥ 3
1. Churn	-	x	x	-	-	x
2. Not Churn	x	-	-	x	x	-
Contributing rule (s)	R4	R1	R2	R3	R5	R6

Table 32. DT from Table 31 with eliminated anomalies

Additionally, for simplicity of legibility, the columns are arranged in lexicographical order, in which entries at lower rows alternate first. Consequently, a tree structure emerges in the condition entry part of the DT, which led itself to a top-down evaluation procedure: starting at the first row, and then working its way down the table by choosing from the relevant condition states, it is possible to safely arrive at the set action for a given case. This condition-oriented inspection approach often proves more intuitive, faster, and less prone to human error, than evaluating a set of rules one by one.

As a result, the DT formalism thus allows for easy validation by a practitioner or domain expert of the knowledge extracted by a rule- or tree-based classification

technique (such as C4.5, CART, CHAID) against their own domain knowledge and prior expectations.

Another tool that could be useful in the DT evaluation is decision diagrams (Mues *et al.*, 2004).They are a graph-based representation of discrete functions, accompanied by a set of graph algorithms that implement operations on these functions.

A well-known property that can undermine the conciseness and interpretability of DT is the inherent replication of identical subparts. For example, see the DT representation in Table 33.

1. Service Calls	< 3											≥ 3
2. Day Charge	< 45	≥ 45										-
3. International Charge	-	< 20					≥20 and < 35				≥ 35	-
4. International Plan	-	Yes				No	-				-	-
5. Evening Charge	-	< 25			≥ 2.5	-	< 25			≥ 2.5	-	-
6. Night Charge	-	< 12	≥ 12		-	-	< 12	≥ 12		-	-	-
7. Voice-Mail Plan		-	Yes	No	-	-	-	Yes	No	-	-	-
1. Churn	-	x	x	-	x	-	x	x	-	x	x	x
2. Not Churn	x	-	-	x	-	x	-	-	x	-	.	-

Table 33. DT with replication of subparts

By transforming the DT into a decision diagram, recurring parts are shared through multiple incoming edges, hereby giving an even smaller representation. This is demonstrated in Figure 29. If simplification of visualisation is required, the use of decision diagrams would be a suitable choice.

In this research however, decision diagrams will not be used, as the purpose is the interpretation of each individual rule generated, through the use of domain knowledge. This will be done with the change in the condition order for each individual variable, which will be further discussed in section 7.5.

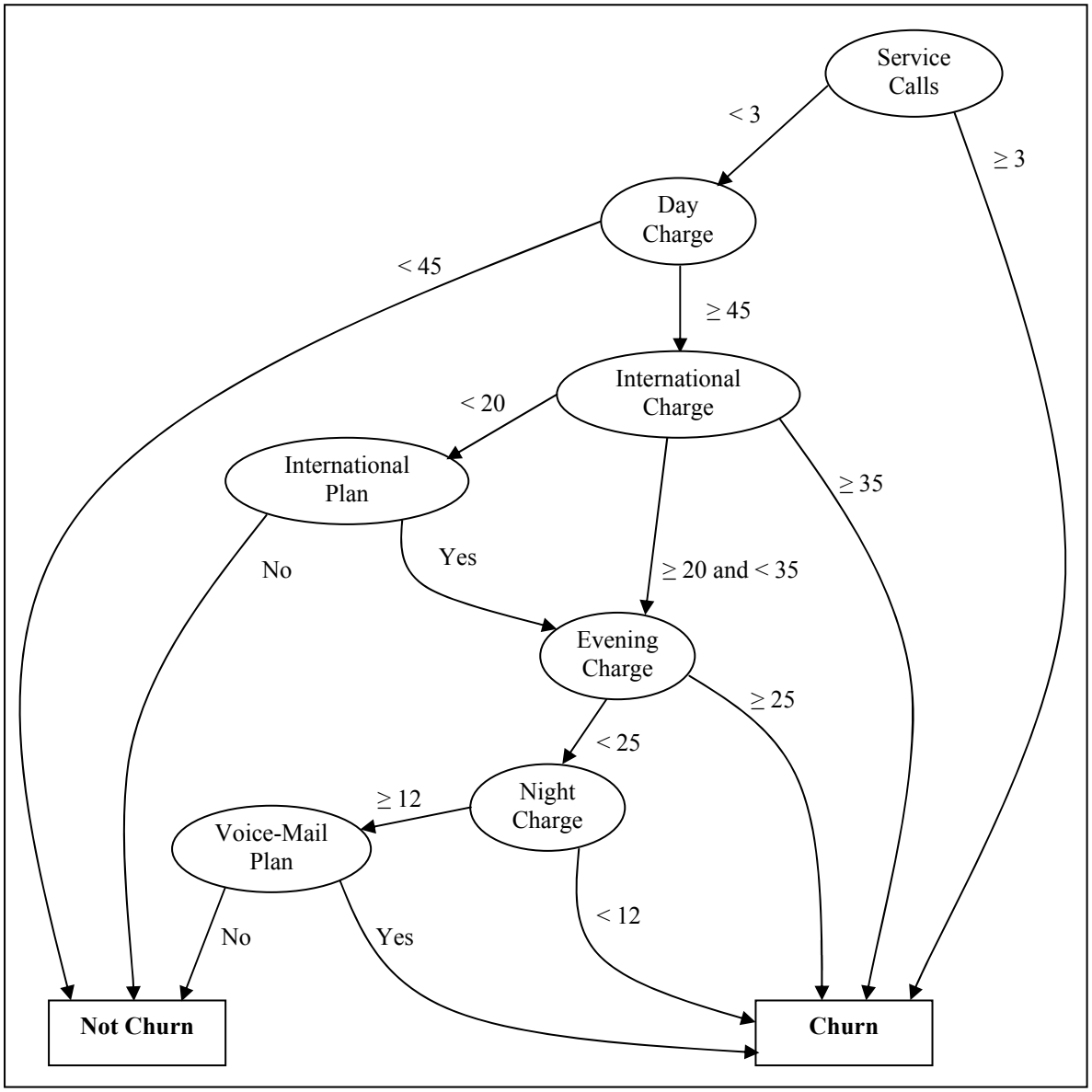

Figure 29. Decision diagram representation of DT in Table 33

7.3. Decision Table Applications in the Literature

Based on the literature on DTs, some of their applications are listed in Table 34.

Vanthienen and Wets (1994) argue that once the DT has been approved by the expert, it could be incorporated into a deployable expert system. In this case, domain knowledge is incorporated in the initial process, when creating the DTs for the generation of knowledge. They used the tool Prologa, which has been developed for the construction of DTs, being able to construct expanded, contracted and minimised DTs. Then, they applied DTs to generate, verify and validate knowledge bases, also transforming the DT's rules into knowledge bases.

Source	Number of Columns (Examples)	Context	Purpose of Application
Vanthienen and Wets (1994)	Expanded: 8 Contracted: 6 Optimised: 4	Generation, verification and validation of knowledge bases.	Use of DTs to generate knowledge, based on previous domain knowledge.
Wets *et al.* (1997)	Worst case: 7,776 Contracted: 722	Verification and validation of knowledge bases.	Use of DTs to model knowledge from extracted rules, based on previous feature selection.
Kohavi and Sommerfield (1998)	No specification of contraction	Knowledge discovery	Use of DTs to facilitate understanding of data.
Vanthienen *et al.* (1998a)	No examples presented	Verification and validation of knowledge bases.	Use of Prologa to verify and validate DTs.
Vanthienen *et al.* (1998b)	No data presented	Verification and validation of knowledge bases.	Use of Prologa to verify and validate inter-tabular DTs.
Baesens *et al.* (2003a)	Expanded: 6,600 Contracted and minimised: 11	Credit scoring	Use of DTs to visualise Neural Network Rules.
Hewett and Leuchner (2003)	Normal: 24 Compressed: 7	Knowledge acquisition	Use of DTs to manage and generate knowledge.
Piramuthu (2004)	Normal: 17 Reduced: 6 Further reduced: 4	Knowledge discovery	Use of feature construction to compact and construct DT without loss.

Table 34. Literature review on decision tables (DT)

DTs provide an alternative method of representing data mining knowledge in a user-friendly way (Baesens *et al.*, 2003a; Wets *et al.*, 1997). Wets *et al.* (1997) use DTs to model knowledge from extracted rules, also using initial feature selection to reduce complexity of stored knowledge and removal of unnecessary features. They then use the DTs for verification and validation of knowledge bases, in order to detect anomalies in a proper way, facilitating their correction. They use the concept of contracted DTs (using Prologa), which makes it possible to reduce considerably the number of rules in their analysis, making them easier to be evaluated. This research will also use contracted DTs, as it can facilitate the interpretability of rules.

Vanthienen *et al.* (1998a) use the Prologa software (it will be described in section 7.4) to construct DTs, in order to aid the verification and validation of knowledge bases, facilitating the knowledge acquisition process in the modelling phase of knowledge based systems' development. Vanthienen *et al.* (1998b) also use Prologa to check for inter-tabular anomalies, which are found in more than one DT that are inter-connected. In both papers they use the concept of contracted DTs, and the DT evaluation is used to prevent anomalies when creating and evaluating the DTs' rules.

Kohavi and Sommerfield (1998) use DTs to facilitate understanding of extracted data, using only a list of key attributes. The DTs are used instead of other data mining methods, such as decision trees and naïve bayes, in order to facilitate interpretation and results. They do not explore the concept of minimising or contracting DTs, nor are DTs used for individual evaluation of rules. In their case, they use two methods called DTMaj (decision table majority) and DTLoc (decision table local), where the first returns the majority of the class in the DT and the second uses the local neighbourhood to return their majority answer. These algorithms were used for classification tasks and for many of the data sets investigated they presented similar accuracy to C4.5 classification trees. When comparing the two algorithms, each of them performed better for some of the data sets, but in overall their performance was very similar. Their approach exemplifies a different application for DTs; however, it is not in accordance with the approach that will be adopted in this research.

Baesens *et al.* (2003a) analyse credit scoring data using neural network rule extraction techniques. However, the resulting rules lack easy interpretation, making it difficult to justify any decision making. As a result, they use DTs to visualise the neural network's

rules, in a format that is easier to comprehend and to be verified by managers. Their approach is in accordance with the one that will be used in this research, where the rules from a data mining model will be transferred to a DT, in order to facilitate understanding and interpretation of the rules.

Hewett and Leuchner (2003) state that DTs represent complex logic in a familiar and understandable way. They use DTs in a knowledge acquisition process to manage previous domain knowledge and then generate knowledge to be used in knowledge-based systems. They use a learning system called SORCER that induces second-order DTs, where an attribute can assume a set of finite values when providing a specific outcome. They use a specific compression approach to minimise the DTs, which could be equivalent to the contraction and minimisation adopted for single-hit DTs. Table 40 in section 7.6 demonstrates that this research will use a similar approach when necessary, minimising the DT with the incorporation of second-order and single-hit DTs in one analysis.

An interesting point to make based on the works above is that DTs can be used for two different purposes. There are methods that can be used to extract DTs from the data and those that are used for representing expected knowledge directly in the form of DTs. In the first category are, for example, the algorithms used by Kohavi and Sommerfield (1998) and Hewett and Leuchner (2003), which are used for knowledge extraction directly from the data. In the representation category is the Prologa software, which is used by Vanthienen and Wets (1994), Vanthienen *et al.* (1998a) and Baesens *et al.* (2003a).

Piramuthu (2004) also proposes a method for extraction of DTs from the raw data. He uses a framework of feature construction to compact and construct DTs, in a way so as to avoid complexity and loss of information. This feature construction builds new features based on existing ones, and these new features are also used in the construction of the DT. Only the main features are kept in the construction. In this case, the DT created will be able to facilitate the extraction of knowledge, as its complexity will be reduced based on a smaller number of features used. They also compact the DT, by eliminating redundant rules and merging some together. Their approach involves the reduction of rules to a minimum, which can be achieved also by the feature selection, combining different attributes in one row. This would be useful if the amount of original

rules is too large; if only a few rules exist (seven rules for their example), it would be better to keep the compacted rules, in order to avoid complexity of interpretation.

All these previous research papers present some key elements that need to be considered when using DTs. Regarding DTs in this research, domain knowledge will be evaluated after the construction of DTs. Feature selection (in line as much as possible with the domain knowledge) is performed beforehand (as explained in Chapter 6), in order to reduce complexity and provide more understandable results. DTs will then be used to aid the domain knowledge evaluation of data mining methods; in this specific case, the analysis of the decision tree model.

7.4. Prologa - Software for the Decision Table Analysis

The Prologa software was used to construct the DTs for the rules extracted in the empirical analysis (section 7.6). Prologa is an interactive design tool for computer-supported construction and manipulation of DT's (Vanthienen and Dries, 1994). Prologa acquires and verifies knowledge in the form of DTs.

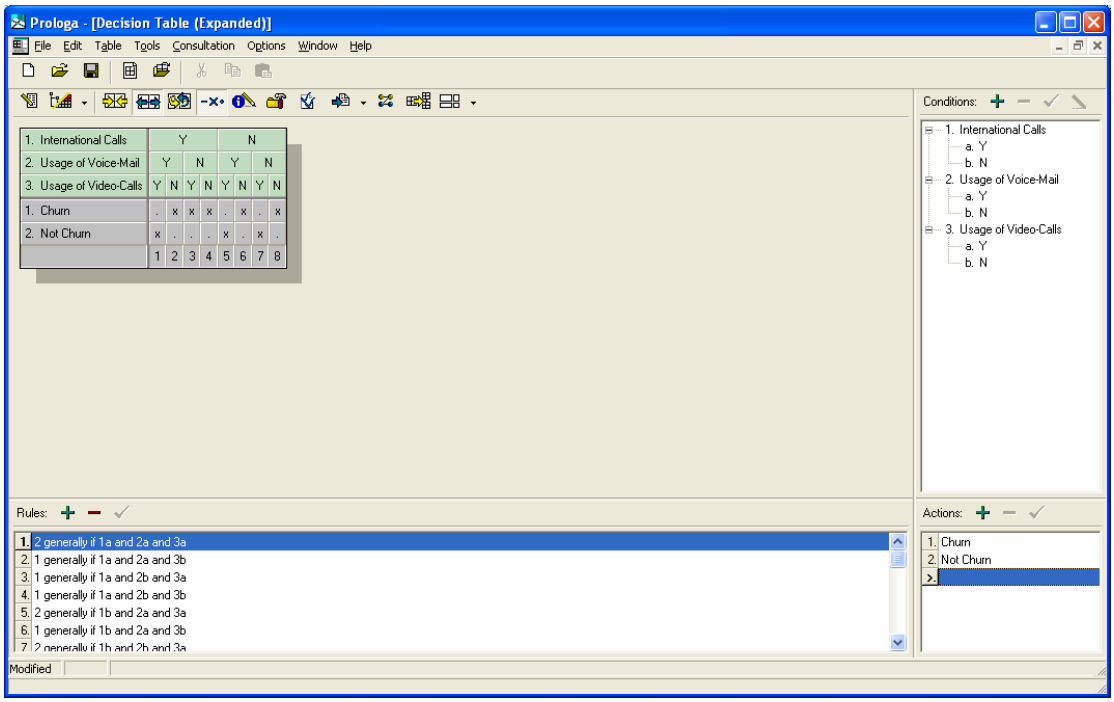

Figure 30. Expanded DT in Prologa from Figure 28

It allows the evaluation of expanded and contracted DTs (in accordance with Figure 30 and Figure 31), as well as easy manipulation of the condition subjects, which are essential for the domain knowledge analysis (see Figure 32).

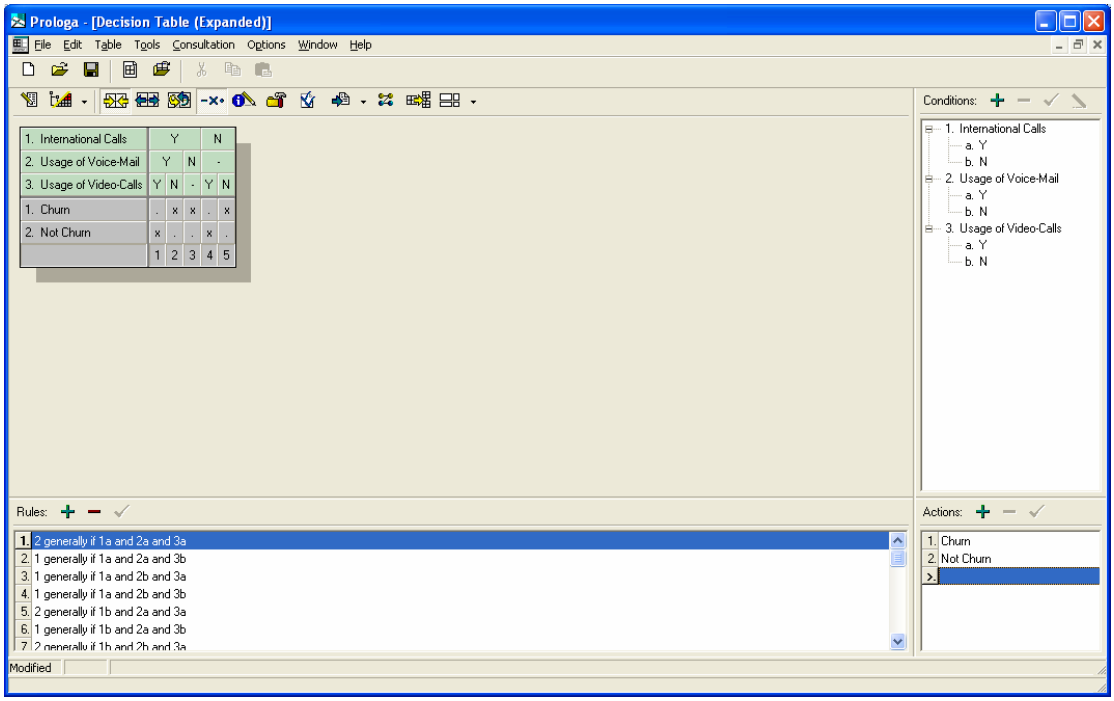

Figure 31. Contracted DT in Prologa from Figure 30

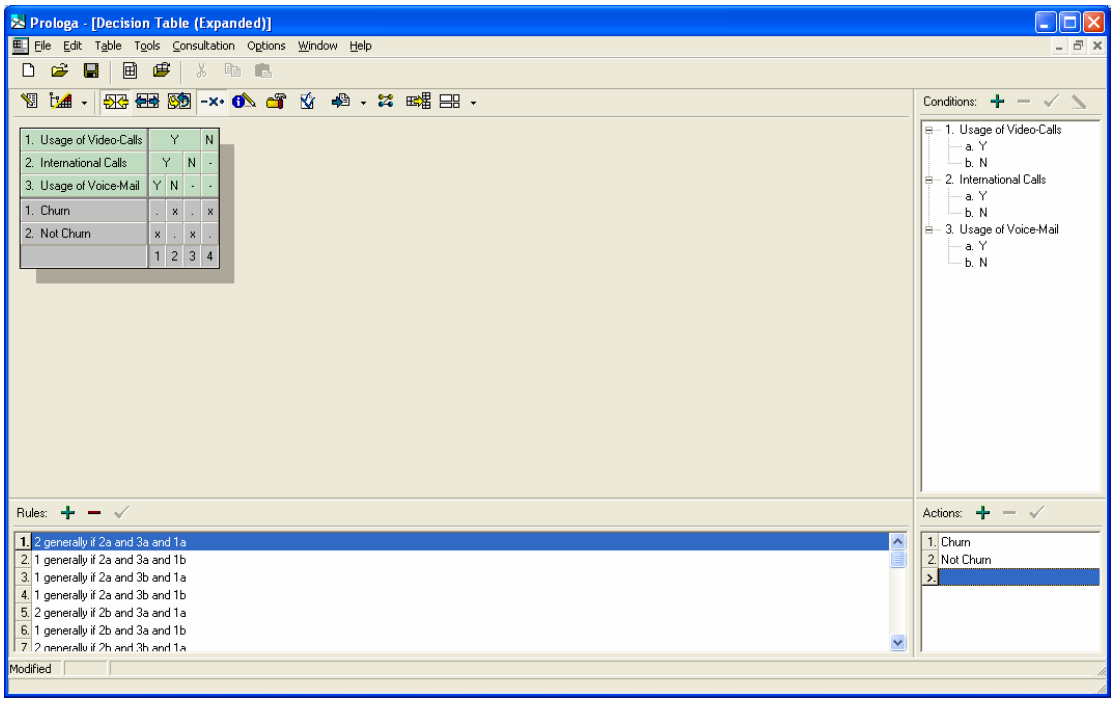

Figure 32. Changing of sequence of condition subject from Figure 31

This can be seen in Figure 32, where the condition subject “International Calls” was moved down in the condition order, with the purpose to provide a further minimised

DT. More information about this software can be found at the website (http://www.econ.kuleuven.ac.be/Prologa/, accessed 3 September 2007).

7.5. Integrating Domain Knowledge with Decision Tables

Domain knowledge was applied by evaluating the rules generated from the decision tree model, which were then transferred to the DT. The DT was used to check for expected effects of variables one-by-one, evaluating their monotonicity by moving them down in the condition order. This was done based on the algorithm provided in Figure 33. This algorithm allows verifying if the rules extracted from the DT are in agreement with the domain knowledge applied.

For each condition subject CS_i (i = 1, ..., c, where c is the number of condition subjects):
  If a domain constraint D_i has been defined over CS_i then
    Begin
      Move CS_i to last position in condition order;
      Construct contracted decision table;
      For each adjacent group of columns having identical condition entries for condition rows 1, ..., c-1, and having condition entries for CS_i that are different from '-', flag action value changes that violate domain constraint D_i;
    End;

Figure 33. Algorithm to investigate monotonicity of a DT

Following this pseudo-code, the rules can be reorganised to provide a better set of easier to understand and more accurate rules. The domain knowledge analysis is done following each one of the rules, and the results then compared to see if any inconsistencies were generated. In case any inconsistence is found, the problem is treated and the rules re-evaluated.

Two methods were defined to treat the inconsistencies. Firstly, after identifying the counter-intuitive rules, the action entries would be changed in the DT. Secondly, after identifying the counter-intuitive rules, instead of simply changing the action entries in the DT, the condition term would be removed altogether, and the table would be contracted accordingly. This can be motivated as follows. Suppose there is a rule ‘if service calls < 3 then churn’, which is not intuitive, then there is no motivation why this cut-off would remain when changing the inequality of signs. So, instead of having two

condition entries < 3 and ≥ 3, a contracted entry '-' would exist, and the predicted class would be taken as the majority class for that subsample. Both approaches will be applied and their results discussed in subsections 7.6.1 and 7.6.2.

7.6. Empirical Evaluation

Both data sets prepared for the logistic regression evaluation are used for the analysis of DT and domain knowledge. The data sets were also evaluated using the SAS 9.1 Solution and Enterprise Miner. For the decision tree assessment, both training sets were further divided into training and validation sets, keeping the original proportion of churners and non-churners. The validation set was used for pruning, in order to avoid overfitting of the tree.

After scoring the results of both data sets, the performance metrics are shown in Table 35 (these values were previously demonstrated in Table 13, page 86). Again, to analyse the performance of the model, the measurements used were the AUC, and the CA, sensitivity and specificity assuming the KS statistic as the base for cut-off.

Data Set	Metric	Training Set (%)	Test Set (%)
Telecom1	CA	89.44	90.64
	Sensitivity	71.61	72.34
	Specificity	92.38	93.64
	AUC	84.21	84.94
Telecom2	CA	60.58	52.30
	Sensitivity	68.20	74.44
	Specificity	52.96	51.89
	AUC	64.20	64.25

Table 35. Decision tree performance measurements

Once more, with traditional churn evaluation, the results obtained with the decision tree model, if better than other models, would automatically be selected for any analysis. However, the purpose of this research is to find a model with good predictive ability, but also making sure that the resulting variables are compliant with the domain knowledge. As a result, the outcome from the decision tree model was re-evaluated for both data sets.

7.6.1. Telecom1 – Domain Knowledge Evaluation

The tree model resulting from the analysis of Telecom1 is showed in Figure 34. The tree was reduced, limiting the number of leaves, in order to obtain a more understandable tree. This limitation was done based on the misclassification rates, where only the most predictive attributes were chosen, with a minimal increase on misclassification. The idea is to have the lowest misclassification rate. A better classification would not be useful if it could not be understood and if the tree overfitted. This tree gives the performance estimates in Table 35.

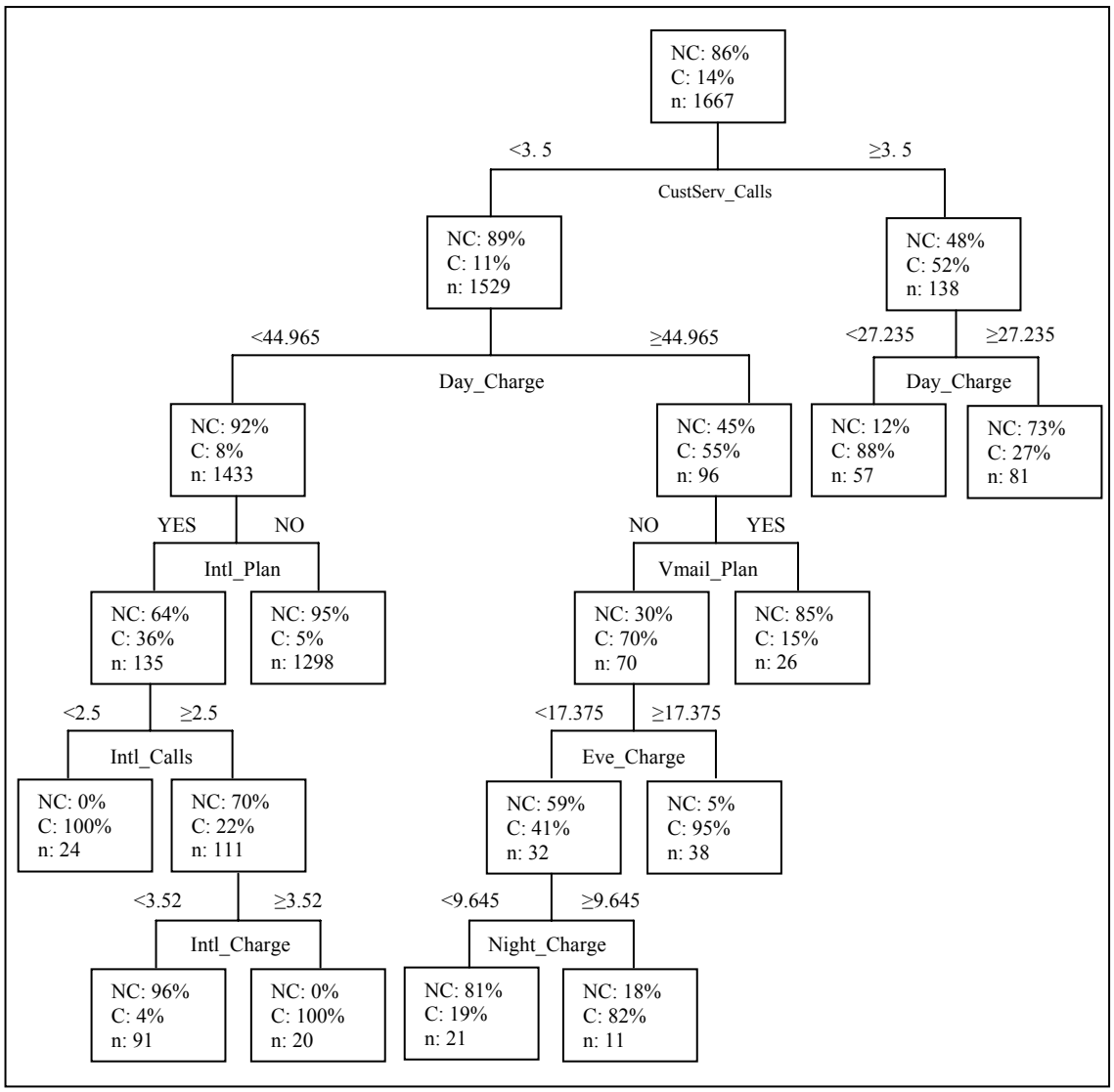

Figure 34. Decision tree result – Telecom1

Consequently, the validation set was used to control the generalisation of rules resulting from the analysis, avoiding that the decision tree characterises too much detail in the training data. Therefore, the tree was pruned based on the balance of misclassification

rates between training and validation sets, and the number of variables was chosen based on this evaluation. As a result, the most predictive variables were kept in the tree, which is the purpose of this analysis.

To facilitate examining and validating its rules, the decision tree was first converted to a DT (see Table 36), which was then checked for violations against domain knowledge using the procedure described in Figure 33 (section 7.5).

1. CustServ_Calls	< 3.5								≥ 3.5	
2. Day_Charge	< 44.965				≥ 44.965				< 27.235	≥ 27.235
3. Intl_Plan	Yes			No	-				-	-
4. Intl_Calls	< 2.5	≥ 2.5		-	-				-	-
5. Intl_Charge	-	< 3.52	≥ 3.52	-	-				-	-
6. Vmail_Plan	-	-	-	-	Yes	No			-	-
7. Eve_Charge	-	-	-	-	-	< 17.375		≥ 17.375	-	-
8. Night_Charge	-	-	-	-	-	< 9.645	≥ 9.645	-	-	-
1. Churn	x	.	x	.	.	.	x	x	x	.
2. Not Churn	.	x	.	x	x	x	.	.	.	x

Table 36. DT rules from the decision tree model (Figure 34)

The DT was used to check for expected effects of variables one-by-one, moving them down in the condition order. Assuming the same approach regarding its relationship with the target variable as in Chapter 6, counter-intuitive rules were identified for the variable "CustServ_Calls", as specified by the arrows in Table 37. It showed that in some parts of the table, if there is an increase in customer service calls, the prediction will change at some point from churn to not churn instead of vice-versa. However, domain knowledge suggests that the probability of churning should increase rather than decrease with an increase in customer calls. To remedy this, two different techniques were applied, change of action entries and term removal (as described in section 7.5).

Applying the first option means one can also directly change the action entries to make the DT in line with domain knowledge. In this case, the DT was used to change the action entries for the variable "CustServ_Calls", providing consistent rules all through the DT. The resulting DT can be seen in Table 38.

Counter-intuitive rules were also identified for the variable "Day_Charge". It shows that in some parts of the table, if there is a decrease in the amount the customer pays, the prediction will change at some point from not churn to churn instead of vice-versa.

However, domain knowledge suggests that the probability of churning should decrease rather than increase with a decrease in the amount paid by the customer. Nevertheless, as discussed previously in this chapter and in Chapter 5, not all the rules need and should be changed. The idea is to provide an understandable way of exploring the outcome of the technique, getting good and compliant results from it, but there will always be exceptions in the analysis. In this specific case, changing the values for the variable "Day_Charge" would generate too many rules and would also interfere in the monotonicity of other variables. This is an example where it is not recommended to change the rules. Finally, with the variables in general presenting a more intuitive set of values, the final set of rules (see Table 38) follows the idea of using the DT to manipulate the rules according to the domain knowledge.

With the second option, instead of directly changing the action entries to make the DT in line with domain knowledge, the condition term would be removed altogether. As stated in the methodology section, instead of having two condition entries < 3.5 and ≥ 3.5 for "CustServ_Calls" where counter-intuitive rules exist, the condition entries were combined into one (indicated by '-'), and the predicted class was taken as the majority class for that subsample. The resulting DT can be seen in Table 39.

To obtain the confidence for these new rules, the training set was used, applying each rule to all records. For example, suppose rule one says, if "A" and "B" then churn = yes, and if it is found that one-hundred customers for whom "A" and "B" is true, but only eighty are churners; then this rule has a confidence of 80%. However, if another rule says, If "C" and "B" then churn = no, then the confidence will be "1 – confidence", since the target in this case is not churn. A simpler way to explain this is that, as the target variable is churn, the confidence can be obtained by dividing the amount of real churners (classified by the rule) by the total amount of customers (classified by the rule, independently of being churners or not). The results in both cases will be the same.

After obtaining the new confidences, these were then applied to both training and test sets, in order to obtain the CA, sensitivity and specificity based on the KS statistic, and the AUC. These measurements are presented in Table 44 in section 7.7, showing the amended measures against the original measures.

1. Day_Charge	< 27.235						(≥ 27.235) - (< 44.965)						≥ 44.965					
2. Intl_Plan	Yes				No		Yes					No	-					
3. Intl_Calls	< 2.5	≥ 2.5			-		< 2.5		≥ 2.5			-	-					
4. Intl_Charge	-	< 3.52		≥ 3.52	-		-		< 3.52		≥ 3.52	-	-					
5. Vmail_Plan	-	-		-	-		-		-		-	-	Yes	No				
6. Eve_Charge	-	-		-	-		-		-		-	-	-	< 17.375			≥ 17.375	
7. Night_Charge	-	-		-	-		-		-		-	-	-	< 9.645		≥ 9.645	-	
8. CustServ_Calls	-	< 3.5	≥ 3.5	-	< 3.5	≥ 3.5	< 3.5	≥ 3.5	-	< 3.5	≥ 3.5	-	-	-	< 3.5	≥ 3.5	< 3.5	≥ 3.5
1. Churn	x	.	x	x	.	x	x ←	.	.	x →	.	.	.	.	x →	.	x →	.
2. Not Churn	.	x	.	.	x	.	. →	x	x	←	x	x	x	x	. ←	x	. ←	x

Table 37. Table 36 reordered – variable "CustServ_Calls" moved to last row

1. Day_Charge	< 27.235						(≥ 27.235) - (< 44.965)						≥ 44.965					
2. Intl_Plan	Yes				No		Yes					No	-					
3. Intl_Calls	< 2.5	≥ 2.5			-		< 2.5		≥ 2.5			-	-					
4. Intl_Charge	-	< 3.52		≥ 3.52	-		-		< 3.52		≥ 3.52	-	-					
5. Vmail_Plan	-	-		-	-		-		-		-	-	Yes	No				
6. Eve_Charge	-	-		-	-		-		-		-	-	-	< 17.375			≥ 17.375	
7. Night_Charge	-	-		-	-		-		-		-	-	-	< 9.645		≥ 9.645	-	
8. CustServ_Calls	-	< 3.5	≥ 3.5	-	< 3.5	≥ 3.5	< 3.5	≥ 3.5	-	< 3.5	≥ 3.5	-	-	-	< 3.5	≥ 3.5	< 3.5	≥ 3.5
1. Churn	x	.	x	x	.	x	-	x	.	-	x	.	.	.	-	x	-	x
2. Not Churn	.	x	.	.	x	.	x	-	x	x	-	x	x	x	x	-	x	-

Table 38. DT rules with changed action entries for variable "CustServ_Calls" – Telecom1

1. Day_Charge	< 27.235						(≥ 27.235) and (< 44.965)				≥ 44.965			
2. Intl_Plan	Yes				No		Yes			No	-			
3. Intl_Calls	< 2.5	≥ 2.5			-		< 2.5	≥ 2.5		-	-			
4. Intl_Charge	-	< 3.52		≥ 3.52	-		-	< 3.52	≥ 3.52	-	-			
5. Vmail_Plan	-	-		-	-		-	-	-	-	Yes	No		
6. Eve_Charge	-	-		-	-		-	-	-	-	-	< 17.375		≥ 17.375
7. Night_Charge	-	-		-	-		-	-	-	-	-	< 9.645	≥ 9.645	-
8. CustServ_Calls	-	< 3.5	≥ 3.5	-	< 3.5	≥ 3.5	-	-	-	-	-	-	-	-
1. Churn	x	.	x	x	.	x	x	.	x	.	.	.	x	x
2. Not Churn	.	x	.	.	x	.	.	x	.	x	x	x	.	.

Table 39. DT rules with term removed for variable "Cust_Serv_Calls" – Telecom1

7.6.2. Telecom2 – Domain Knowledge Evaluation

The tree model resulting from the analysis of Telecom2 is shown in Figure 35, making it possible to evaluate and extract the decision rules. The rules were then transferred to a DT to facilitate the examination and validation of the rules (see Table 40). It was also useful to visualise that with the contracted table, some rules are hidden by the model, facilitating the analysis.

1. eqpdays	< 467.5			≥ 467.5			
2. hnd_price	G3		G2, G1	-			
3. hnd_webcap	WC,NA	WCMB	-	WCMB,WC	NA		
4. avgmou	-	-	-	-	< 22.425		≥ 22.425
5. kid3_5	-	-	-	-	Y,UK	U	-
1. Churn	x	.	.	x	x	.	x
2. Not Churn	.	x	x	.	.	x	.

Table 40. DT rules from the decision tree model (Figure 35)

The DT analysis again revealed counter-intuitive rules, for example, when investigating the rules relating to the variable "avgmou". Table 41 shows the conflicting rules, where, conversely to what was expected from the domain knowledge, the customer becomes more likely to churn if there is an increase in the average usage.

1. eqpdays	< 467.5			≥ 467.5			
2. hnd_price	G3		G2, G1	-			
3. hnd_webcap	WC,NA	WCMB	-	WCMB,WC	NA		
4. kid3_5	-	-	-	-	Y,UK	U	
5. avgmou	-	-	-	-	-	< 22.425	≥ 22.425
1. Churn	x	.	.	x	x	↑ .	↓ x
2. Not Churn	.	x	x	.	.	↑ x	↓ .

Table 41. Table 40 reordered – variable "avgmou" moved to last row

To remedy these situations, the options of change of action entries and term removal were applied. The option of changing action entries proved successful in providing consistent rules for the variable "avgmou". The resulting DT can be seen in Table 42. The final set of rules were more in accordance with the domain knowledge, where the variables are presenting values that are more related to what is expected of them.

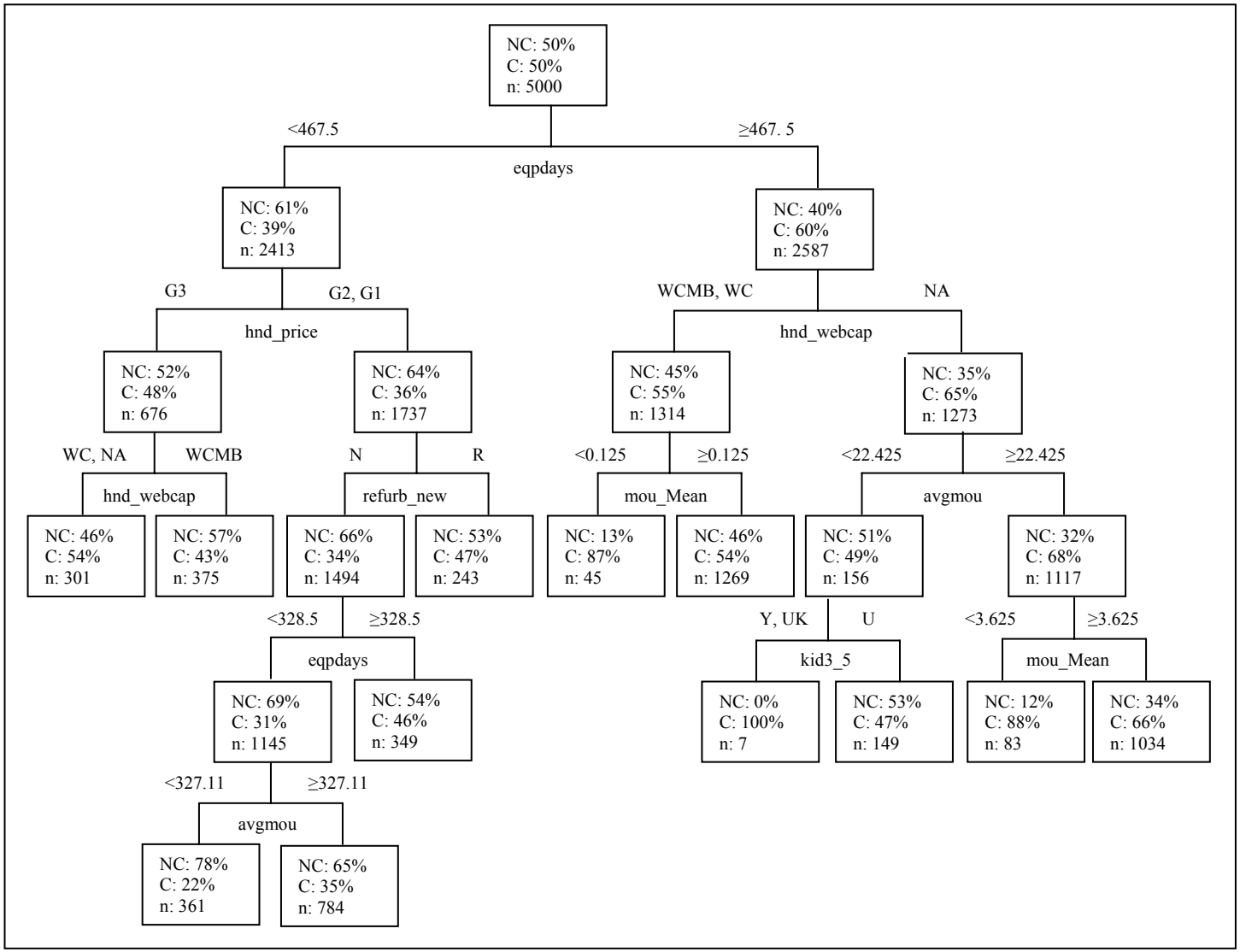

Figure 35. Decision tree result – Telecom2

1. eqpdays	< 467.5			≥ 467.5			
2. hnd_price	G3		G2, G1	-			
3. hnd_webcap	WC,NA	WCMB	-	WCMB,WC	NA		
4. kid3_5	-	-	-	-	Y,UK	U	
5. avgmou	-	-	-	-	-	< 22.425	≥ 22.425
1. Churn	x	.	.	x	x	x	.
2. Not Churn	.	x	x	.	.	.	x

Table 42. DT rules with changed action entries for variable "avgmou" – Telecom2

The analysis also identified counter-intuitive rules when investigating the variable "eqpdays" (age of handset). It shows that in some parts of the table, if there is an increase in the value of "eqpdays", the prediction will change at one point from churn to not churn instead of vice-versa. However, domain knowledge suggests that the probability of churning should increase rather than decrease with an increase in the age of the mobile equipment. Nevertheless, in the same context as with Telecom1, not all the rules need and should be changed. This would be the exception in this data set analysis. In this case, again, changing the values for the variable "eqpdays" would generate too many rules and would also interfere in the monotonicity of other variables. This is an example where the value of the variable is desired, but to make it mandatory would not be a simple and coherent task. The results would become complex, making it difficult to make the model interpretable.

Following the condition term removal approach, instead of directly changing the action entries to make the DT in line with domain knowledge, the term would be removed altogether. Again, instead of having two condition entries < 22.425 and ≥ 22.425 for the variable "avgmou" where counter-intuitive rules exist, they were replaced with a single '-' entry, and the predicted class was taken as the majority class for that subsample. The resulting DT can be seen in Table 43. In this case, the DT presented is a lot simpler, and the variable "eqpdays" now presents coherent values.

1. eqpdays	< 467.5			≥ 467.5
2. hnd_price	G3		G2, G1	-
3. hnd_webcap	WC, NA	WCMB	-	-
1. Churn	x	.	.	x
2. Not Churn	.	x	x	.

Table 43. DT rules with term removed for variable "avgmou" – Telecom2

In the end, with the variables generally presenting a more intuitive set of values, the new rules were then extracted (following the same procedure applied to Telecom1, in order obtain the new confidences) and applied to both training and test sets. As with Telecom1, the final performance measurements based on both approaches are reported in Table 44.

7.7. Final Results

From the final results in Table 44, it can be concluded that the performance measurements for the decision tree presented very good results when the domain knowledge was applied. The performances named "Action1" are related to the change of action entries procedure; the performances named "Action2" are related to the term removal procedure.

Data Set	Metric	Training Set (%)			Test Set (%)		
		Original	Action1	Action2	Original	Action1	Action2
Telecom1	CA	89.44	88.42	91.96	90.63	89.26	93.46
	Sensitivity	71.61	67.37	65.68	72.34	66.38	65.53
	Specificity	92.38	91.89	96.30	93.64	93.01	98.04
	AUC	84.21	85.79	86.85	84.94	84.78	86.35
Telecom2	CA	60.58	60.90	60.72	52.30	52.28	52.28
	Sensitivity	68.20	65.68	68.48	74.44	74.44	74.44
	Specificity	52.96	56.12	52.96	51.89	51.87	51.87
	AUC	64.20	63.44	61.44	64.25	64.44	64.09

Table 44. Original and amended decision tree performance measures

Regarding both training sets (Telecom1 and Telecom2), the DeLong *et al.* (1988) test showed that there was significant difference between the models (see Figure 36 and Figure 37). However for Telecom2, there was no significant difference between the original and action1 models, but both models were significantly different from the action2 model. After making the necessary changes to satisfy monotonicity constraints, the decision tree for Telecom1 even demonstrated a small improvement in the AUC performance, and using the DeLong *et al.* (1988) test, the difference was considered significant when comparing the original model against action2.

AUC and DeLong Comparison at 95% Confidence Intervals

	AUC
Original - DTREE1	0.8421
Action1 - DTREE2	0.8579
Action2 - DTREE3	0.8685

Contrast Coefficients

	DTREE1	DTREE2	DTREE3
Row1	1	-1	0
Row2	1	0	-1
Row3	0	1	-1

Tests and 95% Confidence Intervals for Contrast Rows

	Estimate	Std Error	Chi-square	P-value
Row1	-0.0158	0.0122	1.6621	0.1973
Row2	-0.0264	0.0093	8.0675	0.0045
Row3	-0.0106	0.0076	1.9609	0.1614

Overall P-value 0.0055

Figure 36. DeLong evaluation of training set for decision tree – Telecom1

AUC and DeLong Comparison at 95% Confidence Intervals

	AUC
Original - DTREE1	0.6420
Action1 - DTREE2	0.6344
Action2 - DTREE3	0.6144

Contrast Coefficients

	DTREE1	DTREE2	DTREE3
Row1	1	-1	0
Row2	1	0	-1
Row3	0	1	-1

Tests and 95% Confidence Intervals for Contrast Rows

	Estimate	Std Error	Chi-square	P-value
Row1	0.0076	0.0039	3.8021	0.0512
Row2	0.0277	0.0039	50.8616	<.0001
Row3	0.0200	0.0028	50.6569	<.0001

Overall P-value <.0001

Figure 37. DeLong evaluation of training set for decision tree – Telecom2

For both test sets, it was found that the incorporation of domain knowledge did not have an adverse effect on performance. Instead, after making the necessary changes to satisfy monotonicity constraints, the decision tree for Telecom2 even showed a small improvement in performance with the change of action entries (action1 model), but this performance increase was not significant according to the DeLong *et al.* (1988) test (see Figure 38).

AUC and DeLong Comparison at 95% Confidence Intervals

	AUC
Original - DTREE1	0.6425
Action1 - DTREE2	0.6445
Action2 - DTREE3	0.6409

Contrast Coefficients

	DTREE1	DTREE2	DTREE3
Row1	1	-1	0
Row2	1	0	-1
Row3	0	1	-1

Tests and 95% Confidence Intervals for Contrast Rows

	Estimate	Std Error	Chi-square	P-value
Row1	-0.0020	0.0138	0.0207	0.8856
Row2	0.0016	0.0138	0.0136	0.9072
Row3	0.0036	0.0104	0.1192	0.7299

Overall P-value 0.9420

Figure 38. DeLong evaluation of test set for decision tree – Telecom2

For Telecom1 test set, there was no improvement in performance regarding the action1 model, but the performance difference was also not significant according to the DeLong *et al.* (1988) test (see Figure 39), when comparing with the original model. The only significant difference at 95% level was between the two models changed by domain knowledge, but there was no significant difference if a 99% significance level was chosen.

AUC and DeLong Comparison at 95% Confidence Intervals

	AUC
Original - DTREE1	0.8494
Action1 - DTREE2	0.8478
Action2 - DTREE3	0.8635

Contrast Coefficients

	DTREE1	DTREE2	DTREE3
Row1	1	-1	0
Row2	1	0	-1
Row3	0	1	-1

Tests and 95% Confidence Intervals for Contrast Rows

	Estimate	Std Error	Chi-square	P-value
Row1	0.0016	0.0106	0.0231	0.8791
Row2	-0.0141	0.0093	2.3152	0.1281
Row3	-0.0157	0.0062	6.5109	0.0107

Overall P-value 0.0176

Figure 39. DeLong evaluation of test set for decision tree – Telecom1

For both data sets, the performance differences were not significant according to the DeLong *et al.* (1988) test, hence showing that the proposed procedure can produce an intuitive model without sacrificing performance.

For the decision tree, Table 45 and Table 46 show the performance metrics for the test set of both Telecom1 and Telecom2, based on the KS statistic, assuming a cut-off of 0.5 and also using the churn proportion as base for the cut-off. In the last case, because the tree only provides a limited number of scores, it might be that more (or less) of the 14.1% of the observations are predicted as churners for Telecom1. The same applies to Telecom2.

	Metric (%)	KS Statistic as cut-off base	0.5 cut-off	Churn proportion as base for cut-off
Before Domain Knowledge	CA	90.63	92.98	90.70
	Sensitivity	72.34	54.47	65.96
	Specificity	93.64	99.30	94.76
After Domain Knowledge (Action1)	CA	89.26	91.24	88.96
	Sensitivity	66.38	40.00	62.13
	Specificity	93.01	99.65	93.36
After Domain Knowledge (Action2)	CA	93.46	93.46	92.32
	Sensitivity	65.53	58.72	65.96
	Specificity	98.04	99.16	96.64

Table 45. Decision tree performance measures for different cut-offs on the test set – Telecom1

For Telecom1 (see Table 45), the models do not present big variations in performance as noted for logistic regression. However, if the 0.5 cut-off was chosen, it is possible to notice a decrease in sensitivity (with a higher imbalance against specificity), which means that the target churners would be classified worse than if the KS statistic is used as a base for cut-off.

Similar to the logistic regression analysis, Telecom2 still presents reasonable balance between sensitivity and specificity at the 0.5 cut-off (see Table 46). Nevertheless, when the churn proportion is used as base for the cut-off, there is a significant variation between the models' performance. The more imbalance there is between sensitivity and specificity corresponds to better CA, favouring the classification of non-churners, which

represents the majority class in the sample data. This difference is explained by the fact that more than 1.8% of the highest scores are predicted as churners for action1 and action2, due to the limited number of scores.

	Metric (%)	KS Statistic as cut-off base	0.5 cut-off	Churn proportion as base for cut-off
Before Domain Knowledge	CA	52.30	52.30	95.92
	Sensitivity	74.44	74.44	4.44
	Specificity	51.89	51.89	97.60
After Domain Knowledge (Action1)	CA	52.28	54.56	84.02
	Sensitivity	74.44	70.00	25.56
	Specificity	51.87	54.28	85.09
After Domain Knowledge (Action2)	CA	52.28	52.28	57.14
	Sensitivity	74.44	74.44	66.67
	Specificity	51.87	51.87	56.96

Table 46. Decision tree performance measures for different cut-offs on the test set – Telecom2

7.8. Conclusions

In this chapter, it was shown how to make data mining models developed for churn prediction more understandable, easier to interpret and compliant with domain knowledge. To achieve this, it was demonstrated how the monotonicity analysis of decision tables can be used to check whether the knowledge contained in data mining models, in this case decision trees, is in line with domain knowledge, and how to correct any inconsistencies found.

In conclusion, this research demonstrates that the variables presented are coherent and showing the expected relationship with the target variable, according to the relevant domain knowledge, with no significant loss of performance for the decision tree models. It illustrates that the use of domain knowledge in data analysis can provide a useful tool when understanding and defining strategies related to the relationship with customers.

Chapter 8. Customer Lifetime Value Analysis and Domain Knowledge Integration

8.1. Introduction

The two previous chapters examined the use of logistic regression and decision trees for churn evaluation, in conjunction with the application of domain knowledge to make the models more understandable and acceptable for decision making. In this chapter, the CLV calculation is demonstrated, taking into consideration the domain knowledge evaluation.

To achieve this, linear regression, decision trees and neural networks will be used for the benchmarking analysis of the data set under evaluation. The sign evaluation will be used in the linear regression model, to compare the original model result with the amended model. A classification task will then be applied to the target variable, in order to more easily verify the domain knowledge in the decision tree model. Conclusions will then be extracted from the empirical results.

8.2. Data Set Description and Preparation

The objective in this research is to estimate the future CLV for individuals using past purchase behaviour information.

For this analysis, the DMEF3 data set will be used. This data set was made available by the Direct Marketing Educational Foundation (DMEF) for academic research and teaching, and it was provided for this research by Edward Malthouse. As a consequence, to keep the standard and to progress with the evaluation of this data set, the assumptions made by Malthouse and Blattberg (2005) were kept in this research.

The DMEF3 data set is from a long-time specialty catalogue company that mails both full-line and seasonal catalogues to its customer base. The data set is a random sample of 106,284 customers who have bought before from the company and were being considered for a mailing in the autumn of 1995. The data set has twelve years of purchase history until 31st of July, 1995. It is from a real retail consumer catalogue

company, which contains a long time series and is available to all researchers from the DMEF.

The approach adopted for this research is similar to the one used by Malthouse and Blattberg (2005), using a study design that "turns back the clock". With this approach, "now" is defined as 1st of August, 1990 and the universe selected are all customers who were on file before this date. This gives a sample of 41,669 customers, with six-year base and target periods. These observations are randomly assigned to training and test sets of approximately equal size.

Based on some additional pre-processing, from these 41,669 customers, a further 2,441 customers were eliminated as they did not have any purchase on the base period, only on the target period. As a result, a final sample of 39,228 customers was used for analysis.

In terms of variable selection, variables that were not clear indicators of past behaviour and could include future information (for the six-year target period) were eliminated from the analysis, with a remaining total of forty variables, including first purchase amount, and indicators of product classes and sub-classes. Some variables were also created or transformed, which is the case for the RFM (recency, frequency and monetary) variables, and time on file.

To maintain the standards established by Malthouse and Blattberg (2005), the square root and logarithm transformations were applied to all amount and count variables, in order to obtain more symmetric distributions. Also, the influence of outliers of untransformed count and amount variables is reduced with 1% Winsorization, which indicates that values greater than the 99th percentile are set equal to the 99th percentile. This would bring the final total of variables under investigation to seventy-seven predictor variables, with a target variable, CLV6.

The net contribution used in the calculation of CLV6 corresponds to the variables totsal1, …, totsal6 used in a cumulative way to calculate CLV over the six-year target period. The discount rate in this case is also assumed to be equal to 15%. The CLV6 equation is then described as:

$$CLV6 = \frac{totsal6}{1.15} + \frac{totsal5}{1.15^2} + \frac{totsal4}{1.15^3} + \frac{totsal3}{1.15^4} + \frac{totsal2}{1.15^5} + \frac{totsal1}{1.15^6}$$

By inspecting the distribution of the target variable CLV6, it was possible to visualise that the distribution was right skewed, illustrating that most values are small, with less large values. This is demonstrated in the histogram presented in Figure 40. The histogram gives an indication of the shape of the distribution and whether the data are distributed symmetrically. In this case, it demonstrates that the customers for this specific retailer tend to have a smaller lifetime value, with only a few high value customers, which represents the general pattern in the retailing industry.

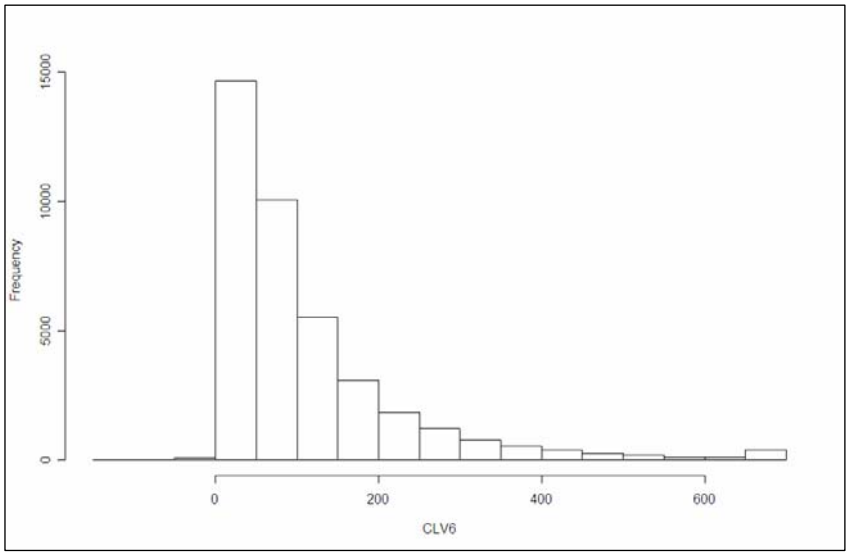

Figure 40. Histogram showing right skewed CLV6 variable

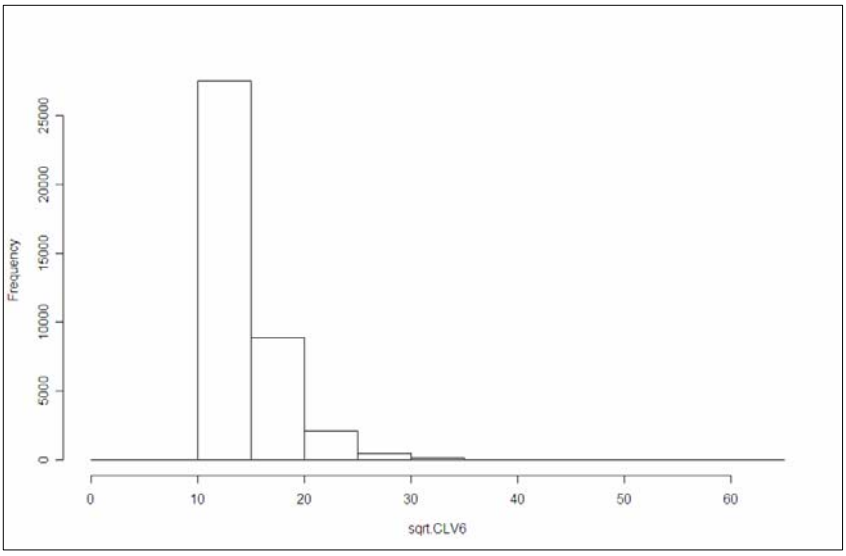

Figure 41. Histogram showing CLV6 variable transformed by the square root

In this case, the variable CLV6 was transformed based on the square root and logarithm, in order to symmetrise its distribution, increasing the density of observations in the right tail and reducing the influence of outliers. This is demonstrated in Figure 41 and Figure 42. The purpose of these transformations was to evaluate if there would be a significant difference in performance when using these transformed or original target variables. As a result, the best variable would be chosen for prediction.

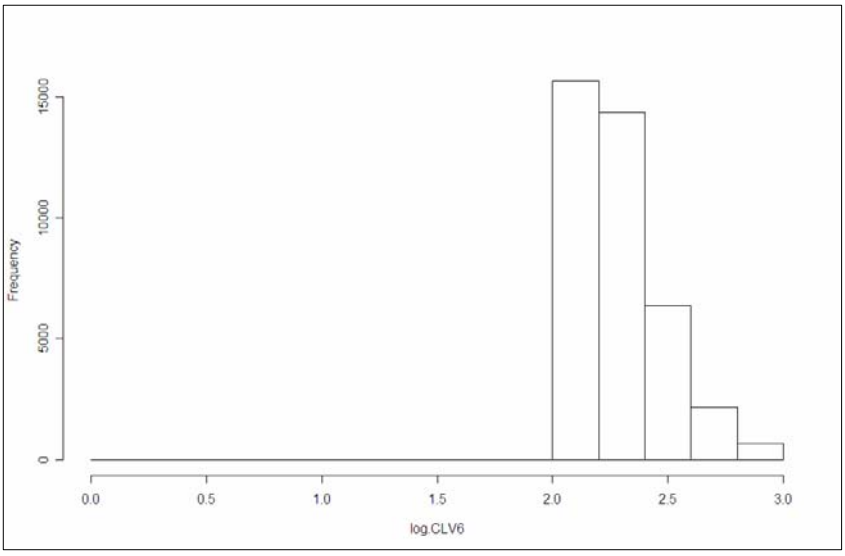

Figure 42. Histogram showing CLV6 variable transformed by the log function

A description of the variables under analysis is shown in Table 47.

Variable	Description
CONVSALE	Number of sales made when converting to the company.
Sqrtconvsale	Variable convsale transformed by the square root function.
Logconvsale	Variable convsale transformed by the logarithm function.
Recency	Number of months since last order.
Sqrtrec	Variable recency transformed by the square root function.
Logrec	Variable recency transformed by the logarithm function.
dol1 - dol6	Corresponds to the sales from year-one to year-six in a cumulative way.

Sqrtdol1 - sqrtdol6	Variables dol1 to dol6 transformed by the square root function.
logdol1 - logdol6	Variables dol1 to dol6 transformed by the logarithm function.
ord1 - ord6	Corresponds to the orders made from year-one to year-six in a cumulative way.
Sqrtord1 - sqrtord6	Variables ord1 to ord6 transformed by the square root function.
logord1 - logord6	Variables ord1 to ord6 transformed by the logarithm function.
tof	Time on file in days, based on the day since the customer converted to the company.
Sqrttof	Variable tof transformed by the square root function.
Logtof	Variable tof transformed by the logarithm function.
fstcls1 - fstcls7	Binary variable, indicating if first order was from product class 1, 2, 3, 4, 5, 6 or 7 (seven variables).
CNVCAT1-CNVCAT7, CNVCAT11, CNVCAT13, CNVCAT15, CNVCAT18, CNVCAT22-CNVCAT26, CNVCAT31-CNVCAT33	Binary variable, indicating if the first purchase was made in one of these categories (nineteen categories – nineteen variables).
xx	Binary variable, indicating if there were orders in year-one and year-two, at least.
xxx	Binary variable, indicating if there were orders in year-one, year-two and year-three, at least.
aos	Average order sale over the six-year base period.
sqrtaos	Variable aos transformed by the square root function.
logaos	Variable aos transformed by the logarithm function.
aosmiss	Binary variable, indicating if there were no orders in the six-year base period.

Table 47. Predictive variables for DMEF3

To evaluate the best target variable, the linear regression model was used. The regression evaluation will be described in subsection 8.4.1. The variable chosen will then be used as the model target variable for the three data mining techniques under

investigation. The data will then be ready, with all the variables manipulated and prepared for analysis.

8.3. Performance Metrics for Continuous Target variable

As the variable under investigation in this session is not binary, the performance metrics adopted in previous chapters cannot be used. As a result, different measures need to be selected for the performance evaluation of the data mining models.

The mean squared error (MSE) is a commonly used measure of performance for a continuous prediction (Witten and Frank, 2005). It is represented by the sum of all squared prediction errors, divided by the total number of instances under investigation, as denoted below:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - z_i)^2}{n}$$

where n represents the number of instances, y_i represents the actual values and z_i represents the predicted values. The difference between y_i and z_i represents the residual or predicted error.

The root mean squared error (RMSE) is used to give the MSE measure the same dimension as the predicted value. It is represented as:

$$RMSE = \sqrt{MSE}$$

In the case of MSE and RMSE, the smaller their values the better the performance of the model selected. This is done assuming that the variable under investigation in all models is represented on the same scale.

The R-squared, also known as coefficient of determination, indicates how good the fit of a model is. In linear regression, it indicates how well the regression line approximates the real data points. The bigger the value (between 0 and 1), the better is the fit. It is defined by:

$$R^2 = 1 - \frac{SSE}{SST}$$

where SSE (sum of squared error) is:

$$SSE = \sum_{i=1}^{n} (y_i - z_i)^2$$

and SST (total sum of squares) is defined as:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad .$$

In this case, $\bar{y}$ represents the mean of the target variable.

Lastly, the correlation coefficient measures the correlation between the observed and predicted values of the target variable (Witten and Frank, 2005). It is a different measure from the RMSE and MSE, as it is scale independent, which means that the error does not change if, for example, all predictions are multiplied by a constant factor or left unchanged. For the linear regression model, the R-squared measure (which is also scale independent) is obtained through the square of the correlation coefficient.

In conclusion, most of the literature on CLV does not compare between different techniques for its prediction. These studies usually focus on one method and try to generate better results based on that method. This research focussed on R-squared, RMSE and correlation between actual and estimated values. They were used to compare the performance for the data mining models under investigation.

8.4. Empirical Analysis

For the CLV analysis, three data mining techniques were used for benchmarking evaluation: linear regression, decision trees and neural networks. Some of these techniques were also chosen for the churn analysis and the idea was to maintain a standard of evaluation. As the variable under investigation in this case is continuous and

not binary, linear regression was used instead of logistic regression; also, in a first instance, regression trees were used instead of classification trees. The descriptions of the analyses are shown in the subsections below.

8.4.1. Linear Regression Application

Linear regression is a type of regression analysis in which data are modelled by a function which is a linear combination of the model parameters and depends on one or more independent variables (Witten and Frank, 2005; Moore and McCabe, 2006). When the outcome of an analysis and all the predictive variables are numeric, linear regression is a natural choice for estimation.

As explained in Chapter 4, linear regression expresses the target variable as a linear combination based on the predictive variables that are relevant for the target variable. For m variables and i = 1,…, n individuals, the linear regression formula can be represented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m$$

where:

- y is the target variable for each individual i.
- β_0 is the intercept.
- β_m represents the weight given to the specific variable x_j (j = 1, … m) associated to it.
- x_j represents the predictive variables for each individual i, from which y is to be predicted.

The multiple linear regression method chooses the coefficients weights based upon the least mean-squared difference between actual and predicted values. This is done based upon the n individuals in the sample under evaluation. The model function represents a straight line and the results of data fitting are subject to statistical analysis. If the data exhibits nonlinear dependency, the best-fitting straight line will be found (Witten and

Frank, 2005). As this line may not fit well, transformations can be used to facilitate the analysis (Malthouse and Blattberg, 2005).

As mentioned in section 8.2, the target variable CLV6 was transformed based on the square root and logarithm in order to create more symmetry in its distribution and to provide a better model fit. The model was then analysed with the original CLV6 variable and with the transformed variables based on the square root and logarithm, sqrtclv6 and logclv6 respectively. This was done following the same approach as Malthouse and Blattberg (2005).

The purpose was to evaluate if there would be a significant difference in performance and the best transformation would be chosen for prediction. The results obtained from the stepwise linear regression are shown below, showing the performance measured base on R-squared and RMSE (Table 48). This analysis was done based on the most predictive variables with a significance level of 1%. The performance was also measured based on the correlation between the target and predicted values.

	Metrics	Target: Sqrtclv6	Target: logclv6	Target: CLV6
Training	R-squared	0.3077	0.2950	**0.3155**
	Correlation	0.5547	0.5431	**0.5617**
	RMSE	111.0769	96.4715	**95.0551**
Test	R-squared	0.3058	0.2846	**0.3062**
	Correlation	0.5530	0.5335	**0.5535**
	RMSE	113.4324	99.4449	**97.9346**

Table 48. R-squared, correlation and RMSE for CLV6, sqrtclv6 and logclv6 as target variables

CLV6 is the variable of interest, not the square root or logarithm. So, after using the square root and log for prediction, their results were converted back to CLV for performance measurement.

The values for the correlation (and consequently for the R-squared) are higher for the stepwise regression using the CLV6 target variable and are selected in bold. In this case, RMSE could also be used for comparison. If the target values were kept in the original form, assuming different scales for each of the analysis, the values of RMSE would be different. The R-squared and correlation are generalised measures, independent of the

variation of scale for the variables under investigation, being a natural choice for comparison.

Based on the above analysis, CLV6 was then chosen as the target variable for the linear regression and it was also used for the regression tree and neural network models.

8.4.2. Use of Regression Trees to Predict Customer Lifetime Value

For the regression tree analysis, the training set used in the linear regression analysis was further divided into training (9,831 customers) and validation (9,830 customers) sets. For the purpose of evaluation and comparison between the models, the test set was kept the same. The resulting performance measures are shown in Table 49.

	R-squared	Correlation	RMSE
Training	0.3381	0.5815	95.2971
Test	0.2833	0.5335	99.5372

Table 49. Performance measures for the regression tree model

The validation set was used to avoid the tree overfitting. The tree model was then chosen based on the smaller mean squared error for training and validation sets, also taking into consideration the possibility of overfitting. The resulting graph is depicted in Figure 43.

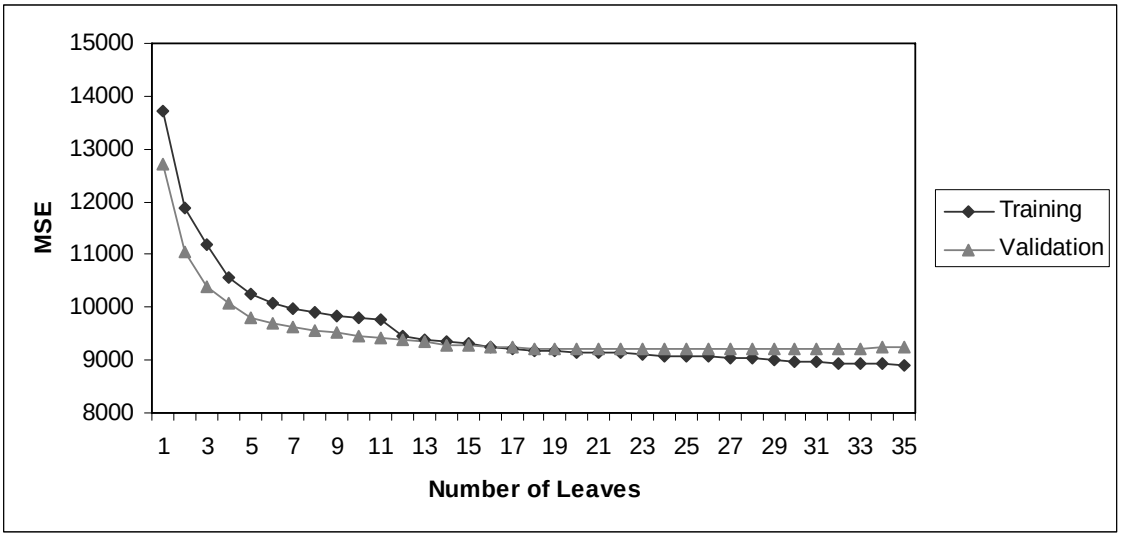

Figure 43. Early stopping graph using MSE as the error criterion

The generated tree contained sixteen main predictor variables, which would be considered a reasonable number for estimation. The number of leaves in the tree was twenty-five, generating a relatively wide tree, which represents a considerable number of rules for evaluation. As the purpose of this subsection is to generate the best tree model for benchmarking against other data mining techniques, this original result was kept. Further changes in the tree analysis will be made in subsection 8.4.6 for the incorporation of domain knowledge and better interpretability of the model.

8.4.3. Neural Network Analysis

For the neural network (NN) modelling, as the target variable is continuous, the linear transfer function was also used in this analysis, but only in the output layer. The linear combination function was used to combine the weights and the values, feeding them into the hidden neuron and then added the bias value, which acts like an intercept. The weights and the bias were iteratively adjusted in order to minimise the error function, which was calculated based on the comparison between the predicted and target values.

The linear activation function, also called identity function, was used to transform the value obtained from the combination function in the output layer. In this case, it does not change the value of the argument, and its range is potentially limitless. As the model used was a MLP, a sigmoid activation function was used in the hidden layer, in order to model nonlinear patterns.

Using a similar approach to the one adopted in Chapter 4, the NN model was estimated using 1, 2, 4 and 6 hidden neurons, in order to define the best selection for the NN model. The performance on the validation set is shown in Table 50.

Measure	1 Hidden Neuron (HN)	2HN	4HN	6HN
R-squared	**0.3165**	0.3021	0.2833	0.2936
Correlation	**0.5640**	0.5499	0.5362	0.5423
RMSE	**92.9880**	93.9645	95.2202	94.5310

Table 50. Performance selection for NN model

The resulting models were then compared to evaluate the performance on the validation set. This was done based on the model performance metrics in comparison with the

other resulting NN models. It was then concluded that the NN model with one hidden neuron performed better. Note that the lower the number of hidden neurons, the more linear the model is. This model was then chosen for comparison against the other data mining techniques investigated. The resulting performance measures for training and test sets are shown in Table 51.

	R-squared	Correlation	RMSE
Training	0.3521	0.5936	94.2841
Test	0.3127	0.5594	97.4738

Table 51. Performance measures for the NN model

Another important aspect when evaluating the NN model is to visualise the weights, which can be used for interpretation. The inputs that have the biggest weights are the most important for the model. This can be visualised in the sample graph in Figure 44. This graph only shows some of the predictive variables.

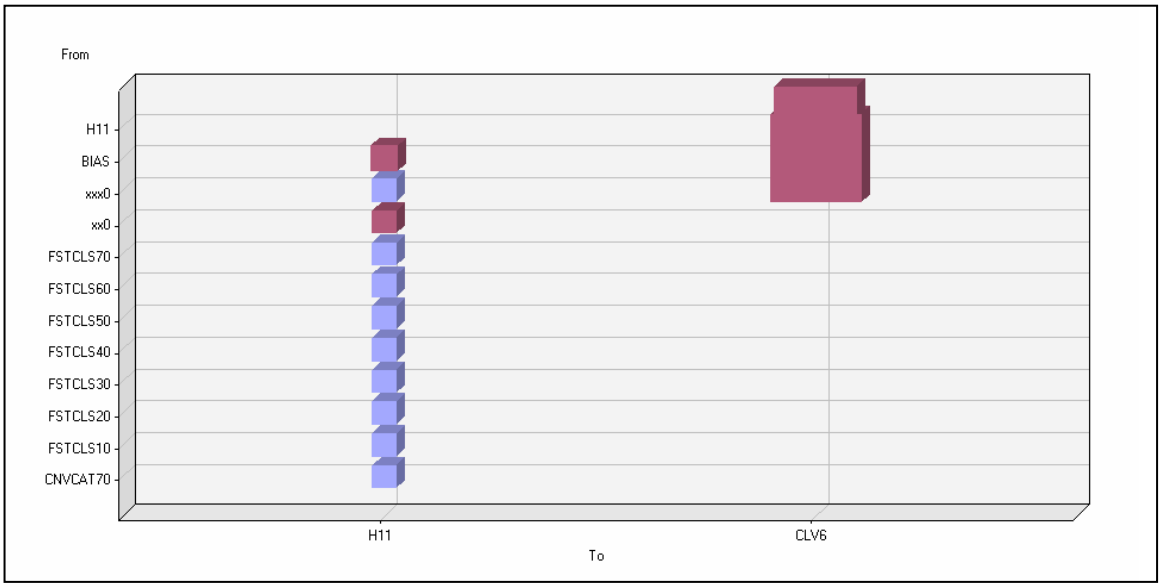

Figure 44. Graph plotting some of the weights for the NN model

This graph, also known as Hinton diagram (Despagne and Massart, 1998; Matignon, 2005), plots the size of the weights for each one of the variables connected to the hidden neuron, and also the general effect that this neuron will have on the target variable. In this case, the darker colour represents positive weight and the lighter colour represents negative weight. The size of the blocks is supposed to vary in accordance with the weight value.

As this graph is not differentiating between the sizes of the weights (the values are very similar) and there were too many variables in the plot (the NN model uses all the predictive variables in its prediction), their values were analysed independently, as shown in Table 52.

From	To	Weight (–)	From	To	Weight (+)
sqrtord1	H11	-8.59478	BIAS	CLV6	322.7766
FSTCLS6	H11	-3.30718	H11	CLV6	283.2308
FSTCLS1	H11	-3.29624	BIAS	H11	15.64021
FSTCLS3	H11	-3.29342	logord1	H11	5.128891
FSTCLS5	H11	-3.27371	ord1	H11	1.455177
FSTCLS2	H11	-3.26402	sqrtord3	H11	1.077027
FSTCLS4	H11	-3.24821	sqrttof	H11	0.887854
FSTCLS7	H11	-3.2394	logdol6	H11	0.828543
xxx	H11	-2.33715	sqrtdol3	H11	0.62684
sqrtrec	H11	-0.94901	sqrtdol5	H11	0.55134
sqrtord2	H11	-0.68065	sqrtord4	H11	0.475994
Logdol3	H11	-0.62604	logord2	H11	0.417193
Logord3	H11	-0.61559	logrec	H11	0.414253
sqrtord5	H11	-0.57358	logdol2	H11	0.324577
logtof	H11	-0.55069	logord5	H11	0.273238
sqrtdol6	H11	-0.38922	logdol1	H11	0.269683
aosmiss	H11	-0.38641	dol6	H11	0.214741
Ord3	H11	-0.3796	ord6	H11	0.207027
tof	H11	-0.32174	Recency	H11	0.204138
sqrtord6	H11	-0.31472	ord2	H11	0.199378
sqrtdol4	H11	-0.30933	Xx	H11	0.143036
sqrtdol2	H11	-0.2616	dol4	H11	0.141856
Logord4	H11	-0.21133	CNVCAT2	H11	0.134744
Dol5	H11	-0.20654	sqrtconvsale	H11	0.118462
CNVCAT3	H11	-0.20529	logaos	H11	0.109615
Logord6	H11	-0.1405	ord5	H11	0.082572
Dol3	H11	-0.13839	CNVCAT13	H11	0.06105
Logdol5	H11	-0.12497	dol2	H11	0.051657
sqrtaos	H11	-0.11408	logdol4	H11	0.028381
CNVCAT7	H11	-0.09311	CNVCAT11	H11	0.01917
Ord4	H11	-0.08878			
CNVCAT32	H11	-0.08742			
aos	H11	-0.06463			
Dol1	H11	-0.05848			
CONVSALE	H11	-0.05326			
CNVCAT31	H11	-0.04451			
Logconvsale	H11	-0.04432			
CNVCAT25	H11	-0.04341			
CNVCAT4	H11	-0.04031			
sqrtdol1	H11	-0.0209			
CNVCAT5	H11	-0.01919			
CNVCAT15	H11	-0.01883			
CNVCAT24	H11	-0.00931			
CNVCAT6	H11	-0.00739			
CNVCAT1	H11	-0.00265			

Table 52. Weights for the predictive variables in the NN model

From this list, the weight selection was used for comparison of performance, but not to apply domain knowledge as the sign definition in a NN model has a complex way of being defined. The variable selection (the ones with the highest values, independently of sign) was used only to compare the performance with the original model and choose the most similar. The purpose was to experiment with different numbers of variables.

Choosing a different number of variables based on their weight values, the number of variables increased and the performance was compared against the original model. Table 53 shows these performance measurements for 11, 17, 22 and 30 variables, and the original model.

	Metrics	Original	11 Var.	17 Var.	22 Var.	30 Var.
Training	R-squared	0.3521	0.1552	0.3335	0.3377	0.3396
	Correlation	0.5936	0.3969	0.5786	0.5826	0.5864
	RMSE	94.2841	107.6622	95.6240	95.3283	95.1956
Test	R-squared	0.3127	0.1361	0.3099	0.3132	0.3154
	Correlation	0.5594	0.3690	0.5568	0.5598	0.5634
	RMSE	97.4738	109.2824	97.6736	97.4394	97.2773

Table 53. Performance measures for the NN model for different numbers of variables

From this table, one can conclude that with only a few variables, even being the most predictive ones, the model performed a lot worse that the original. However, as the number of predictive variables increased, the performance of the new models became much nearer to the original model, but without the need to use too many variables. If the test set is taken into consideration, the model with fewer variables performed better than the original model. This analysis was done to demonstrate that the model can produce good performance without the need to use all available variables.

8.4.4. Performance Comparison

The performance measurements for the three data mining techniques are demonstrated in Table 54.

Based on these performance measures, it is possible to conclude that the NN model had a better performance than the linear regression and regression tree models, in both training and test sets. The model generated by the NN algorithm showed a bigger difference in performance for the training model, but the performance on the test set was very similar to the linear regression model.

	Metrics	Linear R.	RTrees	NN
Training	R-squared	0.3155	0.3381	**0.3521**
	Correlation	0.5617	0.5815	**0.5936**
	RMSE	95.0551	95.2971	**94.2841**
Test	R-squared	0.3062	0.2833	**0.3127**
	Correlation	0.5535	0.5335	**0.5594**
	RMSE	97.9346	99.5372	**97.4738**

Table 54. Performance measures for data mining techniques

The regression tree model showed a better performance than the linear regression model in terms of R-squared and correlation for the training set, but the RMSE was worse. However, the linear regression model performed better than the regression tree model for the test set. The linear regression model showed a similar performance in terms of training and test sets evaluation, especially when taking into consideration the R-squared and correlation metrics.

In conclusion, the models had a relatively similar performance, which was reasonable, especially taking into consideration the type of data under investigation. However the lower values of R-squared indicate that it would not be possible to accurately predict future behaviour of customers based on their past purchase behaviour. It works as a good indicator, but does not present a high predictive accuracy.

Therefore, subsections 8.4.5 and 8.4.6 will explore the possibility of using domain knowledge to make the models more understandable and to visualise the variables that would have a greater impact on the CLV estimation. This would facilitate the development of strategies to target many customers at a time, enabling the possibility of increasing CLV in general.

8.4.5. Linear Regression Application and Evaluation of Signs

After the selection of the target variable and with the performance metrics already defined for the stepwise regression and compared against the other techniques, the focus of the analysis now changes to the evaluation of signs of the linear regression model. The purpose was to apply domain knowledge in the evaluation of a continuous variable and evaluate its performance against the original model. Based on that, the stepwise linear regression coefficients are shown in Table 55.

Parameter	Expected	Estimate	Wald chi-square	P-value
Intercept	+/-	165.47585	1,146.31	<.0001
sqrtdol6	+	11.08822	337.66	<.0001
dol1	+	0.40314	251.28	<.0001
sqrtrec	-	-27.42346	200.86	<.0001
FSTCLS7	+/-	-12.22741	53.79	<.0001
dol3	+	0.13761	50.56	<.0001
logord6	+	**-38.21880**	49.12	<.0001
sqrtaos	+	**-9.42598**	45.40	<.0001
Recency	-	**0.95466**	21.48	<.0001
logaos	+	11.40303	21.29	<.0001
logord5	+	**-44.80276**	19.71	<.0001
sqrtord5	+	28.57316	15.64	<.0001
FSTCLS6	+/-	-7.97229	9.91	0.0016
CNVCAT5	+/-	-6.66437	7.87	0.0050

Table 55. Analysis of maximum likelihood coefficient estimates – DMEF3

In this analysis, the domain knowledge was also expressed in the "Expected" Column, showing the expected sign for each variable. For example, the variables "dol1" and "dol3" suggest that they have a positive effect on CLV: if the dollar amount in these periods increases, the CLV also increases.

The signs of some variables were not in accordance with the expectations. This is the case with the variables "logord6", "logord5", and "sqrtaos", for which one would expect that an increase in the number of orders and amount spent would represent an increase in CLV rather than a decrease. Also, the variable "recency" suggests that it influences

CLV negatively: if the time since last purchase increases, the chance of reduction in CLV increases; however, the results were presenting a positive influence. These contradictory results are highlighted (bold) in Table 55.

Hence, these four variables were eliminated one by one, and the stepwise linear regression procedure was rerun, also comparing the performance measurements. This analysis was done following the pseudo-code presented in Figure 22, Chapter 6.

After evaluating the signs and eliminating any contradictory variables, the linear regression results presented coherent variables that were more in line with the domain knowledge analysis (see Table 56). The performance measurements are then presented in Table 57, comparing the original measurements with the ones obtained after the domain knowledge evaluation.

Parameter	Estimate	Wald chi-square	P-value
Intercept	160.43583	1,303.76	<.0001
dol6	0.21386	433.00	<.0001
dol1	0.40893	255.65	<.0001
Aos	0.34530	53.53	<.0001
dol3	0.13473	52.47	<.0001
FSTCLS7	-11.36908	46.96	<.0001
sqrtrec	-11.10011	32.35	<.0001
logrec	-16.90383	15.63	<.0001
FSTCLS6	-8.32352	10.81	0.0010
CNVCAT5	-7.30877	9.39	0.0022

Table 56. Final analysis of maximum likelihood coefficient estimates – DMEF3

	Metrics	Original	Amended
Training set	R-square	0.3155	0.3111
	Correlation	0.5617	0.5578
	RMSE	95.0551	95.36357
Test set	R-square	0.3062	0.3009
	Correlation	0.5535	0.5485
	RMSE	97.9346	98.30727

Table 57. Original and amended linear regression performance measures – DMEF3

From Table 57, it can be concluded that the performance in general of the linear regression model stayed relatively stable in the analysis of both training and test sets, with only small variations in the performance values. The results, however, are more interpretable, making the model more acceptable and useful for strategy definition.

8.4.6. Mapping to a Classification Task for Domain Knowledge Evaluation

The purpose of this subsection was to make a classification task, such that the domain knowledge in the decision tree model could be more easily verified. To proceed with this, the first step was to discretise CLV into three groups: small, medium, and high.

CLV was ordered from low to high and divided into three equal groups based on the frequency, with 13,076 customers in each group. These groups were then put forward for the same test, training and validation sets used for the regression tree in subsection 8.4.2. With the target variable transformed into categorical values, the classification tree was then obtained. The resulting tree is shown in Figure 45.

After obtaining the tree model and calculating its performance (see Table 62, subsection 8.4.7), it was necessary to analyse the influence of domain knowledge for the interpretability of the results. As a result, the classification tree was then transformed into a decision table (DT), and the rules obtained were analysed using the same procedure as with the churn case in Chapter 7 (Figure 33). The resulting DT is shown in Table 58.

In this scenario, it would be difficult to evaluate the domain knowledge, as the purpose would be to see if the variables would make CLV increase or decrease, therefore the "medium" value would make things less straightforward. One of the options would be to eliminate the customers that have medium value CLV and only evaluate the possibility of small and high CLV. This would make it easier to evaluate the DT, but it would not take into consideration all the cases.

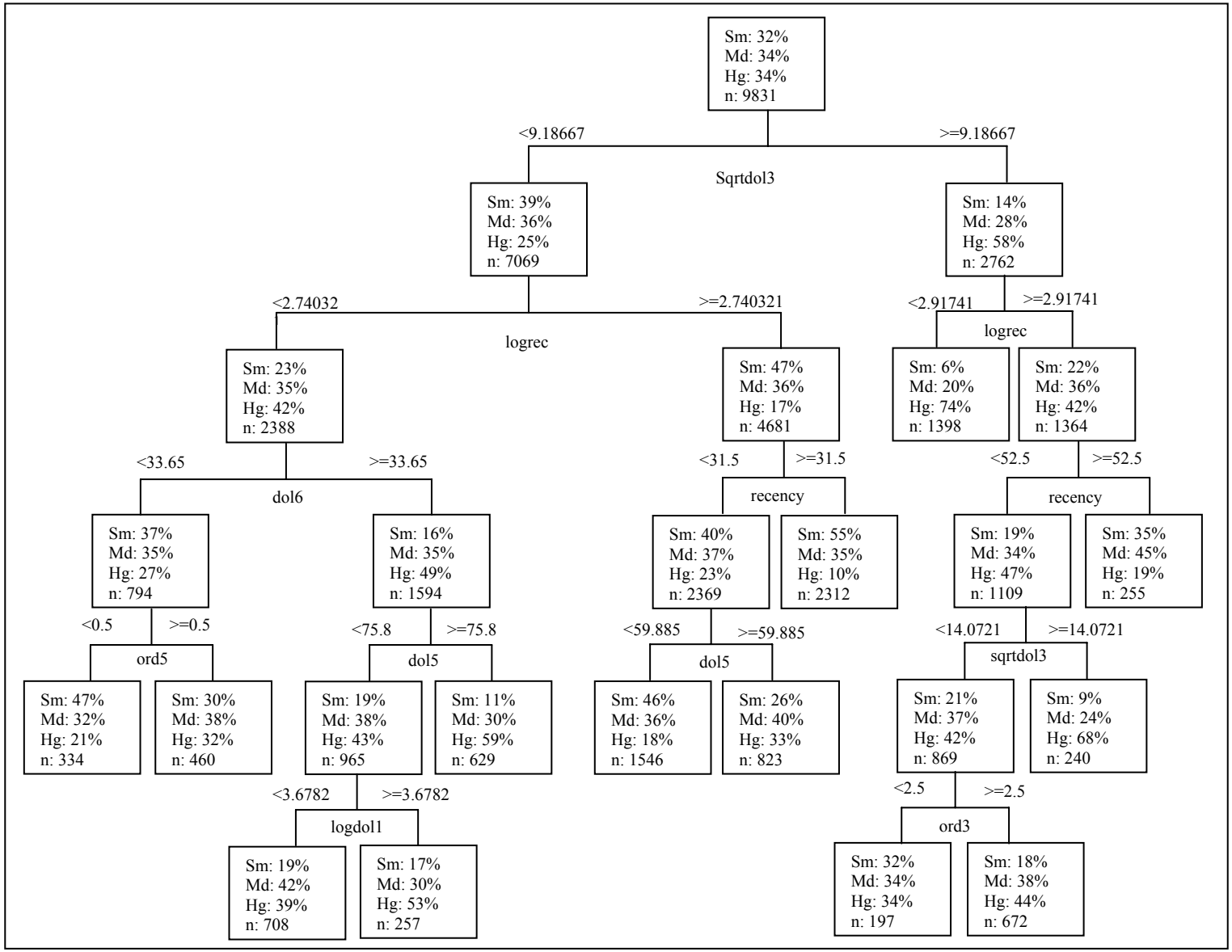

Figure 45. Decision tree model for DMEF3 with categorical target variable

Hence, the medium values were kept and if necessary to change the rules based on the domain knowledge, the original probabilities would be used. For example, if there are two leaf nodes with the following class distributions: first node has low = 60%, medium = 30%, high = 10%, the majority prediction says low; second node has low = 30%, medium = 20%, high = 50%, the majority prediction says high. Suppose that this is a counter-intuitive direction and the rules need to be changed. For the first node it would change to either medium or high, depending on the proportions. Hence, in this case, it is medium since 30% is greater than 10%. For the second node, it would change to small, since medium = 20% is smaller than small = 30%.

The DT analysis did not reveal any counter-intuitive rules for the variable "logdol1". Table 58 shows that an increase in the dollar amount spent by the customer in a certain period would indicate an increase in CLV, changing from medium to high. The table was then reordered (see Table 59) and the domain evaluation was executed for the variable "dol6", where, also in accordance with the domain knowledge, an increase in the dollar amount spent by the customer would indicate an increase in CLV, changing from medium to high, small to medium and small to high.

With regards to the variable "recency", Table 60 shows that this variable indicates that if the time since last purchase increases, the chance of reduction in CLV increases: from medium to small or from high to medium. This finding is intuitive and in accordance with domain knowledge.

This analysis was done for all the variables with all presenting intuitive rules, except for the variable "logrec": in some parts of the tables the prediction will change from medium or high to small instead of vice-versa. Nevertheless, in the same context as with the churn analysis, not all the rules need and should be changed. This means that changing the values for the variable "logrec" would generate too many rules and would also interfere in the monotonicity of other variables. The results would become too complex, proving difficult to make the model interpretable, invalidating the purpose of using domain knowledge for model adaptation. As a result, with the majority of conditions presenting intuitive rules, both approaches of changing action entries and condition term removal were not applied. The set of rules extracted from the original model were already in accordance with the domain knowledge. The final performance measurements are reported in Table 62 (subsection 8.4.7) for both training and test sets.

1. sqrtdol3	< 9.18667								≥ 9.18667		
2. logrec	< 2.74032					≥ 2.74032			< 2.91741	≥ 2.91741	
3. dol6	< 33.65		≥ 33.65			-			-	-	
4. recency	-		-			< 31.5		≥ 31.5	-	< 52.5	≥ 52.5
5. ord5	< 0.5	≥ 0.5	-			-		-	-	-	-
6. dol5	-	-	< 75.8		≥ 75.8	< 59.885	≥ 59.885	-	-	-	-
7. logdol1	-	-	< 3.6782	≥ 3.6782	-	-	-	-	-	-	-
1. small	x	.	.	.	.	x	.	x	.	.	.
2. medium	.	x	x	.	.	.	x	.	.	.	x
3. high	.	.	.	x	x	.	.	.	x	x	.

Table 58. Decision rules transferred from Figure 45

1. sqrtdol3	< 9.18667										
2. logrec	< 2.74032										
3. recency	-										
4. dol5	< 75.8							≥ 75.8			
5. ord5	< 0.5				≥ 0.5			< 0.5		≥ 0.5	
6. logdol1	< 3.6782		≥ 3.6782		< 3.6782	≥ 3.6782		-		-	
7. dol6	< 33.65	≥ 33.65	< 33.65	≥ 33.65	-	< 33.65	≥ 33.65	< 33.65	≥ 33.65	< 33.65	≥ 33.65
1. small	x	.	x	.	.	.	.	x	.	.	.
2. medium	.	x	.	.	x	x	.	.	.	x	.
3. high	.	.	.	x	.	.	x	.	x	.	x

Table 59. Table 58 reordered – variable "dol6" moved to last row (partial table)

1. sqrtdol3	< 9.18667								≥ 9.18667		
2. logrec	< 2.74032					≥ 2.74032			< 2.91741	≥ 2.91741	
3. dol6	< 33.65		≥ 33.65			-			-	-	
4. ord5	< 0.5	≥ 0.5	-			-			-	-	
5. dol5	-	-	< 75.8		≥ 75.8	< 59.885	≥ 59.885		-	-	
6. logdol1	-	-	< 3.6782	≥ 3.6782	-	-	-		-	-	
7. recency	-	-	-	-	-	-	< 31.5	≥ 31.5	-	< 52.5	≥ 52.5
1. small	x	.	.	.	.	x	.	x	.	.	.
2. medium	.	x	x	.	.	.	x	.	.	.	x
3. high	.	.	.	x	x	.	.	.	x	x	.

Table 60. Table 59 reordered – variable "recency" moved to last row

8.4.7. Performance Metrics and Results for the Classification Task Model

To measure the performance of the decision tree model with the three classes (small, medium and high), a prediction matrix was created to evaluate the performance, which is shown in Table 61.

Predicted	**Actual value**		
	1	2	3
1	a	d	g
2	b	e	h
3	c	f	i

Table 61. Prediction matrix for the performance evaluation of the tree model in Figure 45

Based on this matrix, the accuracy of predicting small (*CAs*), medium (*CAm*), high (*CAh*) and the overall accuracy (*CA*) are expressed as:

$$CAs = \frac{a}{a+b+c}$$

$$CAm = \frac{e}{d+e+f}$$

$$CAh = \frac{i}{g+h+i}$$

$$CA = \frac{a+e+i}{a+b+c+d+e+f+g+h+i}$$

Other methods were also used to evaluate the model's performance: Pearson correlation, Spearman rank-order correlation, Kendall's tau and Kruskal's gamma (Brown and Benedetti, 1977; Sheskin, 2004).

The Pearson correlation coefficient uses the scores specified from the predicted values in its definition. It has the range $-1 \leq \rho \leq 1$. The Pearson correlation coefficient is computed as:

$$\rho = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where n is the number of customers, x_i is the target value, y_i is the predicted value, $\bar{x}$ and $\bar{y}$ are the mean of the x_i and y_i, respectively.

In the case of ordinal data, which is the type under investigation in this subsection, Spearman rank-order correlation, Kendall's tau and Kruskal's gamma are more appropriate. The Spearman rank-order correlation is a nonparametric measure of association based on the ranks of the predicted and original values. It also has the range $-1 \leq \rho_s \leq 1$. It can be computed as:

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

where n is the number of customers and d_i is the difference between the rankings. The Spearman rank-order correlation is equivalent to the Pearson correlation on ranks instead of the scored values. The formula above assumes no ties, which means no equal ranks in either column under evaluation. In case of ties, which is the case for this data set, the formula used will be the one for the Pearson correlation and the average ranks are used.

Another performance metric is Kendall's tau statistic (Kendall, 1970). It is based on the number of concordant and discordant pairs of observations, using a correction for tied pairs (pairs of rows that have equal values of X or equal values of Y). Kendall's tau also ranges from -1 to 1. It can be calculated as:

$$Kendall's\ tau = \frac{A - B}{\frac{1}{2}n(n-1)}$$

where A represents the number of concordant pairs and B represents the number of discordant pairs. The denominator represents the number of possible pairs for n

customers. Two customers are then said to be concordant if the customer who is higher rated (scored) by X, is also higher rated (scored) by Y, and are discordant if the customer higher rated (scored) by X is lower rated (scored) by Y. The two cases are neither concordant nor discordant if they are tied on X or Y or both.

When there are a large number of ties, the formula's denominator changes:

$$Kendall's\ tau_b = \frac{A - B}{\sqrt{[\frac{1}{2}n(n-1) - \sum_{i=1}^{t}\frac{1}{2}t_i(t_i-1)][\frac{1}{2}n(n-1) - \sum_{i=1}^{u}\frac{1}{2}u_i(u_i-1)]}}$$

where t_i and u_i are the number of observations (customers) tied in a particular score of X and Y, respectively. If there are no ties, the two formulas above are assumed to be the same.

These three methods are metrics of correlation, measuring the strength of the relationship between two variables. For them, -1 represents perfect disagreement and +1 represents perfect agreement between the variables. Increasing values imply increasing agreement between the scores.

The Kruskal's gamma metric is similar to the Kendall's tau metric, but it is based only on the number of concordant and discordant pairs of observations (Goodman and Kruskal, 1979). It ignores tied pairs and it is also appropriate only when both variables lie on an ordinal scale. It also ranges from -1 to 1. Gamma is defined as follows:

$$Gamma = \frac{A - B}{A + B}$$

where A represents the number of concordant pairs and B represents the number of discordant pairs.

Gamma is +1 if there are no discordant pairs, -1 if there are no concordant pairs, and 0 if there are equal numbers of concordant and discordant pairs. If the two variables are independent, then the estimator of gamma tends to be close to zero.

The performance measures for the classification tree model are shown in Table 62 for both training and test sets.

Metrics	Training Set (%)	Test Set (%)
CA*s*	68.28	63.86
Ca*m*	29.72	28.38
CA*h*	59.03	57.03
CA	52.13	49.90
Pearson correlation	47.09	42.26
Spearman correlation	47.10	42.20
Kendall's tau	42.43	37.88
Kruskal's gamma	60.51	54.57

Table 62. Classification tree performance measures for training and test sets

Based on this table, it is possible to visualise that both training and test sets presented variations in their performance values. Because the medium values were worst classified in both training and test sets, the CA of the model was not as good as one would expect. As a result, any decrease in performance could be considered not good for the model validity. However, considering the target variable under investigation and the results from the other models using the continuous target value, it would be expected that the performance values would not be too high. Also, Pearson correlation, Spearman rank-order correlation, Kendall's tau and Kruskal's gamma demonstrated a relatively good association between the predicted and target values.

The most interesting aspect in this analysis was that, as the model was analysed in accordance with domain knowledge, the results are more interpretable, making the model acceptable and useful for strategy definition. For example, if the purpose was to focus on customers with high CLV, these were reasonably classified and strategies could be defined focussing on the key predictive variables. These strategies would, consequently, affect some of the small and medium customers.

As a result, the use of domain knowledge would make it possible to evaluate the precise influence of some variables, making it possible to define strategies specifically or generally directed to each group of customers.

8.5. Conclusions

In this chapter, linear regression, decision trees and neural networks were applied for the benchmarking analysis of the CLV calculation. All techniques performed relatively well, especially considering the type of data under investigation, but neural networks proved to have the superior performance. Domain knowledge was then used in order to evaluate the models' usability and power when applied to a continuous or ordinal target variable.

As with the use of logistic regression for churn evaluation, the use of domain knowledge for CLV prediction helps to ensure that the variables presented in the linear regression model are showing the expected relationship with the target variable, with only limited loss of predictive power.

Then, it was explained how to discretise CLV into groups for the domain knowledge analysis in the form of a classification tree model. The data set investigation showed that the application of domain knowledge helped to ensure the understanding of variables and their impact in different group of customers. This would make it possible to define strategies appropriate for each group.

In summary, the overall performances for both linear regression and classification tree were not very high, but they are considered good for the type of target variable under investigation. When considering the application of domain knowledge to the linear regression model, the decrease in performance was fairly small. Also, the target variable adaptation for the classification tree model was of importance, in order to make it possible to use decision tables in the domain knowledge evaluation.

Chapter 9. Monitoring and Backtesting Churn Models

9.1. Introduction

During and after extracting a data mining model for use, it is necessary to have constant monitoring of the process, which would involve qualitative monitoring during development and to maintain consistency, as well as quantitative monitoring for validity evaluation and comparison.

This monitoring is necessary to guarantee that a model developed in the past is still appropriate for use in the future. That is why it is necessary to keep control of the whole process of extracting the model, as well as comparing it against other models and periodically evaluating it using more recent population data.

The purpose of this chapter is to adapt a methodology of backtesting, commonly applied in the context of credit risk, creating a framework for churn models monitoring. First, the monitoring approaches will be discussed. Next, the methodology for backtesting will be adapted for churn models. To illustrate the approach, it will then be applied to an out-of-time data set, comparing the results against the initial training model and initial scoring data set.

9.2. Monitoring Approaches for Model Assessment

After developing a model for prediction, it is necessary to have in place validation processes to constantly monitor the quality of the models in use. This monitoring can be done from qualitative and quantitative perspectives (Castermans *et al.*, 2007).

Where qualitative monitoring is concerned, data quality, model design, and documentation, are key elements to be taken into consideration, as well as the involvement of senior management to support the use and implementation of the model. Section 4.9 in Chapter 4 explored the improvement of churn models based on the data, so this aspect will not be further investigated in this chapter. It should be noted here though that the qualitative aspect is clearly essential for the development of a good and acceptable model.

From a quantitative monitoring perspective, benchmarking and backtesting are the key validation methods. Benchmarking compares the outcomes of a data mining model with a second model or more. This is done in order to assess the consistency of the estimated model parameters and outcomes. This process was applied in Chapter 4, when evaluating the performance of four data mining techniques in three different data sets.

The backtesting procedure is the focus of this chapter. It is related to contrasting ex-post observed reality with ex-ante made predictions. It presents challenges in its determination as different sources of variation can affect the data, for example, sample variation, external and internal effects (e.g. macro-economy and strategy change, respectively). For example, differences in variable values or the emergence of new relevant variables in out-of-time data sets can cause barriers to the use of backtesting procedures, indicating that the reference model would already not be adequate for future evaluation. Another important aspect is that data can be difficult to obtain, making it difficult to validate the model. Assuming data is available though, it is possible to use backtesting procedures to evaluate the model. Backtesting is widely used for credit scoring models, being a key requirement in a Basel II setting (Van Gestel, 2005; Castermans *et al.*, 2007; Van Gestel *et al.*, 2007). In this chapter, the aim is to discuss some of the tests used for credit scoring, and show how they can be incorporated into a churn backtesting procedure.

As discussed by Castermans *et al.* (2007), backtesting procedures take into consideration three key aspects of a model: calibration, discrimination and stability. Firstly, calibration of churn segments refers to comparing the actual probability of churn with the predicted values. Two measures that can be used for this evaluation are the Hosmer-Lemeshow test and the Binomial test.

The Hosmer-Lemeshow test (Hosmer and Lemeshow, 1989) is based on grouping of the estimated probabilities. The test statistic is defined as follows:

$$T = \sum_{i=1}^{k} \frac{(n_i p_i - \theta_i)^2}{n_i p_i (1 - p_i)}$$

where n_i is the number of customers in group i, p_i is the calibrated average estimated churn probability of group i, θ_i is the number of observed churners in group i and k is the number of churn segments.

The Hosmer-Lemeshow test converges towards a chi-square distribution with k-2 degrees of freedom. The test tells whether observed churn rates over the various segments significantly differs from the predicted values. Large values of T and small p-values indicate that there is difference between predicted and observed churn rates.

The binomial test (Sheskin, 2004; Moore and McCabe, 2006) contrasts the calibrated average estimated probability of a group against the observed churn rate, using the following hypothesis test:

H_0: the calibrated average estimated churn probability p of a group is correct
H_A: p is underestimated

When assuming that churners occur independently and H_0 (the null hypothesis) is true, the number of churners follows a standard normal distribution for growing n. Using similar parameters as for the Hosmer-Lemeshow test, the binomial test is then defined as:

$$z = \frac{\frac{\theta}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0,1)$$

where n is the number of customers in a specific group, p is the calibrated average estimated churn probability of that specific group and θ is the number of observed churners in that group. Given a significance level α, H_0 is then rejected if the observed churn rate in the group is greater than the specified critical value k^*, calculated as follows:

$$k^* = \Phi^{-1}(\alpha)\sqrt{\frac{p(1-p)}{n}} + p$$

where Φ^{-1} denotes the inverse standard normal distribution.

Whereas the former calibration tests refer to the calibration of probabilities, the discrimination evaluation of the model measures how well the model provides an ordinal ranking of the risk measure considered. In the case of churn prediction, this boils down to verifying how well the model distinguishes between churners and non-churners. The AUC performance measure may for example be used to evaluate the discrimination power. The DeLong *et al.* (1988) test is also used to compare if there is a significant difference between the AUC of the models. CA, sensitivity and specificity were also presented, as a way of identifying variations in classification. These performance metrics were further discussed in Chapter 4. Two other measures that can be used for discrimination analysis are the Kruskal's gamma and the Kendall's tau statistic. They are used to express how similar two rankings are, in this case, observed churn rates versus estimated probabilities. These methods were further investigated in Chapter 8.

To better differentiate between calibration and discrimination, discrimination corresponds to the rank ordering, meaning that churners should get high scores and non-churners should have low scores (or vice-versa, depending on which one is the target variable). Calibration testing means checking whether for a segment of customers for which 5% are predicted to churn, it turns out that indeed actually 5% churn.

In the churn context, discrimination is the most relevant aspect. Calibration is arguably somewhat less relevant for churn analysis, but it is very relevant though in other contexts, such as credit scoring. This is due to the fact that for churn, the main concern is in assigning high scores to churners (or vice-versa), not in assigning specific probabilities that a customer is going to churn. For credit scoring, however, Basel II requires the exact probabilities of a segment of customers going into default.

The third backtesting procedure involves the evaluation of population/data stability. Stability backtesting is used to check whether internal or external environmental changes will impact the predictive model in an unfavourable way. A change of business strategy, exploration of new market segments, and changes in organisational structure are examples of internal changes that can impact the model. As a result, it is necessary to check whether the population on which the model is currently being used (the out-of-time sample) is similar to the population that was used to develop the model. It can also be compared to a population observed in a different time period. If there is a difference

between the populations, it is necessary to verify the stability of individual variables used in the model.

To check the population stability, a stability index (SI) can be used to detect shifts in the churn segments. It can be computed as (Thomas *et al.*, 2002; Castermans *et al.*, 2007):

$$SI = \sum_{i=1}^{m} (R_i - O_i) \ln \frac{R_i}{O_i}$$

where m is the number of score segments or probability ranges for churn, R_i represents the percentage of the population (customers) in class i in year $t-1$ or at the time of model building, and O_i represents the percentage of the population in class i in year t. The higher the value of SI, the more substantial is the shift in the population. The rules of thumb below are generally used:

- $SI \leq 0.10$ → No significant shift.
- $0.10 < SI \leq 0.25$ → Minor shift.
- $SI > 0.25$ → Significant shift.

The rules above are independent of m. If the analysis of population stability detects significant shifts, it is necessary to further investigate which variables are responsible for these shifts. SI can then be calculated for each variable individually and the outcome may indicate which variables have changed and should be monitored. In this case, m is the number of classes or categories of the variable under investigation. The use of these rules may be too rigid, so it is necessary to also use expert intuition when interpreting them.

The SI is related to the Kullback-Leibler divergence (Kullback and Leibler, 1951; Kullback, 1959). The Kullback-Leibler divergence is calculated based on two probability distributions (for example, P and Q) of a discrete variable, in order to evaluate how much the distributions differ from each other. In the discrete case it can be obtained as follows:

$$KLD(P,Q) = \sum_{i} (P(i) - Q(i)) \log \frac{P(i)}{Q(i)}$$

If the distributions P and Q were of a continuous variable, the summation would be substituted by:

$$KLD(P,Q) = \int (p(x) - q(x)) \log \frac{p(x)}{q(x)} dx$$

where p and q are the densities of P and Q. Regardless of the type of variable under investigation, this divergence is a measure of the difficulty in discriminating between P and Q. This means that the smaller the divergence value, the more similar the values of P and Q are.

An important point to be taken into consideration is that, as with the Kullback-Leibler divergence, the SI has a symmetric distribution. This indicates that, by reversing the role of the two populations in the definition of the measure, the resulting SI is the same.

For the SI calculation, a table is then constructed showing the population distribution across different segments, which can be represented by classes in a categorical variable or intervals if it is calculated based on a continuous variable. The divergence between the percentages of observations allocated to each one of the segments is going to define how similar the populations are. This is demonstrated in detail in subsection 9.4.2.

These methods discussed in this section are incorporated in a framework for churn backtesting in section 9.3 and they are then applied in the empirical evaluation, section 9.4.

9.3. Methodology for Backtesting Churn Models

Taking as a starting point the techniques discussed above, the framework shown in Figure 46 was then defined to be used for backtesting churn models.

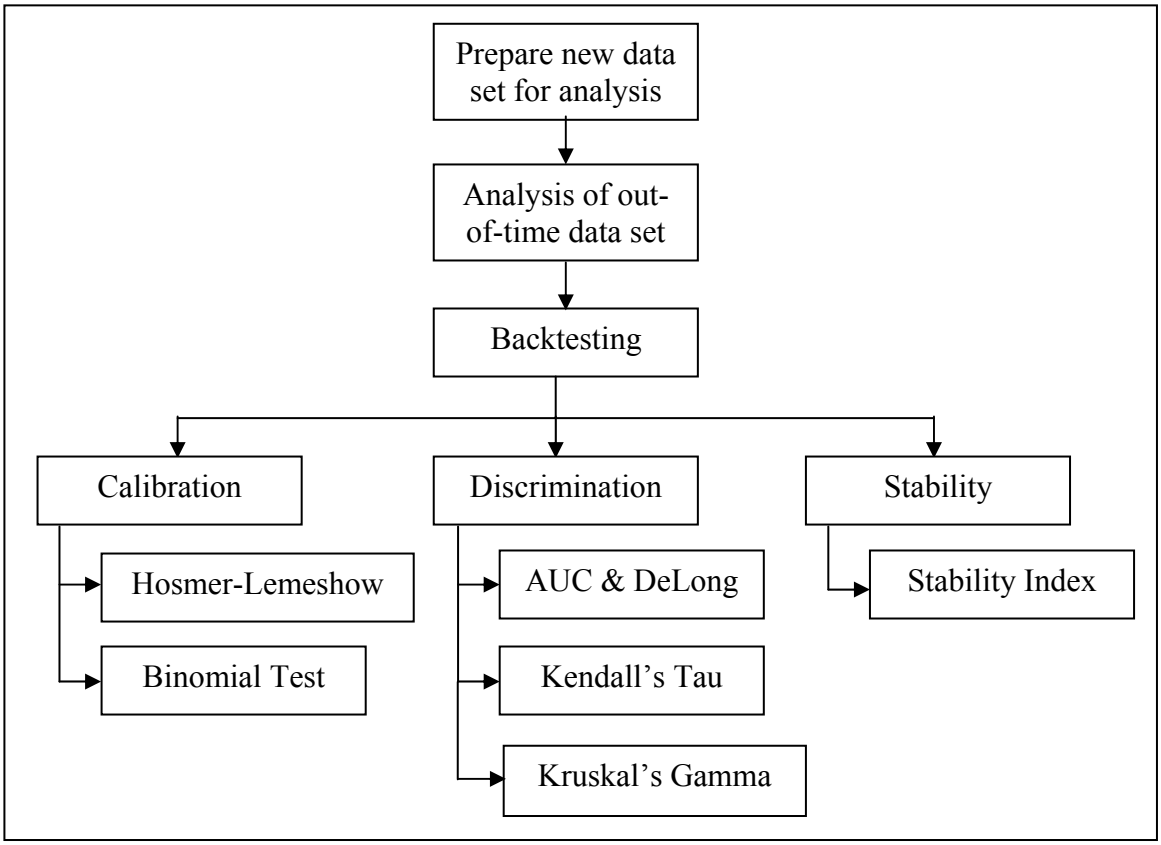

Figure 46. Backtesting framework for churn models

From this framework, it is possible to follow all the steps used in analysing an out-of-time data set, from its preparation and analysis, to its comparison against the original data set used to train the model:

- **Step 1:** Prepare new data set for analysis: missing values, outliers, transformations and manipulation of variables based on the transformations made on training and test sets. For example, if any variable was coarse classified into smaller groups in the training and test sets, this same coarse classification must be applied in the new data set.

- **Step 2:** Analysis of out-of-time data set: this involves scoring this data set using the model generated from the training set, and obtaining the performance measures (CA, sensitivity, specificity and AUC), using the same definitions as with the initial data set.

- **Step 3:** Backtesting – Discrimination: verifies how well the model distinguishes between churners and non-churners. In this case, the use of the DeLong *et al.* (1988) test to compare the AUC was the main aspect of the evaluation. The

monitoring of Kruskal's gamma and Kendall's tau statistics will also be used for the discrimination backtesting.

- **Step 4:** Backtesting – Stability: to test if changes in the population could adversely impact the predictive model. The use of the stability index is the key element in checking the stability of the model.

- **Step 5:** Backtesting – Calibration: to evaluate the actual probability of churning in comparison with the predicted values, the Hosmer-Lemeshow and the Binomial tests are used.

These results will then indicate if the model is appropriate for future application or if the model has to be adapted for any occurring changes.

9.4. Empirical Application of Backtesting Framework

In this section, the framework presented in Figure 46 was used to evaluate an out-of-time sample from one of the churn data sets previously evaluated in Chapter 4, Chapter 6 and Chapter 7. This data set is described in subsection 9.4.1 and the backtesting analysis takes place in subsection 9.4.2.

9.4.1. Data Set Description

As discussed in Chapter 4, the Telecom2 data set is a telecom data set used in the churn tournament 2003, organised by Duke University. This data set was used for evaluating various prediction models.

For this analysis, a distinct sample of 5,000 customers was selected from the out-of-time data set, with a proportion of 1.8% churners, as defined in the population data set. The variable selection and manipulation used on training and test sets was also assumed for the out-of-time sample. The data set was then ready for analysis.

9.4.2. Data Set Analysis

The out-of-time sample was also evaluated using SAS 9.1 and Enterprise Miner. After scoring the results of the out-of-time sample, the performance measures for the three data sets are shown in Table 63. When analysing the performance of the model, the measures used were the AUC, and the CA, sensitivity and specificity on training and test sets, assuming the KS statistic as the basis for cut-off, as described in previous chapters.

This data set was scored based on the resulting models from logistic regression and decision trees, which were manipulated based on domain knowledge. The purpose was to visualise how these models would behave with out-of-time data sets.

From the CA, sensitivity and specificity point of view there was no major difference in the performance values, keeping a similar variation in values. This is especially true when comparing the values for the test set and the out-of-time sample. In this case, the CA even demonstrated a slight improvement for the out-of-time sample in both logistic regression and decision tree models.

Data Set	Measurement	Training (%)	Test (%)	Out-of-time (%)
LR	CA	61.28	64.01	65.15
	Sensitivity	64.58	64.44	61.36
	Specificity	57.99	64.00	65.22
	AUC	64.56	65.15	65.87
DTree	CA	60.90	52.28	57.50
	Sensitivity	65.68	74.44	71.11
	Specificity	56.12	51.87	57.25
	AUC	63.44	64.44	64.87

Table 63. Performance metrics for Telecom2 data sets

Table 64 and Table 65 show the AUC values for both logistic regression and decision tree models, respectively. These tables are used for backtesting discrimination power of the models. It can be seen that the AUC values are again very similar when comparing training, test and out-of-time scores, with out-of-time score even showing a small improvement in value.

	# Obs	# Churners	AUC
Training Set	10,000	5,000	64.56
Test Set	5,000	90	65.15
Out-of-time Sample	5,000	90	65.87

Table 64. Monitoring churn on logistic regression model using AUC

	# Obs	# Churners	AUC
Training Set	5,000	2,500	63.44
Test Set	5,000	90	64.44
Out-of-time Sample	5,000	90	64.87

Table 65. Monitoring churn on decision tree model using AUC

The DeLong *et al.* (1988) test showed that there was no significant difference between the original and amended models, as demonstrated in Figure 47 for the decision tree model. This demonstrates that the discriminating power of the model still applies to the out-of-time sample.

```
AUC and DeLong Comparison at 95% Confidence Intervals

                                AUC
                  TRAINING     0.6354
                  TEST         0.6403
                  OUT-OF-TIME  0.6489

               Contrast Coefficients
                       TRAIN       TEST      FUTURE
             Row1          1         -1          0
             Row2          1          0         -1
             Row3          0          1         -1

Tests and 95% Confidence Intervals for Contrast Rows
          Estimate Std Error Chi-square  P-value
   Row1   -0.0049    0.0093   0.2838     0.5942
   Row2   -0.0136    0.0098   1.9289     0.1649
   Row3   -0.0086    0.0049   3.0998     0.0783

            Overall P-value = 0.1510
```

Figure 47. DeLong evaluation of training, test and out-of-time samples for decision tree – Telecom2

Kruskal's gamma and the Kendall's tau statistic were then used, also with the purpose to evaluate the discrimination power of the models. The results are displayed in Table 66.

Model	Data sets	Kendall's tau	Kruskal's gamma
Logistic Regression	Training Set	0.2059	0.2911
	Test Set	0.0571	0.3030
	Out-of-time Sample	0.0591	0.3173
Decision Tree	Training Set	0.2196	0.3509
	Test Set	0.0642	0.3708
	Out-of-time Sample	0.0679	0.4003

Table 66. More discrimination metrics for training, test and out-of-time samples

From the results in Table 66, it can be seen that there are more concordant pairs than discordant pairs in all data sets evaluated. The Kendall's tau statistic measures the agreement between the observed and estimated values. Its values for both data mining models show that there is no strong agreement between the churn probabilities. Nevertheless, as their results are very similar for test set and out-of-time sample, they support the use of the model for the out-of-time sample, taking the test set under consideration. The higher values for the training set could be due to the fact that there are more churners in this oversampled data set.

Taking the Kendall's tau formula for no ties:

$$Kendall's\ tau = \frac{A - B}{\frac{1}{2}n(n-1)}$$

It is correct to assume that for no ties, $\frac{1}{2}n(n-1) = A + B$ (Kendall, 1970; Gibbons, 1993).

In this case, the Kendall's tau formula would be assumed the same as the Kruskal's gamma:

$$Gamma = \frac{A - B}{A + B}$$

The difference lies on the fact that Kendall's tau uses the total amount of possible pairs in the denominator. As a result, it uses a correction for tied pairs (pairs of rows that have equal observed values (X) or equal estimated values (Y)) when the number of tied pairs

is large. This could be one of the reasons for the large change in the Kendall's tau statistic values (the amount of tied pairs).

Using the formula below (as described in subsection 8.4.7):

$$Kendall's\ tau_b = \frac{A - B}{\sqrt{[\frac{1}{2}n(n-1) - \sum_{i=1}^{t}\frac{1}{2}t_i(t_i-1)][\frac{1}{2}n(n-1) - \sum_{i=1}^{u}\frac{1}{2}u_i(u_i-1)]}}$$

t_i and u_i are the number of observations tied in a particular score of X and Y, respectively. For example, for the logistic regression model with no ties on the predicted value Y (it is a continuous probability distribution) and with the value of A and B, for the training set with the same amount of churners and non-churners, the formula would be represented as:

$$\tau_{training} = \frac{16074949 - 8826219}{\sqrt{[\frac{1}{2}10000(10000-1) - \left[\frac{1}{2}5000(5000-1) + \frac{1}{2}5000(5000-1)\right]] * [\frac{1}{2}10000(10000-1)]}} = 0.21$$

For the test set with a small proportion of churners, the formula is represented as:

$$\tau_{test} = \frac{286034 - 153006}{\sqrt{[\frac{1}{2}5000(5000-1) - \left[\frac{1}{2}4910(4910-1) + \frac{1}{2}90(90-1)\right]] * [\frac{1}{2}5000(5000-1)]}} = 0.057$$

For the out-of-time same, also with a small proportion of churners:

$$\tau_{out} = \frac{284567 - 147478}{\sqrt{[\frac{1}{2}5000(5000-1) - \left[\frac{1}{2}4910(4910-1) + \frac{1}{2}90(90-1)\right]] * [\frac{1}{2}5000(5000-1)]}} = 0.059$$

The values presented are approximated. The purpose is to demonstrate the impact of the sample proportions when calculating this statistic. It is possible to notice that the values for the test set and out-of-time sample are significantly smaller than the value for the training set. The disproportion in the amount of tied pairs on the data sets makes the denominator in the Kendall's tau formula for tied pairs increase more for the test and out-of-time data sets (in comparison to its reduction in the numerator) than for the training set, reducing their statistic value. This shows that Kendall's tau was affected by

the large class imbalance in the test set and out-of-time sample and by the high amount of tied pairs, demonstrating that it was not appropriate in this context.

The Kruskal's gamma performs a similar analysis to the Kendall's tau statistic, but it does not take into consideration tied pairs, only using the number of concordant and discordant pairs in its calculation. Sheskin (2004) recommends the use of Kruskal's gamma instead of Kendall's tau statistic when there are many ties in the data sets.

Sheskin (2004) also argues that as the number of ties increase, the value of Kruskal's gamma will become increasingly larger relative to the absolute value of Kendall's tau, as demonstrated in Table 66. This argument is supported by Gibbons (1993), who goes further and states that if ties exist, the maximum value of Kendall's tau will no longer be equal to one (1). This is because ties are not counted in calculating A and B, hence the total number of pairs is no longer $\frac{1}{2}n(n-1)$, while Kruskal's gamma can still achieve the absolute value of one (1). This is true because the denominator of Kruskal's gamma is equal to the number of pairs that are either concordant or discordant, not including ties.

As a result, Kruskal's gamma is not affected by the class imbalance in the data sets, demonstrating a more even result in terms of its value regarding training, test and out-of-time data sets in this research. The values above zero indicate that there are more concordant than discordant pairs and the results indicate that there is reasonable agreement between the observed and estimated values.

Next, the stability index (SI) was used to measure the population stability. This is done in terms of training set against test set, training set against out-of-time sample, and test set against out-of-time sample. The training set is the reference population, the test set is the population at year *t-1* and the out-of-time sample is the population at year *t*.

First this analysis is done for the target variable, churn. As the data set under investigation had an oversampled training set (50% of churners) and both test and out-of-time data sets have the same amount of churners (1.8%), this general stability check would not work when test set or out-of-time sample are compared against the training set. As explained in Chapter 4, the proportion of churners was oversampled in order to

give the predictive model a better capability of detecting discriminating patterns, but the test and out-of-time data sets were not oversampled to provide more realistic evaluation sets, according to a specified monthly churn rate of 1.8%.

Instead of evaluating the two classes "churn" and "not churn", the score range was discretised to create the churn segments. For the decision tree model, the segments were the scores used for estimation. For the logistic regression model, the classes of rating were based on an interval, as the predicted values range in a continuous way. Table 67 and Table 68 show the segments for the decision three and logistic regression models, respectively, with the particular amount of customers (in percentage) that were classified in each one of the segments.

Segment - DTree	Training Set	Test Set (at $t-1$)	Out-of-time (at t)
A	34.74%	42.70%	47.02%
B	7.50%	8.70%	9.70%
C	2.98%	2.44%	2.20%
D	32.30%	31.06%	30.20%
E	22.48%	15.10%	10.86%
SI reference year		0.0491	0.1311
SI previous year			0.0197

Table 67. Stability index calculation for ranges in the decision tree model

Segment - LR	Training Set	Test Set (at $t-1$)	Out-of-time (at t)
A	6.96%	10.45%	5.69%
B	17.84%	22.78%	21.55%
C	23.65%	25.17%	31.86%
D	25.16%	21.22%	17.42%
E	23.06%	17.77%	16.66%
F	3.33%	2.63%	6.81%
SI reference year		0.0493	0.1083
SI previous year			0.0934

Table 68. Stability index calculation for ranges in the logistic regression model

Both tables show that there were no significant shifts between training and test sets and between test and out-of-time data sets, but there was minor shift regarding training set

and out-of-time sample. This shift shows that the predictive model may not be applicable for the out-of-time sample. Note that care must be taken when interpreting these results. In this analysis, the shifts may happen based on the difference in class distribution between training and test sets, and between training set and out-of-time sample.

If there are any significant shifts in the population, as maybe indicated by the analysis above, it is advisable to calculate the stability for all the variables used in the logistic regression and decision tree models.

To demonstrate how the values for the stability index were obtained, the details of the calculations are shown for one of the variables. For the variable "refurb_new", the data details and results are shown in Table 69.

refurb_new	Training Set (a)	Test Set (at *t* – 1) (b)	Out-of-time (at *t*) (c)
N	84.11%	83.96%	85.42%
R	15.89%	16.04%	14.58%
SI reference year		0.0000167	0.0013
SI previous year			0.0016

Table 69. Stability index calculation for "refurb_new" variable

To obtain the values of SI for Table 69, the calculations are demonstrated below.

For training and test sets:

$$SI = \sum_{i=1}^{2} (a_i - b_i) \ln \frac{a_i}{b_i} = 0.0000027 + 0.000014 = \underline{\mathbf{0.0000167}}$$

- $(0.8411 - 0.8396)\ln\frac{0.8411}{0.8396} = 0.0000027$
- $(0.1589 - 0.1604)\ln\frac{0.1589}{0.1604} = 0.000014$

For training set and out-of-time sample:

$$SI = \sum_{i=1}^{2} (a_i - c_i) \ln \frac{a_i}{c_i} = 0.000202 + 0.001127 = \underline{\mathbf{0.0013}}$$

- $(0.8411 - 0.8542)\ln\frac{0.8411}{0.8542} = 0.000202$
- $(0.1589 - 0.1458)\ln\frac{0.1589}{0.1458} = 0.001127$

For test set and out-of-time sample:

$$SI = \sum_{i=1}^{2}(b_i - c_i)\ln\frac{b_i}{c_i} = 0.000252 + 0.001393 = \underline{\mathbf{0.0016}}$$

- $(0.8396 - 0.8542)\ln\frac{0.8396}{0.8542} = 0.000252$
- $(0.1604 - 0.1458)\ln\frac{0.1604}{0.1458} = 0.001393$

From these results, it is demonstrated that there is no significant shift in this variable, which means that it is not necessary to monitor the changes on it.

For the second variable, "eqpdays", the stability results are shown in Table 70.

Eqpdays	Training Set	Test Set	Out-of-time
< 200	21.55%	28.10%	28.64%
200 ≤ eqpdays < 600	35.11%	37.42%	41.08%
600 ≤ eqpdays < 1200	36.97%	29.00%	16.96%
≥ 1200	6.37%	5.48%	13.32%
SI reference year		0.04	0.36
SI previous year			0.14

Table 70. Stability index calculation for "eqpdays" variable

The results in Table 70 show that there is no significant shift when comparing the variable "eqpdays" between test and training sets. There is a minor shift between out-of-time and test sets, and a major shift between out-of-time and training sets. This result indicates that this variable would require close monitoring. The analysis then continued for the remaining predictive variables, with those variables presenting a significant shift being taken into consideration first, when results are aversely affected.

The Hosmer-Lemeshow test was then used for calibration. Following Hosmer and Lemeshow (1989), the data set was ordered by the estimated probabilities, where the first group contains the smallest estimated probabilities and the last group contains the largest estimated probabilities. Ranges were adopted for all three data sets when evaluating the decision tree and the logistic regression model. For the decision tree model, the ranges were based on the confidences values used for estimation. For the logistic regression model, the classes of segments were created based on an interval, as the values of the predicted values range in a continuous way. As a result, four degrees of freedom (k-2 = 6-2 = 4) and three degrees of freedom (k-2 = 5-2 = 3) were assumed for the logistic regression and the decision tree models, respectively, where k is the number of segments being evaluated.

An important aspect to be taken into consideration is that, as the churners were oversampled in the training set to create a 50-50 split between churners and non-churners, but oversampling was not undertaken in creating the test and out-of-time data sets, it was necessary to use an adjustment procedure. Saerens *et al.* (2002) describe an adjustment procedure for when the population prior churn probabilities are known. The method is a simple application of Bayes' theorem, where an adjustment factor is used to calculate the calibrated average estimated churn probability for each one of the segments. In simple terms, it is done as follows:

- The original training set has the following churn probabilities: p*(C) = 0.018 and p*(NC) = 0.982
- The sample training set has the following churn probabilities: p(C) = 0.5 and p(NC) = 0.5
- For each segment X of the training set, the calibrated average estimated churn probability p(C|X) needs to be adjusted. Following Bayes' rule, p*(C|X) and p*(NC|X) are then defined as:

$$\mathrm{p}^*(\mathrm{C} \mid \mathrm{X}) = \frac{\frac{p^*(C)}{p(C)} p(C \mid X)}{\mathrm{a}} \quad \text{and} \quad \mathrm{p}^*(\mathrm{NC} \mid \mathrm{X}) = \frac{\frac{p^*(NC)}{p(NC)} p(NC \mid X)}{\mathrm{a}}$$

in such a way that p*(C|X) + p*(NC|X) = 1. As a result, the factor "a" is defined as:

$$a = \frac{p^*(C)}{p(C)} p(C \mid X) + \frac{p^*(NC)}{p(NC)} p(NC \mid X)$$

The adjusted churn probabilities (represented by p^*) were then used for the Hosmer-Lemeshow and binomial tests. From Table 71, it is possible to notice that the higher p-values indicate that there is no significant difference between predicted and observed values. It is possible to verify that the p-values are higher for the out-of-time sample for both logistic regression and decision tree models, which could represent a better fit of the models when comparing with the test set.

	Results	Decision Tree	Logistic Regression
Test Set	*T*	4.3791	5.1828
	p-value	0.2233	0.2691
Out-of-time Sample	*T*	2.6461	2.2684
	p-value	0.4495	0.6865

Table 71. Hosmer-Lemeshow calculation for training, test and out-of-time samples

The Binomial test was then used to evaluate each one of the segments individually. The results are presented from Table 72 to Table 75.

Segment	*p**	θ/*n*	*k** *(at 95%)*	*z*	P-value
A	0.007	0.0058	0.013	-0.3400	0.6331
B	0.009	0.0132	0.0136	1.5004	0.0668
C	0.015	0.0096	0.0206	-1.5850	0.9435
D	0.025	0.0265	0.0329	0.3052	0.3801
E	0.029	0.0305	0.0383	0.2615	0.3969
F	0.045	0.0382	0.0748	-0.3772	0.6470

Table 72. Binomial test for calibration backtesting of logistic regression model – test set

Segment	*p**	θ/*n*	*k** *(at 95%)*	*z*	P-value
A	0.007	0.0106	0.0151	0.7203	0.2357
B	0.009	0.0065	0.0137	-0.8664	0.8069
C	0.015	0.0126	0.02	-0.7972	0.7873
D	0.025	0.0218	0.0337	-0.5972	0.7248
E	0.029	0.0288	0.0386	-0.0264	0.5105
F	0.045	0.0441	0.0635	-0.0785	0.5313

Table 73. Binomial test for calibration backtesting of logistic regression model – out-of-time sample

For the logistic regression model, Table 72 and Table 73 indicate that the model is not significantly underestimating the churners in the test and out-of-time samples. In both data sets, the observed churn rate for each segment (θ/*n*) is smaller than the critical value (*k**). Also, based on the binomial analysis (*z*), the p-values are higher than 0.05, indicating that *p** was not underestimated at the 95% significance level, providing good results in terms of calibration.

Segment	*p**	θ/*n*	*k** *(at 95%)*	*z*	P-value
A	0.01	0.0075	0.0135	-1.1637	0.8777
B	0.014	0.0161	0.0232	0.3714	0.3552
C	0.016	0.0328	0.0347	1.4777	0.0697
D	0.025	0.0258	0.0315	0.1910	0.4243
E	0.036	0.0305	0.0472	-0.8166	0.7929

Table 74. Binomial test for calibration backtesting of decision tree model – test set

Segment	*p**	θ/*n*	*k** *(at 95%)*	*z*	P-value
A	0.01	0.0098	0.0134	-0.1037	0.5413
B	0.014	0.0062	0.0228	-1.4648	0.9285
C	0.016	0.0182	0.0357	0.1824	0.4276
D	0.025	0.0265	0.0316	0.3709	0.3554
E	0.036	0.0405	0.0491	0.5648	0.2861

Table 75. Binomial test for calibration backtesting of decision tree model – out-of-time sample

From Table 74 and Table 75, it is possible to verify that the decision tree model also does not significantly underestimate churn for all the segments in both test and out-of-

time data sets, demonstrating good results for the calibration backtesting at the 95% significance level.

In summary, the discrimination performance measures indicate that the model performs better than random selection, and the stability and calibration also demonstrated good results. Special attention should be given to the outcomes from the model, with constant monitoring and maybe the development or adaptation of the models, when needed. This will be further discussed in the next section.

9.5. Action Plans

As discussed previously, it is necessary to have constant monitoring during and after model extraction, in order to maintain consistency and validity. When monitoring a model for use in an out-of-time data set, this procedure will have the purpose of guaranteeing that the model under evaluation is still applicable in the future.

To achieve this, it is necessary to maintain data quality, for example by using recent data for the model development and having continuous data updates, in order to support the model usage for current customers. Another aspect is to keep track of the model design, having specific information, such as the period when the model was developed, what kind of data was used and how the sample was constructed.

Maintaining this level of control requires the use of proper documentation. This should be of easy understanding, in order to facilitate the smooth transition if a new team was to continue development or production of the existing predictive model. As a result, all the steps of the model development and monitoring should be effectively recorded.

After these aspects are checked and corrected accordingly, it is necessary to use the backtesting procedures to evaluate if the model is still adequate. If the model becomes outdated, for example, because of differences in data content and variables definition, the most recommended option would be to generate a new model for this out-of-time data set. This model could then be compared with the old reference model, in order to identify the main differences, and proper documentation would follow. This new model would then become the reference model for future evaluations, but the old model could be kept for consultation if needed.

An interesting option would be the use of the original model as the basis for the development of a new model. This means that the variables that presented good results in the stability backtesting would be ensured in the new model, also allowing the model to select other variables that would be relevant. Again, this new model would become the reference model for future evaluations and monitoring would be required for its use with out-of-time data sets.

9.6. Conclusions

This chapter explored the use of backtesting procedures to evaluate a model generated from churn analysis. Based on existing procedures for backtesting credit scoring problems, the methods were analysed and modified to be appropriate in a churn context. As a result, a new framework for backtesting churn models was proposed. This was then applied on an out-of-time data set, to compare its results with the ones generated by the reference data set (training set) and score data set (test set).

The results were stable and coherent, presenting expected values when taking into consideration discrimination and stability backtesting. The backtesting calibration also showed that the models built earlier generalise well to a later time period, suggesting that, although arguably less relevant in the churn setting than it is in a credit scoring context, its analysis can provide the foundation to make the model acceptable for future evaluation. The results then emphasised the applicability of the backtesting procedures.

Chapter 10. Conclusions and Further Research

10.1. Introduction

During the writing of this thesis, it was discussed how to develop models that could incorporate domain knowledge into the data mining process, when investigating CLV and churn. In the next section, an overview of the thesis will be provided, based on the discussion of each one of the chapters presented in this work. Finally, some issues for further research will be outlined, emphasising some points that could add value to the results from this research.

10.2. Review of Chapters and Major Conclusions

Throughout this research it is emphasised that the use of data mining is key to evaluate churn and CLV. Their analyses are becoming very important to companies, but it is necessary to make the data mining models more understandable and compliant with domain knowledge. As emphasised before, the purpose is not only to identify which customers are more valuable or will churn, but to help companies identify the main elements in their data that can contribute positively or negatively to the relationship with the customer. The company could use the final information regarding the variables to define strategies that would be beneficial for the whole population of customers, instead of targeting only specific customers.

With that in mind, Chapter 2 and Chapter 5 are used as key literature reviews to evaluate CLV and churn, and domain knowledge, respectively. In Chapter 2, the concepts of churn and CLV were examined, considering how they have been investigated in previous research, and important conclusions about their theoretical aspects were drawn. One of the key findings, which apply to both churn and CLV evaluation, was that the data available will indicate what kind of analysis will be done. Also, attention to the variables and simplicity in the calculation make it possible to obtain reliable results, avoiding the risk of being misled or using parameters that are not adequate. As emphasised in Chapter 2, this research had no intention of invalidating previous methods of investigating CLV and churn. The purpose was to develop models for their prediction that could be more understandable and easier to interpret.

Taking this into consideration, the use of domain knowledge worked as a key factor in the data mining analysis. It was then necessary to understand and explore how domain knowledge is applied in data evaluation and analysis, and how its importance is perceived when obtaining the results. Chapter 5 explored the domain knowledge concept, demonstrating that it can be used in a variety of situations and its use applies from the pre-processing stage, to the data analysis and evaluation of results. It can assume many forms such as data preparation, insertion of rules, and monotonicity constraints. For this research, the monotonicity approach was used. A methodology was then defined to incorporate domain knowledge into the data mining analysis, using an intuitive and practical approach that facilitates the interpretability of the models.

To support the application of domain knowledge when developing understandable models for churn and CLV, a qualitative survey was executed. This was undertaken in order to evaluate if and how companies implement their customer analysis, investigating how key concepts such as churn, CLV, domain knowledge, data mining and decision tables, are perceived by companies. This resulted in acquiring a fresh view on how the concepts of CLV and churn are being approached in practice by companies. In this case, it was possible to conclude that when evaluating different companies, with different sizes and sectors, no great variation was found in their attitude towards customers. However, depending on their level of customer research, this would influence their ability of better implementing their customer metrics.

Some key findings from the survey were that, a) CLV is viewed as important, but not evaluated by the majority of companies interviewed; b) churn is also important, but most of the companies emphasise the difficulty of predicting it. From the three companies that are evaluating churn, two of them assume that the churn prediction is used as an element in the CLV calculation, demonstrating the importance of integrating these two elements; c) domain knowledge is a key element in the data analysis, where the idea of using common sense leads to a more detailed evaluation, using the knowledge acquired to guide and influence the analysis; d) data mining is more in use by companies that are developing or implementing a deeper customer evaluation, which includes churn and CLV predictions.

From this, it was concluded that there is always the chance the customer is going to leave, and the companies are aware of that. What they are trying to do is to define methods that can help them identify which customers to focus on, which ones have a greater propensity to leave and which ones are going to stay. However, this is not an easy task and not all the companies are prepared for this type of analysis. As a result, domain knowledge could be used in their analysis, and to do that, they need to understand their data, prepare it well and make sure that the results presented are coherent and useful.

So, before applying domain knowledge to the final models, a benchmarking study was executed in Chapter 4, exploring some key data mining techniques and their performance when evaluating churn. The data mining techniques used were briefly explained, but the focus was on the evaluation of results. In this chapter, the use of four data mining techniques (logistic regression, decision trees, k-nearest neighbour and neural networks) was explored on three different churn data sets. The purpose was to set a benchmarking of how the techniques perform for the evaluation of churn, defining the two techniques that were going to be further explored in Chapter 6 and Chapter 7.

It was concluded that decision trees and logistic regression are well-known and widely used techniques for churn and general prediction problems. These two techniques were chosen to be further investigated based on their predictive characteristics and facility of interpretation and understanding. Neural networks could also be used, especially as they presented the best results, but the complexity added would make them difficult to interpret. K-nearest neighbours presented the worst performance measurement, and as a result, this technique was excluded from further investigation.

Also in Chapter 4, it was emphasised that independently of the data mining technique used, it is necessary to improve the data quality, through, for example, the evaluation of missing values, outliers, and the use of data definition and data enhancement. These are ways of analysing and improving the quality of the data under investigation. The use of domain knowledge at this point can also prove itself useful, as it provides insights into the pre-processing stage, increasing the reliability of the data collected.

As a consequence of the findings from these previous chapters, logistic regression and decision trees were further investigated, in order to make their results more compliant

with the knowledge in the company. The empirical analysis of two telecom data sets demonstrated how domain knowledge can be integrated as part of the data mining process when predicting churn, through, firstly, the evaluation of signs in the logistic regression in Chapter 6 and secondly, the analysis of rules' monotonicity in decision tables in Chapter 7.

In terms of signs, Chapter 6 explored the concept of wrong sign, evaluating the causes and how it is perceived in the literature. Then, it was investigated how to evaluate wrong signs in a logistic regression model, through the use of a method developed to assist the enforcement of correct signs in combination with the logistic regression evaluation. This method would implement the integration of domain knowledge in the wrong sign evaluation. It was then concluded that logistic regression is a suitable choice of classifier for integrating domain knowledge into the model, as in the analysis of both data sets the model's performance stayed relatively stable with the introduction of domain constraints. The variables presented in both models were showing the expected relationship with the target variable, with only limited loss of predictive power.

Decision tables were explained in Chapter 7 and used to simplify and support the interpretation of rules generated by the decision tree model, with the purpose of analysing if the results presented were compliant with domain knowledge. An algorithm was used to check whether the knowledge contained in a decision table was in accordance with the domain knowledge available. It was then concluded that the variables presented in the final model were coherent and showing the expected relationship with the target variable, with no significant loss for the decision tree models.

An empirical evaluation of CLV was also performed in Chapter 8, using a DMEF data set originated from a catalogue company, in order to evaluate the CLV results through the use of domain knowledge. First, a benchmarking analysis was done, comparing linear regression, regression trees and neural networks, in order to evaluate the models generated when evaluating a continuous target variable. The results demonstrated that even though neural networks performed better, all three techniques performed relatively similar and well, especially considering the type of data under investigation. Then, domain knowledge was applied to the linear regression model, with the purpose to evaluate if its use in the context of a continuous or ordinal target variable was applicable

and presented valid results. The validity was then confirmed, and as with the use of logistic regression for churn evaluation, the use of domain knowledge for CLV prediction helped to ensure that the variables presented in the linear regression model were showing the expected relationship with the target variable, with only limited loss of predictive power. A classification task was then applied to the target variable, in order to more easily verify the domain knowledge in the decision tree model. The results demonstrated that this was a useful approach, and regarding the data set under investigation, it showed that the application of domain knowledge helped to ensure the understanding of variables and their impact in different groups. The adaptation of the target variable based on the classification task was important, in order to make it possible to use decision tables in the domain knowledge evaluation.

In the end, to maintain consistency, and to proceed with a valid evaluation and comparison of results, it is necessary to have constant monitoring of a model. This monitoring is necessary to guarantee that a model developed in the past is still appropriate for use in the future. Chapter 9 was then used to adapt a methodology of backtesting, creating a framework for churn model monitoring. First, the monitoring approaches were discussed and the methodology for backtesting adapted for churn modelling. Then, this methodology was applied to an out-of-time sample from one of the churn models previously investigated in Chapter 6 and Chapter 7, Telecom2, comparing the results with the initial training model and test set.

The results from Chapter 9 were stable and coherent, presenting expected values when taking into consideration discrimination and stability backtesting. The backtesting calibration also performed well in this context, working as a support in assessing the acceptability of the model for future evaluation. These results represent an important contribution from this research.

10.3. Issues for Further Research

Based on the conclusions listed in section 10.2, it is possible to identify some challenging issues that could be further investigated in future research.

Based on the evaluation of CLV, further topics of interest would be the calculation of CLV based on both finite and infinite time horizon, also applying different values of

discount rates, and based on a sensitivity analysis evaluating how close these values would be to the reality of the company. For this analysis to be done, it would be necessary to have a good source of data from a long period of time, in order to be able to compare the results from the infinite time horizon approach to CLV. Also, a retention rate or churn probability needs to be incorporated into the calculation, especially taking into consideration an infinite time horizon approach. This emphasises the need to integrate churn and CLV. In this research, these two elements were not put together, based on the data available and type of investigation done. On a further investigation of CLV, its integration with churn would represent an essential aspect.

Regarding the churn evaluation, in the first data set (Telecom 1) there was no specification of time period of investigation. Also, the variables were all static; dynamic variables would maybe show different behaviours, for example, an increase or decrease in usage could generate different effects. The second data set (Telecom 2) had a better structured set of variables, which may explain some of the difference in performance results. Further research could be done in order to apply this approach to different data sets from the same industry, to analyse if the behaviour of the variables would be similar. Also, the application of domain knowledge could go further in different industries, which could show its potential as a tool to support managerial and strategic decisions.

In terms of the qualitative approach, it would be an interesting approach to extend the interviews to more companies in different sectors, but keeping a number of these companies in the same sector. This would be useful to compare not only how companies in different sectors analyse and understand these key concepts, but also how companies from the same group implement their customers' evaluation. Focus groups within some of the companies would also be useful, as the information could flow more easily, adding value to the research results and also distributing the knowledge within the company. The results could then facilitate and instigate the implementation of customer analysis in the decision making process.

The domain knowledge application in the data mining modelling was of relevant importance in the churn and CLV evaluation. The evaluation of signs in the logistic and linear regression models and the evaluation of rules from the decision tree model (through the use of decision tables) demonstrated that there was limited loss of

performance. This indicates that the domain knowledge application is a valid and useful approach in identifying the key variables that are closely related to the target variable, and making sure that these variables are demonstrating their expected effects. The focus in this research was on logistic regression, linear regression and decision trees for the domain knowledge incorporation, but it could be extended to other data mining techniques, for example, for the k-nearest neighbours and neural networks models. This could be done in order to evaluate how these models could become more interpretable and easy to understand, in order for them to be acceptable for the definition of strategies based on the selected variables.

In terms of backtesting churn models, the framework presented is of significant validity. However, it would be interesting to apply it in different churn data sets, as this would solidify the methodology and add value to the procedures established. Also, although some procedures for backtesting were identified in the literature, further investigation could be done in order to incorporate different procedures, evaluating if they would add more value and keep the consistency of the generated churn model.

Finally, it would also be interesting to develop a backtesting procedure to evaluate CLV models, especially taking into consideration that out-of-time data would be available for validation of the framework. In any case, a constant control and understanding of the data under investigation is essential in the development of any churn or CLV model.

Figure 48 represents the adaptation of Figure 1 (presented in Chapter 2), demonstrating how the research developed itself and which factors have not yet been explored.

From this revised figure, it is possible to see the additional elements, represented by the domain knowledge and backtesting links. The dotted lines represent the parts of the framework that were not implemented in this research; nevertheless, the connection between CLV and churn was extensively discussed. The segmentation link presented in Figure 1 was eliminated, as it would be considered part of the strategy approach when focussing on an individual, groups or the whole population. Again, this would be a constant cycle, where the feedback from the strategies together with the measurements of CLV and churn allow for the evaluation of the strategy used, recalculating CLV and churn. In case the old model is no longer applicable, which is verified through the backtesting procedure, a new model is developed, making the whole process happen

again with the application of domain knowledge. This occurs in order to define the best strategies to reduce churn and increase the value of customers.

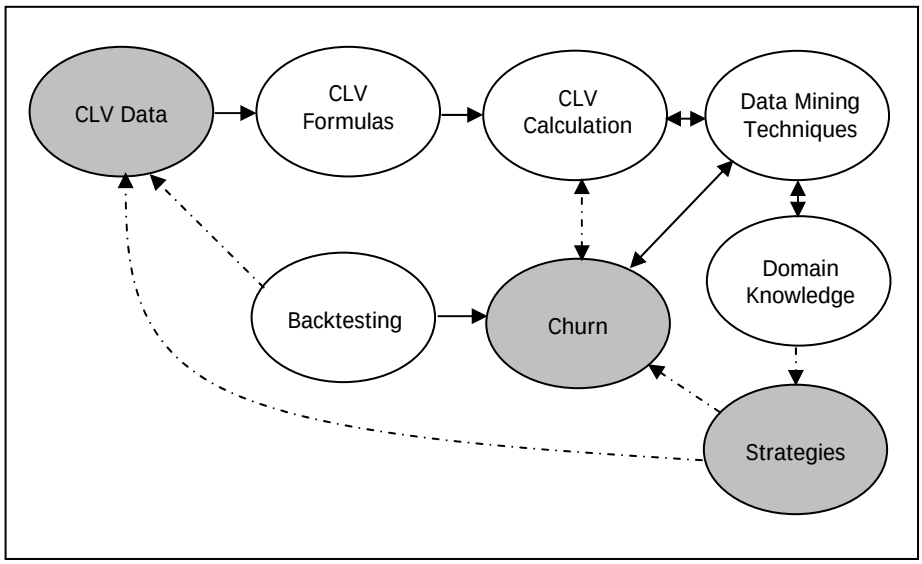

Figure 48. New CLV and churn research framework

References

ALMEIDA, P. and TORGO, L., 2001. The Use of Domain Knowledge in Feature Construction for Financial Time Series Prediction. *Lecture Notes in Computer Science: 10th Portuguese Conference on Artificial Intelligence,* 2258, 29-39.

AHN, J-H., HAN, S-P. and LEE, Y-S., 2006. Customer churn analysis: churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications Policy,* 30(10-11), 552-568.

ALLISON, P.D., 2001. *Logistic regression using the SAS system: theory and application.* Cary, NC: SAS Institute Inc.

ALONSO, F., CARAÇA-VALENTE, J.P., GONZÁLEZ, A.L. and MONTES, C., 2002. Combining expert knowledge and data mining in a medical diagnosis domain. *Expert Systems with Applications,* 23(4), 367-375.

ALTENDORF, E., RESTIFICAR, E. and DIETTERICH, T., 2005. Learning from sparse data by exploiting monotonicity constraints. *Proceedings 21st conf on Uncertainty in Artif Int: Edinburgh, Scotland.*

ANAND, S.S., BELL, D.A. and HUGHES, J.G., 1995. The role of domain knowledge in data mining. *Proceedings of the fourth international conference on Information and knowledge management,* Baltimore, Maryland, United States, pp. 37 – 43.

AWERBUCH, S. and DEEHAN, W., 1995. Do consumers discount the future correctly? A market-based valuation of residential fuel switching. *Energy Policy,* 23(1), 57-69.

BABUSIAUX, D. and PIERRU, A., 2001. Capital budgeting, investment project valuation and financing mix: methodological proposals. *European Journal of Operational Research,* 135(2), 326-337.

BAESENS, B. VIAENE, S., VAN DEN POEL, D., VANTHIENEN, J. and DEDENE, G., 2002. Bayesian Neural Network Learning for Repeat Purchase Modelling in Direct Marketing. *European Journal of Operational Research,* 138(1), 191-211.

BAESENS, B., SETIONO, R., MUES, C. and VANTHIENEN, J., 2003a. Using neural network rule extraction and decision tables for credit-risk evaluation. *Management Science,* 49(3), 312-329.

BAESENS, B., VAN GESTEL, T., VIAENE, S., STEPANOVA, M., SUYKENS, J. and VANTHIENEN, J., 2003b. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society,* 54(6), 627-635.

BARZILAY, O. and BRAILOVSKY, V. L., 1999. On domain knowledge and feature selection using support vector machine. *Pattern Recognition Letters,* 20(5), 475-484.

BEJAR, J., CORTES, U., SANGUEESA, R. and POCH, M., 1997. Experiments with domain knowledge in knowledge discovery. *Proceedings of the International conference on the Practical application of knowledge discovery and data mining,* London, 1, 65-78.

BEN-DAVID, A., 1995. Monotonicity maintenance in information-theoretic machine learning algorithms. *Machine Learning,* 19(1), 29–43.

BERGER, P.D. and NASR, N.I., 1998. Customer lifetime value: marketing models and applications. *Journal of Interactive Marketing,* 12(1), 17-30.

BERGER, P.D., WEINBERG, B. and HANNA, R.C., 2003. Customer lifetime value determination and strategic implications for a cruise-ship company. *Journal of Database Marketing & Customer Strategy Management,* 11(1), 40–52.

BERTI-ÉQUILLE, L., 2007. Data quality awareness: a case study for cost optimal association rule mining. *Knowledge and Information Systems,* 11(2), 191-215.

BISHOP, C.M., 1995. *Neural networks for pattern recognition.* Oxford University Press.

BLATTBERG, R.C. and DEIGHTON, J., 1996. Manage marketing by the customer equity test. *Harvard Business Review,* (July–August), 136–144.

BRAGA, A.C. and OLIVEIRA, P., 2003. Diagnostic analysis based on ROC curves: theory and applications in medicine. *International Journal of Health Care Quality Assurance,* 16(4), 191-198.

BREIMAN, L., FRIEDMAN, J., STONE, C.J. and OLSHEN, R.A., 1984. *Classification and regression trees.* Boca Raton, FL: Chapman & Hall.

BROWN, M.B. and BENEDETTI, J.K., 1977. Sampling Behavior of Tests for Correlation in Two-Way Contingency Tables. *Journal of the American Statistical Association*, 72, 309 - 315.

BUCKINX, W. and VAN DEN POEL, D., 2005. Customer base analysis: partial defection of behaviorally-loyal clients in a noncontractual FMCG retail setting. *European Journal of Operational Research,* 164(1), 252-268.

BUREZ, J. and VAN DEN POEL, D., 2007. CRM at a pay-TV company: using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications,* 32(2), 277-288.

CASTERMANS, G., MARTENS, D., VAN GESTEL, T., BART HAMERS, B. and BAESENS, B., 2007. An Overview and Framework for PD Backtesting and Benchmarking. *Credit Scoring and Credit Control X, Edinburgh (UK), July 30th 2007.*

CHERNICK, M.R. and FRIIS, R.H., 2003. *Introductory biostatistics for the health sciences.* Wiley-Interscience.

COUSSEMENT, K. and VAN DEN POEL, D., 2008a. Churn prediction in subscription services: an application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications,* 34 (1), 313-327.

COUSSEMENT, K. and VAN DEN POEL, D., 2008b. Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. *Expert Systems with Applications*, *In Press, Corrected Proof*, *Available online 17 July 2008.*

CRONE, S.F., LESSMANN, S. and STAHLBOCK, R., 2006. The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. *European Journal of Operational Research,* 173(3), 781-800.

DANIEL, W.W., 2005. *Biostatistics: A foundation for analysis in the health sciences.* 8th Edition, USA: John Wiley & Sons Inc.

DELONG, E.R., DELONG, D. M. and CLARKE-PEARSON, D. L., 1988. Comparing the areas under two or more correlated receiver operating curves: A nonparametric approach. *Biometrics,* 44(3), 837-845.

DESAI, V.S., CROOK, J.N. and OVERSTREET JR., G.A., 1996. A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research,* 95, 24-37.

DESPAGNE, D. and MASSART, D-L., 1998. Variable selection for neural networks in multivariate calibration. *Chemometrics and Intelligent Laboratory Systems*, 40(2), 145-163.

DJOKO, S., COOK, D.J. and HOLDER, L.B., 1997. An empirical study of domain knowledge and its benefits to substructure discovery. *IEEE Transactions on Knowledge and Data Engineering,* 9(4), 575 – 586.

DRYE, T., WETHERILL, G. and PINNOCK, A., 2001. When are customers in the market? Applying survival analysis to marketing challenges. *Journal of Targeting, Measurement, and Analysis for Marketing,* 10(2), 179–188.

DWYER, F.R., 1997. Customer lifetime valuation to support marketing decision making. *Journal of Direct Marketing,* 11(4), 6–13.

EMHJELLEN, M. and ALAOUZE, C.M., 2002. Project valuation when there are two cashflow streams. *Energy Economics,* 24(5), 455-467.

FADER, P.S., HARDIE, B.G.S. and LEE, K.L., 2005. RFM and CLV: using iso-value curves for customer base analysis. *Journal of Marketing Research,* 42(4), 415-430.

FADER, P.S., HARDIE, B.G.S. and JERATH, K. (2007). Estimating CLV using aggregated data: the Tuscan lifestyles case revisited. *Available at SSRN: http://ssrn.com/abstract=930745*

FEELDERS, A. and PARDOEL, M., 2003. Pruning for monotone classification trees. *In: Advanced in intelligent data analysis V: Springer,* 2810, 1–12.

GIBBONS, J.D., 1993. *Nonparametric measures of association.* Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-091. Newbury Park, CA: Sage.

GIUDICI, P., 2003. *Applied data mining: statistical methods for business and industry.* England: Wiley.

GLADY, N., BAESENS, B. and CROUX, C., 2006. Modeling customer loyalty using customer lifetime value. *Available at SSRN: http://ssrn.com/abstract=968584*

GOODMAN, L.A. and KRUSKAL, W.H., 1979. Measures of Association for Cross Classifications. New York: Springer-Verlag.

GUPTA, S. AND LEHMANN, D.R., 2003. Customers as assets. *Journal of Interactive Marketing,* 17(1), 9–24.

GUPTA, S., LEHMANN, D. and STUART, J., 2004. Valuing customers. *Journal of Marketing Research,* 41, 7–18.

HADDEN, J., TIWARI, A., ROY, R., and RUTA, D., 2007. Computer assisted customer churn management: state-of-the-art and future trends. *Computers & Operations Research,* 34(10), 2902-2917.

HAND, D., MANNILA, H. and SMYTH, P., 2001. *Principles of data mining.* Massachusetts: The MIT Press.

HANLEY, J. A. and MCNEIL, B. J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology,* 143(1), 29–36.

HANLEY, J. A. and MCNEIL, B. J., 1983. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology,* 148, 839-843.

HARPER, P.R. and WINSLETT, D.J., 2006. Classification trees: A possible method for maternity risk grouping. *European Journal of Operational Research,* 169 (1), 146-156.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J., 2001. *The elements of statistical learning, data mining, inference, and prediction.* Springer.

HENLEY, W.E. and HAND, D.J., 1997. Construction of a k-nearest neighbour credit-scoring system. *IMA - Journal of Management Mathematics,* 8(4), 305-321.

HEWETT, R. and LEUCHNER, J., 2003. Restructuring decision tables for elucidation of knowledge. *Data & Knowledge Engineering,* 46(3), 271-290.

HOSMER, D.W. and LEMESHOW, S., 1989. *Applied Logistic Regression***.** New York: John Wiley & Sons, Inc.

HWANG, H., JUNG, T. and SUH, E., 2004. An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. *Expert Systems with Applications,* 26(2), 181-188.

JACOBS, F.A., JOHNSTON, W. and KOTCHETOVA, N., 2001. Customer profitability: prospective vs. retrospective approaches in a business-to-business setting. *Industrial Marketing Management,* 30(4), 353-363.

JAIN, D. and SINGH, S.S., 2002. Customer lifetime value research in marketing: a review and future directions. *Journal of Interactive Marketing,* 16(2), 34-46.

KASS, G.V., 1980. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics,* 29(2), 119-127.

KEAVENEY, S.M., 1995. Customer switching behavior in service industries: An exploratory study. *Journal of Marketing,* 59(2), 71–82.

KENDALL, M., 1970. *Rank Correlation Methods*. 4th edition, London: Charles Griffin & Co.

KENNEDY, P., 2002. Oh No! I Got the Wrong Sign! What Should I Do?. *Discussion Papers,* Canada: Department of Economics, Simon Fraser University.

KIM, S., SHIN, K.S. and PARK, K., 2005. An application of support vector machines for customer churn analysis: credit card case. *ICNC 2005, Lecture Notes in Computer Science,* 3611, 636-647.

KIM, S., JUNG, T., SUH, E. and HWANG, H., 2006. Customer segmentation and strategy development based on customer lifetime value: a case study. *Expert Systems with Applications,* 31(1), 101-107.

KOHAVI, R. and SOMMERFIELD, D., 1998. Targeting business users with decision table classifiers. *In Proc. KDD '98: 4th Intl. Conf. on Knowledge Discovery and Data Mining,* New York City, pp. 249-253.

KOPANAS, I., AVOURIS, N.M. and DASKALAKI, S., 2002. The role of domain knowledge in a large scale data mining project. *In: Methods and Applications of Artificial Intelligence: Proceedings for the Second Hellenic Conference on AI,* SETN 2002. Thessaloniki, Greece, April 11-12, 2002.

KULLBACK, S. and LEIBLER, R.A., 1951. On information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79-86.

KULLBACK, S., 1959. Information Theory and Statistics. New York: John Wiley & Sons, Inc.

KUMAR, V., 2006. CLV: the databased approach. *Journal of Relationship Marketing,* 5 (2/3), 7-35.

KUMAR, V., RAMANI, G. and BOHLING, T., 2004. Customer lifetime value approaches and best practice applications. *Journal of Interactive Marketing,* 18(3), 60-72.

KUMAR, V., SHAH, D. and VENKATESAN, R., 2006. Managing retailer profitability—one customer at a time!. *Journal of Retailing,* 82(4), 277-294.

LARIVIÈRE, B. and VAN DEN POEL, D., 2004. Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: the case of financial services. *Expert Systems with Applications,* 27(2), 277–285.

LAROSE, D.T., 2005. *Discovering knowledge in data: an introduction to data mining.* New Jersey: Wiley.

MACSKASSY, S.A. and PROVOST, F., 2003. A simple relational classifier. *Proceedings of the Multi-Relational Data Mining Workshop (MRDM) at the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

MALTHOUSE, E.C. and BLATTBERG, R.C., 2005. Can we predict customer lifetime value?. *Journal of Interactive Marketing,* 19(1), 2-16.

MATIGNON, R., 2005. *Neural Network Modeling using SAS Enterprise Miner.* UK: AuthorHouse Inc.

MASAND, B., DATTA, P., MANI, D.R. and LI, B., 1999. CHAMP: a prototype for automated cellular churn prediction. *Data Mining and Knowledge Discovery,* 3(2), 219-225.

MASON, J., 2002. *Qualitative researching.* 2nd Edition, London: Sage Publications Ltd.

MASON, C.H., 2003. Tuscan lifestyles: assessing customer lifetime value. *Journal of Interactive Marketing,* 17 (4), 54-60.

MARTENS, D., DE BACKER, M., HAESEN, R., BAESENS, B., MUES, C. and VANTHIENEN, J., 2006. Ant-Based approach to the knowledge fusion problem. *In:* Dorigo M *et al.* (eds). *ANTS Workshop 2006, Lecture Notes in Computer Science,* 4150: 84–95.

MATHEUS, C.J., CHAN, P.K. and PIATETSKY-SHAPIRO, G., 1993. Systems for knowledge discovery in databases. *IEEE Transactions on Knowledge and Data Engineering,* 5(6), 903-913.

MAXHAM, J.G., 2001. Service recovery's influence on consumer satisfaction, positive word-of-mouth, and purchase intentions. *Journal of Business Research,* 54(1), 11–24.

MCCLEAN, S., SCOTNEY, B. and SHAPCOTT, M., 2000. Incorporating domain knowledge into attribute-oriented data mining. *International Journal of Intelligent Systems,* 15(6), 535-548.

MCDOUGALL, D., WYNER, G. and VAZDAUSKAS, D., 1997. Customer valuation as a foundation for growth. *Managing Service Quality,* 7(1), 5-11.

MCCLURE, B., 2003. Investors need a good WACC. *Available at: http://www.investopedia.com/articles/fundamental/03/061103.asp on 28/06/2006.*

MELTZER, M., 2002. Are your customers profitable?. *Customer Management Zone,* August 2002.

MOORE, D.S. and MCCABE, G.P., 2006. *Introduction to the practice of statistics.* 5th Edition, New York: W.H. Freeman and Company.

MUES, C., BAESENS, B., FILES, C.M., VANTHIENEN, J., 2004. Decision Diagrams in Machine Learning: an Empirical Study on Real-Life Credit-Risk Data. *Expert Systems with Applications,* 27(2), 257-264.

MULLET, G., 1976. Why regression coefficients have the wrong sign. *Journal of Quality Technology,* 8(3), 121-126.

NATH, S.V. and BEHARA, R.S., 2003. Customer churn analysis in the wireless industry: a data mining approach. *Proceedings - Annual Meeting of the Decision Sciences Institute,* 2003, 505-510.

NESLIN, S.A., GUPTA, S., KAMAKURA, W., LU, J. and MASON, C.H., 2006. Defection detection: measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research,* 43(2), 204-211.

NEVILLE, P.G., 1999. *Decision trees for predictive modeling.* SAS Institute Inc.

PIRAMUTHU, S., 2004. Feature construction for reduction of tabular knowledge-based systems. *Information Sciences,* 168 (1-4), 201-215.

PFEIFER, P.E. and CARRAWAY, R.L., 2000. Modeling customer relationships as markov chains. *Journal of Interactive Marketing,* 14(2), 43–55.

PFEIFER, P.E. and BANG, H., 2005. Non-parametric estimation of mean customer lifetime value. *Journal of Interactive Marketing,* 19(4), 48-66.

PRINZIE, A. and VAN DEN POEL, D., 2006. Exploiting randomness for feature selection in multinomial logit: a CRM cross-sell application. *ICDM 2006,* pp.310-323.

QUINLAN, J.R., 1993. *C4.5: programs for machine learning.* San Francisco, CA: Morgan Kaufmann.

RAJAGOPALAN, B. and ISKEN, M.W., 2001. Exploiting data preparation to enhance mining and knowledge discovery. *IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews,* 31(4), 460-467.

REICHHELD, F., 1996. *The loyalty effect.* Boston: Harvard Business School Press.

REICHHELD, F.F., MARKEY Jr, R.G. and HOPTON, C., 2000. The loyalty effect – the relationship between loyalty and profits. *European Business Journal,* 12(3), 134-139.

REINARTZ, W.J. and KUMAR, V., 2000. On the profitability of long-life customers in a noncontractual setting: an empirical investigation and implications for marketing. *Journal of Marketing,* 64(4), 17–35.

REINARTZ, W. and KUMAR, V., 2003. The impact of customer relationship characteristics on profitable lifetime duration. *Journal of Marketing,* 67(1), 77–99.

RIPLEY, B.D., 1996. *Pattern recognition and neural networks.* Cambridge University Press.

ROSSET, S., NEUMANN, E., EICK, U. and VATNIK, N., 2003. Customer lifetime value models for decision support. *Data Mining and Knowledge Discovery,* 7(3), 321-339.

RUD, O.P., 2001. *Data mining cookbook: modelling data for marketing, risk, and customer relationship management.* New York: John Wiley & Sons, Inc.

RUST, R.T., ZEITHAML, V.A. and LEMON, K.N., 2000. *Driving customer equity, how customer lifetime value is reshaping corporate strategy,* New York: The Free Press.

RUST, R., LEMON, K. and ZEITHAML, V., 2004. Return on marketing: using customer equity to focus marketing strategy. *Journal of Marketing,* 68, 109–127.

RYALS, L.J., 2002. Are your customers worth more than money?. *Journal of Retailing and Consumer Services,* 9(5), 241-251.

RYALS, L., 2003. Making customers pay: measuring and managing customer risk and returns. *Journal of Strategic Marketing,* 11(3), 165-175.

RYALS, L.J. and KNOX, S., 2005. Measuring risk-adjusted customer lifetime value and its impact on relationship marketing strategies and shareholder value. *European Journal of Marketing,* 39(5/6), 456-472.

RYALS, L.J. and KNOX, S., 2007. Measuring and managing customer relationship risk in business markets. *Industrial Marketing Management,* 36(6), 823-833.

SAERENS, M., LATINNE, P. and DECAESTECKER, C., 2002. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural Computation,* 14(1), 21–41.

SAS HELP AND DOCUMENTATION. HTML Help for SAS 9.1 for Windows and SAS Enterprise Miner Solution, version 4.3.

SCHMITTLEIN, D.C., MORRISON, D.G. and COLOMBO, R., 1987. Counting your customers: who are they and what will they do next. *Management Science,* 33(1), 1–24.

SHESKIN, D.J., 2004. *Handbook of Parametric and Nonparametric Statistical Procedures*. 3rd Edition, USA: Chapman & Hall/CRC.

SILL, J., 1998. Monotonic networks. *In:* Advances in Neural Information Processing Systems: *The MIT Press*, 10, 661-667.

SILVERMAN, D., 2005. *Doing qualitative research.* 2nd Edition, London: Sage Publications Ltd.

SINHA, A.P. and ZHAO, H., 2008. Incorporating domain knowledge into data mining classifiers: An application in indirect lending. *Decision Support Systems*, *In Press, Accepted Manuscript*, *Available online 16 July 2008.*

SLACK, R., 1999. Discount rates for project appraisal. *Available at ACCA: http://www.acca.org.uk/publications/studentaccountant/36945 on 28/06/2006.*

SO, Y., 1995. *A tutorial on logistic regression.* Cary, NC: SAS Institute Inc.

SOLNICK, S. J. and HEMENWAY, D., 1992. Complaints and disenrollment at a health maintenance organization. *The Journal of Consumer Affairs,* 26(1), 90–103.

STAHL, H.K., MATZLER, K. and HINTERHUBER, H.H., 2003. Linking customer lifetime value with shareholder value. *Industrial Marketing Management,* 32(4), 267-279.

TAX, S.S., BROWN, S.W. and CHANDRASHEKARAN, M., 1998. Customer evaluations of service complaint experiences: Implications for relationship marketing. *Journal of Marketing,* 62(April), 60–76.

THOMAS, L.C., EDELMAN, D. and CROOK, J., 2002. *Credit scoring and its applications.* Philadelphia, PA: SIAM – Monographs on Mathematical Modeling and Computation.

THOMAS, L.C., 2007. Measuring the discrimination quality of suites of scorecards: ROCs ginis, bounds and segmentation. *Credit Scoring and Credit Control Conference (CSCC X),* August 2007, Edinburgh.

VAN DEN POEL, D. and LARIVIÈRE, B., 2004. Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research,* 157(1), 196–217.

VAN ERKEL, A.R. and PATTYNAMA, P.M.Th., 1998. Receiver operating characteristic (ROC) analysis: basic principles and applications in radiology. *European Journal of Radiology,* 27, 88-94.

VAN GESTEL, T., 2005. Transversal backtesting guidelines for PD, LGD and CCF models. Basel II and Credit Risk Modelling Workshop, University of Southampton.

VAN GESTEL, T., MARTENS, D., BAESENS, B., FEREMANS, D., HUYSMANS, J. and VANTHIENEN, J., 2007. Forecasting and analyzing insurance companies' ratings. *International Journal of Forecasting,* 23(3), 513–529.

VANTHIENEN, J. and DRIES, E., 1994. Illustration of a decision table tool for specifying and implementing knowledge based systems. *International Journal on Artificial Intelligence Tools,* 3(2), 267-288.

VANTHIENEN, J. and WETS, G., 1994. From decision tables to expert system shells. *Data and Knowledge Engineering,* 13(3), 265-282.

VANTHIENEN, J., MUES, C. and AERTS, A., 1998a. An illustration of verification and validation in the modelling phase of kbs development. *Data and Knowledge Engineering,* 27(3), 337-352.

VANTHIENEN, J., MUES, C., WETS, G. and DELAERE, K., 1998b. A tool-supported approach to inter-tabular verification. *Expert Systems with Applications,* 15 (3-4), 277-285.

VENKATESAN, R. and KUMAR, V., 2004. A customer lifetime value framework for customer selection and resource allocation strategy. *Journal of Marketing,* 68, 106–125.

VELIKOVA, M. and DANIELS, H., 2004. Decision trees for monotone price models. *Computational Management Science,* 1(3–4), 231–244.

VELIKOVA, M., DANIELS, H. and FEELDERS, A., 2006. Solving partially monotone problems with neural networks. *Proceedings of World Academy of Science, Engineering and Technology,* 12, 82–87.

VERSTRAETEN, G., 2005. Issues in Predictive Modeling of Individual Customer Behavior: Applications in Targeted Marketing and Consumer Credit Scoring. PhD Thesis, Faculty of Economics and Business Administration, Ghent University.

WEI, C-P. and CHIU, I-T., 2002. Turning telecommunications call details to churn prediction: a data mining approach. *Expert Systems with Applications,* 23 (2), 103-112.

WETS, G., VANTHIENEN, J. and PIRAMUTHU, S., 1997. Extending a tabular knowledge based framework with feature selection. *Expert Systems with Applications,* 13, 109-119.

WILKIE, A.D., 2004. Measures for comparing scoring systems. *In:* L.C. THOMAS, D. EDELMAN and J. CROOK, eds. *Readings in credit scoring: Recent developments, advances, and aims,* pp. 51-62. Oxford: Oxford University Press.

WITTEN, I.H. and FRANK, E., 2005. *Data mining: practical machine learning tools and techniques.* 2nd Edition, San Francisco: Morgan Kaufmann.

WYNN, G.W. and CRAWFORD, J.C., 2001. *Data mining: a concept of customer relationship management.* New Orleans, Louisiana: Academy of Collegiate Marketing Educators.

YARDLEY, L., 2000. Dilemmas in qualitative health research. *Psychology and Health,* 15, 215-228.

YU, W., JUDA, D.N. and SIVAKUMAR, S.C., 2005. A churn-strategy alignment model for managers in mobile telecom. *Proceedings of the 3rd Annual Communication Networks and Services Research Conference (CNSR'05),* pg. 48-53.

YU, T., SIMONOFF, J. S. and STOKES, D., 2007. Incorporating prior domain knowledge into a kernel based feature selection algorithm. *Lecture Notes in Computer Science: Advances in Knowledge Discovery and Data Mining,* 4426, 1064-1071.

ZHAO, Y., LI, B., LI, X., LIU, W. and REN, S., 2005. Customer churn prediction using improved one-class support vector machine. *ADMA 2005, Lecture Notes in Artificial Intelligence,* 3584, 300-306.