Correction Sheet

UNIVERSITY OF SOUTHAMPTON

Abstract

SCHOOL OF MANAGEMENT

Doctor of Philosophy

CONSUMER DEBT RECOVERY MODELS INCORPORATING ECONOMIC AND OPERATIONAL EFFECTS

By Angela Moore

This research compares in-house and third party recovery processes, including Loss Given Default (LGD) models for in-house and third party and looks at advanced LGD models required for the Basel II Capital Accord, including LGD models using payment patterns, economic variables and individual characteristics.

The in-house LGD models include using economic variables as well as individual characteristics. The Basel regulations require lenders to use economic conditions as part of the model. The data set for the in-house modelling covers the recession during the 1990's and recovery, this makes it ideal for including economic variables.

Once a debtor defaults on a loan the majority will try to pay back what they can in instalments. These debtors often stop paying again and again, causing the collector to renegotiate the instalments. These sequences of instalment patterns are referred to as payment patterns in this thesis, where the patterns being the stop-start payments, which can potentially go on for years.

Using individual and economic characteristics in a regression analysis to estimate the size of each payment using and the length and number of the payment patterns. These payment patterns can be used to predict LGD. This approach is completely new and novel but has great potential. This approach is far more flexible than other models because it can be used to not only calculate the final LGD but also the LGD at any given time. This approach can be used to help lenders not only estimate the final LGD but also assess the effects of collections' policy; different write off policies, and selling prices.

Table of Contents

List of Figures

List of Tables

Declaration of Authorship

I, Angela Moore declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

“Consumer Debt Recovery Models Incorporating Economic And Operational Effects”

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;

2. 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;

3. Where I have consulted the published work of others, this is always clearly attributed;

4. 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;

5. I have acknowledged all main sources of help;

6. 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;

7. Either none of this work has been published before submission, or parts of this work have been published as:

    a. LC. Thomas, A. Matuszyk, A. Moore, “Comparing debt characteristics and LGD models for different collections polices”, *International Journal of Forecasting*

Signed: ………………………………………………………………………………

Date: …………………………………………………………………………

Acknowledgements

I dedicate this thesis to my family, my husband Nicholas who has been patient, understanding and loving. To my parents and sister Lucy for their unconditional support and love.

I would like to express my gratitude and many thanks to Professor Lyn Thomas for all his help, untiring patience and guidance.

I'd like to express my gratitude to the University of Southampton for funding my Ph.D. studies.

Special thanks go to Ania for all the chats and help while we worked together.

Words cannot express what you have all helped me to achieve and my happiness at finally completing.

Chapter 1 Introduction

Consumer credit in the UK is on the rise. This thesis looks at consumer debt recovery models, for defaulted debt. This chapter discusses the problem of Loss Given Default (LGD) and how this research models recovery rate to predict LGD.

In 2002, the outstanding consumer credit for the UK was over £150 trillion (American trillion), a threefold increase since 1992 [25]. By 2009 it was over £200 trillion [29]. The Bank of England in 2010 reported that loss given default rates have been increasing over the past three years [58] and the reported change was increasing for almost all quarters. Default rates were also on the rise during the same time period.

Given that the UK debt industry is so large and that defaults are increasing it is astounding how little research has been done into the Loss Given Default (LGD). LGD is the loss incurred by a financial institution when a debtor defaults on a loan, given as the fraction of Exposure At Default (EAD). LGD usually has a value between 0 and 1 where 0 means the balance is fully recovered and 1 means total loss. LGD needs to be estimated accurately because it is used to calculate the expected financial loss of a loan, which is required under the advent of the Basel regulations [7]. LGD can also help determine the appropriate collection policy to maximise their potential revenue from defaulted debtors. This revenue could come from either the collections department after default or from the sale of debt to a third party. For example, if the LGD is estimated to be close to 1 it may be more profitable to sell off the debt quickly to a third party thereby eliminating the collection costs and allowing the collections department to concentrate on the more profitable defaults with expected LGDs closer to 0.

What LGD modelling had been done was mainly in the corporate lending market where LGD was needed as part of the more sophisticated bond pricing formulae. In the consumer debt market, modelling LGD is not something that had really been addressed until the advent of the Basel regulations. The Basel II Capital Accord [7] allows banks the opportunity to estimate LGD using their own models with the advanced Internal Ratings Based (IRB) approach.

Since 2006 there have been some papers modelling consumer LGD, however this research is still in its infancy and the problems with estimating LGD are vast. To put the problem into perspective, there is not even a common definition of default, some use six months, others only three months, of money overdue. There is also no set time period over which LGD is calculated. The UK law states that a loan cannot be collected if there have been no payments or written acknowledgement for over six years [40]. However as some of the data in this thesis shows debtors can pay on and off for years, meaning that the lifetime of the loan can stretch for decades. This adds a difficult complication onto the models, as the loan could have an immensely long lifetime after default. Another complication is the recovery process itself, which is entirely determined by the lender. The lender can choose when, and if, to sell off the debt, farm out the debt to a collections agency to collect the debt for them or collect in-house. So far none of the consumer LGD research has focused on the drivers of recovery in the collections process, particularly for collections by debt recovery agents, where there have usually been several previous attempts to recover the consumer's debt.

This research is original because (1) it compares in-house and third party recovery processes, (2) compares the actual LGD for in-house and third party and models for predicting the LGD, (3) looks at advanced LGD models required for the Basel II Capital Accord [7], (4) creates LGD models using payment patterns not just individual characteristics.

Chapter 3 discusses the differences between debt that is collected in-house and debt that is collected by a third party. The two collections mediums have a variety of differences, including; debt age, information available and collection processes. This is because the 'easier' debt is collected first by the in-house collectors. So only the debt which has proven difficult to collect, is passed on to the third party. Also included is a comparison of debt collection models for predicting LGD for in-house and agency collections over a similar time period.

Chapter 4 is focused on improving the third party collection predictions. The models used in chapter 3 were designed to compare third party LGD models with in-house models. Therefore these predictions were improved by a more detailed analysis by splitting the debtors into groups based on debt amounts and then modelled using regression. Again these results although not impressive are competitive with industry results.

Chapter 5 discusses the in-house LGD models and improves them by including economic variables. The Basel regulations require lenders to use economic conditions as part of the model. For most LGD models this causes problems since until the credit crunch the UK had been enjoying a relatively uneventful economic situation for the last fifteen years. This means that most of the data being modelled was collected during that period. The data set for the in-house modelling however comes from the recession during the 1990's and the relatively uneventful period after. This makes it ideal for including economic variables. The improvements to LGD models by including economic variables are demonstrated here.

All of the models discussed are based on calculating the final LGD or the LGD after a predetermined time period. This is similar to other new models being developed. However chapter 6 discusses the advantages of a revolutionary LGD modelling approach. Once a debtor defaults on a loan they do not behave the same way as a non-defaulted debtor. Some pay back all of their debt in one go, others never pay back anything but the majority pay back what they can with instalments. These instalments are discussed with the collector, and often the lender describes these debtors as being "cured". However these "cured" debtors do not stay "cured," they stop paying again and again, causing the collector to renegotiate the instalments time and again. These sequences of instalment patterns are referred to as payment patterns in this thesis, where the patterns being the stop-start payments, which can potentially go on for years.

Chapter 6 uses these payment patterns to predict LGD by using regression to estimate the size of each payment using individual and economic characteristics and the length and number of the payment patterns. This approach is completely new and novel but has great potential. This approach is far more flexible than other models because it can be used to not only calculate the final LGD but also the LGD at any given time. This approach can be used to help lenders not only estimate the final LGD but also assess the affects of collections' policy; different write off policies, selling prices, etc.

The model discussed in chapter 6 only uses the individual and economic characteristics to calculate the size of each payment. This could be enhanced in future work by using the individual and economic characteristics to calculate the length and number of the payment patterns. The economic variables could also be improved to include not only the variables at default but also future predicted variables during the lifetime of the loan after default.

Chapter 2: Literature Review

2.1 Introduction

This chapter first discusses the limited literature on LGD for both corporate and consumer lending. Then there is a discussion on the practical ways in which debt is recovered. Although the debt recovery techniques are not strictly part of a literature review, it is still useful to record the different actions available to the lender during the recovery process. Finally this chapter reviews the literature on the different techniques available for building LGD models.

The Diners Club issued the first credit cards in 1950, which were used to pay for food in restaurants anywhere that the Diners Club Card was accepted. Technically this was a charge card not a credit card because the entire balance had to be paid when the user was billed. American Express issued the first real credit card in 1958 followed by BankAmericard (now Visa) later that year. [20] The first credit card to be issued in the UK was the Barclaycard owned by Barclays Bank in 1966. [57] Over the last ten years, there has been a rapid rise in the popularity of plastic cards. The credit card industry is booming. In July 2004, the UK broke through the symbolic £1 trillion barrier of outstanding debt for the first time. [11] By 2006 nearly a third of all consumer spending was on plastic cards. [10] In June 2007 there were 66 million cards in the UK making 157.3 million transactions that month with a value of £12.3 billion. [20] During the first quarter of 2010, UK banks and building societies wrote off £2.13bn of which £1.25bn was credit card debt. [50]

Unfortunately with increased credit card use, many consumers fail to pay back the debt. There are many factors contributing to customer delinquency. These include poor financial management skills, the economy and ease of access to loans and credit cards. When a debtor becomes delinquent for 180 days (FSA definition) then the loan is considered to be in default.

There is no standard definition of default. For the New Basel Accord, a default is considered to have occurred when either or both of the following criteria have been met: [8]

- The lender considers the debtor to be unable to repay their credit obligations in full
- The debtor is more than 90 days in arrears on any credit obligation to the lender

When this happens most lenders will try to collect the debt in-house. However some companies use outside agents or will just sell off the debt. If the lender's collection department is unable to collect the debt, then they may also decide to use a collection agency or just sell off the debt. The debt can be passed on several times, and can be collected up to six years after the last payment was made. [40] The amount of debt passed to debt collection agencies, exceeds £5 billion per annum. [26]

Under the Limitations Act 1980, for unsecured loans the limitation period is 6 years. If the debtor acknowledges the debt in writing or pays an instalment within the original limitation period, then the time limit begins again from the date of acknowledgement or the date of payment. If the creditor does not contact the debtor for 6 or more years, then the debtor may be able to claim that the outstanding debt is Statute Barred under this Limitations Act, where Statute Barred means that the creditor cannot use the legal system to enforce payment.

Relevant literature in the area of debt collection is very limited. Bennett et al [16] looks at the validation of LGD models, and Chin and Kotak [23] discusses using rule-based engines. In corporate default, there has been a growing literature on building regression based models to determine the drivers of recovery rate, see for example the book edited by Altman, Resti and Sironi [4]. Though there is no collection process. Data on the way banks collected debts from small and medium sized firms was used in Dermine and Neto de Carvalho [28] to build a regression model of how much was recovered. On the consumer debt side, there is very little modelling literature. Bower et al [18] were looking at charged off credit card accounts being collected via an automated call centre. They found that by using a prioritisation model to arrange accounts based on probability of contact and value of account they were able to improve recovery rates. Makuch et al [42] looked at managing the delinquency in the US economy using linear programming to optimise the allocation of resources within collections. However there has been no work on the drivers of recovery in the collection process of consumer debt, particularly for collections by debt recovery agents, where there have usually been several previous attempts to recover the consumer’s debt.

This chapter will look at all the relevant literature for predicted Loss Given Default (LGD), Recovery Rates (RR) where RR=1-LGD for consumer debt and discuss the techniques available to the collector once the debtor has defaulted and the mathematical techniques to predict LGD and RR. LGD is defined as the ratio of losses to exposure at default and therefore usually has a value between 1 and 0 where, 1 indicates that no money was recovered and 0 indicates all of the debt was recovered.

2.2 LGD Corporate Borrowing

The New Basel Accord allows a bank to calculate credit risk capital requirements according to either of two approaches: a standardized approach which uses agency ratings for risk-weighting assets and Internal Ratings Based (IRB) approach which allows a bank to use internal estimates of components of credit risk to calculate credit risk capital. To use the IRB, the institution needs to develop methods to estimate the following components of their loan portfolio:

- PD (probability of default in the next 12 months);
- LGD (loss given default);
- EAD (expected exposure at default).

Modelling PD, the probability of default has been the objective of credit scoring systems for fifty years but modelling LGD is not something that had really been addressed in consumer credit until the advent of the Basel regulations. EAD is the expected amount outstanding at default and the expected loss (EL) at default: EL = EAD*PD*LGD.

What LGD modelling has been done was mainly in the corporate lending market, where LGD (or Recovery Rate (RR), where LGD = 1-RR) was needed as part of the more sophisticated bond pricing formulae. Even there, until the mid nineties LGD was assumed to be a deterministic value obtained from a historical analysis of bond losses or banks worked it out through experience [5]. Only when it was recognised that LGD was needed for the pricing formula for non-defaulted risky bonds were models of LGD developed.

Bruche and González-Aguado [21] look at the time-series behaviour of default probabilities and recovery rates distributions using an econometric model. They state that the time-variation in recovery rate distributions does amplify risk, but that this effect is much smaller than the contribution of the time variation in default probabilities to systematic risk. Also their results indicate that default rates and recovery rates are more tightly related to each other than to macroeconomic variables. They found that credit downturns do not perfectly aligned with recessions; they start earlier and last longer.

To determine the average LGD for a portfolio, there are three approaches available: dollar-weighting, default-weighting and time-weighting. However since the LGD distribution is "bimodal" (two-humped), the average can be very misleading. [4]

These market values, or implied market values, of Loss Given Default were then used to build regression models that related LGD to some relevant factors;

- Seniority of the debt,
- Country of issue,
- Size of issue,
- Size of firm,
- Industrial sector of firm,
- Economic conditions.

The need for such models was identified by Altman and Kishore [3] and reviews of some of the models are given in several recent books ([4], [27], [30]).

2.3 LGD Personal Borrowing

Corporate LGD modelling is not appropriate for consumer credit LGD models since there is no continuous pricing of the debt as is the case on the bond market. The Basel Accord [8] suggests using implied historic LGD as one approach in determining LGD for retail portfolios. The realised losses (RL) per unit amount loaned in a segment of the portfolio is identified and then if default probability (PD) for that segment can be estimated, one can calculate LGD since RL=LGD*PD. One difficulty with this approach is that it is accounting losses that are often recorded and not the actual economic losses, which should include the collection costs and any repayments after a write-off. Also since LGD must be estimated at the segment level of the portfolio, if not at the individual loan level there is often insufficient data in some segments to make a robust estimation.

The alternative method suggested in the Basel Accord is to model the collections. Dermine and Neto de Carvalho [28] used such data of bank loans to small and medium sized firms in Portugal. Small and medium sized companies falls under the retail (or consumer) segment of the Basel Accord. Dermine et al [28] used a regression approach, in the form of log-log regression to estimate the data.

In 2006 Lucas [41] suggested the idea of using the collection process to model LGD mortgages. The collection process was split into whether the property was repossessed and the loss if there was repossession. So a scorecard was built to estimate the probability of repossession and then a model used to estimate the sale value of the house that is actually realised at the time of sale. Qi & Yang [44] used linear regression to model LGD in mortgages. They observed that LGD could be explained by the loan characteristics; the nature of the underlying property, and variables measuring the default, foreclosure and settlement process. The most important factor they found is the current loan-to-value ratio.

Leow et al [39] model mortgage LGD by using the probability of repossession multiplied by a haircut model to predict LGD. They found that the two stage model was more affective at accurately reflecting the LGD distribution. They also tried using macroeconomic variables to predict LGD but found that while they were significant they had very little effect on improving the predictive performance of the model.

Bower et al [18] found that by using a prioritisation model arranges accounts based on probability of contact and value of account they were able to improve recovery rate via an automated call centre.

Allred et al [2] looked at dynamic data-driven decision making tools and procedures to evaluate all aspects of credit card operations, specifically, bankruptcy, fraud, and collections. For collections the most important predictor variables were based on past payments, specifically time since last payment and frequency of payments. They found a positive relationship between the number of past payments and the probability of future payments. Other significant variables were initial balance, balance remaining, frequency of calls made, and frequency of contacting the right party.

Qi and Zhaoa [45] report regression results from four parametric methods; ordinary least squares regression, fractional response regression, the inverse Gaussian, and inverse Gaussian with beta transformation, and two non-parametric methods; regression tree and neural network They found that the non-parametric methods outperform the parametric methods in terms of model fit and predictive accuracy.

Bastos [9] looked at forecasting LGD on bank loans using parametric fractional regression and a nonparametric regression tree models. The nonparametric model gave better results over shorter time periods, of 12 and 24 months. The parametric regression model was better at predicting for longer time periods.

Thomas and Zhang [54] modelled recovery rates and recovery amounts, for unsecured consumer loans using linear regression and survival analysis models. They found that in all cases, the models were better at modelling recovery rate and that using this estimate the recovery amount, was more effective than modelling the recovery amount directly. Linear regression achieved a higher R^2 value and Spearman rank coefficient than the survival analysis models for modelling recovery rate.

Querci [46] sought to explain LGD geographic location, loan type, workout process length and borrower characteristics. Querci found that borrower characteristics gave the best results but concluded that none of them could fully explain LGD.

Thomas et al [53] pointed out that one of the problems with LGD modelling for unsecured credit is that the outcome depends not only on the ability and the willingness of the debtor to repay but also on the decisions by the lender. They used a decision tree approach to model the strategic level decisions of a lender of whether to collect in-house, through an agent or to sell off the debt to a third party. They also suggested that LGD estimates for one type of collection might be built using mixture distributions. Caselli et al [22] used data from an Italian bank's in-house collection process to show that economic effects are important in LGD values. Bellotti and Crook [13] also looked at using economic variables as well as loan and borrower characteristics in a regression approach to LGD for in-house collection while Somers and Whittaker [47] suggested using quantile regression to estimate LGD, but in all cases the resultant models had R^2 values between 0.05 and 0.2. It seems estimating LGD is a difficult problem.

2.4 Debt Recovery Techniques

The debt recovery techniques discussed here were based upon observations made whilst the author attended a Debt Collection Techniques course run by a debt collections agency.

The debt collection agency works out of London. Their primary method of debt collection is telephone with written communication in support. The telephone is used because it can lead to fast recovery of debt, as it is a direct line of communication with the debtor and can result in a payment from the first conversation. The telephone is also very cost effective compared to face-to-face communication but is just as personal. There is also the element of surprise and the debtor and collector can negotiate to achieve a mutual satisfactory result.

There are two objectives for every call. The primary objective is to obtain payment in full or at least a partial payment and an arrangement to pay the rest. The secondary objective is to obtain further information on the debtor to improve negotiations.

The collector has a range of “tradables” in their arsenal to negotiate with. They can threaten the debtor with legal action; this can result in an increase in debt due to charges. They can give the debtor a discount or offer to remove their interest charges. They can repair or damage the debtor’s credit rating thus making it easier or harder to obtain future credit.

The debtor has four options for dealing with the debt:

1. Pay in full (ask for a deal or just pay)
2. Pay part of the debt to avoid legal action or additional charges
3. Set up a payment plan (e.g. £10 per week)
4. Deny the debt or don’t make any payment

The collector has seven options for dealing with the debt:

1. Ask (persuade) the debtor to pay in full
2. Ask (persuade) the debtor to pay part of the debt
3. Ask (persuade) the debtor for an arrangement to pay the debt
4. Write off the debt or send it back to the bank (Recourse)
5. Start County Court proceedings
6. Start Bankruptcy proceedings
7. Add additional charges and interest

2.4.1 Payment In Full

Getting the debtor to pay the debt in full is the ideal solution for the collector because it is the most cost effective use of the collector’s time and also helps cash flow. It is also a good result for the debtor because it is the least hassle and gives them peace of mind, the debtor may also be able to cut a deal and get a discount.

Trying to persuade the debtor of the virtues of this is never easy. The collector can use the threat of additional charges, starting county court proceedings or even bankruptcy, which are all discussed in more detail further on. The collector could also try to encourage them with discounts (e.g. 10% discount if you pay by the end of the day), reduce or remove previous charges, or offer to repair or damage the debtor's credit rating thus making it easier or harder to obtain future credit.

2.4.2 Part Payment

If the debtor pays part of the debt, even if it is only a pound, then the debtor has admitted to the debt. This means that legally the collector has a full six years to collect the remainder of the debt. So the collector should always try to get some sort of payment out of the debtor. The collector can try to persuade the debtor with discounts (since you can't afford to pay the whole debt today, how about paying £1000 today and we will arrange a payment plan for the rest, and I'll knock £100 off your debt), reduced or removed previous charges, add charges, threaten legal action etc.

The collector should try to find out as much about the debtor as possible, especially if the debtor claims that they cannot afford to pay the debt. If the debtor claims to be working, then finding out their income, expenditure and job can help the collector assess how much they can afford and if they are lying. If the debtor claims not to be working, then finding out what sort of benefits they are on provides the collector with their income and again finding out their expenditure can also lead to finding other sources of income. Finding out about other people they can ask to lend them the money to stop further charges can help, e.g. spouse, parent, child, or friend.

The debtor can try to use part payment to cut a deal e.g. I'll pay £100 today, if you remove the interest charges for the last six months.

Part payments can be used in conjunction with payment plans, i.e. paying part of the debt today and then setting up a payment scheme for the remainder.

2.4.3 Payment Plan

If the debtor is unable to pay the full amount but can afford to contribute on a weekly or monthly payment plan, then this is a good solution for both parties. The debtor has the peace of mind that they are paying off the debt at a rate they can afford and will avoid further inconvenience, charges and legal proceedings. The collector can rely on a steady income from the debtor and if they fail to pay all of their payments, then they can sue the debtor for the arrears. This has the advantage that it is cheaper to sue for part of the debt. The debtor can be sued again for the debt. Since it costs the debtor each time they are sued and causes inconvenience when the county court judgements are enforced then it is in the debtor's best interest to ensure they stick to the payment plan.

Again the collector should try to find out as much as possible about the debtor to be able to assess what is the correct rate at which the debtor should pay them back. If the collector asks for too little, then they are not helping their cash flow and it will take longer to remove the debt from the books. On the other hand if they ask for too much and the debtor cannot afford to pay, then the debtor will fall behind again on their payments, leading to additional time and resources being spent on suing the debtor or trying to find an alternative payment plan.

2.4.4 Recourse: Returning Debt to the Bank

There are several situations when the third party can return the debt to the bank from which it was bought: this is called Recourse. Typical conditions where this can apply are:

- If the debtor is dead (depending on when he died)
- If the debt is disputed
- If the debt does not meet the conditions under which it was bought e.g. already been through the legal process
- There is the question of fraud
- If the account holder is in prison (depending on when they were sentenced)

If the debt cannot be recoursed, and it is unlikely that the collector will be able to collect the debt, then the third party can decide to write off the debt. It is very unlikely that the debts will be written off because then third party will have lost the money unlike in-house collectors who then may sell off the debt to third parties.

2.4.5 County Court Judgements

If the debtor refuses to pay any of the debt, then the collector can start county court proceedings. Before you can start county court proceedings there are some pre-action protocols, which were introduced by the Woolf reforms of 1999 [35] to reduce the amount of unnecessary court action. Before starting proceedings you must first try to seek a resolution, (i.e. try to set up a payment scheme) and send a solicitor's letter stating that if the debtor does not pay then they will be taken to court.

If the debtor still does not pay then the collector can fill out a county court claim form for either the full amount or part of the debt. If the collector sues for part of the debt then they may persuade the debtor to pay the rest of the debt and avoid further inconvenience. If they don't then they can be sued again for the remainder of the debt. From the date the county court judgement is issued (day the form is completed), there is a pause of five days while it is being processed and then the debtor/defendant has 14 days to respond.

The debtor/defendant can:

- Ignore it (judgement by default)
- Admit it and make an offer which is accepted (judgement by admission)
- Admit it and make an offer which is not accepted (judgement by admission)
- Admit part of it
- Defend it and/or counter claim
- Pay it
- Ask for a further 14 days to respond

There are four types of judgement:

- Judgement by default – the defendant ignores the county court claim
- Judgement by admission – the defendant admits to the debt
- Judgement by determination – the judge decides the outcome
- Summary judgement – the judge decides that the defence is invalid.

In most of the cases that the debt collectors are involved with, the debtor/defendant will ignore it and hence judgement will be made by default. The debtor will then have a further 28 days to pay the debt before judgment is entered (49 days since the date of issue). Once the judgement is entered it will be on record for 6 years even if the debtor pays back the debt. Once the judgement is entered onto record it will have a very negative effect on their credit rating.

Once a judgement has been made and before it has been entered onto record, the judgement can be enforced. Typical types of enforcement are:

- Order to Obtain Information
- Warrant of Execution
- Attachment of Earnings Order
- Third Party Debt Order
- Charging Order

Any or all of these can be used on the debtor/defendant.

2.4.6 Order to Obtain Information

This is an oral examination in court. The defendant must attend the court and is then asked a series of questions by the collectors. If the defendant does not attend the court twice, then they will be in contempt of court, which can mean being arrested. The collector can ask the defendant any question they want, e.g. the start of their bank accounts, home life, income, expenses, property and if the defendant tells a lie they are committing perjury, which is also an arrestable offence.

2.4.7 Warrant of Execution

This is basically sending the bailiffs round. They can take property that has a value of up to 6 times the debt. The bailiffs can come round up to three times in the course of six months; all of the goods are then sold at public auction. If the goods are sold for more than the debt, the remaining money goes back to the defendant, and they are left with the very expensive cost of replacing everything that has been sold. If the goods do not cover the cost of the debt then the bailiffs can be sent around again or another enforcement can be used.

2.4.8 Attachment of Earnings Order

If the defendant is employed, then the court can contact their employer and have money taken out of their wages regularly until the debt is paid off. This causes an irritation to the employer because each month they have to inform the court that the debtor is still employed by them and arrange for the payments to be removed form their wages. This is not likely to do the debtor's career much good, for instance if some redundancies are coming up which person is more likely to be sacked, a person with or without an attachment to earning. If the debtor is made redundant then they must inform other companies they are applying to that they have an attachment to earnings. Technically you can put an attachment of earnings onto statutory sick, maternity and paternity pay. The court determines the percentage of earnings, which is paid.

2.4.9 Third Party Debt Order

This freezes the debtor's bank account for six weeks and any money, which is in the bank account on the day it is frozen, can be used to pay the debt. When the bank account is frozen, money can still be paid into the bank account but no money can be taken out. Therefore direct debits, standing orders and checks will not be paid. This will result in the bank charging for the payments not being made. This means that the debtor will be facing additional charges, none payment of direct debits which could result in the termination of goods and services e.g. if the insurance is not paid then it may become invalid, and will have all the money in the bank account removed. This can only be used on one bank account at a time. If this enforcement is used in conjunction with an Order to Obtain Information, then the collector could find out which bank account has the most money in it and when the next payment is being made and therefore freeze that bank account on the day the next payment is to be made. This enforcement can be used several times to ensure the debt is collected.

2.4.10 Charging Order

A Charging Order is where a charge is put on the debtor's property. This property can be a house, land, or stocks and shares. Once a charge is put on the property, then the debtor has to pay when they sell the property. The court can also force the sale of the property but this is not done very often in the case of the debtor's home.

2.4.11 Charges

Each time a County Court Judgement is made there are additional costs which must be met by the debtor if they lose the judgement.

2.4.12 Statutory Interest

Statutory Interest is charge on the debt as stated in section 69 of the County Courts Act 1984 [36]. Interest is charged at a rate of 8% per annum between the default date and the date of issue. To calculate the statuary interest:

Interest =No. of days (Issue Date-Default Date) * Daily Rate of Interest (Balance*8%/365)

The interest is added to the principal of the debt.

2.4.13 Bankruptcy

Bankruptcy Law has been around for many centuries. The first act was passed in England in 1542; the most recent acts are the Insolvency Act of 1986 [51] and the Enterprise Act of 2002. The aims of bankruptcy are twofold: to free the individual from pressure of creditors and to ensure that all of the assets are distributed fairly among the creditors. The individual (debtor) or the creditors may start bankruptcy proceedings. However it is the Courts who are officially responsible for making an order against the individual. All of the individual's assets then fall under the control of the Trustee appointed, whose responsibility is to get all of the assets to the creditors. Hence ensuring the creditors can no longer bother the individual. In the Insolvency Act of 1986 [51] the amount of time you remained bankrupt depended on what you owed: it was 2 years for less than £20,000 and 3 years for more. The Enterprise Act of 2002 changed the negating factor from the amount owed to whether or not the individual is considered to be "Reckless" and "Negligent". If bankruptcy was not their fault then they are given a fresh start in 12 months. However if they are found to be “Reckless” and “Negligent” then it is between 2 and 15 years, with a limit of three years to release the equity in their home.

When a bankruptcy order has been made, you must:

- Comply with the Official Receiver’s request to provide information about your financial affairs;
- Give the Official Receiver a full list of your assets and details of what you owe and to whom;
- Look after and then hand over your assets to the Official Receiver together with all your books, records, bank statements, insurance policies and other papers relating to your property and financial affairs;
- Tell your trustee about assets and increases in income you obtain during your bankruptcy;
- Stop using your bank, building society, credit card and similar accounts straightaway;
- Not obtain credit of £500 or more from any person without first disclosing the fact that you are bankrupt;
- You may also have to go to court and explain why you are in debt. If you do not co-operate, you could be arrested.

Bankruptcy will affect you in many ways, including;

- You will no longer control your assets as these will be sold to pay your debts
- Your home may have to be sold to go towards paying your debts. Any increase in value after being made bankrupt does not belong to you
- Contributions from your salary can be deducted to pay your debts for up to 3 years
- If you are self-employed, your business is normally closed down and any employees are dismissed

It is a Criminal offence during bankruptcy to;

- Obtain credit of over £500;
- Start a new business;
- Become a Company Director, either formally or informally;
- Not disclose that you are an undischarged bankrupt

After being discharged, there will still be certain restrictions, such as;

- Any increase in the value of your home will not belong to you
- You cannot obtain overdraft facilities
- Any excess funds over and above reasonable living expenses can be claimed to pay off your debts
- Inheritance and assets can be affected many years after discharge

2.4.14 Summary of Debt Recovery Techniques

Debt collectors conduct their business over the telephone. This would appear to be mainly from the point of view of cost efficiency and the safety of their staff. From discussions with collectors, a lot of their debtors are the sort, who would use a face to face meeting as a chance to show off to their friends. They use bullying tactics to belittle the collector and have no intention of paying. Over the phone they do not have this audience and so can be more responsive.

The ideal outcome for the collector is to receive payment in full on the first call. In order to achieve this they will often offer discounts or threaten legal action. If they discover that the debtor is unable to pay the full amount then the fall back positions are to set up a payment plan and to receive part of the payment in advance. In order to set up the correct payment plan, the collector should find out as much information as possible from the debtor.

If the debtor refuses to pay or doesn’t make the agreed payments then the agency will start County Court Proceedings. This will most likely result in a warrant of execution to retrieve the debt.

The author did not observe any of the collection telephone calls but did hear several of the role playing exercises, where experienced debt collectors pretended to be the debtors, and the collectors had to try to get the money out of them. The debtors were trying to be deliberately difficult sometimes not even admitting to their own name. It was very impressive the way they tried to coax payments out of them.

2.5 Statistical Techniques

2.5.1 Logistic Regression

Logistic regression is a generalised linear model used when the target variable can take only two possible values. The distributions of LGD have large spikes at LGD=1 and LGD=0. The logistic regressions in this thesis will be used to predict if the LGD will be zero or one as applicable.

The logistic function is given by:

$$f(x) = \frac{e^x}{e^x + 1} = \frac{1}{1 + e^{-x}}$$

The logistic function can take in any input value from negative infinity to positive infinity, yet the output values are confined to between 0 and 1. The variable *x* is a measure of the total contribution of all of the independent variables used in the model, defined as:

$$x = \beta_0 + \beta_1 z_1 + +\beta_2 z_2 + ... + \beta_k z_k$$

where β_0 is the intercept, $\beta_1, \beta_2, \beta_3, \ldots, \beta_k$ are the regression coefficients of $z_1, z_2, z_3, \ldots, z_k$ respectively. Each regression coefficient describes the size of the contribution of that variable. A positive regression coefficient indicates that the associated variable increases the probability of the outcome, while a negative indicates a decrease in the probability of the outcome. The size of the regression coefficient indicates how strongly the variable influences the probability of the outcome.

Logistic regression is used to express the relationship in the form of a probability between one or more independent variables and a binary response variable. The main approaches are:

- Forward selection, which involves starting with no variables in the model, trying out the variables one by one and including them if they are 'statistically significant'.
- Backward elimination, which involves starting with all candidate variables and testing them one by one for statistical significance, deleting any that are not significant.

- Stepwise methods are a combination of the above, testing at each stage for variables to be included or excluded.

2.5.2 Linear Regression

Linear regression is modelling the relationship between a scalar variable *y* and one or more variables denoted *x*. The model depends linearly on the unknown parameters to be estimated from the data. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of *y* given *x*.

Thus the model takes the form of:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i$$

where *i=1,…,n,* ε_i is an unobserved random variable which adds noise to the linear relationship.

Linear regression is often used for modelling LGD. To name but a few, Hillebrand [35], Huang and Oosterlee [36],Thomas et al [53] all used linear regression models for predicting LGD.

The linear regression in the thesis was all done within SAS where the model is fitted by least-squares. The output includes the two-tailed significance probability for all variables, which was used to determine significance of the variable (<0.05). The model output also quotes the root Mean Squared Error (MSE) and the R-squared values for the model that were used to determine the goodness of fit of the model. However once a model was created, the R-squared values quoted were based on the explained variance method on the holdout sample, which is explained later in this chapter.

One of the problems with using linear regression to predict LGD is that LGD was not normally distributed, and linear regression assumes a normally distributed variable. This meant that when the linear regression models were applied, the predicted results clustered around the mean. This model predicted a normally distributed LGD with a very small variance. Therefore the target values (LGD) had to be transformed into a normal distribution before linear regression could be performed. Two of the distributions used were beta and lognormal. This meant that the target values were transformed using beta or lognormal, then the regression analysis run, and the predicted results transformed back using the inverse transformation. One example of a non-linear regression is the commercial product LossCalc [34] that is based on the fact that the LGD distribution should be approximated by a beta distribution.

2.5.3 Box-Cox method

Box-Cox changes non normal distributions to a closer approximation of a normal distribution. The Box–Cox, or power-normal distribution, is the distribution of a random variable *X* for which the Box–Cox transformation on *X* follows a truncated normal distribution. It is a continuous probability distribution having probability density function (pdf) given by

$$f(y) = \frac{1}{(1 - I(f < 0) - \operatorname{sgn}(f)\Phi(0, m, \sqrt{s}))\sqrt{2\pi s^2}} \exp\left\{-\frac{1}{2s^2}\left(\frac{y^f}{f} - m\right)^2\right\}$$

for y > 0, where m is the location parameter of the distribution, s is the dispersion, *ƒ* is the family parameter, *I* is the indicator function, *Φ* is the cumulative distribution function of the standard normal distribution, and sgn is the signum function. [19]

2.5.4 Trend lines

A trend line represents a trend; the long-term movement in time series data after other components have been accounted for. It tells whether a particular data set has increased or decreased over the period of time. The trend line's position and slope is calculated using statistical techniques such as linear regression. Typically they are just straight lines, although some variations use higher degree polynomials depending on the degree of curvature desired in the line.

The trend line function in Excel was used in this thesis to determine the rudimentary shape of graphs in order to select the best models for describing the data.

2.5.5 Weight Of Evidence (WOE)

In the WOE approach the target variable is changed to a binary variable by asking: is the LGD value above or below the mean LGD value? Each characteristic is then split into deciles and the ratio of above mean to below mean in each group is assessed. This means that the percentage above the mean in each group is divided by the percentage below the mean and adjacent groups with similar odds are combined. Thus each characteristic is divided into the appropriate number of "bins", each consisting of one or more neighbouring deciles.

Generally, if N_a and N_b are the total number of data points with LGD values above or below the mean and $n_a(i)$ $n_b(i)$ are the number in bin i with LGD values above or below the mean. The bin is given the value:

$$Weight_i = \log\left(\frac{n_a(i)}{n_b(i)} \Big/ \frac{N_a}{N_b} \right)$$

Once the weights are calculated they can then be used to either calculate the Information Value described below to calculate if the variable is useful or not. Or in a regression model to predict if the debtor will be good or bad by using the modified variables and the binary variable of: is the LGD value above or below the mean LGD value.

The information value is determined by

$$InformationValue = \sum_{i=1}^{n} \left(\frac{n_a(i)}{N_a} - \frac{n_b(i)}{N_b} \right) Weight_i$$

where n is the number of bins

The higher the information value the more useful the variable is to determine the outcome.

2.5.6 R-squared

R^2 is the coefficient of determination; the proportion of variability in the data set, which is accounted for by the statistical model.[1] R^2 can vary from 0 to 1 where the value of 1 indicates that the statistical model perfectly fits the data and 0 indicates that there is no relationship between the model and the data.

There are several different definitions of R^2 depending upon the context. With linear regression R^2 is simply the square of the sample correlation coefficient between the outcomes and their predicted values. In this thesis there are two types of R-squared quoted. One is the reading from the PROC REG when a linear regression was performed in SAS. The other is calculated on the holdout sample in terms of explained variance.

When a regression is performed in SAS, an R^2 is reported as part of the output. This R^2 is calculated by using the likelihood-ratio statistic (G^2) where the probability distribution of the test statistic is approximated by a chi-square distribution for testing the null hypothesis that all covariants have a coefficient of 0. R^2 is calculated by:

$$R^2 = 1 - \exp\left(-\frac{G^2}{m}\right)$$

Where *m* is the sample size. [36]

The PROC REG R^2 is used as an initial test to determine the best linear regression model. However once selected model then had its R^2 calculated using explained variance on the holdout sample.

The explained variance method of R^2 was calculated as followed. The observed data set values (y_i) and the model's predicted values ($\hat{y}_i$) are used as follows:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Where $\bar{y} = \frac{1}{n}\sum_i^n y_i$ and *n* is the sample size.

The R^2 statistic will give some idea about the goodness of fit of a model.

2.5.7 Errors

The mean square error (MSE) quantifies the difference between an estimate and the observed value. MSE corresponds to the expected value of the squared error loss or quadratic loss. The error is the amount the estimate differs from the observed value. The difference can occur because of randomness or because the model doesn't account for all information to produce a more accurate estimate. [38] Taking the square root of the MSE yields the root mean squared error (RMSE), which has the same units as the observed values. For an unbiased estimator, the RMSE is the square root of the variance, also known as the standard error. MSE is calculated by: $MSE = E\left[(\hat{y}_i - y_i)^2\right]$.

A MSE of 0 means that the model predicts the observed values with perfect accuracy. The main disadvantage to the MSE is that, like the variance, it is heavily weighted towards outliers. [17] Since each of the terms is squared, larger errors are more heavily weighted than smaller ones. The root MSE is quoted from the SAS results for linear regression as another indicator of the goodness of fit of the models.

2.5.8 Spearman's Rank

Spearman's Rank measures the difference in ranking rather than in the values of the predicted and observed data sets. It is used to compare the predicted ranks of the debtors' LGD with their observed ranks. This is useful since some of the models can be used not to determine the value of the debt but to assess which will be the worst debtors and best debtors at repaying their debts. This allows the collections department to allocate resources accordingly and improve their recovery rate like in the Bower et al [18] study.

R^2 can be used to determine the goodness of fit for the models. However since two-stage models are used in this thesis as well as linear regression models R^2 is not as useful for the individual stages. An alternative is the Spearman Rank Correlation, which is a non-parametric measure of correlation. The real LGD observed results and the predicted LGD results are converted to ranks, and the differences d_i, between the ranks of each observation and prediction are calculated.

The debtor's predicted rank was based on their predicted LGD result; descending. The debtor's real observed rank was based on their observed LGD result; descending. The differences d_i between their predicted rank and real observed rank are used to calculate the Spearman Rank Correlation coefficient. However there are many tied ranks (share the same rank) since there are several debtors with an observed or predicted LGD of 0. When a rank is tied; all associated ranks are assigned the mean of the tied ranks. Tied ranks also means that the classic Pearson's Correlation coefficient has to be used instead of the abbreviated Spearman Rank Correlation coefficient.

$$\rho = \frac{n\left(\sum x_i y_i\right) - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{n\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}\sqrt{n\left(\sum y_i^2\right) - \left(\sum y_i\right)^2}}$$

Where

ρ = Pearson's Correlation coefficient

x_i = real observation rank

y_i = predicted rank

n = sample size

If there are no tied ranks, then ρ is given by:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}.$$

2.6 Summary

There is a body of literature on LGD modelling for corporate loans, mainly because LGD is vital factor in the pricing of risky bonds. The literature for unsecured consumer credit is much sparser and it was only following the advent of the new Basel Accord [8] in 2007 that there has been any concentrated attempt by practitioners and academics to model LGD for consumer debt. Most of the research uses are linear regression to predict LGD.

This chapter also describes the debt collection techniques employed by a third party. So far most of the LGD models do not include collection variables.

Chapter 3: Comparison Of In-house Collections & Third Party Collections

Once a loan has defaulted the lender can choose to collect the debt in-house, pass it along to a debt collections agency to collect or sell off the debt to a third party. This chapter discusses the differences between debt that is collected in-house and debt that is collected by a third party. The two collection mediums have a variety of differences, including; debt age, information available and collection processes. Also included is a comparison of debt collection models for predicting LGD (Loss Given Default) for in-house and agency collections.

Normally an in-house team belonging to the lender undertakes the first attempt at collections. Such a team will have the information the debtor supplied on application, all the details of the loan and the borrower's repayment performance until default. Although the formal Basel definition in the UK for default is that the debtor is 180 days overdue (unlike most other countries which is 90 days overdue) most lenders will freeze the loan or credit card facilities and undertake recovery measures once the loan is 90 days overdue. A UK financial institution provided the representative data set used for modelling such "in-house" collections. It consisted of 10,000 defaulted consumer loans, which defaulted over a two-year period in the 1990s together with their repayment performance in the collection process. The collection models concentrated only on their performance in the first two years of collections to match the information that was available on the third party collections process. For modelling purposes the data was split into 70% for the training set and 30% for a holdout test set.

The lender can also decide to use a third party to try and collect the defaulted amounts usually on a percentage fee basis so the third party will keep x% of what is collected. Alternatively or sometimes after using agents, the lender can sell the debt to a third party who then has the right to seek recovery of the outstanding debt. Our second data set consisted of such loans, which had been purchased by a third party from several of the UK banks. This data set consisted of the information on 70,000 loans where the outstanding debts varied from £10 to £40,000. These debts were purchased in 2000 and 2001 and so most of the defaults had occurred in the late 1990s. The repayments of the debtors for the first 24 months in this “third party” collections process were available at an individual loan level. Again for modelling purposes the data was split into training and hold out test set in the ratio 70:30.

It is clear when examining the “third party” data that there is less information available on the debtor than was available to the in-house collectors. The details of the debt, including the amount outstanding, when default occurred and when last there was a payment, was available. Also in order to set the purchase price, the history of how many different parties had sought to collect the debt is reported. There was some information available about the debtor including details of address and telephone numbers when available, and some demographic information. However there was little information on the default risk scores of the borrower, either application score or behavioural score; on the borrower’s performance before or since default. Thus the data is restricted to the details that were available both in the “in-house” and in the “third party” data sets.

3.1 Collection Strategies

The information available to the in-house collection department is different from the data available to the third party. This has a direct effect on their collection strategies because the in-house collectors with greater knowledge have an interest in saving the debtor. The original lender is initially interested in protecting their relationship with the debtor. Once they believe this relationship cannot continue they are only interested in recovering the money they are owed. The third party has no relationship with the debtor and so from the start is only interested in recovering the money owed. Thus the following sequences of events can be distinguished:

1. Recovery process – internal collection tries to save person
2. Collection process – internal collection tries to save money
3. Collection process – Third Party tries to save money

The actions undertaken by the lender during the recovery or collection process do not differ; only the objective has changed. The main tools used in the in-house recovery process are letters backed up with telephone calls. There are different types of letters and sending them depends on the status of the customers and the characteristics of the debt. The debt sold to the third party will normally be debt that has proven hard for the lender to collect in-house. Since this is the case the distribution for LGD shows that the majority of the debts have not been paid. In fact over 80% of the third party's debts have had no payments made on them at all.

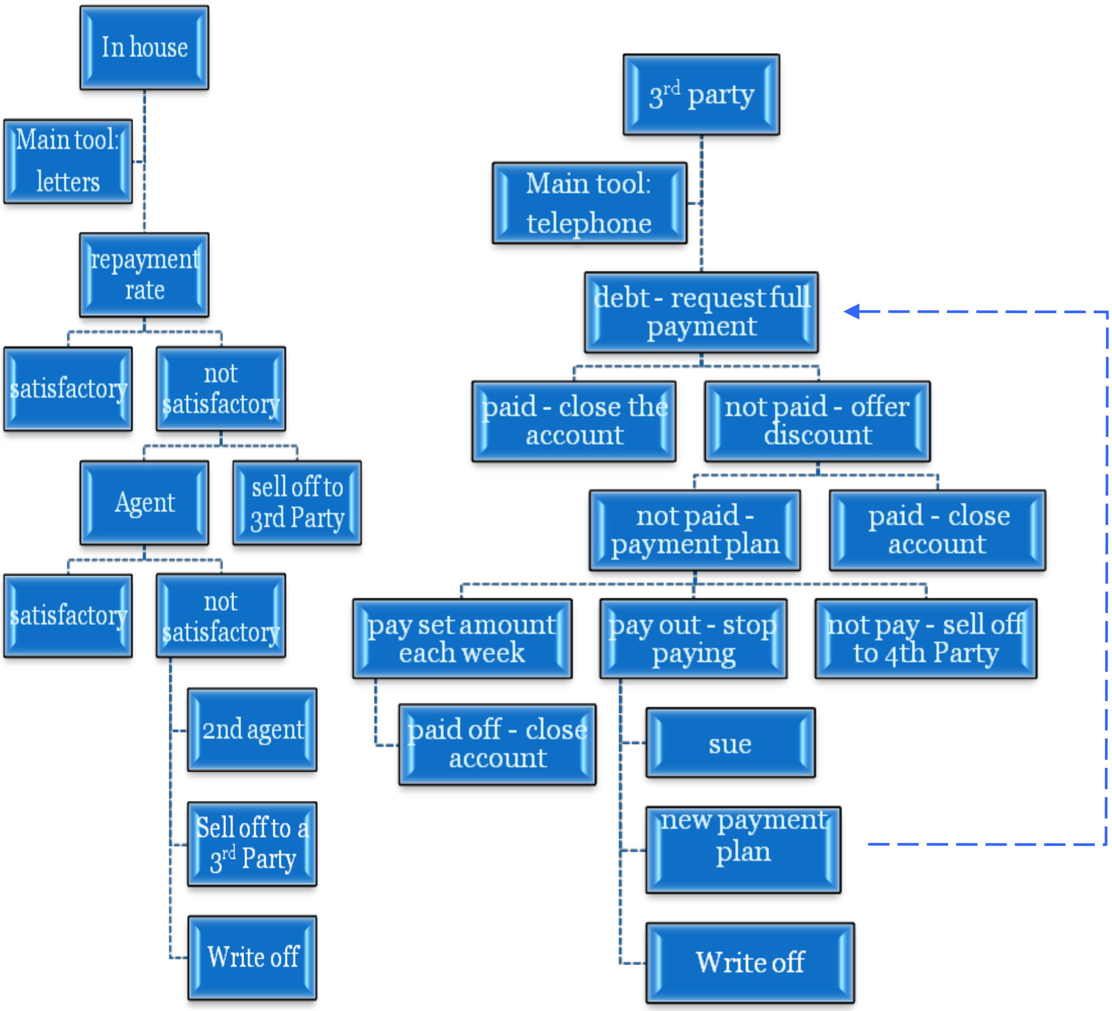


Figure 3.1: Collection trees

Figure 3.1 summaries the decision flows in both in-house and third party when they are collecting defaulted debt. This information was collected through personal correspondence with the in-house lender and third party collector.

In-house collection decisions are different depending upon company policy. Usually, the first step is to send the letters at the beginning of every month. There are different types of letters and sending them depends on the customers. Gentle reminder letters are sent out for one or two missed payments. The longer the customer is not paying the stronger language used in the letter, also old customers are treated in a more polite way than the new customers. If this method is not sufficient the company must use other possible methods: calling the client, paying a visit to the client, trying to set up a payment agreement or find other possible solutions such as rearranging the mortgage, selling their property etc.

The third party uses different methods to achieve their collections. Their primary technique for debt collection is the telephone with written communication in support. The telephone is used because it can lead to fast recovery of debt, as it is a direct line of communication with the debtor and can result in a payment from the first conversation. The telephone is also very cost effective compared to face-to-face communication but is just as personal. There is also the element of surprise and the debtor and collector can negotiate to achieve a mutually satisfactory result.

The ideal outcome for the collector is to receive payment in full on the first call. In order to achieve this they will often offer discounts or threaten legal action. If they discover that the debtor is unable to pay the full amount then the fall back positions are to set up a payment plan and to receive part of the payment in advance. In order to set up the correct payment plan, the collector should find out as much information as possible from the debtor.

If the debtor refuses to pay or doesn’t make the agreed payments then the debt collection Third party will start County Court Proceedings. Chapter 2 covers the different legal options a collector has to force a debtor into paying. This will most likely result in a warrant of execution to retrieve the debt.

When either a third party or in-house collections department takes over an account, they have to decide how to collect the debt. Their first step will be to always collect the full outstanding debt. If debtor pays then they close the account. If not then a discount is offered for a lump sum payment. If the debt is paid then the account is closed, otherwise the payment plan is set up (most likely outcome).

If the full amount is paid at £x per week the account is closed. If the customer pays and stops then the lender will have to decide to either close the account if the total amount paid is satisfactory. If it is not satisfactory; they may try to sue or set up a new payment plan. If the debtors don't pay their payment plan at all then the third party will either sell the debt or close the account.

Factor	In-house data set	3rd Party data set
Main tool	Letter	Telephone
Age of Debt	New	Old
Type of Debt	Personal Loans	Credit Card
Average Debt Amount	£3,609	£562
Percentage Who Paid Back Whole Debt	30%	0.7%
Percentage Who Paid Back Part of the Debt	60%	16.3%
Percentage Who Paid Nothing	10%	83%
Mean value of LGD	0.544	0.95
Collection model	Decision tree model with sub-models	Agent's sub-model
LGD model	2-step model	2-step model
Information available	All details of loan and customer	Restricted data since not original lender

Table 3.1: Summary of in-house and third party data sets

The two data sets are completely different and hence show the two extremes of debt collection. The in-house data is for 10,000 cases over the entire in-house collection lifetime for the majority of the cases. For the third party case study, the data is for a 70,000 of cases over a very short time period. In order to ensure that the data is comparable only the first 2 years after default was used in the in-house data set. Table 3.1 summaries the two data sets used to compare in-house and third party collections.

Even the way the debt is collected is different; the in-house debt is collected via letter [personal correspondence with collectors] (see figure 3.1 and table 3.1), whereas the Third party use mainly telephones to discuss the debtors' personal situation and come up with a collection timetable, which is agreeable to both parties.

The two types of debt are different too, the in-house collections is recovering unsecured personal loan and the third party is collecting bad credit card debt. So not only are the amounts of debt very different but the debtors will have been intending to pay back the debt over different time periods and will have different reasons for taking out the loan in the first place. The lender will have had different checks performed on the debtors before issuing the loans and the original terms of the loan (loan amount, maximum credit limit) may be very different from the situation when the debtor defaults.

3.2 Distribution of LGD

Analysing the in-house distribution of LGD, in Figure 3.2, it can be seen that 30% of the debtors paid in full and so had LGD=0. Less than 10% paid off nothing. For some debtors the LGD value was greater than 1 since fees and legal costs had been added. This is not the case usually in third party collection where almost 90% of the population have LGD=1 (Figure 3.3). It is clear that if more attempts had already been made to collect the debt, then the recovery rate would be lower.

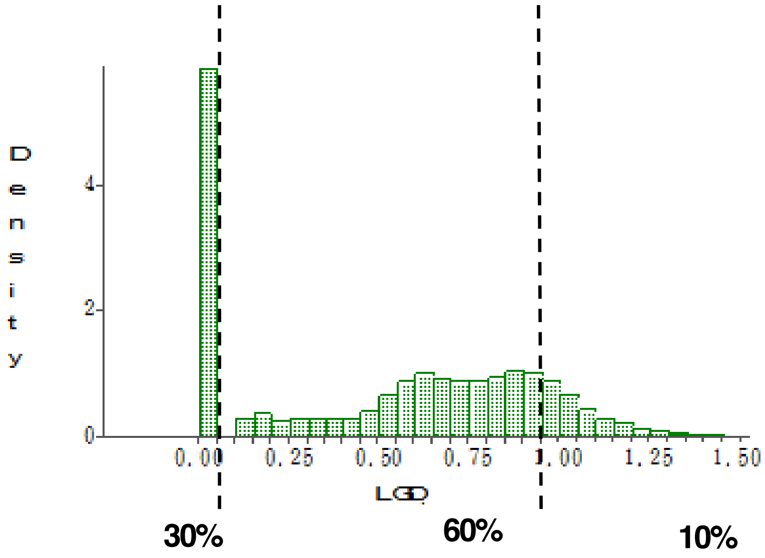


Figure 3.2: Distribution of LGD in the sample for in-house collection (collection for 24months: January1991-December 1992)

Figure 3.3 shows the LGD for the credit card debt collected by the third party. The x-axis shows the LGD, the column above 1 represents the number of debtors who failed to pay back any of their debt hence LGD=1. The column above 0.95 represents all of the debtors who paid back up to 5% of their debt (0.95<=LGD<1). The column above 0 represents all of the debtors who paid back more than 95% of their debt

(0<=LGD<0.05). The y-axis shows the percentage of debtors within each LGD bracket. The majority of the debtors (83%) failed to pay back any of their debt.

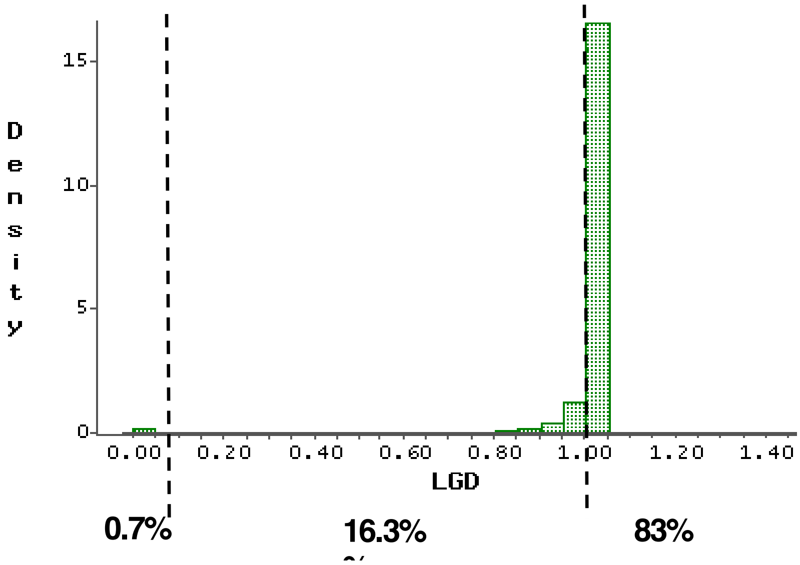


Figure 3.3: Distribution of LGD for credit card debt sold to a third party

The recovery rates or LGD for the two samples are very different. The majority of loans collected in-house have an LGD < 1, whereas the majority of the loans collected by the third party have LGD = 1. There are several factors contributing to this difference. Firstly the debt collected in-house is new debt, no one else has previously tried to collect the debt and they have only recently defaulted at the time of collecting. On the other hand the third party debt is most likely old and has been collected before. This makes it harder for the third party to collect further. Secondly the in-house collection department will have access to more data and that data will have more details. This means that they can look at past behaviour, the original loan details in some cases. They may also have access to data connected with their bank account and income. The third party will not have any of this data. In some cases the debtor may even need to be traced because they have moved or are deliberately trying to hide from the debt collections third party so that they cannot collect the debt.

3.3 Data Available to In-house and Third Party Collectors

As has been mentioned before, the in-house collections department will have very different data available to them as opposed to the third party collectors.

The following data was the data issued by the company in each case study to the author for modelling purposes, so any sensitive data was unavailable. The data discussed in this chapter is information, which could be of use when modelling recovery rates. As was expected the in-house collectors had access to more detailed and accurate data.

Table 3.2 shows a summary of the data accessible in both of the data sets. Contact information was similar with both in-house and agency having access to address, telephone and name of the debtor. However the in-house team would normally have up to date information whilst the agency may have data that is out of date. Telephone numbers given to the agency could include up to eight different numbers for contacting the debtor. However most of these numbers will be no longer valid.

	In-House	Agency
Contact Information	Address Telephone Name	Address Telephone Name
Past Behaviour Information	Number of months in arrears Yearly balance	None
Personal Information	Employment Sex Marital Status Age	Some Employment Title Age
Home Ownership	Yes	Yes
Debt Information	Original Loan Amount Default Amount Default Date	Value of Debt at sale Default Date

Table 3.2, Summary of In-house Versus Agency Debtor Data

The in-house collections case study had detailed information on the debtors past history whereas in contrast the third party had no data at all on what had previously been collected. The default amount was not even known, only the balance at time of sale to the third party. The in-house collections had yearly balances, before and after default, and the number of months they were in arrears for every month on the books. It is from this information that the payment patterns discussed in the following chapter are derived and will be covered in more detail then.

Personal information was about the same for both but there again there was either more detail or more accurate information available to the in-house collectors. The employment data for instance given to the third party was out of date since it came off the original credit card application, or were unusable as the debtor was listed as anything from employed to company name or job name (sale assistant) when the information was available at all. The in-house collectors had all of the different types of employment coded with complete information e.g. unemployed=00, self-employed=01 etc. This meant that the data was useable and available for all debtors even if it wasn't accurate for the whole time period.

Home ownership was another variable, which both third party and in-house collectors had access to. However the in-house data was most likely to be more detailed and accurate since the lender also lent the debtor their mortgage in some cases.

The debt information is also very different; the in-house collectors had access to the original loan data (this could also be a factor of personal loan as opposed to credit cards), like original loan amount and the term of the loan, also if the loan had been increased. Also the collectors knew the default date and amount. In contrast the third party collectors only know the amount when the debt was sold to the third party and the date of the sale, not the original default date and amount.

3.4 Variable Analysis for Third Party

The variables available for analysis are the debtors' titles, country of residence, age, amount of debt, if contactable by telephone, length of time in collections, and home ownership status.

3.4.1 Age

The debtors range in age between 19 and 100, with the majority of debtors in the 25-35 brackets. The data does appear to suggest that the older the debtor is the more likely they are to have a recovery rate greater than zero. Figure 3.4 illustrates the proportion of debtors whose RR is greater than zero in each of the age brackets.

The data was split according to coarse classification not fine classification so as to increase the robustness and cope with any non-monotonic relationship between the recovery rate and the debtor's age. Using coarse classification is advocated for this type of situation by Thomas [52] since the relationship is not monotonic and when split using fine classification many of the groups are sufficiently close to be grouped together. The debtor's age was split to reflect their stage in life, i.e. 18-25 would normally be students and people just starting out in a career and would therefore be on relatively low incomes, have little responsibility (house, family) and have little history of financial independence. This would cause them to react differently from a person over 65 who would most likely have a house, be retired, and be on a fixed pension.

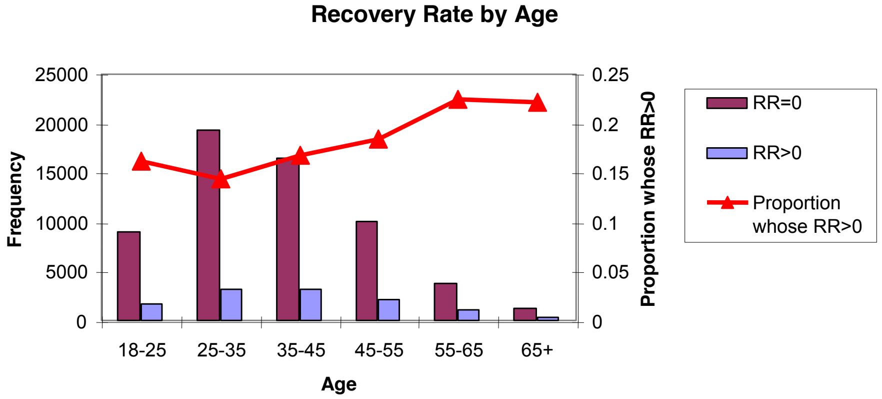


Figure 3.4, Recovery Rate by Age for Third Party

3.4.2 Title

The data included the debtors' title, with five classifications; professional titles (Dr), Mr, Miss, Mrs and Ms, with over 50% of the debtors being men. The debtor's sex and title can be used because this model is not used for determining credit decision. Figure 3.4 illustrates the proportion of debtors whose RR is greater than zero in each of the classifications. As figure 3.4 demonstrates women are more likely to pay something than men and married women are the most likely with 23% of the debtors using the title of Mrs paying something to the third party. What is interesting is that debtors using the title of Dr are least likely to pay anything back.

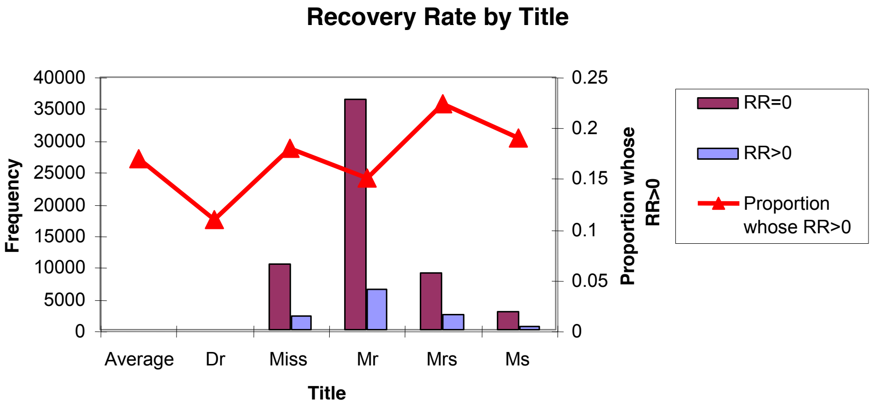


Figure 3.5, Recovery Rate by Title

3.4.3 Homeownership

Homeownership is divided into four classifications; family, solo ownership, joint ownership and tenant. If the debtor is known to reside in a property owned by a member of their family, but not themselves, then their homeownership is classified as Family. If the debtor resides in a property owned solely by them then their homeownership status is Solo. Joint status is recorded if the debtor and another own their residence and Tenant status if they are renting or the details are unknown. The vast majority of the debtors are recorded as Tenants, over 85%.

Figure 3.6 demonstrates that debtors who are classified as Tenants are least likely to pay anything and debtors who reside at a property that is jointly owned appear to be most likely to pay anything back. Presumably this is because not only do they have a property to raise money against, but they also have chattels that could be seized by bailiffs or the other owner could help them to raise the money.

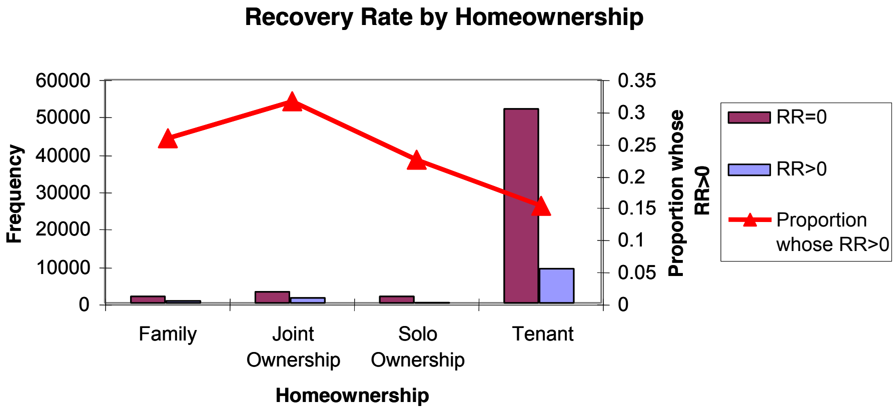


Figure 3.6, Recovery Rate by Homeownership

3.4.4 Country of Residence

Debtors have been divided into four classifications for their country of residence, see figure 3.7; England and Wales, Scotland, Northern Ireland and Foreign. Although the vast majority (over 90%) of the debtors fall into the classification of England and Wales, over 100 debtors reside abroad and they appear to be harder to acquire the debt from.

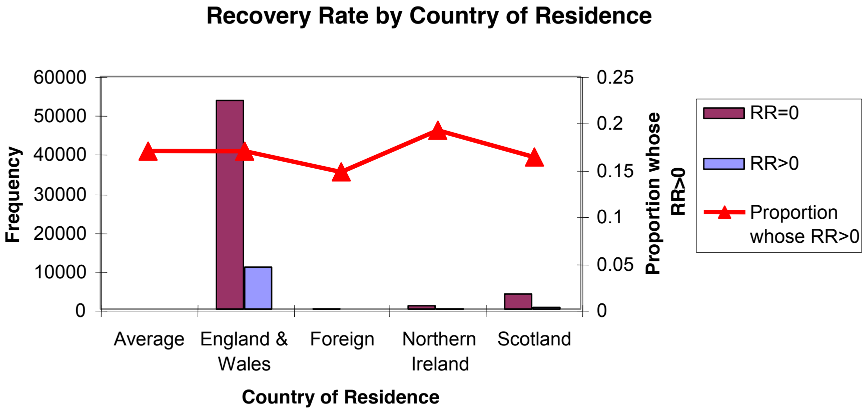


Figure 3.7, Recovery Rate by Country

3.4.5 Debt Amount

The individual debts vary from a few pounds to over £40,000. With the bulk of debtors owing between £500 and £1,000. Figure 3.8 shows that the debt collection agency was especially successful in obtaining money from debtors who owed less than £100, with over 40% of them paying something towards their debt. However there are only 85 debts, which fall into this category.

Debt amount was split by different magnitudes as opposed to using fine classification where the continuous variable would be split into deciles. This is because the range was so large with a majority clustered around £500-£1,000. Also people's behaviour is expected to be grouped, therefore someone who owed £51 would be expected to behave differently to someone who owed £151, the difference being threefold. However if two people owed £1,375, and £1,475, the difference being one hundred pounds as well, they may well react similarly or be treated similarly by the collector. The collector in this case did react differently to those who owed over £500 and those who owed less, which was discovered during personal correspondence with the company. Therefore it made sense to split the bins at this point.

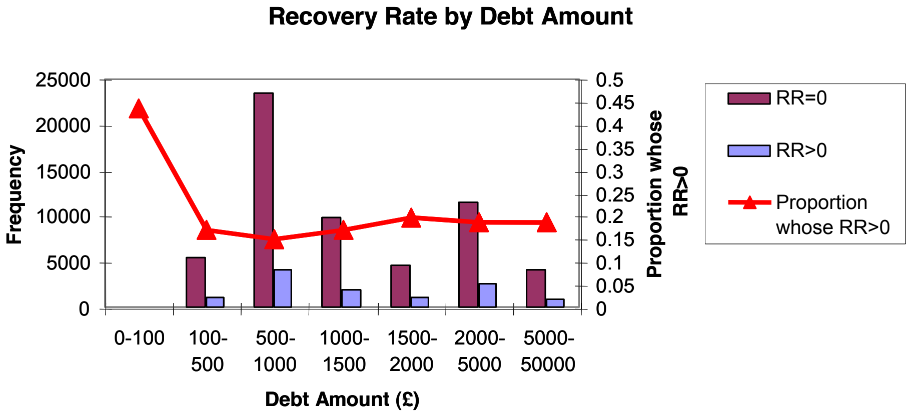


Figure 3.8, Recovery Rate by Debt Amount

3.4.6 Telephone Information

The data included which telephone numbers for the debtors were still active; they had up to five numbers for the debtors, which could include a mobile or work number. Figure 3.9 illustrates the number of active telephone numbers for the debtors and proportion who have a recovery rate of greater than zero. As would be expected the collection agency was least able to obtain money from the debtors, which had no telephone numbers. Having either a work or a mobile number increased the proportion of debtors paying back part of their debt.

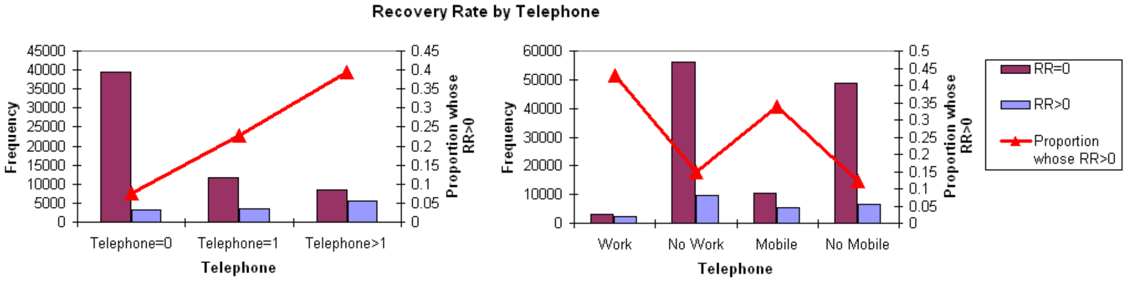


Figure 3.9, Recovery Rate by Telephone

Telephone bins were split at 0, 1 and greater active contact numbers available to the third party collectors. Having no active contact telephone number for a debtor would have a considerable negative impact on the company's ability to communicate with them. Having more than one number may also have been a factor but the difference between two and three contact numbers did not necessarily have any impact at all.

The type of phone might have an impact on the debtor's ability to pay, e.g. a work number implies they are in work and a mobile number implies they can pay some sort of contract with the phone company. This was why they were selected as variables.

3.4.7 Time in Collections

The third party bought the debt over a 20-month period. With the majority of the debt bought in the last eight months. There were two different sets of loans being collected; one is of significantly better quality of debt than the other. The debt collected in Set A is of a lower quality than Set B so they are both shown separately in figures 3.10 and 3.11 respectively.

Set A was old debt, which had been previously collected by other debt agents. Therefore it is harder to collect because others have already tried and failed. The debt on Set B was bought directly from the lender after it had been through their collections department but had not been given to any other agent to collect. This makes a significant difference to the quality of the debt.

As figures 3.10 and 3.11 show, the longer the debt has been with the collections agency, the more likely it is that the debtors will pay back something to the third party. The better quality debt in set B means that the third party is able to collect the debtor more quickly.

The data from the third party is a snapshot at one point in time. However the third party buys the debts over a twenty-month period. Therefore figures 3.10 and 3.11, show not how long the debt has been with the debtor but how long the debt has been with the third party.

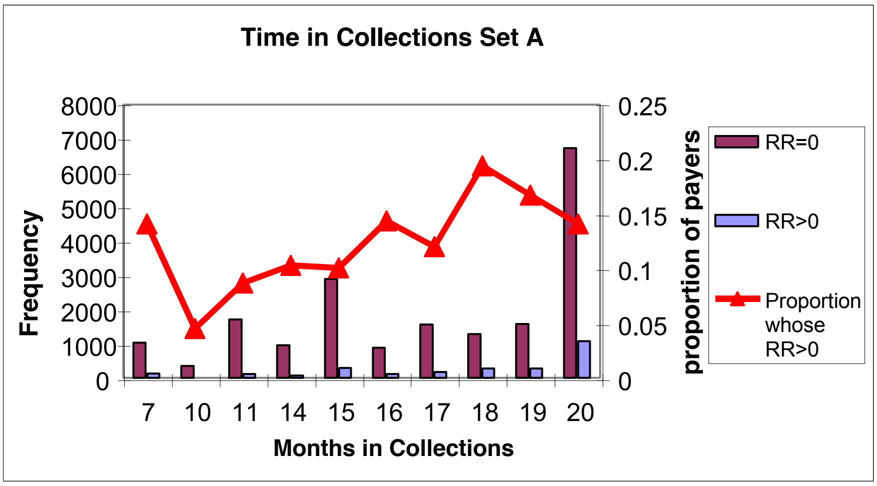


Figure 3.10, Ratio of non-payers to payers by number of months on the books for Set A

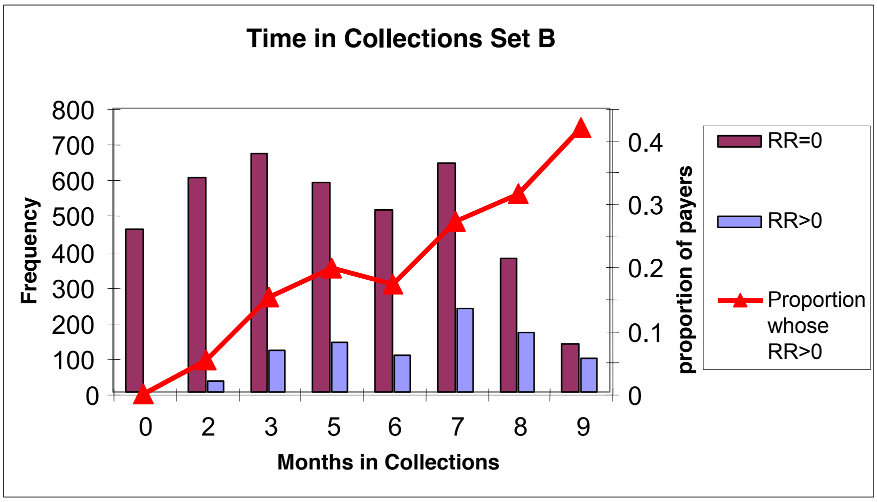


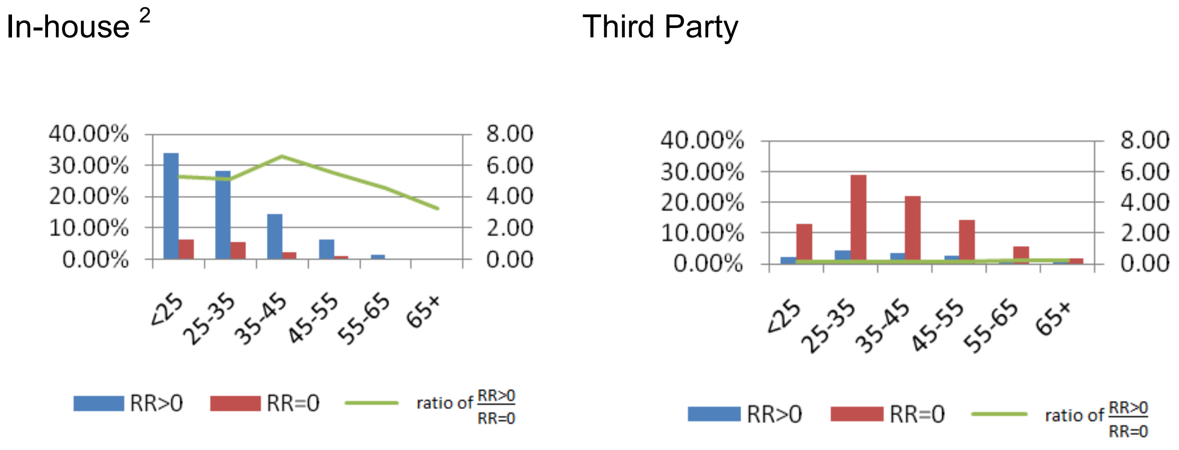
Figure 3.11, Ratio of payers to non-payers for set B by the number of months on the books

3.5 Analysis of the common variables

Both the in-house and the third party data sets have some common variables. These are: age, amount of debt and residential status[1]. This section compares how the distributions of these variables affect the debtors paying back part of their debt.

3.5.1 Age

The majority of debtors from the in-house data set, are in the “<25” and “25-35” bins, the smallest number of debtors are in “65+” bin. Majority of the customers from third party data set are in the “25-35” and “35-45” bins. In the third party case, the trend of the proportion of payer to non-payers is stable, but slightly increasing for the last two bins. Whereas the in-house case, the higher proportion of payer to non-payers is in the “35-45” bin, thereafter the older debtor the lower the proportion of payer to non-payers.


In-house [2]

Third Party

Figure 3.12: RR distribution by age for in-house collection and third party collection

3.5.2 Residential status

Homeownership is divided into the following classifications: ‘family’, ‘owner’, ‘joint ownership’, ‘tenant’ and ‘other’. If the debtor is known to reside in a property owned by a member of their family, but not themselves or live with parents, then their homeownership is classified as ‘family’. If the debtor

[1] Where in-house data set RR<0 due to recovery costs, we made the following assumption: if RR<=0, then RR=0.

[2] Data provided by A. Matuszyk during personal correspondence

resides in a property owned solely by them then their homeownership status is ‘owner’. ‘Joint ownership’ status is recorded if the debtor and another own their residence, ‘tenant’ status if they are renting and finally, ‘other’ if the details are unknown. The vast majority of the debtors in third party data set are recorded as Tenants, over 85%. In the in-house data set, majority of the clients have the Owner status (40%). This can also explain the behaviour of customers. Owners are slightly more likely to pay off the debt whereas tenants belong to the group least likely to pay.

In-house [3] Third Party

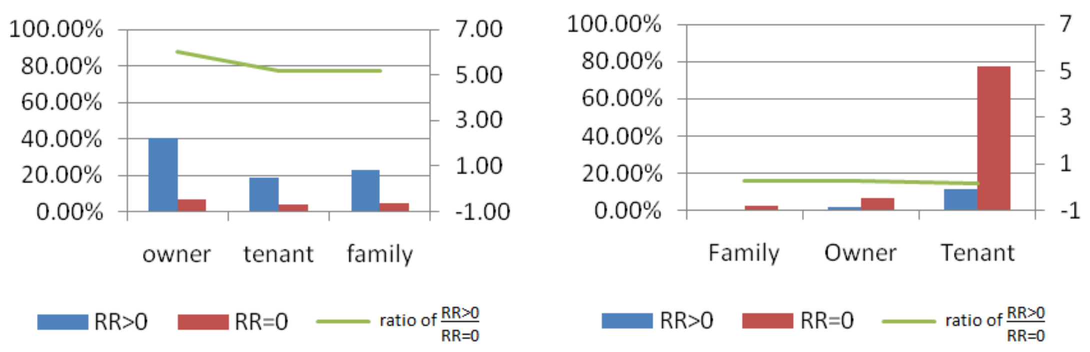
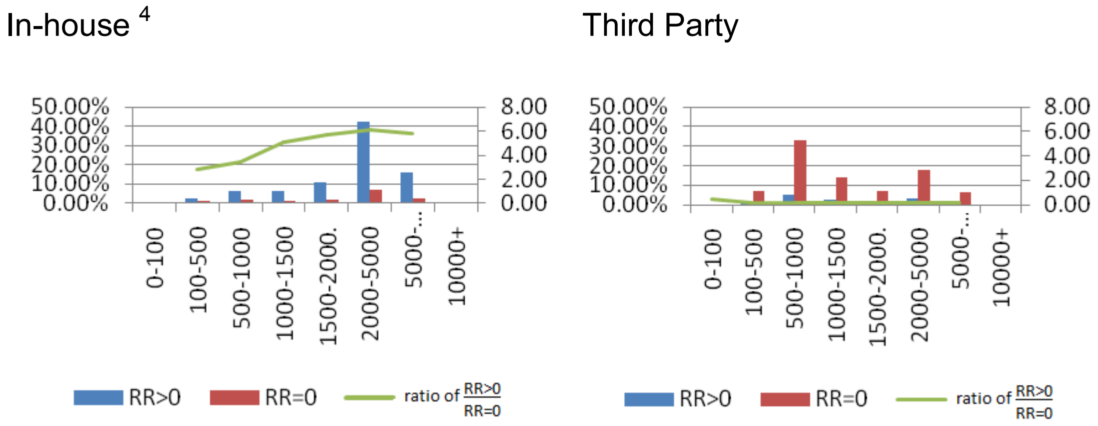


Figure 3.13: RR distribution by homeownership for in-house and third party collection

3.5.3 Debt Amount

The amount of the debt was from a few pounds to £50,000. The variable was divided into eight groups. What is surprising; is that clients, who owe similar amounts in each data set, behave differently. For in-house collection the recovery rate is growing with the amount of debt, in case of Third party the trend is stable with the only exception for the first bucket (£0-£100) where the repayment rate is the highest.

[3] Data provided by A. Matuszyk during personal correspondence



In-house [4]

Third Party

Figure 3.14: RR distribution by debt amount for in-house and third party collection

This analysis demonstrates that some debtor properties like their age, debt amount and residential status have a clear effect on the recovery rate.

3.6 LGD Models

For both the data sets, the models built consisted of two steps. The first step is to estimate the spike in the distributions. So for in-house the split with LGD: LGD≤0 or LGD>0 and LGD=1 or LGD<1 for third party collection. The splits were necessary considering the shape of their respective LGD distributions (Figures 3.2 and 3.3). Logistic regression models were built for both data sets to split them into two groups. The predicted value for those in the first class should be either LGD=0 (In-house) or LGD=1 (third party). For those who paid back part of their debt, the LGD was estimated using a number of different variants of linear regression. These included using ordinary linear regression, applying Beta and log normal transformations to the data before applying regression, the Box-Cox [19] approach to “normalising” the data and using linear regression with Weight Of Evidence (WOE) approach.

[4] Data provided by A. Matuszyk during personal correspondence

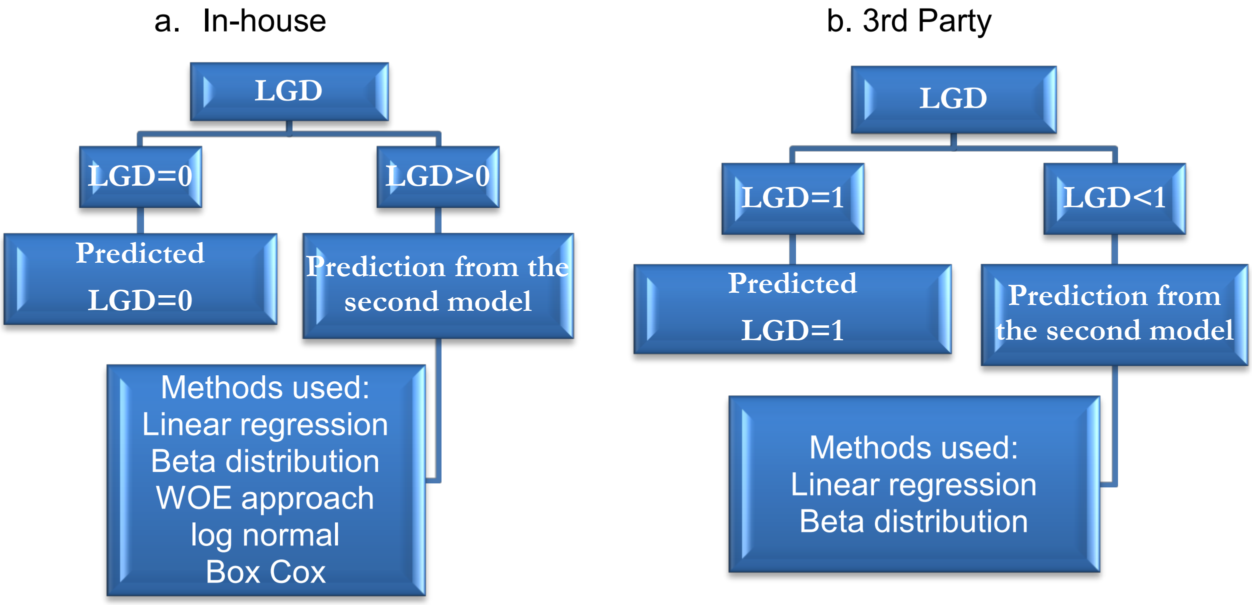

a. In-house
LGD
LGD=0
LGD>0
Predicted LGD=0
Prediction from the second model
Methods used: Linear regression Beta distribution WOE approach log normal Box Cox
b. 3rd Party
LGD
LGD=1
LGD<1
Predicted LGD=1
Prediction from the second model
Methods used: Linear regression Beta distribution

Figure 3.15: LGD models

Table 3.3 contains the variables and results achieved during the LGD modelling for both data sets. As can be seen, different variables were used because of the information available. In-house collections have more data available to them because they have access to the original loan details and behaviour variables from monitoring the loan throughout its lifetime. Whereas the third party is limited to information given by the lender. This information is limited due to lender policy and lack of requirements on the lender to provide useful debtor information.

In-house	3rd Party
1st stage	
LGD=0 versus LGD>0	**LGD=1 versus LGD<1**
The higher the ***loan amount*** the lower the chance of paying off everything	Having a ***work telephone*** number increases the likelihood of paying back part of the debt
The longer the ***lifetime of the loan*** the higher the chance of paying off everything	Having a ***mobile telephone*** number increases the likelihood of paying back part of the debt
The higher the ***application score*** the higher the chance of paying off everything	Having more ***telephone*** numbers increases the likelihood of paying back part of the debt
The more ***time spent in arrears*** during the loan, the higher the chance of paying off everything. However those who were in arrears for more than 2/3 of the time, had a lower chance of paying off everything	Owing ***less than £100*** at default increases the likelihood of paying back part of the debt.
The more the customer was ***in arrears recently*** (in the last 12 months) the higher the chance of paying off everything	

Table 3.3: Variables and results from modelling LGD [5]

[5] In-house data provided by A. Matuszyk during personal correspondence

2nd stage predicting: 0<LGD<1	
LGD>0	**LGD<1**
The higher the ***loan amount*** the higher expected loss rate	The younger the ***debtor's age*** the lower expected loss rate
The higher the ***application score*** the lower expected loss rate	The lower the ***default amount*** owed the lower expected loss rate
The longer the ***lifetime of the loan*** the lower expected loss rate	***Owners*** will have lower expected loss rate
The more the customer was ***in arrears recently*** (in the last 12 months) the lower expected loss rate	Having a ***mobile*** decreases the expected loss rate
The more ***time spent in arrears*** during the loan the lower expected loss rate	Not having a ***contact number*** decreases the expected loss rate

Table 3.3 continued: Variables and results from modelling LGD [6]

Stage one for in-house and third party is focused on different extreme LGD results. The appendix contains a more detailed regression results table for the third party. The contact information was a significant factor in determining who would pay back part of their debt. However where the default amount was separated into bins, not all of the bins were significant. Table A2 shows the 2nd stage linear regression results. All of the variables are significant.

In the 1st stage of the in-house model the concern was with paying off the whole loan whereas for third party the concern was with not paying off any of the loan because this was where the spikes in the LGD distributions were. The in-house model found that the higher the *loan amount* the lower the chance of paying off everything and the third party model found that the higher the *loan amount* the lower the chance of paying off part of the debt. Applicants with a high application score are predicted less likely to default and

[6] In-house data provided by A. Matuszyk during personal correspondence

if they do default the in-house results suggest they are more likely to pay off everything. This suggests the application score recognises the applicant's willingness to pay, which applies both before and after default. A more counterintuitive result is that being in arrears recently increases the chance of paying off completely. Implying that people who have been struggling with debt in their past may cope better with default than those who have never had financial problems. The rest of the in-house model was based on behaviour and application variables, which were unavailable to the third party. Therefore the third party model's variables were more focused on how to contact the debtor i.e. the telephone numbers available.

The second stage model is focused on predicting the LGD between 0 and 1 and trying to fit a distribution. In all cases the models were built in the training set but the results reported are based on the holdout test set. Different methods were tried (see table 3.4), the best method for in-house was weight of evidence with an R^2 of 0.23 and the best method for third party was also the weight of evidence with R^2 of 0.15.

Method	In-house R^2	3rd Party R^2
Box Cox	0.1299	0.0591
Linear regression	0.1337	0.1097
Beta distribution	0.0832	0.1161
Log Normal transformation	0.1347	0.0729
WOE approach	0.2274	0.1496

Table 3.4: Comparison of the results for the 2nd stage models [7]

Table 3.4 shows the fits of the different approaches used in both data sets with R^2 value. It can be noticed that R^2 values are not very different and in

[7] In-house data provided by A. Matuszyk during personal correspondence

both cases not very high. These results suggest that LGD values seem difficult to forecast. All of the models for third party and in-house, except for weight of evidence, gave a narrow distribution focused around the mean. Only weight of evidence gave a distribution covering the whole range 0-1 for which the LGD observed results covered.

The Mean Squared Error (MSE) results for WOE approach were 0.193 for the in-house and 0.195 for the third party.

The linear regression model did not use any transformation of the target variable. In the Box Cox, Beta and lognormal models, the target variable was transformed using Box Cox transformation, the Beta distribution and natural logarithms respectively. Then linear regression was applied and the results transformed back. These transformations were applied because linear regression assumes a normal distribution. However the recovery rates were not normally distributed. These approaches are covered in more detail in chapter 2.

The variables used by the in-house model and the third party models are again very different due to the information available. The in-house collections were privy to application and behaviour variables whereas the third party were limited to personal variables and contact information. Yet despite these different variables and the greater information held in-house the results of the models are very similar. Both the linear regression and the beta distribution models gave R^2 values around 0.1, where the predicted results were a poor representation of the observed results since in all cases the predictions were clustered around the means.

3.7 Summary

Although both analysed data sets are about debt recovery, the information available in each case is quite different and the average recovery rate varied from 5% to 46%. The two-stage model is appropriate for both, even though the spikes are at opposite ends of the LGD distribution. All of this is not surprising because third party debt will usually go through several collection processes, so by definition must be harder to collect.

Both sides can use these models to determine the price at which to buy a debt. The third party model gives an indication of recovery rate so the third party can set an internal upper limit for the price of buying the debt. For the in-house collection; the question is how much more would they get by keeping the debt in their collection process for some further time? To get a feel for this one needs to estimate RR in the next year using the information on the borrower and the amount already recovered which will be covered in chapter 5.

What is remarkable about the models discussed in this chapter is that despite the in-house data set being more detailed, the goodness of fit for both was very similar. This is despite the third party model focusing on contact details and very few personal details including age and homeownership. Whereas the in-house model focused more on loan characteristics; loan amount, time spent in arrears, lifetime of the loan. Models for predicting LGD for both in-house and third party will be covered in more detail later on in this thesis.

Chapter 4: Predicting Third Party Collections

The last chapter focused on comparing the in-house and third party collections. This chapter is focused on improving the third party collection predictions. Since third party predictions were poorer in comparison to the weight of evidence in-house results (chapter 3) therefore the predictions might be improved by a more detailed analysis. With this in mind the debtors were grouped and then modelled using regression. Again the models were split into two stages.

The data assessed in this chapter is the same as the data for third party in chapter three. The data is a single dump of all the debt being collected by the third party. So the information is a single snap shot of the debts bought by the third party over a period of twenty months. Because the data is a single dump, one data set for all debtors, then the time in collections is different for each debtor. However the debt comes from two different sources where the older debt is of a poorer quality than the newer debt.

This data limitation is the motivation behind the modelling methodology in this chapter. The length of time in collections should have a positive relationship with recovery rates, since the more time in collections means that the agency have longer to recover the debt. However with this data set, while there was this positive relationship within the two types of debt, overall this relationship did not exist across the data set, because the newer debt was of a superior quality. Therefore the size of the debt was used to group the debt due to the data limitation.

The poor quality of the data set also means that it cannot be used to accurately predict LGD, therefore this chapter looks at predicting if the debtor would pay back anything rather than predicting the recovery rate. The main reason for this is that the data set did not contain any history of the payments made, just the overall amount recovered. Therefore no set time period could be used to predict recovery rate. With 83% failing to pay back anything, the first stage of the model had to be predicting if the debtor would pay back. This is also the rational for splitting the debt into older and younger than six months for predicting if the debtors would start to pay back their debt. Since there was no historical record of when the debtors started to pay back, six months was used to ensure that the third party would have adequate time to start the collections process without limiting the data set.

4.1 Grouping the Debtors

Since debt amount is a significant factor in all of the previous analyses, and is known up front of the collection process it is an excellent factor to distinguish between different types of debt. Debt amount was separated into four groups and then each group was analysed separately. The summary of the split is in table 4.1.

	Range of debt value	Number of Debtors	Number of Debtors who paid
Small	£0< Debt Value<£750	20620	3271
Medium	£750≤ Debt Value<£1,000	13638	2008
Large	£1,000≤ Debt Value<£2,000	17872	3232
Extra Large	£2,000≤ Debt Value	19556	3680

Table 4.1, Summary of grouping debt by value

Once the debtors were grouped the regression results were different from the results previously found. Assuming a beta distribution, for the linear regression model, resulted in a poorer model than using the ordinary linear regression model. The R-squared values were in some cases improved. In particular the R-squared values for the models predicting for debtors who fell into the category of owing a small amount of debt, were all an improvement on the models predicting for all of the debtors.

Debt Value	Small	Medium	Large	Ex-large	All Debtors
Logistic Regression					
Root MSE	0.33064	0.32651	0.35251	0.36208	0.34518
R-Squared	0.185	0.1489	0.1565	0.189	0.1579
Logistic Regression (on books minimum 6 months)					
Root MSE	0.35618	0.3478	0.36605	0.36439	0.36091
R-Squared	0.2124	0.1848	0.1934	0.2034	0.1945
Linear Regression Model (Recovery Rate)					
Root MSE	0.33542	0.30312	0.28724	0.25945	0.32288
R-Squared	0.2343	0.1503	0.1284	0.0797	0.117

Table 4.2, Summary of regression models with debt grouped by value

Table 4.2, summarises the regression models' results when the debtors are grouped by the value of their debt. These results are from the training data to assess how goodness of fit of the models. R^2 is calculated by using the likelihood-ratio statistic. All of the models use the same variables. These variables are:

- Is the debtor aged between 18-25 (Age25)
- Is the debtor aged between 25-35 (Age35)
- Is the debtor aged between 35-45 (Age45)
- Is the debtor aged between 45-55 (Age55)
- Reference category for age is if the debtor is aged over 55
- The amount of debt owed (Debt Value)
- Does the debtor have one or more active telephone numbers (One or more Telephones)
- Reference category is if the debtor has no active telephone numbers and therefore the third party has no way to contact the debtor via telephone
- Does the debtor have an active mobile number (Mobile Telephone)
- Reference category is if the debtor has no active mobile number known to the third party
- Does the debtor have an active work number (Work Telephone)
- Reference category is if the debtor has no active work number known to the third party

- Number of telephone numbers (No. of Telephones)
- Does the debtor reside in a residence owned by a family member (Family Home)
- Does the debtor reside in a residence jointly owned by them and another (Joint Ownership)
- Does the debtor reside in a residence owned by them solely (Solo Ownership)
- Reference category is if the debtor is a tenant or their residence status is unknown
- If the debtor is female (Female)
- Reference category is if the debtor is male
- If the debtor has been in the third party's collection process for less than six months (In collections <6 months)
- If the debtor has been in the third party's collection process for between six and twelve months (6<collections <12 months)
- Reference category is if the debtor has been in the third party's collection process for longer than twelve months

Table 4.2 shows the results for three sets of prediction models. The top sets of results are for a logistic regression model to predict who will pay back part of their debt; this is the equivalent to 1st stage in the models discussed in chapter 3. The middle sets of results are for a logistic regression model (1st stage) on debtors who have been in collections for longer than six months. The final sets of results are for a linear regression model (equivalent to the 2nd stage in chapter 3) to predict the recovery rate for those debtors who have started to pay back their debt. In each type of model, the debtors were first of all separated by the value of the debt they owed. Then the results of the regression models were compared to the regression models for all of the debtors for that regression model.

The Logistic Regression results (top of table 4.2) are for all of the debtors, modelling who will pay back part of their debt. As can be seen from table 4.2, debtors who owed less than £750 (small) have the best prediction model. The other groups of debtors (medium, large, extra large) had a worse prediction model than the prediction model for all debtors.

The Logistic Regression Model for debtors who had been in the third party's collection process for a minimum of six months is the middle set of results in table 4.2. The reason that this regression model was tested, is that the variable "debtor had been in collections for less than six months" was significant in all cases and the estimate was negative. This shows that those debtors who had been in collections for less than six months, when this data was collected, are less likely to pay back any of their debt. This was assumed to be an operational issue, in that the third party had not had enough time to collect money from the debtor. Some may view six months to be too long. However looking at figure 3.11, six months gave good results for debtors paying back part of their debt so it was used as a variable here. These results for the regression models show an improvement on the regression models using the entire set of debtors. This is most likely due to the fact that, the third party will have probably contacted the debtor within the first six months and collected some money from them if the debtor is willing and able to pay. Again debtors who owe less than £750 (small) have the best prediction model.

The results for the linear regression model (2nd stage) for predicting the recovery rate on debtors who have started paying are displayed at the bottom of table 4.2. The linear regression model was not separated into debtors who have been in collections for more than six months because the reason for the separation of the logistic regression model was to allow time for the third party to contact the debtor and collect money from them. Since only debtors who have paid back part of their debt are included in the linear regression model, the third party has evidently already had sufficient time to contact the debtor and arrange for payment. The prediction models are more accurate when the debtors are separated into the debt value groups before modelling. Only those debtors who owed more than £2,000 (extra large) had a worse prediction model than the model for all debtors. Again debtors who owe less than £750 (small) have the best prediction model.

As can be seen in table 4.2, separating the debtors by the debt amount owed before modelling, improved the prediction models in some cases. Specifically in the case of debtors who owed a small amount (less than £750); the prediction models were all an improvement on the models using all of the debtors regardless of debt amount. Separating out those debtors who had been in collections for less than six months also improved the logistic regression model to predict who would pay back part of their debt. By separating the debtors by the debt amount before modelling, the linear regression models for predicting the debtors' recovery rate were improved in all cases, except those debtors who owed more than £2,000.

4.2 Model

The models all used the same variables, but the resulting parameter estimates were different for each sub group. Tables 4.3 to 4.14 show the results of the regressions.

4.2.1 Small Debts

Table 4.3 gives the results for the logistic regression for small debts to estimate if a debtor will pay back part of their debt. As the table 4.3 shows, only the variable of Solo Ownership (does the debtor reside in a residence owned by them solely) is not significant. In regards to age the results show that if the debtor is over 55 then they are more likely to pay back part of their debt than if they are younger. This is indicated because all of the variables shown are negative and the reference category for age is over 55, thereby the parameter estimate is positive. The higher the estimate of the parameter; the more likely that a debtor with that characteristic, will pay back part of their debt. This result corroborates the results shown in figure 3.4 in chapter 3, which show that debtors over 55 were more likely to pay back part of their debt.

In table 4.3, the Parameter Estimate of Debt Value (value of debt at time of sale) is negative too, which is interesting because for medium and large debts, as show in tables 4.4 and 4.5, the opposite is true. This indicates that the larger the amount owed, the less likely the debtor will be to pay back part of their debt. These results bear out the results in figure 3.8 in chapter 3, where the debtor was more likely to pay back part of their debt for debts owing less than £100, after this the proportion paying back part of their debt fell and then rose after the amount owed was £1,000. It then started to fall again after £2,000; hence in table 4.6 the parameter estimate for debt amount is again negative.

Table 4.3 shows that the more contact telephone numbers available to the collectors the more likely the debtor would pay back part of their debt. Especially if one of those numbers was a work telephone, indicating that they were still employed.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	1	0.21026	0.01358	15.49	<.0001
Age25	1	-0.03909	0.0119	-3.28	0.001
Age35	1	-0.05269	0.01138	-4.63	<.0001
Age45	1	-0.04881	0.01186	-4.12	<.0001
Age55	1	-0.03106	0.01312	-2.37	0.0179
Debt Value	1	-0.00008906	1.52E-05	-5.86	<.0001
One or more Telephones	1	0.08355	0.01176	7.11	<.0001
Mobile Telephone	1	0.03152	0.00989	3.19	0.0014
Work Telephone	1	0.09285	0.01325	7.01	<.0001
Female	1	0.04729	0.00551	8.59	<.0001
No. of Telephones	1	0.10187	0.00693	14.69	<.0001
Family Home	1	0.04903	0.01652	2.97	0.003
Joint Ownership	1	0.0496	0.01591	3.12	0.0018
Solo Ownership	1	-0.017	0.02045	-0.83	0.4057
In collections <6 months	1	-0.2014	0.00717	-28.07	<.0001
6<collections <12 months	1	-0.0276	0.00714	-3.86	0.0001

Table 4.3, Logistic regression results (1st stage) for small debts

If the debtor was female then they were more likely to pay back part of their debt then if they were male. Again this is substantiated in figure 3.5 in chapter

3 where debtors with female titles (Miss and Mrs) had a larger proportion paying back part of their debt than their male counterparts.

For home ownership what is interesting is that solo ownership has a negative effect on the results but this is not as significant a result as stated earlier and goes against the results in figure 3.6 chapter 3. However debtors residing in jointly owned or family own residences are the most likely to pay back part of their debt as shown in figure 3.6 in chapter 3.

The longer the debt was in collections and therefore the more time the third party had to act on the debt, then the more likely it is that the debtor will pay back part of their debt, which is the intuitive response expected. This is indicated by the reference category being zero, for the time in the third party's collections being greater than 12 months. This implies a higher coefficient than the other two variables for time in collections, which both have negative coefficients as the last two rows of table 4.3 show. Therefore the probability that a debtor will have a collection rate>0 is higher for debtors who have been in collections for more than 12 months.

4.2.2 Medium Debts

Table 4.4 gives the results for the logistic regression for medium debts to estimate if a debtor will pay back part of their debt. Again the variable of Solo Ownership is not significant, and the probability that a debtor will have a collection rate>0 is higher if the debtor is over 55 than if they are younger. As discussed earlier the debt value coefficient is positive indicating that the higher the debt amount owed the more likely the debtor's collection rate>0. Once more the greater the number of contact telephone numbers available to the collectors the higher the probability that a debtor will have a collection rate>0. Especially if one of those numbers is a work telephone, indicating that they are still employed.

Again if the debtor was female then they were more likely to pay back part of their debt then if they were male. For home ownership, what is interesting, is that solo ownership now has a positive effect on the results but this is not a significant result as stated earlier and supports the results in figure 3.6 chapter 3. Again debtors residing in a jointly owned or family owned residence are the most likely to have a collection rate>0 as shown in figure 3.6 in chapter 3. Also the longer the debt was in collections the more likely the debtor's collection rate>0.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	1	0.07576	0.04304	1.76	0.0784
Age25	1	-0.06202	0.01502	-4.13	<.0001
Age35	1	-0.07533	0.01416	-5.32	<.0001
Age45	1	-0.06233	0.01445	-4.31	<.0001
Age55	1	-0.03374	0.01572	-2.15	0.0319
Debt Value	1	0.00011518	4.65E-05	2.48	0.0132
One or more Telephones	1	0.06397	0.01359	4.71	<.0001
Mobile Telephone	1	0.05173	0.01122	4.61	<.0001
Work Telephone	1	0.119	0.01563	7.61	<.0001
Female	1	0.03811	0.00665	5.73	<.0001
No. of Telephones	1	0.08	0.00767	10.44	<.0001
Family Home	1	0.06706	0.0181	3.7	0.0002
Joint Ownership	1	0.08397	0.01659	5.06	<.0001
Solo Ownership	1	0.02745	0.0202	1.36	0.1743
In collections <6 months	1	-0.17952	0.00916	-19.59	<.0001
6<collections <12 months	1	-0.05366	0.00887	-6.05	<.0001

Table 4.4, Logistic regression results (1st stage) for medium debts

4.2.3 Large Debts

Table 4.5 gives the results for the logistic regression for large debts to estimate if a debtor will pay back part of their debt. As the table shows the two variables of Solo Ownership and if the debtor has been in collections for between six to twelve months, are not significant. There are no other significant changes in the variables' coefficients as in the regression results for medium sized debts.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	1	0.11539	0.02274	5.07	<.0001
Age25	1	-0.05089	0.01579	-3.22	0.0013
Age35	1	-0.06666	0.0135	-4.94	<.0001
Age45	1	-0.06195	0.01351	-4.59	<.0001
Age55	1	-0.03344	0.01469	-2.28	0.0228
Debt Value	1	0.00003502	1.28E-05	2.75	0.006
One or more Telephones	1	0.08168	0.01422	5.74	<.0001
Mobile Telephone	1	0.03658	0.01185	3.09	0.002
Work Telephone	1	0.09993	0.01562	6.4	<.0001
Female	1	0.04569	0.00724	6.31	<.0001
No. of Telephones	1	0.08295	0.00793	10.46	<.0001
Family Home	1	0.05449	0.01806	3.02	0.0026
Joint Ownership	1	0.06662	0.0144	4.62	<.0001
Solo Ownership	1	0.01651	0.01845	0.89	0.3709
In collections <6 months	1	-0.16367	0.00998	-16.41	<.0001
6<collections <12 months	1	-0.01175	0.00974	-1.21	0.2275

Table 4.5, Logistic Regression Results (1st stage) for Large Debts

4.2.4 Extra Large Debts

Table 4.6 gives the results for the logistic regression for extra large debts to estimate if a debtor will pay back part of their debt. As discussed earlier the debt value coefficient is negative indicating that the higher the debt amount owed the less likely the debtor is to pay back part of the debt. The variables of Solo Ownership and if the debtor has been in collections for more than 12 months are now significant. There are no other significant changes in the variables' coefficients as in the regression results for large sized debts.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	1	0.11302	0.01112	10.17	<.0001
Age25	1	-0.02759	0.0148	-1.86	0.0624
Age35	1	-0.04668	0.00981	-4.76	<.0001
Age45	1	-0.02812	0.00933	-3.01	0.0026
Age55	1	-0.022	0.00997	-2.21	0.0274
Debt Value	1	-0.000003	1.02E-06	-2.93	0.0034
One or more Telephones	1	0.0837	0.01139	7.35	<.0001
Mobile Telephone	1	0.06257	0.00943	6.64	<.0001
Work Telephone	1	0.06641	0.01202	5.52	<.0001
Female	1	0.04457	0.00629	7.09	<.0001
No. of Telephones	1	0.05599	0.00612	9.14	<.0001
Family Home	1	0.03507	0.01461	2.4	0.0164
Joint Ownership	1	0.14564	0.00988	14.74	<.0001
Solo Ownership	1	0.0371	0.01251	2.96	0.003
In collections <6 months	1	-0.10901	0.00889	-12.26	<.0001
6<collections <12 months	1	0.04758	0.00844	5.64	<.0001

Table 4.6, Logistic Regression Results (1st stage) for Extra Large Debts

4.2.5 Small Debts Older than 6 Months

Table 4.7 gives the results for the logistic regression for small debts to estimate if a debtor will pay back part of their debt; for debts, which had been in the third party's collection process for longer than 6 months. The reason for producing this regression model is that the variable "if the debtor had been in collections for less than six months" was significant in all cases and the coefficient was negative. This shows that those debtors who had been in collections for less than six months, when this data was collected, are less likely to have paid back any of their debt. This was assumed to be an operational issue, in that the third party had not had enough time to collect money from the debtor.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	1	0.22306	0.01853	12.04	<.0001
Age25	1	-0.03178	0.01642	-1.94	0.0529
Age35	1	-0.05204	0.01558	-3.34	0.0008
Age45	1	-0.04758	0.01627	-2.92	0.0035
Age55	1	-0.03191	0.01825	-1.75	0.0804
Debt Value	1	-0.00013931	2.2E-05	-6.33	<.0001
One or more Telephones	1	0.1759	0.01872	9.39	<.0001
Mobile Telephone	1	0.01744	0.016	1.09	0.2757
Work Telephone	1	0.05305	0.02078	2.55	0.0107
Female	1	0.05188	0.0077	6.73	<.0001
No. of Telephones	1	0.13169	0.01162	11.33	<.0001
Family Home	1	0.06555	0.02677	2.45	0.0143
Joint Ownership	1	0.09527	0.0246	3.87	0.0001
Solo Ownership	1	0.05128	0.03302	1.55	0.1205
6<collections <12 months	1	-0.072	0.00835	-8.63	<.0001

Table 4.7, Logistic regression results (1st stage) for small debts older than 6 months

The results in table 4.7 are very similar to the results in table 4.3 as expected. What is different however is that now the variables of whether the debtor has an active mobile number, aged between 45 and 55, as well as Solo Ownership are not significant. The variable parameter estimates have also changed slightly in most cases, Solo Ownership having the largest change going from negative to positive but this result is not significant.

Modelling the debt that was older than 6 months improved the regression, giving an R^2 of 0.2124 instead of 0.185 for all small debts.

4.2.6 Medium Debts Older than 6 Months

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	1	0.06375	0.05891	1.08	0.2792
Age25	1	-0.0674	0.02075	-3.25	0.0012
Age35	1	-0.08495	0.01944	-4.37	<.0001
Age45	1	-0.07257	0.01994	-3.64	0.0003
Age55	1	-0.01998	0.02168	-0.92	0.3568
Debt Value	1	0.00012244	6.39E-05	1.92	0.0554
One or more Telephones	1	0.13858	0.02068	6.7	<.0001
Mobile Telephone	1	0.03231	0.01729	1.87	0.0617
Work Telephone	1	0.09609	0.02365	4.06	<.0001
Female	1	0.04682	0.00919	5.09	<.0001
No. of Telephones	1	0.09913	0.01202	8.25	<.0001
Family Home	1	0.09173	0.0281	3.26	0.0011
Joint Ownership	1	0.09809	0.02476	3.96	<.0001
Solo Ownership	1	-0.0108	0.03167	-0.34	0.733
6<collections <12 months	1	-0.08765	0.01012	-8.66	<.0001

Table 4.8, Logistic regression results (1st stage) for medium debts older than 6 months

Table 4.8 gives the results for the logistic regression for medium debts to estimate if a debtor will pay back part of their debt; for debts, which had been in the third party's collection process for longer than 6 months.

The results in table 4.8 are very similar to the results in table 4.4 as expected. What is different however is that again the variable of whether the debtor is aged between 45 and 55, has an active mobile, as well as Solo Ownership is now not significant. The debt value is also less significant than before. The variable parameter estimates have also changed slightly in most cases.

Modelling the debt that was older than 6 months improved the regression, giving an R^2 of 0.1848 instead of 0.1489 for all medium debts.

4.2.7 Large Debts Older than 6 Months

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	1	0.12358	0.02457	5.03	<.0001
Age25	1	-0.04838	0.01742	-2.78	0.0055
Age35	1	-0.06945	0.01488	-4.67	<.0001
Age45	1	-0.05789	0.015	-3.86	0.0001
Age55	1	-0.03142	0.01631	-1.93	0.054
Debt Value	1	0.00002259	1.4E-05	1.61	0.1078
One or more Telephones	1	0.15527	0.01741	8.92	<.0001
Mobile Telephone	1	0.03308	0.01477	2.24	0.0251
Work Telephone	1	0.05942	0.01948	3.05	0.0023
Female	1	0.04219	0.00795	5.31	<.0001
No. of Telephones	1	0.10661	0.01006	10.6	<.0001
Family Home	1	0.03682	0.02381	1.55	0.122
Joint Ownership	1	0.09812	0.01759	5.58	<.0001
Solo Ownership	1	0.03457	0.02308	1.5	0.1342
6<collections <12 months	1	-0.0564	0.00895	-6.3	<.0001

Table 4.9, Logistic regression results (1st stage) for large debts older than 6 months

Table 4.9 gives the results for the logistic regression for large debts to estimate if a debtor will pay back part of their debt; for debts, which had been in the third party's collection process for longer than 6 months.

The results in table 4.9 are very similar to the results in table 4.5 as expected. What is different however is that now the variable of whether the debtor resides at the home of a family member and the debt value as well as Solo

Ownership is not significant. The variable of whether the debtor is aged between 45 and 55 is also less significant than before. The variable parameter estimates have also changed slightly in most cases. The number of active telephones now has a greater positive effect than before.

Modelling the debt that was older than 6 months improved the regression, giving an R^2 of 0.1934 instead of 0.1565 for all large debts.

4.2.8 Extra Large Debts Older than 6 Months

Table 4.10 gives the results for the logistic regression for extra large debts to estimate if a debtor will pay back part of their debt, for debts that had been in the third party's collection process for longer than 6 months.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	1	0.10784	0.01348	8	<.0001
Age25	1	-0.00961	0.01937	-0.5	0.6198
Age35	1	-0.04137	0.01273	-3.25	0.0012
Age45	1	-0.02917	0.01231	-2.37	0.0178
Age55	1	-0.01899	0.01331	-1.43	0.1536
Debt Value	1	-0.00000458	1.26E-06	-3.65	0.0003
One or more Telephones	1	0.1324	0.01636	8.09	<.0001
Mobile Telephone	1	0.03008	0.0139	2.16	0.0305
Work Telephone	1	0.04951	0.01798	2.75	0.0059
Female	1	0.05501	0.00829	6.64	<.0001
No. of Telephones	1	0.08906	0.0092	9.68	<.0001
Family Home	1	0.05764	0.0224	2.57	0.0101
Joint Ownership	1	0.18757	0.01489	12.59	<.0001
Solo Ownership	1	0.07342	0.01859	3.95	<.0001
6<collections <12 months	1	0.00248	0.00917	0.27	0.7866

Table 4.10, Logistic regression results (1st stage) for extra large debts older than 6 months

The results in table 4.10 are very similar to the results in table 4.6 as expected. What is different however is that now the variable of whether the debtor is aged between 18 and 25 or 45 and 55 and if the debt has been in collections for longer than 6 months is no longer significant. The variable parameter estimates have also changed slightly in most cases.

Modelling the debt that was older than 6 months improved the regression, giving an R^2 of 0.2034 instead of 0.189 for all extra large debts.

4.2.9 Recovery Rate for Small Debts

Table 4.11 gives the results for the linear regression for small debts to estimate how much of their debt they would repay if they repaid part of the debt. As the table shows the variables of debtors age, if they have a mobile phone, reside in a family home or are female are not significant. In regards to age the results show that if the debtor is younger then they are more likely to pay back more of the debt but these results are not significant.

Debt value coefficient is negative, indicating that the larger the amount owed, the less of the debt the debtor is likely to pay back. Telephones have a more complicated effect on the recovery rate. This model shows that if the debtor has no contact telephone then they pay back more than if they do have a contact telephone. Evidently if the debtor does pay back part of their debt without being contacted by phone then they are more amenable to paying back their debt and therefore pay back more than those contacted by telephone. This result is only reversed if the debtor had at least four active telephones however one or two of those phone numbers would have to be a mobile or work number. Since both of these had a negative effect on the recovery rate, the debtor would really have to have five active telephone numbers to have the same positive result on the estimated recovery rate as if the debtor had no phone number. If the debtor had five contact numbers (maximum on the records) then the collectors must have contacted them numerous times to try out all of the numbers. However having as little as two numbers, provided they were not a mobile or work number, would have a positive effect on the debtor’s estimated recovery rate.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	1	1.01424	0.03593	28.23	<.0001
Age25	1	0.06053	0.03024	2	0.0454
Age35	1	0.03293	0.0288	1.14	0.2529
Age45	1	0.0294	0.0298	0.99	0.3238
Age55	1	-0.02919	0.03266	-0.89	0.3715
Debt Value	1	-0.00074813	4.45E-05	-16.82	<.0001
One or more Telephones	1	-0.0542	0.028	-1.94	0.053
Mobile Telephone	1	-0.01577	0.02107	-0.75	0.4543
Work Telephone	1	-0.09325	0.02368	-3.94	<.0001
Family Home	1	0.04227	0.03528	1.2	0.231
Joint Ownership	1	0.08457	0.03243	2.61	0.0092
Solo Ownership	1	0.13152	0.0503	2.61	0.009
Female	1	-0.0181	0.0152	-1.19	0.2341
In collections <6 months	1	-0.33971	0.02656	-12.79	<.0001
6<collections <12 months	1	-0.16402	0.02087	-7.86	<.0001
No. of Telephones	1	0.02816	0.01435	1.96	0.0499

Table 4.11, Linear regression results (2nd stage) for small debts

In the logistic regression results, having a female debtor improved their probability of their collection rate>0. However in this linear regression, a female debtor has a negative coefficient decreasing their predicted recovery rate in comparison to male debtors.

For home ownership variables, provided the debtor is not a tenant then it had a positive effect on the estimated recovery rate. Debtors with “solo ownership” have the highest coefficient.

The longer the debt was in collections and therefore the more time the collectors had to act on the debt then the higher the recovery rate as would be expected.

4.2.10 Recovery Rate for Medium Debts

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	1	0.69143	0.10164	6.8	<.0001
Age25	1	0.04769	0.03255	1.46	0.1431
Age35	1	0.06688	0.03029	2.21	0.0274
Age45	1	0.02947	0.0306	0.96	0.3357
Age55	1	0.02179	0.03242	0.67	0.5016
Debt Value	1	-0.00024271	0.000111	-2.18	0.0294
One or more Telephones	1	-0.04237	0.0288	-1.47	0.1414
Mobile Telephone	1	0.02553	0.02119	1.2	0.2286
Work Telephone	1	-0.03703	0.02413	-1.53	0.1251
Family Home	1	0.06866	0.03411	2.01	0.0443
Joint Ownership	1	0.0427	0.0301	1.42	0.1563
Solo Ownership	1	0.14391	0.04119	3.49	0.0005
Female	1	-0.04278	0.01588	-2.69	0.0072
In collections <6 months	1	-0.31173	0.02582	-12.08	<.0001
6<collections <12 months	1	-0.22429	0.02179	-10.29	<.0001
No. of Telephones	1	0.00664	0.0135	0.49	0.6229

Table 4.12, Linear regression results (2nd stage) for medium debts

Table 4.12 gives the results for the linear regression for medium debts to estimate how much of their debt they would repay if they repaid part of the debt. As the table shows only the variables of debt values, female and time in collections are significant. In regards to age the results show that if the debtor is younger then they are more likely to pay back more of the debt but these results (with the exception of age 25-35) are not significant.

Debt value parameter estimate is negative, this indicates that the larger the amount owed, the less of the debt the debtor will be likely to pay back. Again telephones have a more complicated effect on the recovery rate but none of the results are significant. The results are similar to those in table 4.11 except that mobile telephones now have a positive effect on recovery rate estimates.

Once more the coefficient for the variable “female” is negative decreasing their predicted recovery rate.

For home ownership, provided the debtor is not a tenant, it had a positive effect on the estimated recovery rate. Only the result for joint ownership is not significant. The longer the debt was in collections then the more the debtor is likely to pay back as expected.

4.2.11 Recovery Rate for Large Debts

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	1	0.49633	0.04092	12.13	<.0001
Age25	1	0.11446	0.02685	4.26	<.0001
Age35	1	0.0592	0.02252	2.63	0.0086
Age45	1	0.08124	0.02214	3.67	0.0002
Age55	1	0.06546	0.02342	2.79	0.0052
Debt Value	1	-0.0001029	0.000023	-4.47	<.0001
One or more Telephones	1	-0.01625	0.02312	-0.7	0.4822
Mobile Telephone	1	0.01692	0.01705	0.99	0.3212
Work Telephone	1	-0.01859	0.01969	-0.94	0.345
Family Home	1	0.0053	0.02714	0.2	0.8453
Joint Ownership	1	0.09187	0.02163	4.25	<.0001
Solo Ownership	1	0.10598	0.03117	3.4	0.0007
Female	1	-0.02818	0.01312	-2.15	0.0319
In collections <6 months	1	-0.26528	0.02123	-12.5	<.0001
6<collections <12 months	1	-0.18705	0.0189	-9.9	<.0001
No. of Telephones	1	0.00198	0.01094	0.18	0.8564

Table 4.13, Linear regression results (2nd stage) for large debts

Table 4.13 gives the results for the linear regression for large debts to estimate how much of their debt they would repay if they repaid part of the debt. As the table shows only the variables of female, telephone numbers and family home are not significant. In regards to age the results show that again if the debtor is under 25 then they are more likely to pay back more of the debt than those who are older.

Debt value parameter estimate is again negative, this indicates that the larger the amount owed, the less of the debt the debtor is likely to pay back. Again telephones have a more complicated effect on the recovery rate but none of the results are significant. The results are similar to those in tables 4.11 and 4.12. Again mobile telephones now have a positive effect on recovery rate estimates. There is little change between the variable coefficients for home ownership, sex and length of time in collections as in the results in table 4.12.

4.2.12 Recovery Rate for Extra Large Debts

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	1	0.33591	0.02486	13.51	<.0001
Age25	1	0.02084	0.02833	0.74	0.462
Age35	1	0.01607	0.01934	0.83	0.406
Age45	1	-0.00099432	0.01746	-0.06	0.9546
Age55	1	-0.00767	0.01825	-0.42	0.6745
Debt Value	1	-0.00001379	2.31E-06	-5.97	<.0001
One or more Telephones	1	0.00211	0.02064	0.1	0.9186
Mobile Telephone	1	0.0145	0.01513	0.96	0.338
Work Telephone	1	-0.00547	0.01731	-0.32	0.752
Family Home	1	0.03901	0.02472	1.58	0.1148
Joint Ownership	1	0.07178	0.01526	4.7	<.0001
Solo Ownership	1	0.14257	0.02228	6.4	<.0001
Female	1	-0.02405	0.01195	-2.01	0.0442
In collections <6 months	1	-0.19734	0.02167	-9.11	<.0001
6<collections <12 months	1	-0.15422	0.02015	-7.65	<.0001
No. of Telephones	1	0.00097223	0.00948	0.1	0.9183

Table 4.14, Linear regression results (2nd stage) for extra large debts

Table 4.14 gives the results for the linear regression for extra large debts to estimate how much of their debt they would repay if they repaid part of the debt. As the table shows only the variables of debt values, female and time in collections are significant. In regards to age the results show that if the debtor is younger then they are more likely to pay back more of the debt. In fact those over the age of 35 had a negative effect on the recovery rate estimate.

Having active telephones have a positive effect on the recovery rate but none of the results are significant. Only work phone numbers have a negative effect on recovery rate estimates. There is little change between the variable coefficients for home ownership, sex and length of time in collections as the results to table 4.13.

4.3 Prediction

The two-stage model was used to predict the recovery rate of the debts. All of the debtors were split into the groups and then divided into test set and training set. The training sets were used to form the models and then the test sets were used to test the models. The following results are based on the test sets.

The test sets' variables were multiplied by the logistic regression model coefficients (for all debtors not just those older than 6-months). Selecting the logit value at which to cut off the payer from the non-payers depends on how the model is to be used. Trying to predict the value of a group of debts means using a cut-off, which ensures the higher percentage of debtors, are correctly classified. Figure 4.1 shows the effects of applying different cut-off values to the logistic regression on the small debts. Figure 4.2, 4.3 and 4.4 shows the effects of applying different cut-off values to the logit from the logistic regression on the medium, large and extra large debts respectively.

Figures 4.1 to 4.4 show the effects of different logit cut-off values in each of the sets of debt. The x-axis shows the logit values, and the y-axis show how many debtors are correctly classified using each cut-off. The blue series indicates the percentage of non-payers that are correctly classified. The red series indicates the percentage of payers that are correctly classified and the green series indicates the total percentage of debtors which are correctly classified.

As would be expected the non-payers correctly assessed increases as the cut-off increases where as the number of payers correctly assessed falls. At a cut-off of 0.4 the highest number of debtors is correctly gauged ~83% since the number of non-payers is greater than the number of payers. Therefore small increases in the number of none-payers correctly assessed has a proportionate effect on the number of debtors correctly assessed but has a large effect on the number of payers correctly assessed.

Using figure 4.1 to figure 4.4s' results indicates that a cut-off of 0.4 would be best because that gives the highest percentage of debtors correctly classified for all four groups.

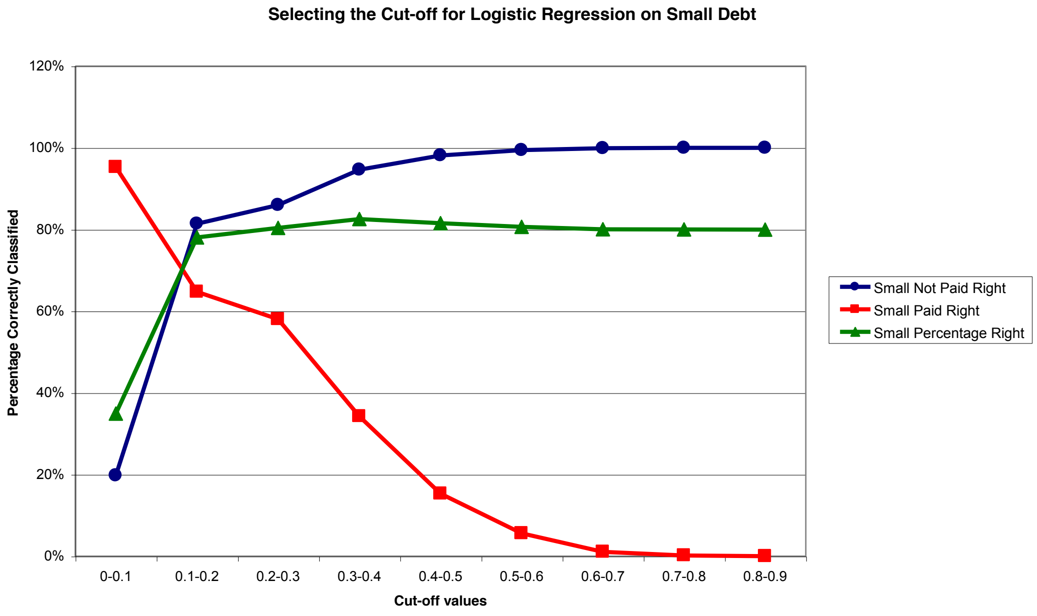


Figure 4.1, Effects of logistic cut-off values on small debts

However since in all groups the non-payers outnumber the payers, many payers are incorrectly classified. Therefore as a model to estimate which debtors will be likely to pay and assess their recovery rate a cut-off of 0.2 would be far more useful. Since at 0.2 approximately 70% of the payers and non-payers were correctly assessed. After 0.2 the proportion of payers correctly assessed fell significantly.

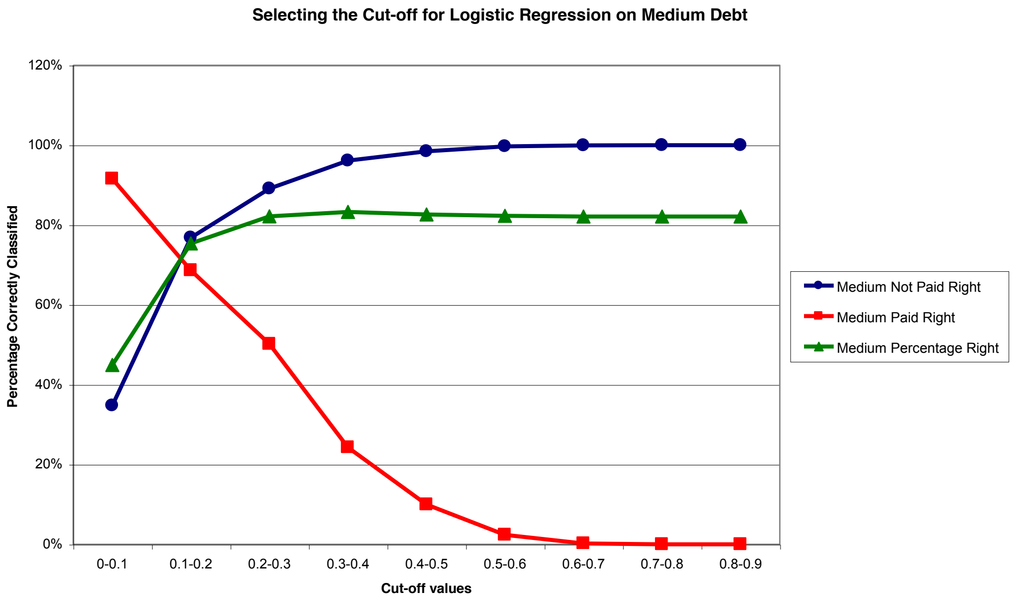


Figure 4.2, Effects of logistic cut-off values on medium debts

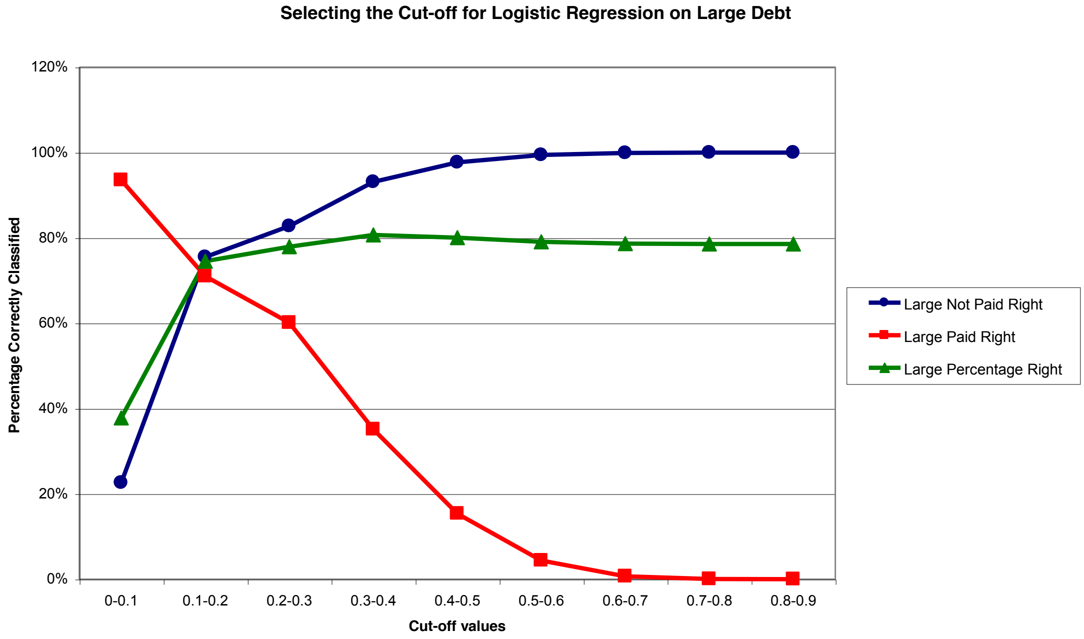


Figure 4.3, Effects of logistic cut-off values on large debts

Hence any debtor with an estimated result greater than 0.2 were assumed to have paid back part of their debt, and therefore passed on to stage 2. The linear regression model was used to estimate the collect rate. The debtor's variables were multiplied by their respective coefficients.

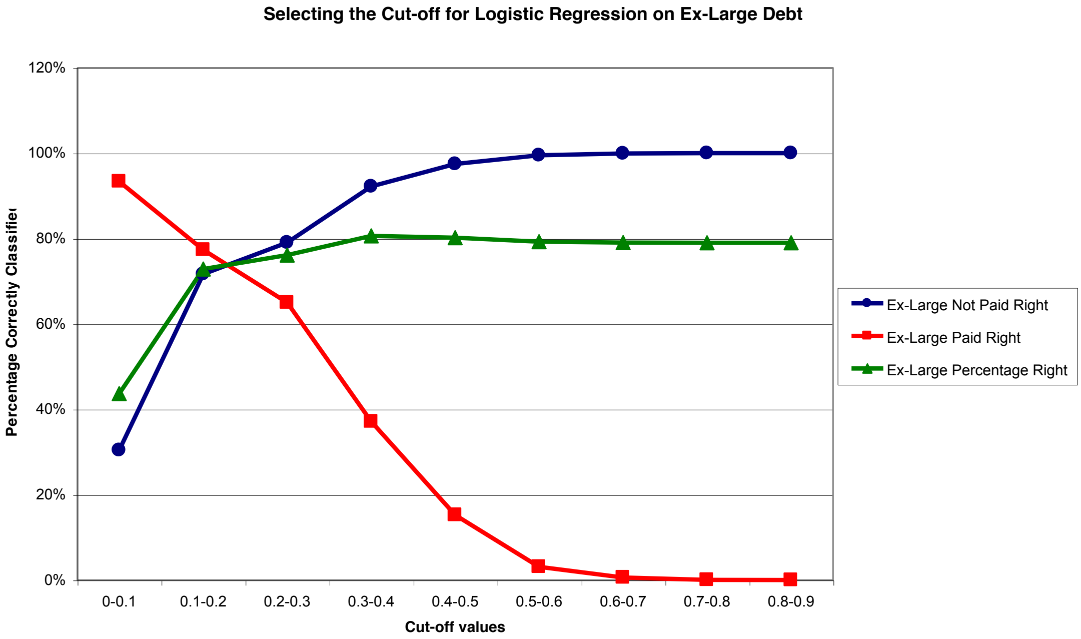


Figure 4.4, Effects of logistic cut-off values on extra large debts

4.3.1 Small Debts

The logistic regression model to predict if the debtor has a collection rate>0 for small debts is based on table 4.3:

$$\alpha = 0.21026 - 0.03909A_{25} - 0.05269A_{35} - 0.04881A_{45} - 0.03106A_{55} - 0.00008906D$$
$$+ 0.8355T_1 + 0.03152T_2 + 0.09285T_3 + 0.10187T_4 + 0.4729S + 0.04903H_F + 0.0496H_J$$
$$- 0.017H_S - 0.2014M_6 - 0.0276M_{12}$$

Where

A_{25}= 1 if the debtor aged between 18-25, 0 otherwise

A_{35}= 1 if the debtor aged between 25-35, 0 otherwise

A_{45}= 1 if the debtor aged between 35-45, 0 otherwise

A_{55}= 1 if the debtor aged between 45-55, 0 otherwise

D= amount of debt owed (£)

T_1= 1 if the collector had one or more active telephone numbers for the debtor, 0 otherwise

T_2= 1 if the collector had an active mobile number for the debtor, 0 otherwise

T_3= 1 if the collector had an active work number for the debtor, 0 otherwise

T_4= number of active telephone numbers the collector had for the debtor

S= 1 if the debtor is female, 0 otherwise

H_F= 1 if the debtor reside in a residence owned by a family member, 0 otherwise

H_J= 1 if the debtor reside in a residence owned jointly by them and another, 0 otherwise

H_S= 1 if the debtor reside in a residence owned by them alone, 0 otherwise

M_6= 1 if the collector has had the debt for less than 6 months, 0 otherwise

M_{12}= 1 if the collector has had the debt for between 6 and 12 months, 0 otherwise

If $\alpha<0.2$ then the debtor is predicted to have a collection rate=0. If $\alpha\geq0.2$ then the debtor's Recovery Rate (RR) is calculated using the results from the linear regression model table 4.11 is as follows:

$$RR = 1.01424 + 0.06053A_{25} + 0.03293A_{35} + 0.0294A_{45} - 0.02919A_{55} - 0.00074813D$$
$$- 0.0542T_1 - 0.01577T_2 - 0.09325T_3 + 0.02816T_4 - 0.0181S + 0.04227H_F + 0.08457H_J$$
$$+ 0.13152H_S - 0.33971M_6 - 0.16402M_{12}$$

Up till now R^2 has been used to determine the goodness of fit for the models. However since this is a two stage model not just a linear regression model that is not as useful as it is for the individual stages. Also the data limitations mean that the results in this chapter will be more useful in collections policy to determine who the best debtors to prioritise are, not as a prediction tool to estimate returns. An alternative to R^2 is the Spearman Rank Correlation, which is a non-parametric measure of correlation. The real collection rate observed results and the predicted collection rate results are converted to ranks, and the differences d_i between the ranks of each observation and prediction are calculated. So the Spearman Rank Correlation is useful in describing how good the predicted ranks are.

The debtor's predicted rank was based on their predicted collection rate result; descending. The debtor's real observed rank was based on their observed collection rate result; descending. The differences d_i between their real observed rank and predicted rank are used to calculate the Spearman Rank Correlation coefficient. However there are many tied ranks (share the same rank) since there are several debtors with an observed or predicted collection rate of 1. When a rank is tied; all associated ranks are assigned the mean of the tied ranks. Tied ranks also mean that the classic Pearson's correlation coefficient has to be used instead of the abbreviated Spearman Rank Correlation coefficient.

$$\rho = \frac{n\left(\sum x_i y_i\right) - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{n\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}\sqrt{n\left(\sum y_i^2\right) - \left(\sum y_i\right)^2}}$$

Where

P = Pearson's correlation coefficient

x_i = real observation rank

y_i = predicted rank

n = sample size

The results for the small debts give a Spearman Rank Correlation coefficient of 0.39, where 0 would indicate no correction between the modelled collection rate and the real collection rate and 1 would indicate perfect correlation.

Small	Real		
Predicted		Paid	Not Paid
	Paid	10%	16%
	Not Paid	5%	68%

Table 4.15, confusion matrix for small debts (1st stage)

Table 4.15 shows the confusion matrix for the results of the model on small debts. As can be seen two thirds of the debts, which were paid, were correctly modelled, and 80% of the debts, which were not paid, were correctly classified. These results agree with the predicted results in figure 4.1.

Table 4.15 illustrates that 10% of the debts were predicted to be paid and were paid. 5% of the debts really had some payment made but were predicted to not be paid. 16% of the debts were predicted to be paid but were not. The majority of the debts, 68% of them, were correctly assessed to not have any payment made.

4.3.2 Medium Debts

The logistic regression model to predict if the debtor has a collection rate>0 for medium debts are based on table 4.4:

$$\alpha = 0.07576 - 0.06202A_{25} - 0.07533A_{35} - 0.06233A_{45} - 0.03374A_{55} + 0.00011518D \\ + 0.06397T_1 + 0.05173T_2 + 0.119T_3 + 0.08T_4 + 0.03811S + 0.06706H_F + 0.08397H_J \\ + 0.02745H_S - 0.17952M_6 - 0.05366M_{12}$$

If $\alpha<0.2$ then the debtor is predicted to have a collection rate=0. If $\alpha\geq0.2$ then the debtor's Recovery Rate (RR) is calculated using the results from the linear regression model table 4.12 is as follows:

$$RR = 0.69143 + 0.04769A_{25} + 0.06688A_{35} + 0.02947A_{45} + 0.02179A_{55} - 0.00024271D \\ - 0.04237T_1 + 0.02553T_2 - 0.03703T_3 + 0.00664T_4 - 0.04278S + 0.06866H_F + 0.0427H_J \\ + 0.14391H_S - 0.31173M_6 - 0.22429M_{12}$$

The results for the medium debts give a Spearman Rank Correlation coefficient of 0.38, where 0 would indicate no correction between the modelled collection rate and the real collection rate and 1 would indicate perfect correlation.

Medium	Real		
Predicted		Paid	Not Paid
	Paid	10%	18%
	Not Paid	4%	67%

Table 4.16, confusion matrix for medium debts (1st stage)

Table 4.16 is the confusion matrix for the results of the model on medium debts. As can be seen 70% of the debts, which were paid, were correctly modelled, and nearly 80% of the debts, which were not paid, were correctly classified. These results agree with the predicted results in figure 4.2.

Table 4.16 illustrates that 10% of the debts were predicted to be paid and were paid. 4% of the debts really had some payment made but were predicted to not be paid. 18% of the debts were predicted to be paid but were not. The majority of the debts, 67% of them, were correctly assessed to not have any payment made.

4.3.3 Large Debts

The logistic regression model to predict if the debtor has a collection rate>0 for large debts is based on table 4.5:

$$\alpha = 0.11539 - 0.05089A_{25} - 0.06666A_{35} - 0.06195A_{45} - 0.033444A_{55} + 0.00003502D$$
$$+ 0.08168T_1 + 0.03658T_2 + 0.09993T_3 + 0.08295T_4 + 0.04569S + 0.05449H_F + 0.06662H_J$$
$$+ 0.01651H_S - 0.16367M_6 - 0.01175M_{12}$$

If α<0.2 then the debtor is predicted to have a collection rate=0. If α≥0.2 then the debtor's Recovery Rate (RR) is calculated using the results from the linear regression model table 4.13 is as follows:

$$RR = 0.49633 + 0.11446A_{25} + 0.0592A_{35} + 0.08124A_{45} + 0.06546A_{55} - 0.0001029D$$
$$- 0.01625T_1 + 0.01692T_2 - 0.01859T_3 + 0.00198T_4 - 0.02818S + 0.0053H_F + 0.09187H_J$$
$$+ 0.10598H_S - 0.26528M_6 - 0.18705M_{12}$$

The results for the large debts give a Spearman Rank Correlation coefficient of 0.38, where 0 would indicate no correction between the modelled collection rate and the real collection rate and 1 would indicate perfect correlation.

Large	Real		
Predicted		Paid	Not Paid
	Paid	13%	22%
	Not Paid	5%	60%

Table 4.17, confusion matrix for large debts (1st stage)

Table 4.17 shows the confusion matrix for the results of the model on large debts. As can be seen 70% of the debts, which were paid, were correctly modelled, and over 70% of the debts, which were not paid, were correctly classified. These results agree with the predicted results in figure 4.3.

4.3.4 Extra Large Debts

The logistic regression model to predict if the debtor has a collection rate>0 for extra large debts is based on table 4.6:

$$\alpha = 0.11302 - 0.02759A_{25} - 0.04668A_{35} - 0.02812A_{45} - 0.022A_{55} - 0.000003D$$
$$+ 0.0837T_1 + 0.06257T_2 + 0.06641T_3 + 0.05599T_4 + 0.04457S + 0.03507H_F + 0.14564H_J$$
$$+ 0.0371H_S - 0.10901M_6 + 0.04758M_{12}$$

If $\alpha<0.2$ then the debtor is predicted to have a collection rate=0. If $\alpha \geq 0.2$ then the debtor's Recovery Rate (RR) is calculated using the results from the linear regression model table 4.14 is as follows:

$$RR = 0.33591 + 0.02084A_{25} + 0.01607A_{35} - 0.00099432A_{45} - 0.00767A_{55} - 0.00001379D$$
$$+ 0.00211T_1 + 0.0145T_2 - 0.00547T_3 + 0.00097223T_4 - 0.02405S + 0.03901H_F$$
$$+ 0.07178H_J + 0.14257H_S - 0.19734M_6 - 0.15422M_{12}$$

The results for the large debts give a Spearman Rank Correlation coefficient of 0.33, where 0 would indicate no correction between the modelled collection rate and the real collection rate and 1 would indicate perfect correlation.

Ex-Large	Real		
Predicted		Paid	Not Paid
	Paid	13%	26%
	Not Paid	5%	56%

Table 4.18, confusion matrix for extra large debts (1st stage)

Table 4.18 shows the confusion matrix for the results of the model on extra large debts. As can be seen 70% of the debts, which were paid, were correctly modelled, and under 70% of the debts, which were not paid, were correctly classified. These results agree with the predicted results in figure 4.4.

4.3.5 All Debts

The results from the logistic and linear regression models must be combined to predict recovery rate. The results for the holdout sample, where 0.2 was used for the logit in the logistic regression model, are as follows:

For small debts model the Spearman Rank Correlation gave a result of 0.39, while the R^2 value was 0.09 and the root MSE was 0.26.

For medium debts model the Spearman Rank Correlation gave a result of 0.38, while the R^2 value was 0.08 and the root MSE was 0.18.

For large debts model the Spearman Rank Correlation gave a result of 0.38, while the R^2 value was 0.05 and the root MSE was 0.18.

For extra large debts model the Spearman Rank Correlation gave a result of 0.33, while the R^2 value was 0.04 and the root MSE was 0.14.

These results show that all of the models are not very good at predicting the returns from the debt. This is partly because of the limitations of the data, and also because predicting accurately what individuals will do results in models with poor R^2. On the other hand the root MSE does improve for larger debts. This is not because the models are improving but because the range of the recovery rates reduces for larger debts and the models reflect this. The holdout sample results are shown in figure 4.5 and the reducing ranges of the recovery rates.

Figure 4.5 illustrates the results of the two-stage model for all debts. The predicted RR is on the y-axis and the real observed RR is recorded along the x-axis. As can be seen, most of the debts had a RR of 0 indicating that no money was recovered. Some of the debts also were predicted to have an RR less than 0, this is a result of the model and not an indication that some of the debts were predicted to incur greater costs than the amounts recovered, since costs were not included in the model. However these results are not displayed in the graph. The 2nd stage of the model was linear regression, which could return a negative recovery rate. This could be fixed so that all results are between 0 and 1, but there were only a few cases (<0.1%) that fell outside this range, so the model was left as is.

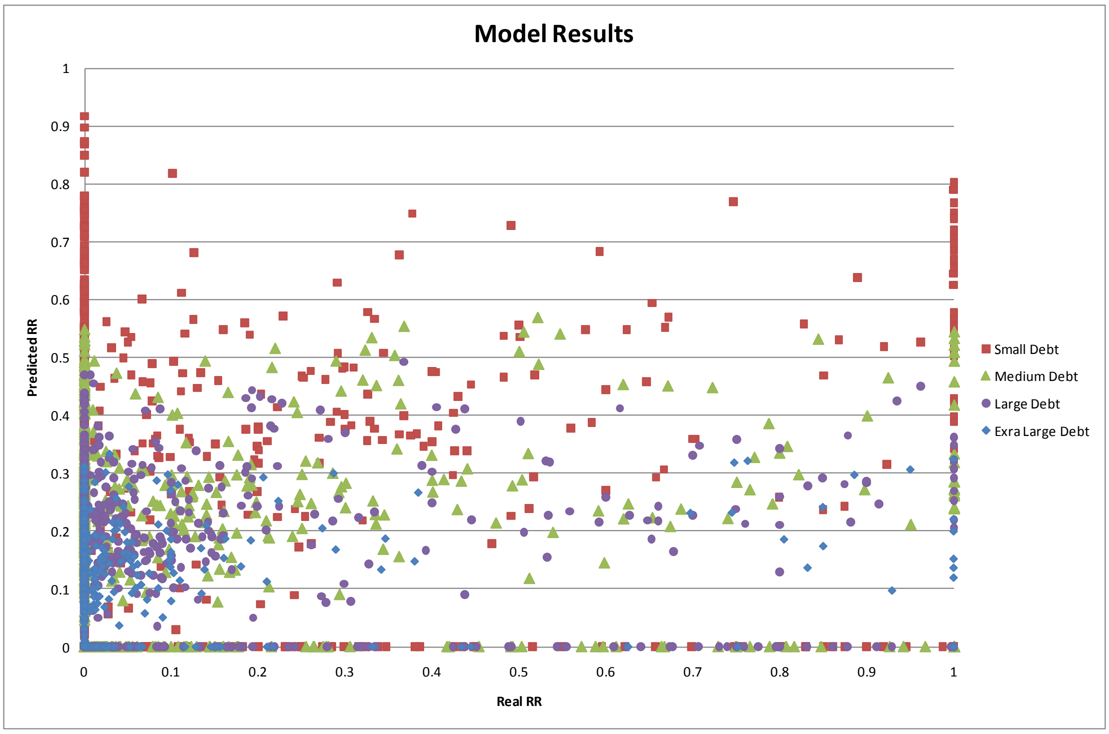


Figure 4.5, Results of the 2-stage model

What is most striking about this graph is that the predicted recovery rate reduces as the debt model's amount increases. For small debt the largest predicted RR is almost 1, but for medium debts the largest predicted RR is 0.6. For large debt the largest prediction is 0.5 and for extra large debt the largest prediction is under 0.4. This is consistent with the real RR results where for small debts 4% completely pay off their debt achieving an RR of one, compared with only 0.6% of extra large debts achieving an RR of one.

Looking at the overall model (for all the debt values) using the two-stage modelling approach to estimate the recovery rate, the R^2 value was 0.08, with a root MSE of 0.20 for the holdout sample. While the models do not give a good recovery rate prediction, they are useful for collections policy to predict who to prioritise and the Spearman Ranks reflect this.

4.4 Summary

This chapter focused on predicting the recovery rate for third party collection over the 20-month time period. By splitting the debtors according to the amount of debt they owe the results of the models were far better than modelling the debtors as a whole. Only predicting the RR for extra large debtors gave a poorer result than the linear regression model in chapter 3. The models for small and medium sized debt even managed to improve on the weight of evidence model.

The model created was a two-stage RR predictor, using logistic regression to predict which debtors would have a RR=0 and which would pay back part of their debt. Those debtors, who achieved a result of 0.2 for their logit and above in the logistic regression model would then, use the linear regression model to predict their RR value; the others would have a predicted RR of 0. Splitting at 0.2 meant that about 70% of debtors who paid and about 70% of debtors who did not pay, were correctly classified.

Waiting until after the debtors had been in collections for at least 6-months gave better results for the logistic regression. That is not to say that the models should only be used after 6-months but rather these models are for predicting the recovery rate after at least 6-months in collections. The results of these models were shown using the Spearman rank correlation, which shows that the model for small debts was the best predictor.

For larger debts their predicted RR was lower than for debts in smaller debt amount models. So for debts larger than £2000 none were predicted to pay back more than 35% of their debt.

Chapter 5: Forward Predicting and Economic Variables

The main objective of this chapter is to show how economic variables effect the LGD predictions. To this end, this chapter will discuss the data set in more detail, including how the default date was determined and cleaning of the data set. Then there is a discussion of the different economic variables and how they changed during the data set's time period. These economic variables are used in models to predict if the debtor's RR=0 and in models to predict the Recovery Rate of debtors at 12-month intervals after default.

The data used in this chapter is from the in-house data set used in chapter 3. The data set is from a UK bank's personal loans book, which defaulted between 1988 and 1999. The lifetime of the loan was recorded between the ends of 1987 to 2003. The data set was very large and disorganised and so it had to be cleaned before it could be used for producing models. One of the problems was that if a debtor took out a loan and then increased the loan amount at a later date the new loan was entered with all of the same variables as the first and so the data could be copied up to four times in the data set. In order to eliminate this only every fifth loan was used to ensure there was no replication to bias the data.

5.1 Default Date

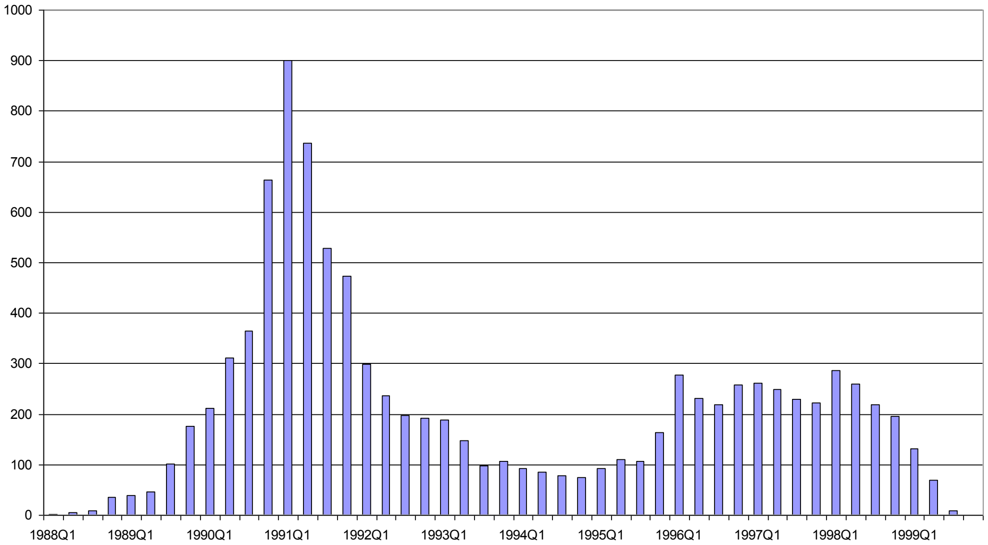


Figure 5.1 default dates

Figure 5.1 shows the number of debtors who defaulted in each quarter recorded in the data set. However getting the default date proved hard to establish as only the "last" default date was recorded. Once a debtor defaulted on the loan this data was recorded, should the debtor resolve this issue at a later date, i.e. pay back the lost arrears and carry on paying the debt off then they were recorded as being cured. Once cured the debtor could then default again. This new default date over wrote the previous default date. Therefore the recorded default date could not be used as a lot of the debtors were recorded as cured on up to three separate occasions. Also there was no information included in the files on how default was determined, whether it was three or six months in arrears.

Therefore to ensure continuity three months in arrears was determined to be default for the purposes of these models. The default month could then be determined because the number of months that each debtor was in arrears was recorded for each month that the loan was outstanding.

Determining how much of the loan was outstanding when they were three months in arrears proved quite complicated. The issues were that after 2001 the outstanding balance was recorded every month but before that the balance was only recorded at the end of every year. So in order to determine how much was paid each month, the amount paid during the year (B_{i-1}-B_i) was divided by the number of months the person paid during the year (P_i). This way the approximate amount paid (a_i) each month, if there was a payment, could be determined for all debtors.

$$a_i = \frac{B_{i-1} - B_i}{P_i}$$

where i is the year, and B_i is the balance outstanding at the end of the year

Another issue to further complicate the matter was that payments were not recorded, either the amount paid or if any payment had been made. Although the number of months the debtor was in arrears was recorded for every month. Therefore this information was used to determine when a debtor paid, based on the number of months they were in arrears.

If the number of months went up then they were evidently not paying. If it went down, then they had paid. If it stayed the same then they were paying only if the amount still outstanding was greater than the number of months of payments still owing. E.g. if the debtor took out a loan for £1000 and agreed to pay £100 per month for ten months to clear the loan. If the debtor then made only one payment of £100, then their months in arrears would be as shown in figure 5.2. Since the debtor can only be a maximum of 9 months in arrears, once he has reached 9 months it will stay at 9 until he either starts to pay or the lender writes off his debt.

Month	1	2	3	4	5	6	7	8	9	10	11
Months in Arrears	0	1	2	3	4	5	6	7	8	9	9

Figure 5.2, Example of months in arrears

Since the maximum number of months that a debtor could be in arrears changes each time a payment is made, it turned a relatively simple problem into a time dependant problem, given that, the maximum number of months in arrears had to be recalculated on a monthly basis determined by the number of payments made. This was further complicated by the fact that once a debtor was in arrears they could pay two or more months worth of payments in one. For example if the debtor in the previous example after paying the first payment, stopped paying for two months then made a double payment in the fourth month but no further payments, this will reduce the maximum number of months in arrears down to seven although only two payments were made.

Month	1	2	3	4	5	6	7	8	9	10	11
Months in Arrears	0	1	2	1	2	3	4	5	6	7	7

Figure 5.3, Example 2 of months in arrears

Now there are several different solutions to this problem and in the above examples it can be seen that the maximum number of months in arrears is reached by the tenth month when the loan was due for full repayment. Therefore this could be used to determine the maximum number of months in arrears. However, there are two problems with this solution. Firstly, the debtor may well start to repay his debt again after the term of the loan has expired again altering the maximum after this date. Secondly the debtors in the case study rarely stopped at just one loan. As has been previously stated, they increased their loan amount on several occasions. So they would take out one loan, start to pay it back, then increase the loan. Since the records of intermediate loans were not included, only the final loan status, the convoluted payment patterns of 10,000 debtors proved difficult to unravel.

For this thesis the maximum number of months in arrears had to be recalculated every month where not only the number of payments had to be included but also double, triple, or larger payments. Once this was determined, a constant number of months in arrears could be correctly classified as a payment or not a payment.

5.2 Economic Indicators

During the period covered by the data, the UK went through a recession and recovery so many of the economic indicators changed radically over this period. Therefore the data is ideal for investigating how economic indicators may influence or predict payment patterns for defaulted personal loans.

Six such indicators of the economy are Consumer Price Index (CPI), Gross Domestic Product (GDP), Interest Rate, Halifax House Prices Index, unemployment and net lending which shall be used throughout this chapter. Figlewski et al [31] used 17 macroeconomic variables when modelling corporate default in the US. These included a consumer price index, GDP, two interest rates, unemployment and some credit variables relating to corporate finance. However because they had so many economic variables that were so closely related they found that many of them had correlations among their macro covariates and so had to eliminate several of them from the study.

Grieb et al [33] found that unemployment leads to a rise in credit card default rates, by looking at time series data to study consumer behaviour, macroeconomic factors, and credit card default between 1981 and 1999. Whitley et al [56] looked at time series in mortgage default rates. They too found that unemployment was related to default rates but in mortgages. Their results showed that the proportion of mortgage loans in at least 6 months arrears were related to mortgage income gearing, unemployment, and loan to value ratio for first time buyers.

Banasik & Crook [6] found that default rates on consumer loans were positively correlated with real disposable income. Their results indicated a relationship between delinquent consumer credit and volume of debt outstanding, optimism and interest rates. They deduced that when people are more optimistic and may intend to borrow in the future they are more careful with their repayments.

Bellotti et al [14] used three economic variables for modelling Loss Given Default (LGD) for retail credit; interest rate, unemployment and earnings for data over the period 1999-2006. The macroeconomic variables were based on their values at time of default.

5.2.1 CPI

The Consumer Price index (CPI) measures the changing prices of a “basket” of goods and services over time within the UK. It is used to estimate the average price of these goods purchased by a household. The percentage change in the CPI is an estimate of inflation. [43]

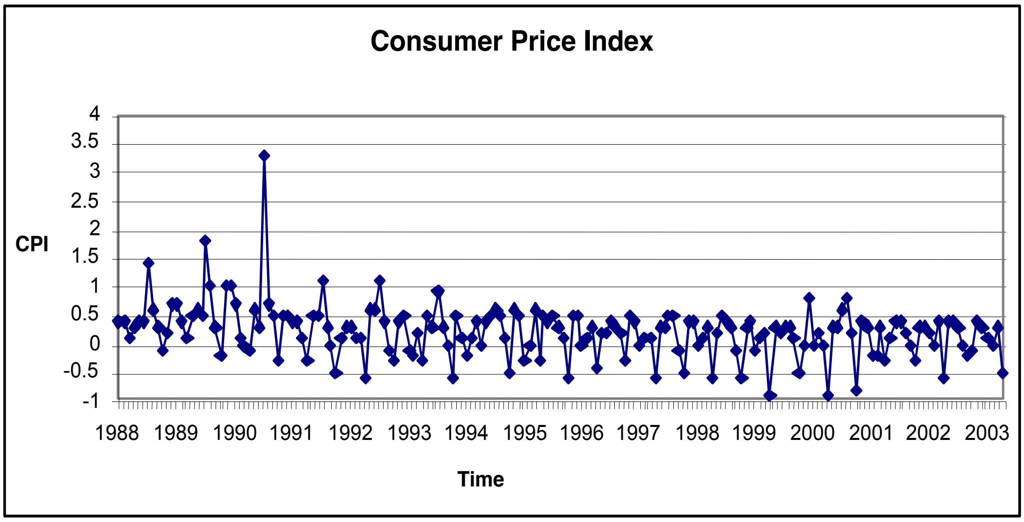


Figure 5.4 percentage change in CPI between 1988 and 2004

Figure 5.4 shows the percentage change in CPI over the period the loan data was collected varies modestly. Before the 1990-1992 recession the percentage change in CPI is higher than afterwards. The data for the percentage change in CPI was collected from the National Statistics Office. The seasonally adjusted CPI was not used because; it would firstly smooth the CPI, which has a small enough variation but also, the whole point in using the percentage change in CPI is to include a variable to show how the debtors' household expenditure affects their willingness to pay. If an individual defaults then it is likely that they are in financial difficulties therefore small changes in their household expenditure could mean the difference between making a payment or not. The seasonally adjusted CPI is useful for determining inflation but not the variation in people's expenditure, which is sought here.

Figlewski [31] used inflation monthly percentage change in the seasonally adjusted Consumer Price Index. They found that inflation was significant and had a positive correlation with corporate default indicating that a rise in inflation suggests a rise in corporate default.

5.2.2 GDP

Gross Domestic Product (GDP) is a basic measure of the country's overall economic output. It is the market value of all goods and services made within a country over one year.

Figure 5.5 shows the percentage change in GDP from the same month the previous year. In view of the fact that there is a recession, a recovery and a boom period during this time GDP varies dramatically. During the recession GDP becomes negative and then swings up to 5% during the recovery. The data for the GDP was collected from the National Statistics Office and uses a moving average to estimate GDP monthly, which was then taken as a percentage change from the same month in the previous year.

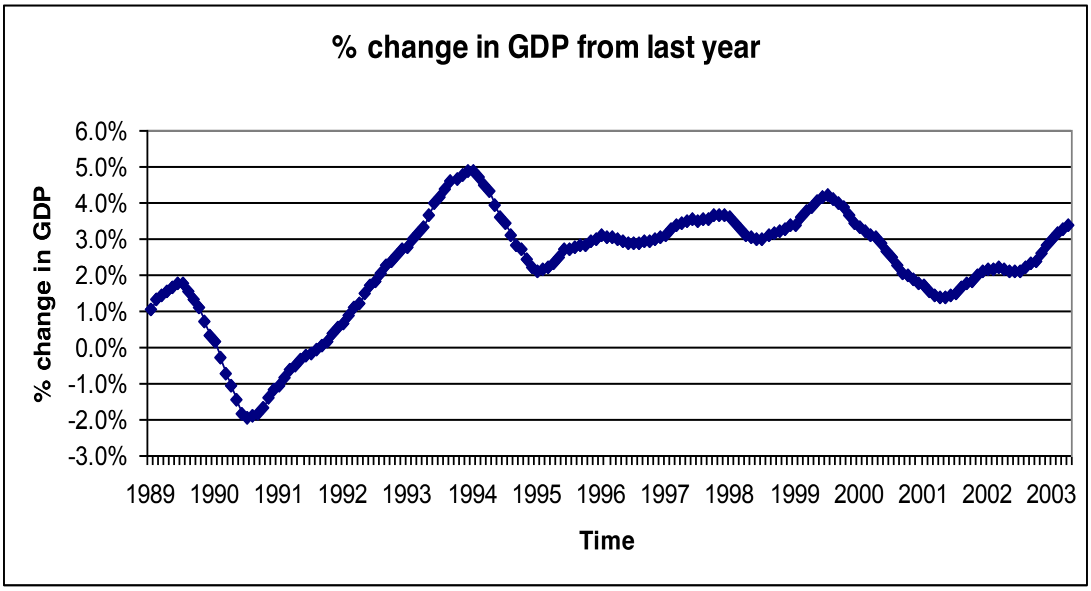


Figure 5.5 GDP between 1989 and 2004

Percentage change in GDP was used because unlike the level of GDP it shows the effect of the recession and recovery clearly whereas the level of GDP just shows a general rise so is really a surrogate for time, and other studies have found that it is insignificant. The percentage change on the previous year was used instead of percentage change on last quarter or from peak, because if a lender wishes to use these models for predicting future recoveries, they will not know what the peak is unlike historical models and any seasonal variation is removed, so you can judge how the economy is really faring.

GDP has been shown in some studies to have an effect on loan defaults. Sullivan found that a fall in GDP growth translates to a rise in default rates across all risk grades. [49] Figlewski [31] also uses “Real GDP actual minus potential” from the U.S. Department of Commerce, and “Real GDP growth”. Finding that “Real GDP growth” was significant however he also found that “Real GDP actual minus potential” was not significant, but both had a negative correlation with corporate default indicating that a rise in GDP suggests a decrease in corporate default.

Bellotti [14] used the UK earnings index (year 2000 = 100) for the whole of the economy including bonuses as a ratio of the retail price index in their models from the UK Office for National Statistics. They found that it was not significant but had a positive correlation with RR indicating that a rise in UK earnings at default predicts a rise in RR indicating that defaulting debtors will pay back more of their debt.

5.2.3 Interest Rate

The Bank of England Base Rate is the interest rate charged by the Bank of England for securing overnight lending. Figure 5.6 shows the fluctuations in the Bank of England Base Rate over the period of the data set. The interest ranges between 15% and 3.5%, a dramatic change. Before the recession the interest rate was higher than afterwards.

As interest rates rise there is a rise in default rates across all risk groups. [49] The Bank of England Base rate was used because many banks use this to determine their own interest rates, especially for variable rate lending.

Figlewski [31] used two variations of interest rates; both were significant and had a positive correlation with corporate default indicating that a rise in interest rates predicts a rise in corporate default.

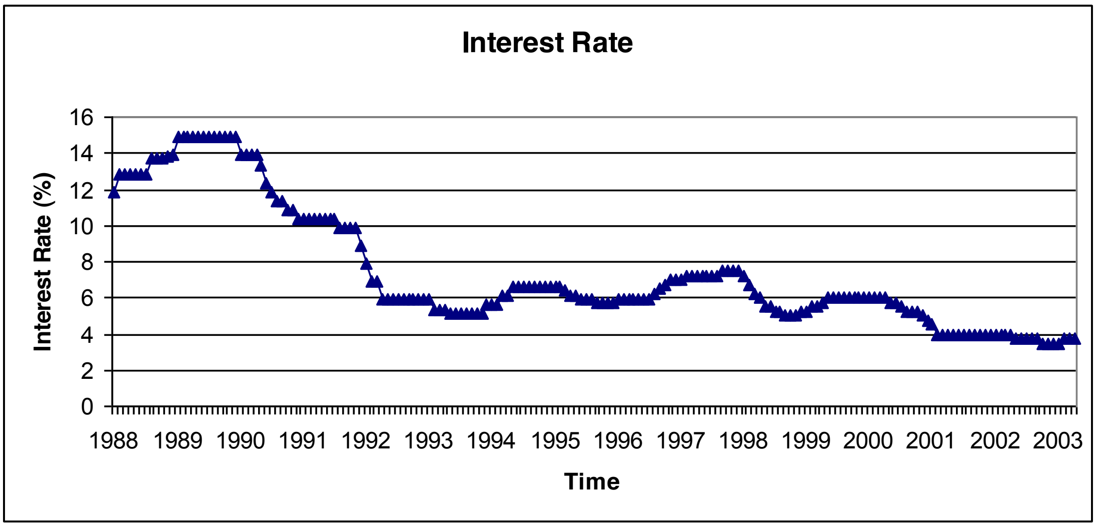


Figure 5.6 Interest Rate between 1988 and 2004

Bellotti [14] used the selected UK retail banks' interest rates in their models from the UK Office for National Statistics. They found that it had a significant and negative correlation with RR indicating that a rise in interest rates at default predicts a fall in RR indicating that defaulting debtors will pay back less of their debt.

5.2.4 Halifax House Price Index

House Price indices have been around in the UK since 1973, initially mortgage providers only collated them, although now government bodies also record them. The Halifax House Price Index was launched in 1984, based on the lending of the UK's largest mortgage lender. It provides the longest unbroken monthly data series in the UK. Therefore it is ideal for assessing changes to the UK's housing market over the time of the data set. Figure 5.7 shows the Halifax House Price Index between 1988 and 2004.

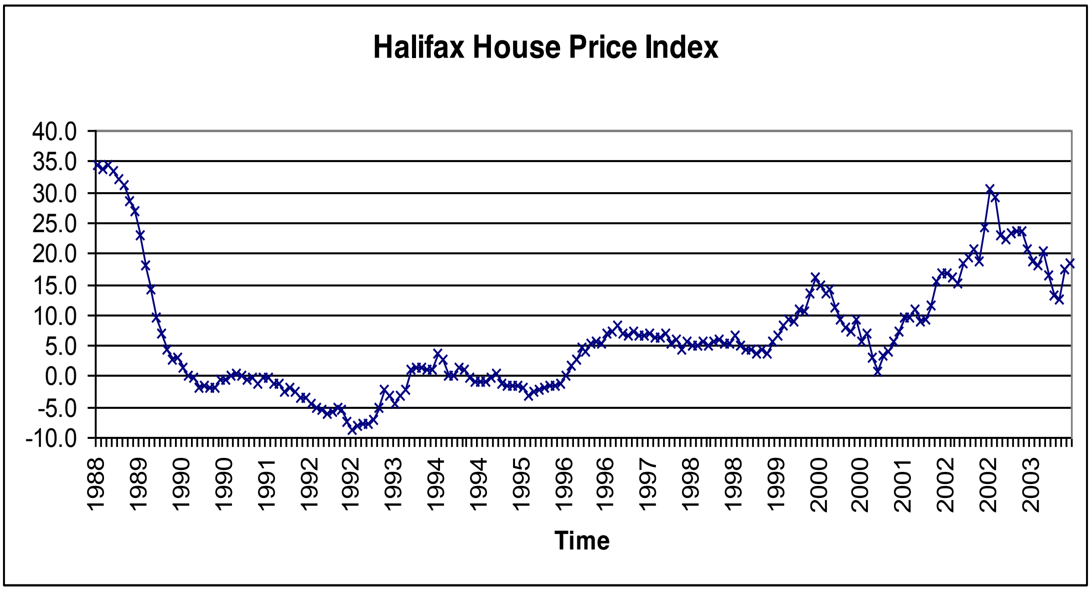


Figure 5.7 Halifax House Price Index between 1988 and 2004 (% change in house price index)

5.2.5 Unemployment

The definition of who are unemployed changes over this period so the unemployment figures for use in this thesis are based on the number of people in the UK not employed divided by the number of people economically active.

$$UnemploymentRate = \frac{(EconomicallyActive - Employed)}{EconomicallyActive}$$

Figure 5.8 shows how unemployment changes over this period. During the recession unemployment rose and then fell during the recovery.

Figlewski [31] examined both the unemployment level and change in the seasonally adjusted monthly civilian Unemployment Rate constructed by the US Bureau of Labour Statistics. However that paper did not use change in unemployment. Unemployment level was significant and had a positive correlation with corporate default indicating that a rise in unemployment suggests a rise in corporate default.

Bellotti [14] used the UK unemployment level measured in thousands of adults (16+) unemployed from the National Statistics Office. They found that it had a significant and negative correlation with RR indicating that a rise in unemployment at default predicts a fall in RR indicating that defaulting debtors will pay back less of their debt.

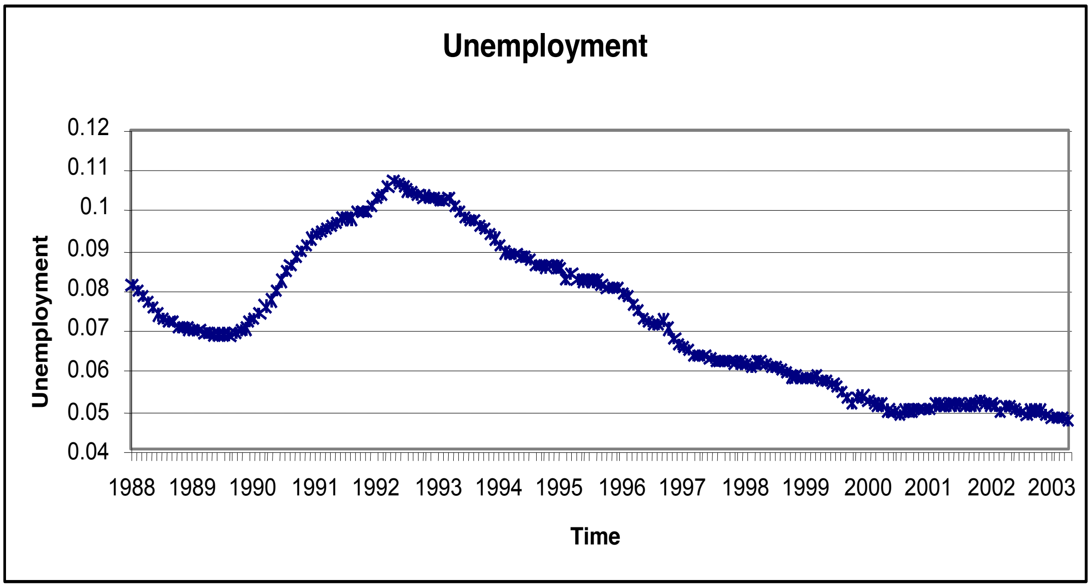


Figure 5.8 Unemployment between 1988 and 2004

5.2.6 Net Lending

Net lending is the total value of loans advanced in the UK less repayments and other adjustments such as written off bad debts. Figure 5.9 shows the net lending over the time period of the data set. As can be seen, net lending fell before the recession and then rose afterwards.

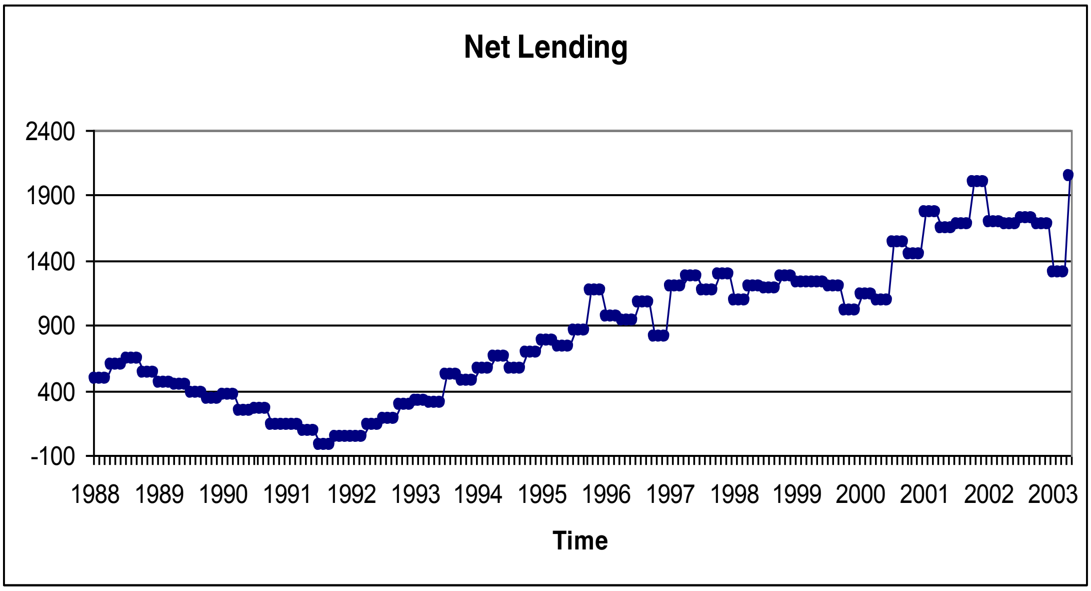


Figure 5.9 Net Lending between 1988 and 2004

5.3 Economic Variables

These six indicators of the economy are Consumer Price index (CPI), Gross Domestic Product (GDP), Interest Rate, Halifax House Prices Index, unemployment and net lending, were used in modelling debt recovery to see if economic variables helped estimate debt recovery rates.

Percentage change in CPI was selected because it estimates the changing cost of living for the debtors. Therefore if they had the same disposable income over time but their cost of living was rising, then they would have less income to spend on repaying their debt. On the other hand if their cost of living were falling then they would have more money to spend on repaying their debt.

GDP was selected as an indication of the UK’s income. If GDP is rising then the debtor’s standard of living might equally be rising since GDP is positively correlated with the standard of living. Therefore if GDP is rising then the debtor may have more money to spend on paying off their debt. So a positive correlation would be expected.

The Bank of England’s base interest rate was selected, because a lot of loans have interest rates tied to this measure or their initial interest rate is determined partially on this rate. Therefore if the Bank of England’s base rate is low, then a variable rate mortgage will be low also, therefore the borrower of this mortgage will be paying less each month and hence have more money to spend on paying back their other loans. Also a debtor could take out a new loan at a lower rate of interest to pay off any previous loans acquired at a higher rate of interest. Therefore a negative correlation could be expected between the interest rate and payments to loans.

The Halifax House Price Index was selected because, if a debtor has a house as an asset, then any increase in the value of this asset could enable them to take out larger loans (mortgages) secured against this asset to pay off other loans. Or the debtor may sell their house at a profit and use the profit to pay off their loans. Therefore increases to the house price index may lead to the debtor paying back more of their loan, provided they have a house. If they do not then it may have the opposite effect.

Unemployment was selected because it indicates the number of people unemployed in the UK. If a debtor becomes unemployed during the course of repaying their loan, the information is either unknown or not recorded within the data set available. Therefore this is the only indicator available for determining unemployment. As unemployment increases, then the probability of the debtor becoming unemployed increases too. If they are unemployed then, rising unemployment will make it harder for them to find new work since there are more people applying for the same jobs. Also if they are not unemployed themselves, rising unemployment means that there could be lots of workers, interested in their jobs, so employers are less likely to pay their current workers high pay rises, if they have lots of workers willing to do the job for less money. Therefore a negative correlation could be expected between the unemployment and payments to loans.

Net lending was selected to show partly how easy it is to acquire loans over this period, for if there are lots of loans being taken out, then it will be easier to acquire credit and therefore the debtors will have more money in their pockets. Alternatively when it is hard to get a loan, net lending will be low, this is shown during the 1991-92 recession. Therefore net lending could have a positive correlation with the debtors' ability to pay their loans.

5.4 Defaults Over Time

The data set covers almost 10,000 loans over nearly 16 years. These loans were first of all taken out before 1999, then they defaulted, and the debtor may start to pay back the loans after default. If they do, then they may pay off the debt entirely, or the debt may be written off if they fail to pay it. Some of the debtors were still trying to pay off their debt at the end of the time period. Figure 5.10 shows the number of defaulted debtors over time. It shows for any calendar time period the number of accounts in the state of default, and how these debtors are split up into paying (blue), paid off (green), written off (yellow) and not paying (red). This graph shows that the number of defaults rapidly increases during the recession and that the percentage of these debtors paying is very small. After the recovery the number of defaults becomes steady and the number of debtors paying increases. Over time the majority of the debtors get written off but some are also paid off. By 2004, 10% are still paying, nearly 18% have paid off and nearly 70% have been written off.

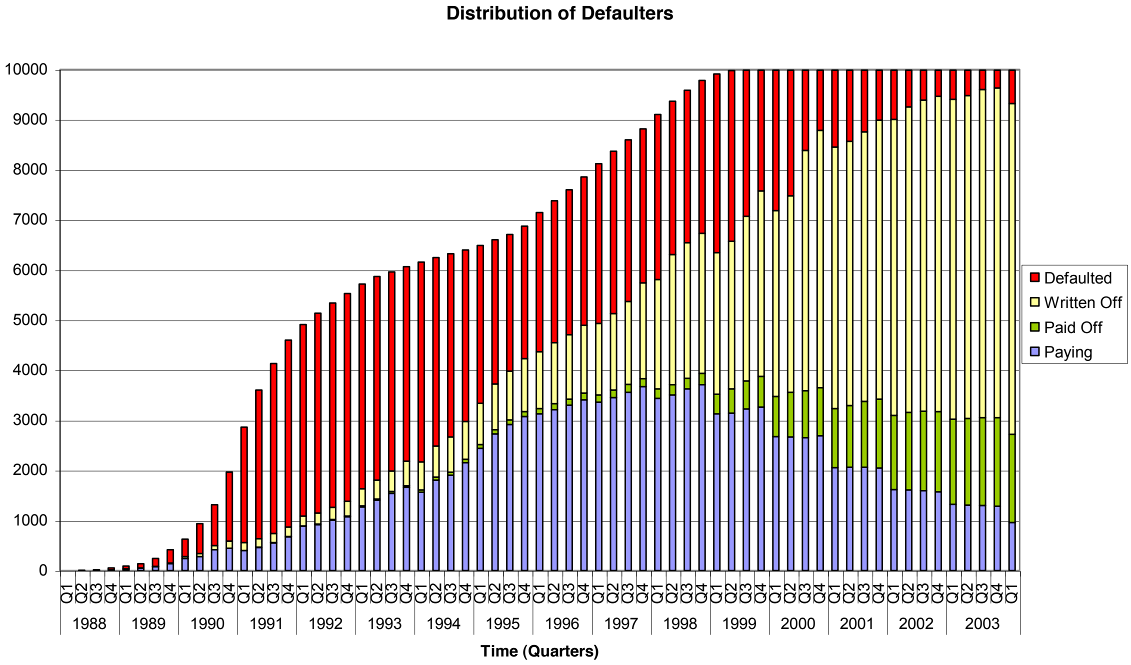


Figure 5.10 Distribution of debtors between 1988 and 2004

The results for comparing the distribution of debtors against economic variables are exploratory. If the percentage who are paying compared with those who are able to pay (whose who have defaulted and have not paid off or been written off) is viewed against all of the following economic variables the results are quite startling.

Figure 5.11 shows percentage change in CPI against the percentage that are paying after default. As can be seen there is almost no correlation between these two variables. Figure 5.13 shows that there is also a moderate correlation between interest rates and the percentage that are paying after default. However there is a negative linear relationship between the two. This is as expected since lower interest rates means that people have more disposable income and can afford to take out larger loans.

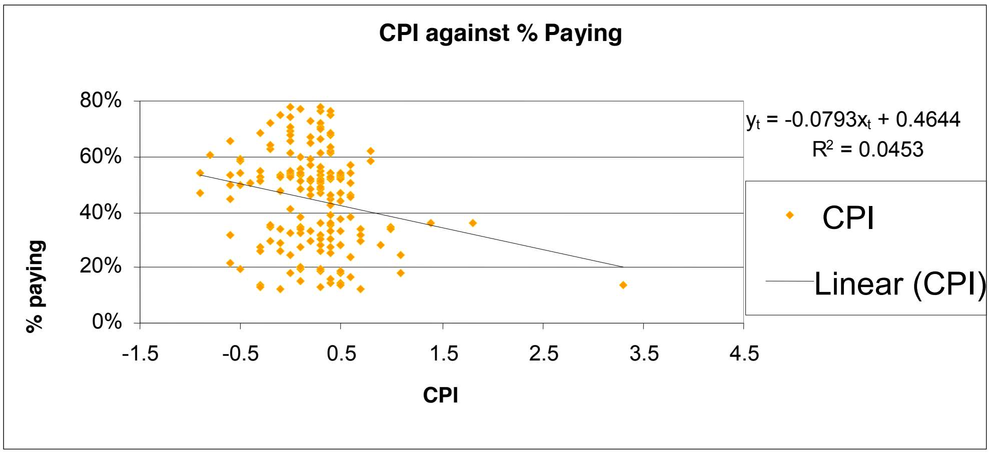


Figure 5.11 percentage change in CPI against the percentage who are paying after default

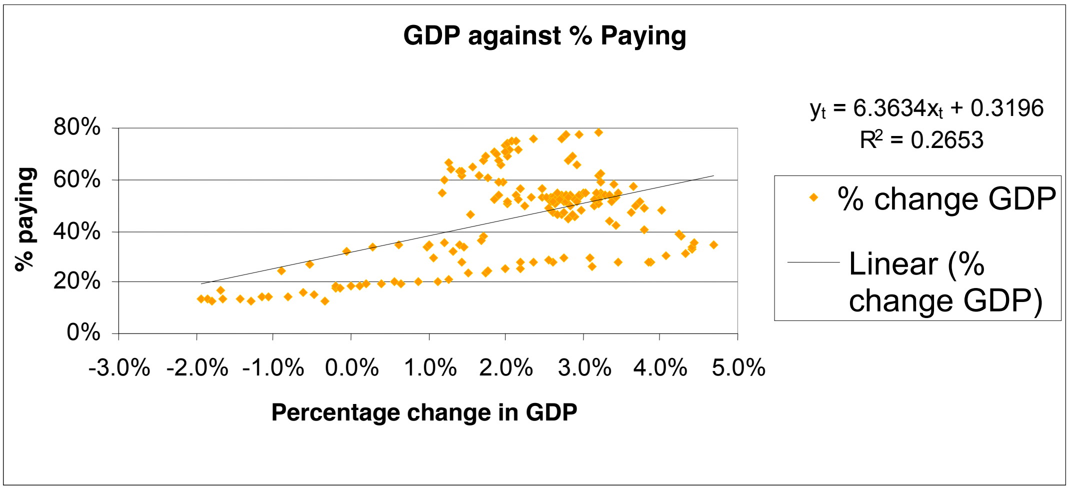


Figure 5.12 GDP against the percentage who are paying after default

Figures 5.12, 5.14 and 5.16 all show strong positive linear relationships between; the percentage of debtors who are paying after default; and GDP; Halifax house price index; and net lending, respectively. The relationship with GDP is as expected because, as GDP rises, so too does people's income indicating that they have more disposable income. The same is true of house prices and lending for if they have more money due to borrowing or the assets increasing in value then they can have more money to spend. If they have more money then they can afford to pay back their outstanding loans. House prices are less strongly correlated because not everyone will be affected equally as not all debtors will own a house.

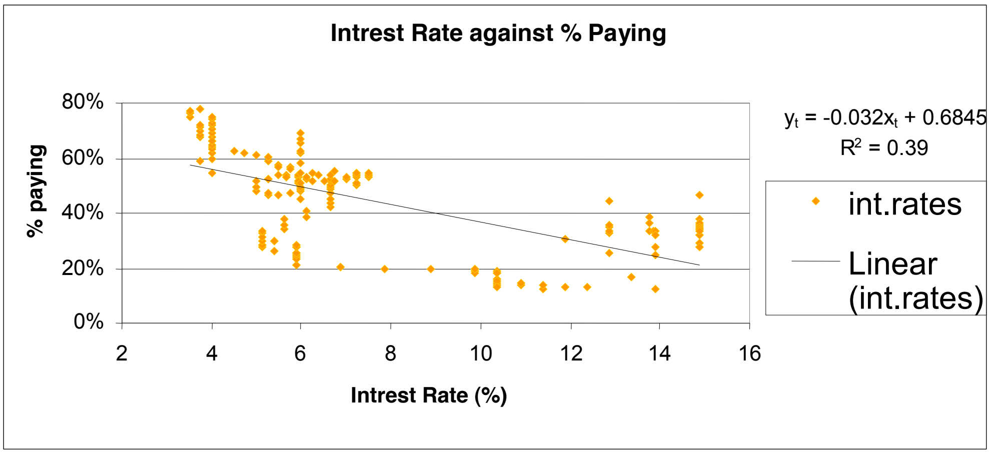


Figure 5.13 Interest Rates against the percentage who are paying after default

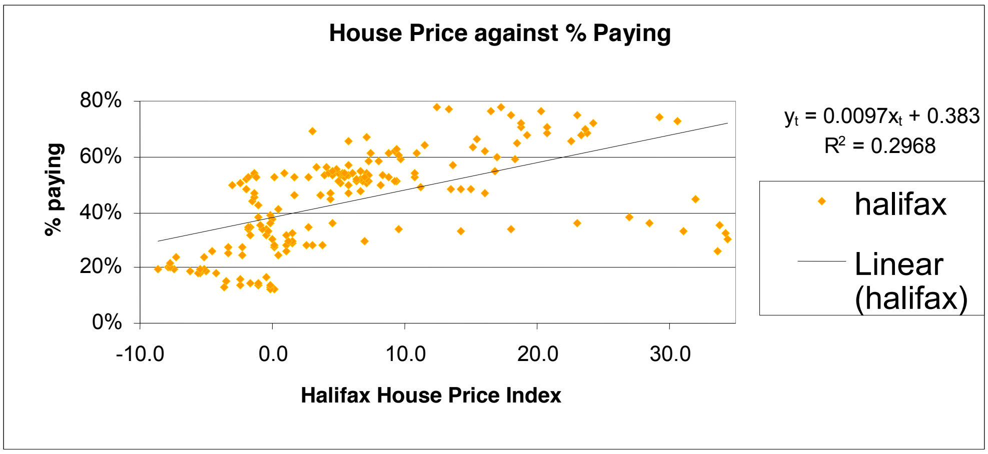


Figure 5.14 Halifax House Price Index against the percentage who are paying after default

Figure 5.15 shows a strong negative linear correlation between unemployment and the percentage paying after default. This result is as expected because, debtors who are unemployed will have less money to spend on paying off their debts, and those employed may receive lower pay rises due to the unemployed lowering wages.

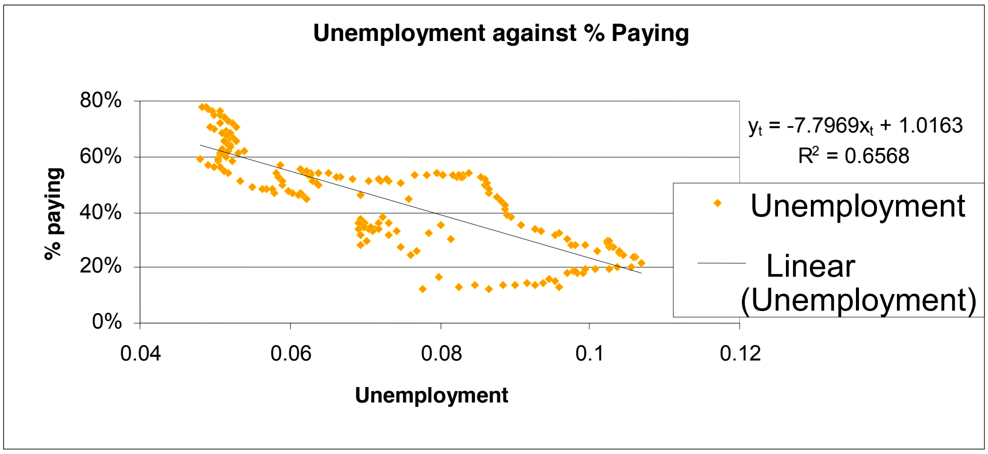


Figure 5.15 Unemployment against the percentage who are paying after default

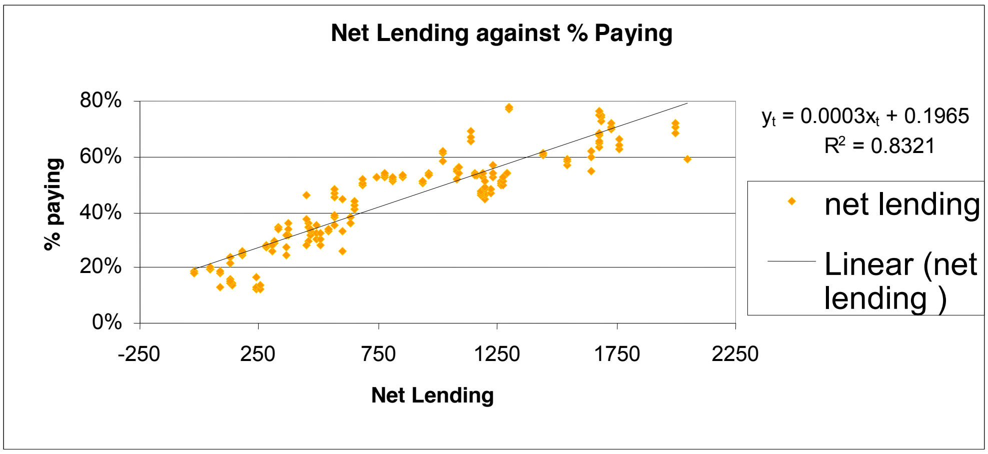


Figure 5.16 Net Lending against the percentage who are paying after default

These economic indicators can be used in a simple linear regression model to determine the percentage of defaulted debtors paying in any month. When using all of the discussed variables, GDP, percentage change in CPI and the Halifax house price index were all found to be insignificant. The model used the first 13years for training data. Then the last 12 months were held as a hold out sample. Interest rates had a Durbin Watson statistic of 0.007 and a t-statistic of -3.1. Unemployment had a Durbin Watson statistic of 0.003 and a t-statistic of -2.4. Net Lending had a Durbin Watson statistic of 0.028 and a t-statistic of 5.9. These Durbin Watson statistics show that there is evidence is positive serial correlation. The holdout sample gave an R^2=0.61 and a Mean Squared Error (MSE) =0.01.

Percentage Paying = 0.481 - 0.0085 Interest Rates -1.97 Unemployment + 0.0002 Net Lending

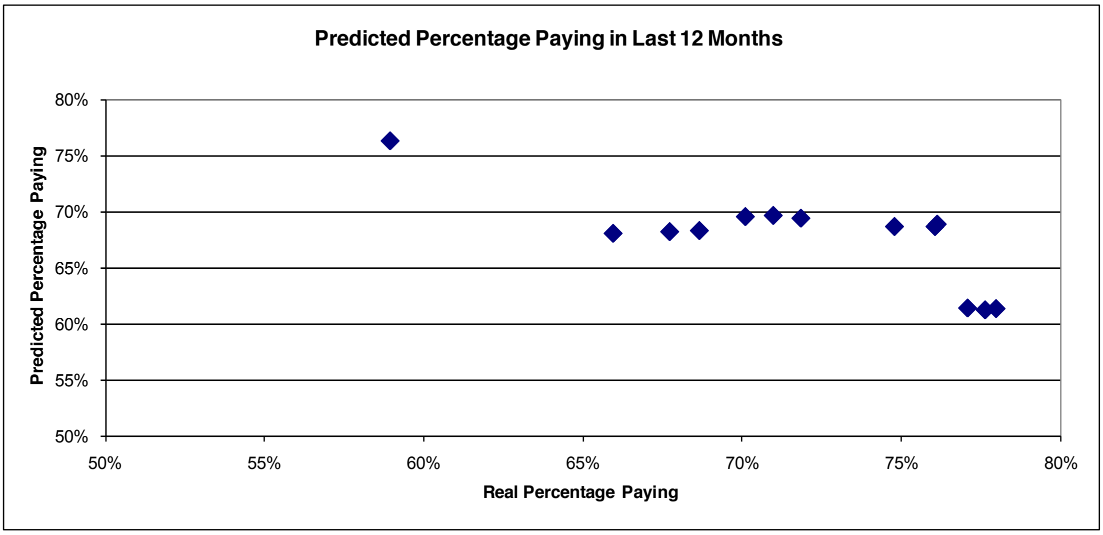


Figure 5.17 Predicting the percentage paying after default in the last 12 months using economic variables

Figure 5.17 shows the results of the regression model for predicting the percentage of defaulted debtors who are paying each month. The results displayed are based on the holdout sample, which shows that the model using all of the significant economic variables. The sample size for these models were not large but covered an interesting period of history of economic volatility. However the results are very close to those predicted by the model and show that the economic variables are very good at predicting the percentage of payers each month.

Predicting the percentage paying is the equivalent of the first stage in the two stage prediction models. Using the economic variable in a more detailed model will be discussed in chapter 6. The next section of this chapter examines models using economic variables to estimate not the final LGD but the LGD for the next 12-months. These models are useful for determining the short-term recovery rates of debtors once they have defaulted and during collections.

Economic variables seem very useful in predicting the percentage of payers, but what past literature has shown, is that they are not that successful in helping to identify who is going to pay or how much an individual will pay.

Because of this, the economic variables have not been added to the models discussed in chapters 3 and 4. Instead the next section looks at how the economic variables can be used to relate to repayments over the first 24 and 36-months of an individual's default. So the question is does knowledge of how the debt has been repaid plus the economic conditions at default give a good indication of how they will pay in the future.

5.5 Recovery Models

When a bad debt defaults the outstanding debt at this time is not the loss given default. More of the debt could be recovered both in-house and by other agencies. This chapter looks at the payment patterns of in-house collections after default has occurred on approximately 10,000 personal loans. The data set is the same as was used for in-house in chapter 3 however the whole data set is used not just the results for the first 2 years.

In previous models the focus was on predicting the final LGD, but when looking at whether to sell the debt or collect in-house; it might be useful to predict what will happen over shorter time periods. The next model is a simple linear regression based on what was collected in the first 12 months in-house to see what would happen in the second 12 months. These models estimate the recovery rate (RR) at 24 months and 36 months after default; RR_{24} and RR_{36} respectively.

$$RR_{24}=0.056+1.2RR_{12}$$

This model had an $R^2=0.58$ and a Root Mean Squared Error (MSE) =0.13.

Expanding the model to see what would happen in the 3rd year gave an $R^2=0.38$ and a Root MSE=0.20:

$$RR_{36}=0.11+1.23RR_{12}$$

Using the above models a lender can make more informed decisions about when to sell and how much to sell for. The reason these results are so superior to the previous models is because there is a dependence on both sides of the equation. RR_{24} and RR_{36} are dependent upon RR_{12} since they cannot be smaller than RR_{12} by definition. This artificially inflates the R^2 results.

These models can be rewritten to calculate the amount recovered in the second year only, the second and third year and the third year only. This eliminates the dependency on the first year's results:

$$RR_{24} - RR_{12} = 0.056 + 0.2RR_{12}$$

Rewriting the model this way to calculate the amount recovered in the second year only giving $R^2=0.05$ and a Root MSE=0.13. The model for estimating the amount recovered in the second and third year gives an $R^2=0.02$ and a Root MSE=0.2. This shows that during the second year the lender can expect to recover 11% of the default amount plus 23% of what was recovered in the first year.

$$RR_{36} - RR_{12} = 0.11 + 0.23RR_{12}$$

Since 5% of the default amount and 20% of what was recovered in the first year was recovered during the second year only another 6% of the default amount can be expected to be recovered during the third year.

As has already been shown in this chapter, the economic environment can have an impact on debtors paying back their debt. Belyaev et al [15] found that when modelling LGD for 12, 24 and 36 months, that some of the models were improved slightly by using economic variables especially linear regression models. The best estimate for the economic environment is to use a binary variable for the year. This means that all debtors who defaulted in similar economic circumstances are grouped together. Using the economic predictors, discussed earlier in this chapter, which are indicators of this economic period and would consequently give a poorer result. Therefore the regression models were recalculated to include the effect of their default year in the model.

$$RR_{24}=0.058 +1.2RR_{12} -0.05D_{88} -0.06D_{89} -0.03D_{90} -0.02D_{91}-0.02D_{92} -0.01D_{93} \\ +0.03D_{95} +0.03D_{96}+0.02D_{97}+0.02D_{98}$$

This model had an $R^2=0.59$ and a Root MSE=0.13. This is a small improvement on the previous model ($R^2=0.58$ to $R^2=0.59$). As this model shows, those who defaulted prior to the recession were estimated to have a poorer recovery rate during their first two years after default than those who default during and after the recession. Since this collection period covers part of the time Britain was in recession and those who defaulted after were collected during Britain's recovery, this is not surprising. This model shows that economic factors have a big impact on how much a lender can expect to collect. In this model a lender could expect to recover nearly 10% more of the debt during the first two years after default if they are collecting during an age of economic prosperity (1995) compared to a period of recession (1989).

The model above is for the first two years after default dependant upon the first year's recovery rate. If we look now at just the second year's recovery rate using the default year the model has an R^2=0.07 and a Root MSE=0.13. It is still a poor model but a definite improvement on the previous model without the default year (R^2=0.05 to 0.07).

$$RR_{24}-RR_{12}=0.058 +0.2RR_{12} -0.05D_{88} -0.06D_{89} -0.03D_{90} -0.02D_{91}-0.01D_{92} - 0.01D_{93} +0.03D_{95} +0.03D_{96}+0.02D_{97}+0.02D_{98}$$

Following on from this model to see what happens in the third year after default. This first model estimates the recovery rate for debtors during the first three years after default based on their recovery rate for the first year and their default year.

$$RR_{36}=0.09 +1.2RR_{12} -0.07D_{88} -0.05D_{89} -0.01D_{90} +0.01D_{93} +0.05D_{94} +0.09D_{95} +0.09D_{96}+0.06D_{97}+0.05D_{98}$$

This model had an R^2=0.4 and a Root MSE=0.2. This is again a small improvement on the previous model (R^2=0.38 to 0.40). Here the economic situation is having an even larger effect on the recovery rate. In this model a lender could expect to recover nearly 16% more of the debt during the first three years after default if they are collecting during an age of economic prosperity compared to a period of recession. Moving on to look at just the second and third year's recovery rate using the default year the model has an R^2=0.05 and a Root MSE=0.2. It is still a poor model but an obvious improvement on the previous model without the default year (R^2=0.02 to R^2=0.05).

$$RR_{36}-RR_{12}=0.09 +0.2RR_{12} -0.07D_{88} -0.05D_{89} -0.01D_{90} +0.01D_{93} +0.05D_{94} +0.09D_{95} +0.09D_{96}+0.06D_{97}+0.05D_{98}$$

Since the improvements were so small it is unproductive to consider modelling using the economic factors discussed earlier since any model would be worse than the models above. However there was an improvement to both models therefore the economic variables do have an effect on recovery rates. Another factor to consider is that the economic indicators will change over time, therefore the problem with regression models is that if only one value is used; should the default dates economic variables be used, or those at the end of the first year in recovery? Or maybe some combination of the two or even a prediction for the next year's economic trends.

5.6 Summary

This chapter has looked at the effects of economic factors in debtors repaying their loans after they have defaulted. The data set used was ideal for testing economic variables on recovery rates, since it covers the loans' history during a recession, recovery and a period of stability.

When looking at the percentage of defaulters who pay back their loans each month, the economic variables were excellent at predicting how many will pay back. In particular, net lending was a very strong indicator. Net lending, GDP and house prices all had a strong positive linear relationship between them and, the percentage of debtors who are paying each month after default. Higher interest rates and unemployment had a negative relationship with the percentage of debtors who are paying each month after default.

When it came to predicting the recovery rates after the first 12 months, the debtors' behaviour during those first 12 months is an indicator for the following 12 and 24 months. On average it appeared that debtors were repaying around 5% of the default balance off each year after the first year.

The results were disappointing in that even when employing dummy years, which are the best economic variables one can hope for, there is little or no improvement on the R^2 values.

In the next chapter when debtors pay and by how much will be covered in greater detail. What is evident is that during the lifetime of a loan, economic conditions can vary wildly, especially as some loans can have debtors repaying even a decade after they have defaulted. This means that using the economic conditions at a certain point, e.g. at default, is not as useful as continuous monitoring within the models. Therefore in the next chapter, survival analysis will be used to predict when debtors repay because the economic conditions for each month can be used to help the predictions.

Chapter 6: Payment Patterns

6.1 Introduction

This chapter looks at the payment patterns of in-house collections after default has occurred on approximately 10,000 personal loans. The data set is the same as was used for in-house in chapter 3 however the whole data set is used not just the results for the first 2 years. The payment patterns are use to estimate RR and LGD.

The data set for this chapter is from a UK bank's personal loans book, which defaulted between 1988 and 1999. The lifetime of the loan was recorded between the ends of 1987 to 2003. Default was taken to be three months in arrears.

6.2 Payment Patterns

When a debtor begins to pay back the debt they could stop the repayments at anytime. After they have stopped again they may restart, and this pattern will continue until the debtor either repays the whole loan or is written off.

Figure 6.1 shows some examples of the actual payment patterns where the red bars are when the debtor is not paying and the green bars are when the debtors are paying. As can be seen from this graph the debtors can go for long periods without paying and then start up again. All of these payment patterns are for after the debtor has defaulted. NP^i is the i^{th} non-payment sequence and P^i is the i^{th} payment sequence.

Some of the debtors never pay back anything more after default as for example Debtor 8 in figure 6.1. Some of the debtors pay back part of their debt but are written off when they stop repaying. Some of the debtors pay back all of their debt and others are still paying back at the end of the observation period.

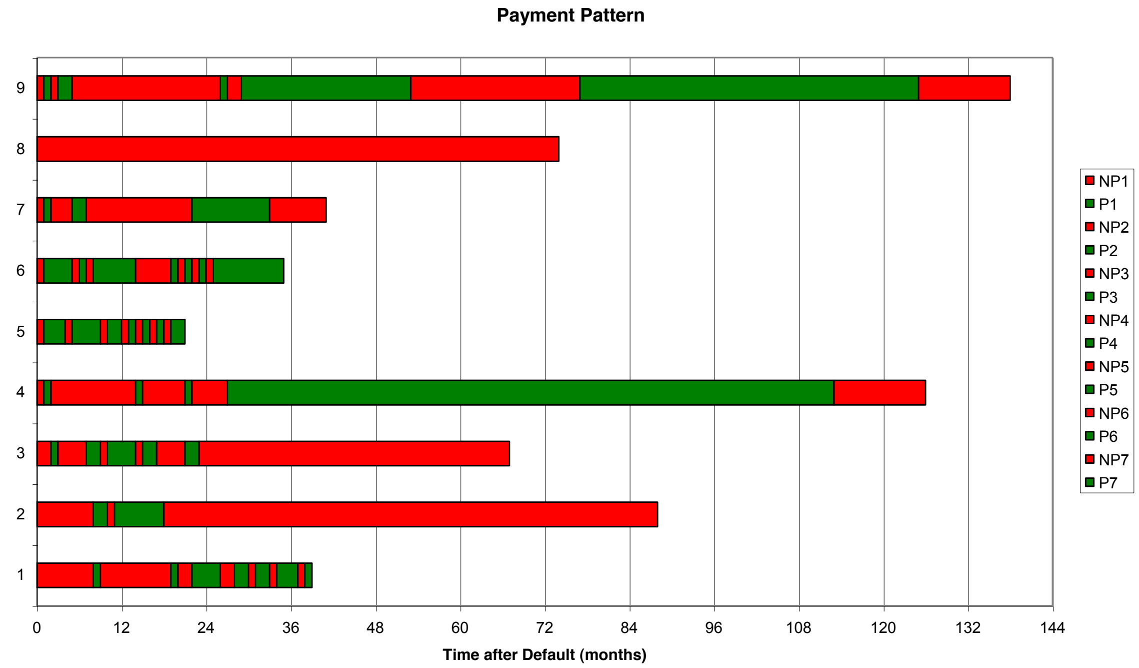


Figure 6.1 Payment Patterns

With regards to payment patterns there are several different aspects, which make up these payment patterns. These shall be separated into the following categories:

- Number of payment sequences (where a sequence is a run of consecutive months of repayment)
- Amount recovered in each payment sequence
- Length of each payment and non payment sequence
- Proportion of default amount recovered in any payment sequences (Recovery Rate)

Each of these categories needs to be considered separately.

6.3 Number of Payment Sequences

When considering how many payment sequences a debtor will participate in, one needs to consider the sequence of paying and the probability of leaving the payment sequence at any point. So all debtors begin in NP^1 (first non payment sequence) since all of the debtors in the data set have defaulted. There are only two ways to leave NP^1, the debtor either has to start paying (P^1) or get written off (W^1). Once the debtor starts paying there are only two ways to leave P^1. The debtor can either stop paying, in which case they enter NP^2 or pay off all of their debt (D^1). So in order to calculate the probability of a debtor entering NP^{i+1} given that they are in NP^i, first calculate the probability of moving to P^i and then the probability of moving to NP^{i+1}.

$$P(NP^{i+1}| NP^i) = P(NP^{i+1}|P^i) * P(P^i| NP^i)$$

	$P(W^i \mid NP^i)$	$P(P^i \mid NP^i)$	$P(D^i \mid P^i)$	$P(NP^{i+1} \mid P^i)$
NP^1	0.273	0.727	0.043	0.957
NP^2	0.163	0.837	0.042	0.958
NP^3	0.138	0.862	0.044	0.956
NP^4	0.113	0.887	0.051	0.949
NP^5	0.122	0.878	0.049	0.951
NP^6	0.105	0.895	0.049	0.951
NP^7	0.097	0.903	0.053	0.947
NP^8	0.089	0.911	0.059	0.941
NP^9	0.104	0.896	0.069	0.931
NP^{10}	0.117	0.883	0.065	0.935

Table 6.1 probability table

Table 6.1 shows the probabilities of moving from one payment state to another. As would be expected the probability of being written off (W^1) is higher in NP^1 and then drops off for subsequent sequences. The probability of paying off the whole debt (D^i) increases with each payment sequence as would be expected since with each payment the debt to be recovered decreases.

The table was calculated by $$P(W^i | NP^i) = \frac{\text{No. of written offs in } NP^i}{\text{No. who reach } NP^i}$$

$$P(P^i|NP^i) = \frac{\text{No. who reach } P^i}{\text{No. who reach } NP^i}$$

Where $P(W^i|NP^i) + P(P^i|NP^i) =1$

And $$P(D^i|P^i) = \frac{\text{No. of paid offs in } P^i}{\text{No. who reach } P^i}$$

$$P(NP^{i+1}|P^i) = \underline{\text{No. who reach } NP^{i+1}}$$

No. who reach P^i

Where $P(D^i|P^i) + P(NP^{i+1}|P^i) = 1$

Therefore the probability of the debt being paid off in the first payment sequence is: $P(D^1) = P(P^1|NP^1)\ P(D^1|P^1) = 0.727 * 0.043 = 0.031$

The probability of reaching the second non-paying sequence is:

$P(NP^2) = P(P^1|NP^1)\ P(NP^2|P^1) = 0.727 * 0.957 = 0.696$

Hence the probability of reaching NP^{11} sequence is given in equation 6.1 below:

$P(NP^{11})$ (eq6.1)

$= P(P^1|NP^1)\ P(NP^2|P^1)\ P(P^2|NP^2)\ P(NP^3|P^2)\ \ldots\ P(P^{10}|NP^{10})\ P(NP^{11}|P^{10}) = 0.727 * 0.957 * 0.837 * 0.958 * \ldots * 0.883 * 0.935 = 0.024$

While there are debtors in the data set that continue on this stop start payment process for up to P^{25}, however the probability of reaching NP^{11} is less than 3%, as can be seen from the equation 6.1 above. Hence the sample sizes become too small to be relied upon so only payment sequences up to P^{10} will be discussed in this thesis.

6.4 Length of Payment Sequence

The length of the payment period is dependant upon when the debtor stops paying after they have started. In the same way the length of the non-payment period is dependant upon when the debtor starts to pay. As could be seen from figure 6.1 the length of any payment period or non-payment period can vary considerably from one month to many years. Due to small sample sizes, the next figures will only cover for up to two years.

Figure 6.2 shows the conditional probability of paying given that the debtor has reached that month without paying for the first six non-payment sequences. As can be seen from this graph the probability of when a debtor starts to pay in the first non-payment sequence (NP^1) is different to when a debtor will start paying in any other sequence. It is also clear that the first nine months are different to the following months. Months 9 to 24 appear to be almost flat at about 0.03. The first nine months resemble a power distribution.

The blips at 12 months and 24 months are caused by the way the data is recorded and are not true spikes. As was discussed in chapter 5, the way the data was collected, meant that if it was unclear when a debtor had made a payment, it was assumed that the debtor paid all year. This assumption causes spikes to occur at twelve-month intervals.

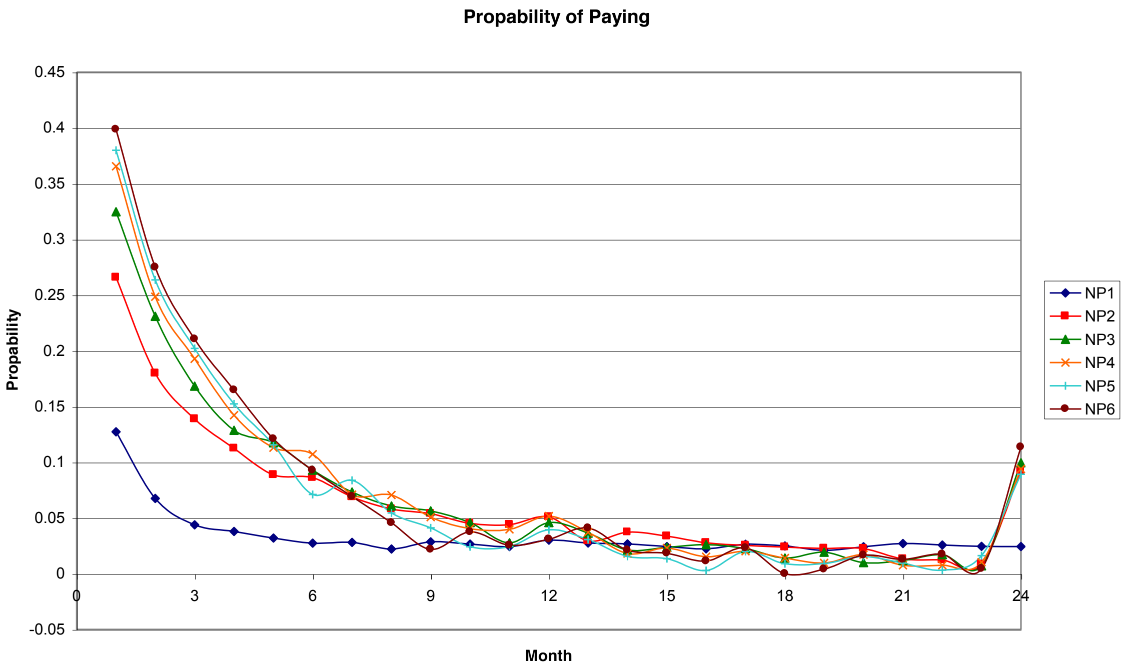


Figure 6.2 Conditional probability of starting to pay given that they reached the month without paying or being written off

This graph shows that the debtor is more likely to start paying again sooner, if they have made some previous payments after default. Also the more times they have started and stopped repayments (i.e. the greater the non-payment sequence) the more likely they are to pay sooner as the curves are stacked in descending order.

Therefore a model for the conditional probability of starting to pay in NP^1 is best expressed as a power function based on the shape of the curves. Different equations were fitted using trend lines and then the model predictions were matched to the real results using an R^2 comparison and the best result was selected. Hence the form $P(P_j|NP_{j-1}) = a\ j^b$ was assumed and then fitted to the values. The data was then split into training and holdout set in the ratio 70:30. Then the model was calculated on the training set using minimum squared errors to gauge the best fit and then the R^2 value was calculated on the holdout sample using the explained variance method.

$P(P_j^1|NP_{j-1}^1)$ = Probability of starting to pay in month j given that they have not started to pay or been written off by month j-1

$= 0.1058\ j^{-0.6729}$ for $j<10$

$= 0.025$ otherwise

This model gives an R^2 value of 0.94.

Since the curves NP^2, NP^3, NP^4, NP^5 and NP^6 are so similar it made sense to use the same curve to estimate all of these cases. Again a power function gave the best match after fitting various forms to the shape of the curves so trend lines were used to estimate $P(P_j|NP_{j-1})= a\ j^b$. The data was split as before and minimum squared error was used to create the model.

A model for the conditional probability of starting to pay for NP^2, NP^3, NP^4, NP^5 and NP^6 is:

$P(P_j^i|NP_{j-1}^i)$, $i>1$ = Probability of starting to pay in month j given that they have not started to pay or been written off by month j-1

$= 0.3017\ j^{-0.7746}$ for $j<12$

$= 0.025$ otherwise

This model gives an R^2 value of 0.86.

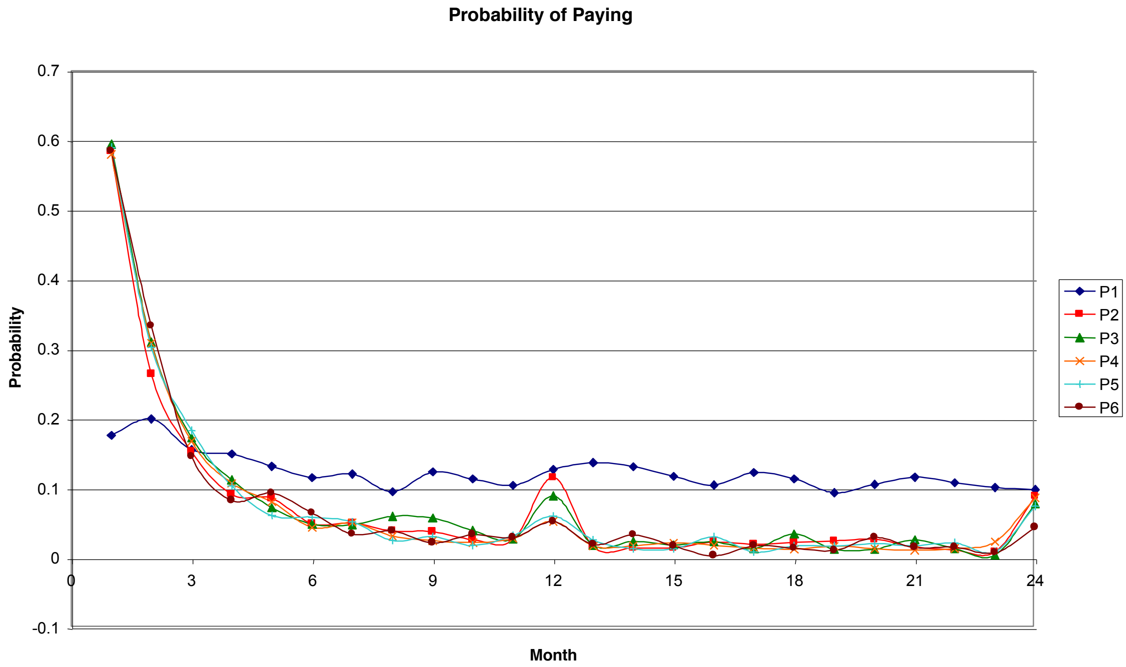


Figure 6.3 Conditional probability of stopping paying given that they reached the month without stopping or paying off the full debt

Figure 6.3 shows the conditional probability of stopping payments for the first six payment sequences given that the debtor has reached that month in the payment sequence without stopping paying. As can be seen from this graph the probability of when a debtor will stop paying in the first payment sequence (P^1) is different to when a debtor will stop paying in any other sequence. It is also clear that the first six months of any payment sequence are different to the following months. In months 6 to 24 the conditional probabilities appear to be almost flat at about 0.11 for P^1 and 0.03 for the other payment sequences. The first six months resemble a linear distribution. The blips at 12 months and 24 months are again caused by the way the data is recorded and are not true spikes. Chapter 5 gives more detail on how the spikes are created, because of the lack of detail within the data set, if it was unclear when a payment was made it was sometimes assumed that the debtor paid for the full year. This assumption causes the spikes at 12 and 24 months, whereas the debtors' payment sequences were probably of a shorter duration.

Therefore a model for the conditional probability of stopping payment in P^1 is expressed using a linear regression since the curve is almost a straight line:

$P(NP_j^1|P_{j-1}^1)$ = Probability of stopping payment in month j given that they have been paying and have not paid off by month j-1

= -0.0128j + 0.2014 for j<7

= 0.11 otherwise

This model gives an R^2 value of 0.98.

Since the curves P^2, P^3, P^4, P^5 and P^6 are so similar it made sense to use the same curve to estimate all of these cases. So again a power function gave the best match for the shape of the curves so trend lines were used to estimate $P(NP_j|P_{j-1})= a\ j^b$.

A model for the conditional probability of stopping payment for P^2, P^3, P^4, P^5 and P^6 is:

$P(NP_j^i|P_{j-1}^i)$, i>1 = Probability of stopping payment in month j given that they have been paying and have not paid off by month j-1

= $0.6193\ j^{-1.32}$ for j<11

= 0.025 otherwise

This model gives an R^2 value of 0.97.

6.5 Amount Recovered in Each Payment Sequence

When considering the amount recovered in each payment sequence one also needs to consider the length of the payment sequence. Since in most cases the debtor has agreed to pay back their debt at a certain rate e.g. £50 per month, then the longer they continue the payment plan, the more will be recovered.

Figure 6.4 shows the conditional probability of amount recovered in £10 segments for each payment sequence given that they do not pay off their debt. As can be seen from this graph there is no discernable pattern. There is a slight downward trend indicating an exponential model may be used but due to the fluctuations in the amount recovered the R^2 value will be very low.

This model to predict the amount recovered in pounds, based on a conditional probability, given that they have paid amount A-10, the model predicts the probability of them paying A (i.e. another £10). The amount for any given sequence of payments is estimated from a linear regression of log P(A) and is as follows:

$P(A) = 0.0658e^{-0.003(A-10)}$ $R^2=0.5$

However since the amount paid back in any payment sequence is dependent upon the length of the payment sequence it may be more stable to consider the average amount repaid in each month during a sequence.

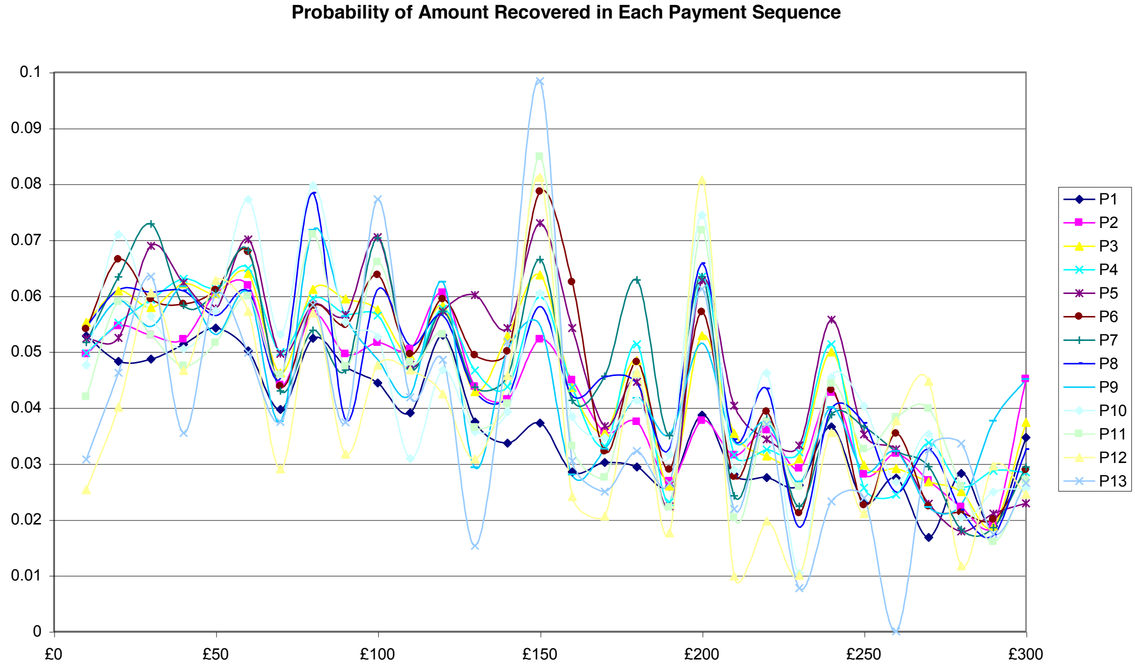


Figure 6.4 Conditional probability of amount recovered in £10 segments for each payment sequence given that they do not pay off their debt

Figure 6.5 shows the mean amount recovered per month during a payment sequence for the first nine payment sequences. Payment sequences after this still followed the same negative relationship with average amount received but become more erratic due to the small sample sizes. There is clearly a very definite exponential negative relationship with average amount received.

A model to predict the average amount recovered per month in pounds for any given sequence of payments is as follows:

$P(A) = 0.116e^{-0.0115A}$ $R^2=0.96$

The curves do not automatically suggest any particular form so various forms were tested using trend lines which were analysed using R^2 to evaluate the goodness of fit of the model to the real values but eventually the log model log P(A) =a+bA gave the best fit.

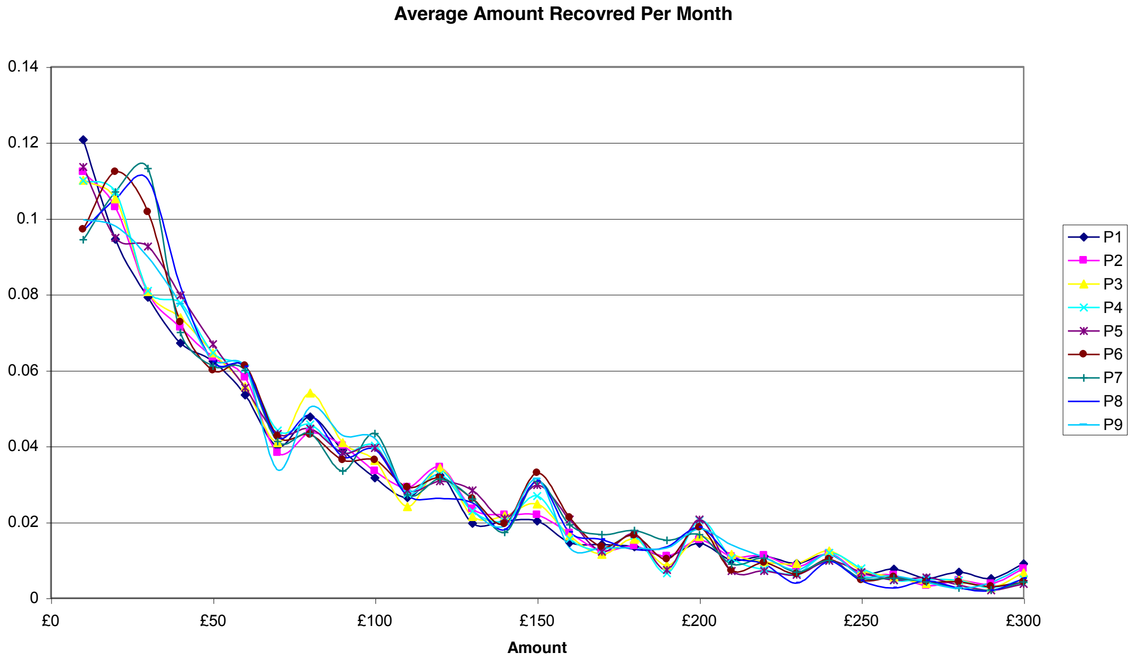


Figure 6.5 Average amount recovered per month during a payment sequence

6.6 Recovery Rate in Each Payment Sequence

Calculating the amount recovered during a payment sequence could be useful for some prediction models however when considering loss given default it can be far more interesting and useful to consider the recovery rate for each sequence rather than the amount recovered. In order to calculate the recovery rate we need to know the amount outstanding and for the following recovery rates the default amount is used for the amount outstanding rather than the amount outstanding at the start of each sequence. For modelling LGD the RR for each month of the payment sequence may be smoother than the RR for the whole payment sequence. This means that the RR per month is independent of sequence length. Therefore both have been modelled.

Using the default amount to determine recovery rate, means that after the first payment sequence the recovery rate cannot equal one. Therefore the total recovery rate is $RR = \sum_{i=1}^{\infty} RR^{i}$ where i is the payment sequence.

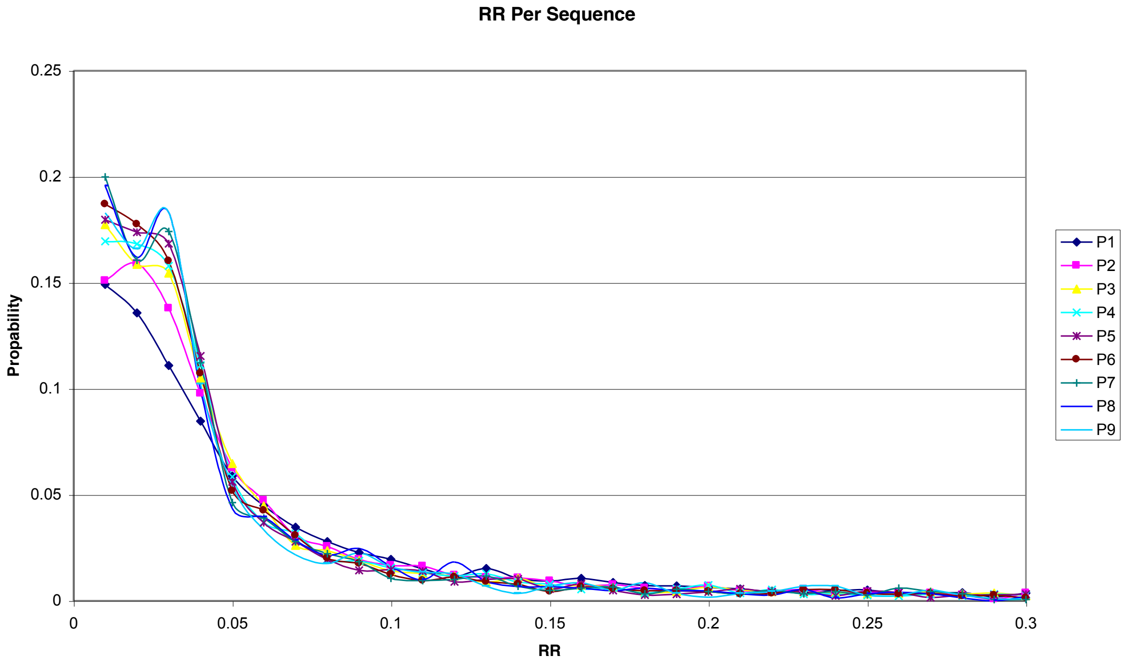


Figure 6.6 Probability of recovery rate per sequence

Figure 6.6 shows the recovery rate for each sequence. All of the sequences follow the same pattern an exponential drop followed by a shallower exponential. The first is slightly different to the rest. Clearly the probability for recovery rate changes at around 0.08 (x-axis) so the following model incorporates this.

P(RR) = Probability that RR proportion of loan at default will be recovered in any payment sequence

$$= 0.3248e^{-33.82RR} \quad \text{for } RR<0.08$$

$$= 0.036e^{-9.6971RR} \quad \text{otherwise}$$

This model gives an R^2 value of 0.92.

Figure 6.7 shows the average recovery rate per month in each sequence. The results are very similar to figure 6.6 but are now smother because they are now independent of the length of the sequence. Therefore the same form has been used for both.

P(RR) = Probability that RR proportion of loan at default will be recovered in month of any payment sequence

$$= 0.4988e^{-41.22RR} \quad \text{for } RR<0.07$$

$$= 0.0339e^{-15.146RR} \quad \text{otherwise}$$

This model gives an R^2 value of 0.97.

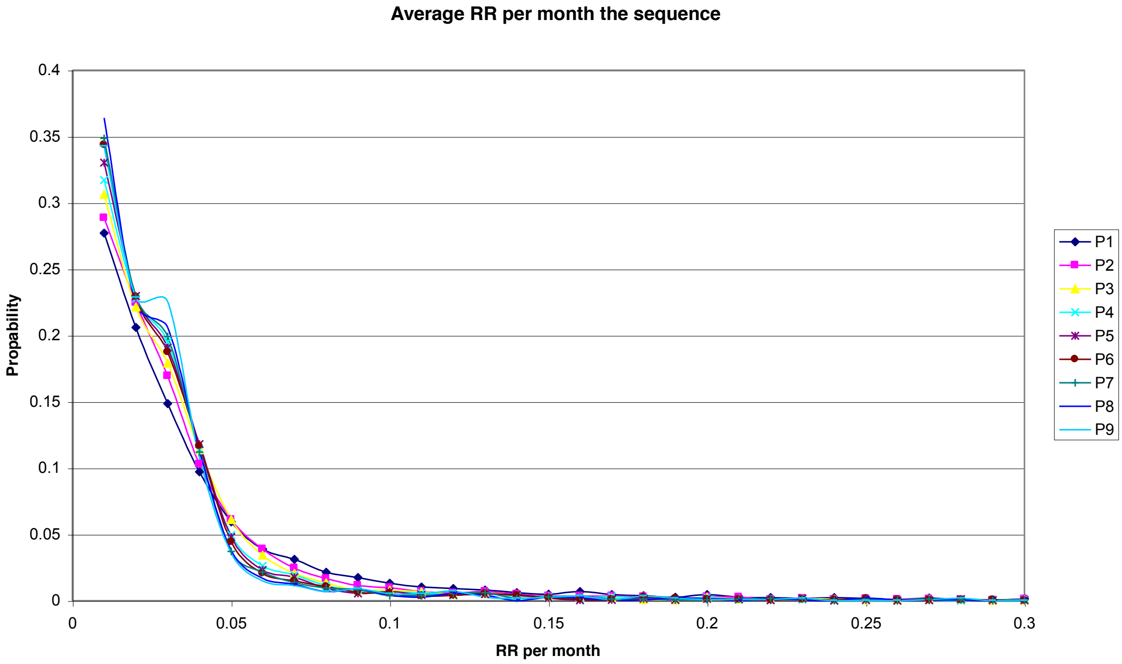


Figure 6.7 Probability of recovery rate per month of each sequence

The next section will look at which individual variables are the best predictors of the average monthly paid amount after default.

6.7 Individual's Payment Pattern

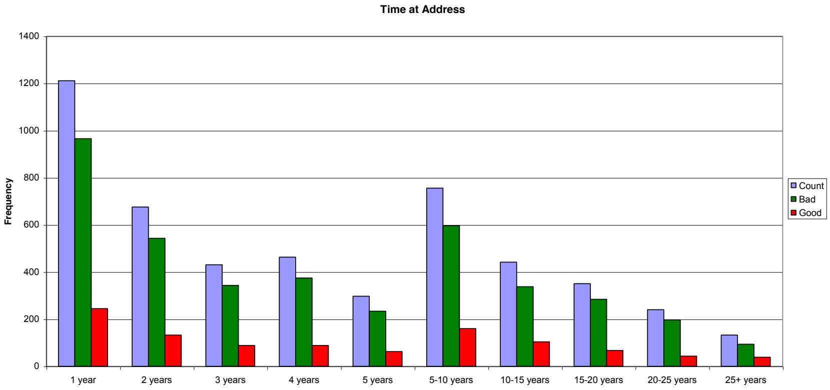
So far this chapter has looked at predicting the payment patterns of a group of debtors, trying to predict how an individual debtor will pay back their debt is more difficult. The same things have to be predicted i.e. amount per sequence, length of sequence, but using individual variables.

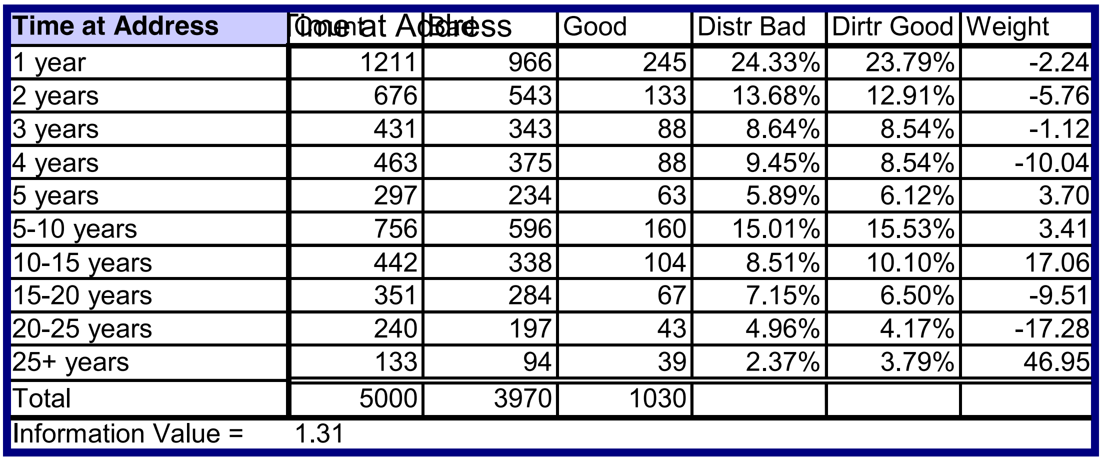
Predicting the amount paid back per sequence by debtors is the aim of this model. In order to achieve this the length of the sequence must be known and the mean amount paid back each month during the sequence, since as figure 6.4 shows trying the predict the total amount recovered in each sequence without reference to its length has too much variation. The length of the sequence can be used to calculate the total amount recovered per sequence by multiply it by the mean amount recovered per month.

Trying to predict the average amount a debtor would pay back each month during their first payment sequence proved to be difficult as regression was resulting in R^2 of 0.04, which even for debt models is very poor. So a decision tree approach was used. To select the variables weight of evidence was used.

6.7.1 Time at Address

Figure 6.8 WOE for Time at Address





Time at Address	Count	Bad	Good	Distr Bad	Dirtr Good	Weight
1 year	1211	966	245	24.33%	23.79%	-2.24
2 years	676	543	133	13.68%	12.91%	-5.76
3 years	431	343	88	8.64%	8.54%	-1.12
4 years	463	375	88	9.45%	8.54%	-10.04
5 years	297	234	63	5.89%	6.12%	3.70
5-10 years	756	596	160	15.01%	15.53%	3.41
10-15 years	442	338	104	8.51%	10.10%	17.06
15-20 years	351	284	67	7.15%	6.50%	-9.51
20-25 years	240	197	43	4.96%	4.17%	-17.28
25+ years	133	94	39	2.37%	3.79%	46.95
Total	5000	3970	1030			
Information Value =	1.31					

Table 6.2 shows the results for Weight of Evidence (WOE) analysis for the variable Time at Address. Time at address is the length of time that a debtor resided in their address when the loan was approved. Count is the number of debtors in each bin, bad; a bad debtor is one who paid back below the mean amount per month in the first payment sequence. A good debtor is therefore one who paid back more than the mean amount per month in the first

payment sequence. “Distr Bad” and “Distr Good” are the percentage of bad or good debtors respectively who fall into each bin. The weight is then determined by

$$Weight = Ln\left(\frac{DistrGood}{DistrBad}\right)\times 100$$

Since the model is trying to determine amount paid by an average debtor the mean is far more appropriate than the median to determine good and bad characteristics. These WOE variables were constructed for the different borrow characteristics and those chosen for the models were the ones with the highest information value.

The information value is determined by

$$InformationValue = \sum_{i=1}^{n}\left(DistrGood_i - DistrBad_i\right)Weight_i$$

where n is the number of bins

The higher the information value the more useful the variable is to determine amount paid per month in the first payment sequence. Figure 6.8 shows the distribution of good and bad debtors for time at address.

Using time in occupation as an example for calculating the information value:

The information value = (0.2194-0.2474)*-11.98+ (0.1621-0.1798)*-10.37 + (0.1282-0.1798)*2.95 + (0.3602-0.3292)*8.99 + (0.1301-0.1191)*8.80 = 0.90

All of the WOE analysis is based on a training set of 5000 debtors who all paid back some money in the first sequence.

6.7.2 Time in Occupation

Table 6.3 shows the results for WOE analysis for the variable Time in Occupation. Time in occupation is the length of time that a debtor was employed in their occupation when the loan was approved. Figure 6.9 shows the distribution of good and bad debtors for time in occupation. The information value for time in occupation shows that the variable is less useful than time at address for determining amount paid in the first sequence each month.

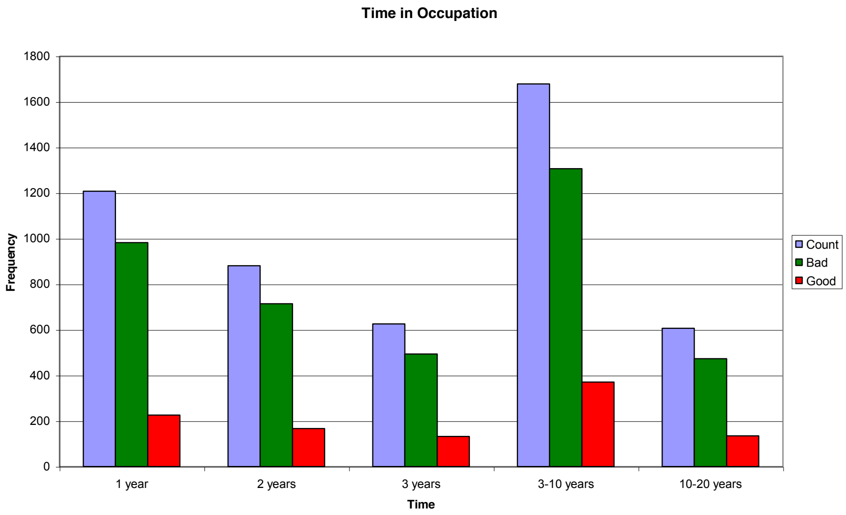


Figure 6.9 WOE for Time in Occupation

Time in Occupation	Count	Bad	Good	Distr Bad	Dirtr Good	Weight
1 year	1208	982	226	24.74%	21.94%	-11.98
2 years	881	714	167	17.98%	16.21%	-10.37
3 years	626	494	132	12.44%	12.82%	2.95
3-10 years	1678	1307	371	32.92%	36.02%	8.99
10-20 years	607	473	134	11.91%	13.01%	8.80
Total	5000	3970	1030			
Information Value =	0.90					

Table 6.3 WOE for Time in Occupation

6.7.3 Default Amount

Default Amount	Count	Bad	Good	Distr Bad	Dirtr Good	Weight
£1,000	360	331	29	8.34%	2.82%	-108.56
£2,000	867	790	77	19.90%	7.48%	-97.90
£3,000	930	807	123	20.33%	11.94%	-53.19
£4,000	772	610	162	15.37%	15.73%	2.33
£5,000	786	606	180	15.26%	17.48%	13.53
£6,000	490	346	144	8.72%	13.98%	47.26
£7,000	267	175	92	4.41%	8.93%	70.62
£8,000	419	262	157	6.60%	15.24%	83.71
+£8,000	109	43	66	1.08%	6.41%	177.77
Total	5000	3970	1030			
Information Value =	45.31					

Table 6.4 WOE for Default Amount

Table 6.4 shows the results for WOE analysis for the variable Default Amount. Default amount is the amount outstanding on the loan when the debtor defaults given that default is three months in arrears. The information value for default shows that the variable is very useful for determining amount paid in the first sequence each month. Figure 6.10 shows the distribution of good and bad debtors for default amount. The figure and table show that the more money owed at default the greater the proportion of debtors who paid back more than the average each month during the first sequence.

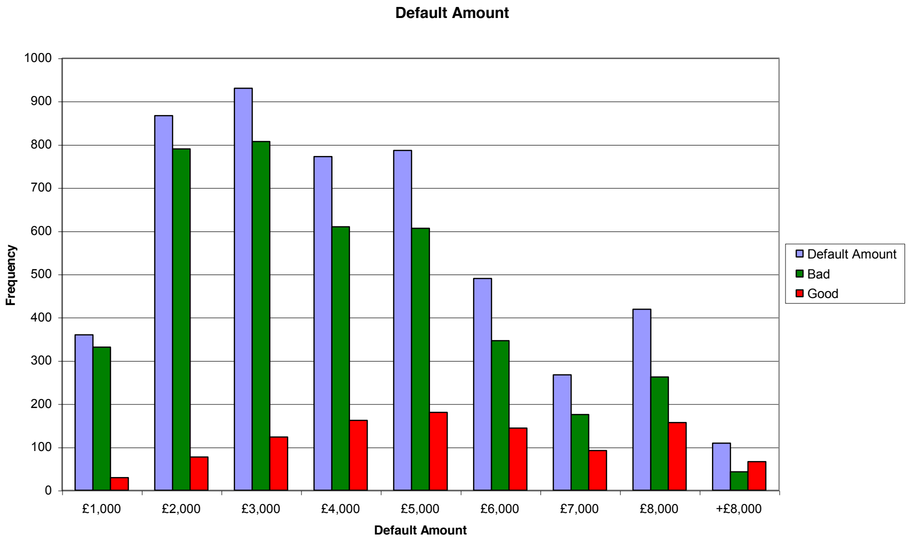


Figure 6.10 WOE for Default Amount

6.7.4 Mortgage

Mortgage	Count	Bad	Good	Distr Bad	Dirtr Good	Weight
Yes	1490	1147	343	28.89%	33.30%	14.20
No	3510	2823	687	71.11%	66.70%	-6.40
Total	5000	3970	1030			
Information Value =	0.91					

Table 6.5 WOE for Mortgage

Table 6.5 shows the results for WOE analysis for the variable Mortgage. Mortgage is whether the debtor took out a mortgage with the lender prior to the loan approval. The information value shows that the variable is poor for determining amount paid in the first sequence each month. Since there are only two bins for this variable there is no accompanying figure.

6.7.5 Married

Married	Count	Bad	Good	Distr Bad	Dirtr Good	Weight
Yes	2287	1858	429	46.80%	41.65%	-11.66
No	2713	2112	601	53.20%	58.35%	9.24
Total	5000	3970	1030			
Information Value =	1.08					

Table 6.6 WOE for Married

Table 6.6 shows the results for WOE analysis for the variable Married. Married is whether the debtor was married at the time of loan approval. The information value shows that the variable is reasonable for determining amount paid in the first sequence each month.

6.7.6 Own Home

Own Home	Count	Bad	Good	Distr Bad	Dirtr Good	Weight
Yes	2511	1986	525	50.03%	50.97%	1.87
No	2489	1984	505	49.97%	49.03%	-1.91
Total	5000	3970	1030			
Information Value =	0.04					

Table 6.7 WOE for Own Home

Table 6.7 shows the results for WOE analysis for the variable Own Home. Own Home is whether the debtor was either the sole owner or had joint ownership of their residence at the time of loan approval. The information value shows that the variable is very poor for determining amount paid in the first sequence each month.

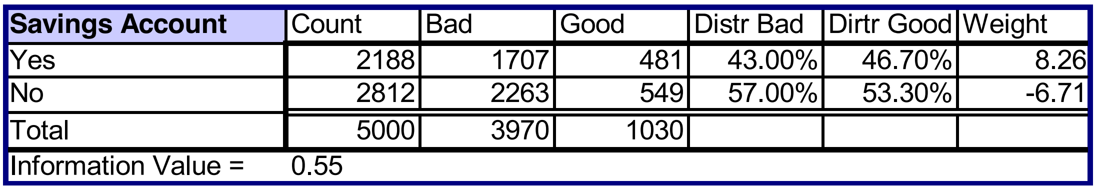
6.7.7 Children

Children	Count	Bad	Good	Distr Bad	Dirtr Good	Weight
0	3222	2540	682	63.98%	66.21%	3.43
1	799	650	149	16.37%	14.47%	-12.38
2	640	499	141	12.57%	13.69%	8.54
+2	339	281	58	7.08%	5.63%	-22.87
Total	5000	3970	1030			
Information Value =	0.74					

Table 6.8 WOE for Children

Table 6.8 shows the results for WOE analysis for the variable Children. Children are the number children the debtor had at the time of loan approval. The information value shows that the variable is poor for determining amount paid in the first sequence each month.

6.7.8 Savings Account

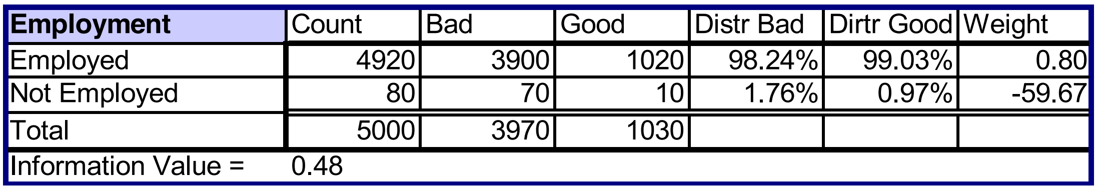


Savings Account	Count	Bad	Good	Distr Bad	Dirtr Good	Weight
Yes	2188	1707	481	43.00%	46.70%	8.26
No	2812	2263	549	57.00%	53.30%	-6.71
Total	5000	3970	1030			
Information Value =	0.55					

Table 6.9 WOE for Savings Account

Table 6.9 shows the results for WOE analysis for the variable Savings Account. Savings Account is whether the debtor had a savings account with the lender at the time of loan approval. The information value shows that the variable is very poor for determining amount paid in the first sequence each month.

6.7.9 Employment



Employment	Count	Bad	Good	Distr Bad	Dirtr Good	Weight
Employed	4920	3900	1020	98.24%	99.03%	0.80
Not Employed	80	70	10	1.76%	0.97%	-59.67
Total	5000	3970	1030			
Information Value =	0.48					

Table 6.10 WOE for Employment

Table 6.10 shows the results for WOE analysis for the variable Employment. Employment is whether the debtor was employed at the time of loan approval. The information value shows that the variable is very poor for determining amount paid in the first sequence each month.

6.7.10 Loan Amount

Loan Amount	Count	Bad	Good	Distr Bad	Dirtr Good	Weight
£1,000	435	407	28	10.25%	2.73%	-132.35
£2,000	921	833	88	20.98%	8.58%	-89.46
£3,000	936	803	133	20.23%	12.96%	-44.49
£4,000	705	550	155	13.85%	15.11%	8.66
£5,000	926	704	222	17.73%	21.64%	19.90
£6,000	329	222	107	5.59%	10.43%	62.32
£7,000	250	160	90	4.03%	8.77%	77.77
£7,000+	498	291	207	7.33%	20.18%	101.25
Total	5000	3970	1030			
Information Value =	48.88					

Table 6.11 WOE for Loan Amount

Table 6.11 shows the results for WOE analysis for the variable Loan Amount. Loan Amount is the amount loaned at the time of loan approval. The information value shows that the variable is very good for determining amount paid in the first sequence each month.

6.7.11 Loan Term

Table 6.12 shows the results for WOE analysis for the variable Loan Term. Loan Term is the original length of the loan at the time of loan approval. The information value shows that the variable is good for determining amount paid in the first sequence each month.

Loan Term	Count	Bad	Good	Distr Bad	Dirtr Good	Weight
2 years	641	567	74	14.28%	7.18%	-68.71
3 years	1121	925	196	23.30%	19.03%	-20.25
4 years	472	363	109	9.14%	10.58%	14.62
5 years	2765	2114	651	53.25%	63.20%	17.14
5 years +	1	1	0	0.03%	0.00%	0.00
Total	5000	3970	1030			
Information Value =	7.66					

Table 6.12 WOE for Loan Term

6.7.12 Application Score

Table 6.13 shows the results for WOE analysis for the variable Application Score. Application Score is the score given to the debtor by the lender at the time of loan approval. The information value shows that the variable is good for determining amount paid in the first sequence each month.

Application Score	Count	Bad	Good	Distr Bad	Dirtr Good	Weight
£180	499	391	108	9.85%	10.53%	6.65
£190	595	484	111	12.19%	10.82%	-11.95
£200	950	778	172	19.60%	16.76%	-15.61
£210	909	724	185	18.24%	18.03%	-1.13
£220	684	532	152	13.40%	14.81%	10.03
£230	561	454	107	11.44%	10.43%	-9.22
£240	304	245	59	6.17%	5.75%	-7.06
240+	494	362	132	9.12%	12.87%	34.43
Total	5000	3970	1030			
Information Value =	2.21					

Table 6.13 WOE for Application Score

The top six variables were Loan Amount (48.88), Default Amount (45.31), Term of Loan (7.66), Application Score (2.21), Time at Address (1.31) and Married (1.08). The top variables were used to create a segmentation tree to determine the amount paid each month during the first sequence.

Loan amount and default amount had the highest weight, however they are both very closely tied and since the amount paid back before default was so small in most cases, they were almost identical for the majority of debtors. Therefore it makes more sense to use the original loan amount and the amount paid back before default. Table 6.14 shows the results for WOE analysis for the variable Amount Paid Before Default.

Amount **Paid Before Default**	Count	Bad	Good	Distr Bad	Dirtr Good	Weight
£0	1530	1190	340	29.97%	33.01%	9.64
£30	1024	892	132	22.47%	12.82%	-56.15
£60	930	806	124	20.30%	12.04%	-52.26
£90	573	461	112	11.61%	10.87%	-6.57
£90+	943	621	322	15.64%	31.26%	69.24
Total	5000	3970	1030			
Information Value =	20.89					

Table 6.14 WOE for Amount Paid Before Default

Amount Paid Before Default is the next highest value after Loan Amount if Default Amount is discounted. Therefore to estimate the average amount paid after default, loan amount, amount paid back before default, and loan term are used for the first branches of the segmentation tree.

6.8 Segmentation Tree to Calculate Repayments After Default

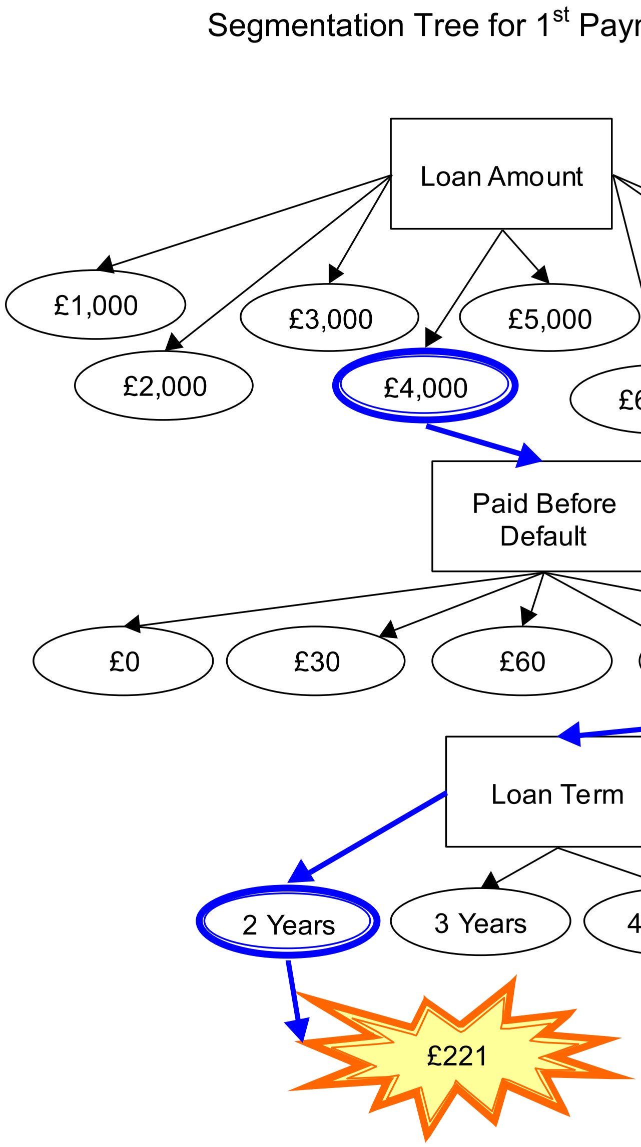


Figure 6.11 example of estimating the average payment amount in the first payment sequence

Figure 6.11 above shows the segmentation tree for determining the average amount paid per month during the first sequence. There are 160 different combinations available. In the figure above one is example is shown where the loan amount was for between £3,000 and £4,000, the debtor paid over £90 before default and the original loan term was for two years. This particular combination predicts that the average amount the debtor will pay per month in the first sequence is £221. Had any of these factors been different than the predicted payment amount would also have been different. For instance if the debtor had not paid anything back before default then the debtor would have been predicted to pay back £215.63 per month.

Count	Amount Paid Before Default					
Loan Amount		£0	£30	£60	£90	£90.00+
	£1000	162	290	283	188	277
	£2000	160	290	205	122	141
	£3000	49	182	208	169	175
	£4000	30	73	98	79	129
	£5000	34	86	142	147	204
	£6000	162	290	283	188	277
	£7000	160	290	205	122	141
	£7,000+	49	182	208	169	175

Table 6.15 Number of training data for the first matrix of segmentation tree

Table 6.15 shows the number of debtors in the training set in each classification. Since just using loan amount and amount paid before default means 40 different bins, and there are only 5000 debtors in the training set means there is only 125 debtor in each bin on average. The mean payment of the debtors in each bin was then used to determine the payment for each bin. Table 6.16 show the mean payments.

Amount	Amount Paid Before Default					
Loan Amount		£0	£30	£60	£90	£90.00+
	£1000	£69.88	£42.41	£39.73	£69.22	£74.83
	£2000	£93.27	£67.89	£68.49	£90.43	£131.73
	£3000	£169.84	£105.79	£93.56	£112.36	£219.47
	£4000	£215.63	£190.10	£143.07	£123.24	£220.76
	£5000	£288.38	£157.02	£227.39	£107.27	£197.99
	£6000	£384.00	£351.27	£194.13	£134.15	£280.22
	£7000	£311.28	£217.15	£293.44	£215.00	£234.01
	£7,000+	£515.51	£590.86	£260.19	£524.20	£432.73

Table 6.16 Mean amount paid by debtors in each bin on average per month in the first payment sequence

The next variable with the highest WOE value was Term of Loan at 7.66. Now since there was only one loan whose term was over 5 years, it makes sense to group that result with the five-year terms. So now there are 160 bins. With only 5000 debts in the training data that is only 31 debts per bin on average.

The next set of tables 6.17 show the number of debtors in the training set within each bin.

Count	Loan Term with £0 paid before default				
Loan Amount		2 years	3 years	4 years	5 years
	£1000	103	35	5	19
	£2000	82	104	15	89
	£3000	31	87	29	136
	£4000	12	42	27	107
	£5000	4	30	24	219
	£6000	1	9	9	86
	£7000	0	4	3	73
	£7,000+	0	4	4	137

Table 6.17a Number of debtors in the training set where the debtor paid nothing before default

Count	Loan Term with £30 paid before default				
Loan Amount		2 years	3 years	4 years	5 years
	£1000	82	42	3	33
	£2000	51	91	29	119
	£3000	5	43	29	128
	£4000	2	26	15	79
	£5000	1	13	7	120
	£6000	0	1	2	40
	£7000	0	3	1	16
	£7,000+	0	2	3	38

Table 6.17b Number of debtors in the training set where the debtor paid £30 before default

Count	Loan Term with £60 paid before default				
Loan Amount		2 years	3 years	4 years	5 years
	£1000	33	12	2	2
	£2000	45	73	19	45
	£3000	14	72	22	100
	£4000	1	30	23	115
	£5000	2	9	21	143
	£6000	0	1	1	47
	£7000	0	0	1	32
	£7,000+	0	2	1	62

Table 6.17c Number of debtors in the training set where the debtor paid £60 before default

Count	Loan Term with £90 paid before default				
Loan Amount		2 years	3 years	4 years	5 years
	£1000	25	3	1	1
	£2000	21	38	5	9
	£3000	13	52	9	24
	£4000	1	23	18	37
	£5000	0	11	16	102
	£6000	0	1	5	46
	£7000	1	2	4	35
	£7,000+	0	3	4	63

Table 6.17d Number of debtors in the training set where the debtor paid £90 before default

Count	Loan Term with more than £90 paid before default				
Loan Amount		2 years	3 years	4 years	5 years
	£1000	29	2	1	2
	£2000	36	34	6	10
	£3000	28	63	13	38
	£4000	9	56	29	53
	£5000	4	63	26	111
	£6000	2	13	15	50
	£7000	0	11	9	55
	£7,000+	3	11	16	145

Table 6.17e Number of debtors in the training set where the debtor paid more than £90 before default

As you can see from tables 6.17 a-e there are several bins with no debtors to use and other bins with very low numbers of debtors in each bin. Therefore when the value was less than 15 debtors in the bin (0.3% of the debtors) just under half the average number of debtors in each bin on average, the value of the bin one branch up the segmentation tree were used. Therefore in tables 6.18 all bins with less than 15 debtors the value from table 6.16 was used instead.

The next set of tables show the final amounts within each bin to determine average payment made during the first sequence.

Amount	Loan Term with £0 paid before default				
Loan Amount		2 years	3 years	4 years	5 years
	£1000	£69.81	£80.94	£69.88	£56.66
	£2000	£99.34	£110.90	£52.07	£74.01
	£3000	£85.90	£210.02	£117.72	£174.39
	£4000	£215.63	£194.09	£150.18	£228.70
	£5000	£288.38	£186.40	£341.33	£300.69
	£6000	£384.00	£384.00	£384.00	£395.51
	£7000	£311.28	£311.28	£311.28	£310.64
	£7,000+	£515.51	£515.51	£515.51	£537.01

Table 6.18a Mean amount paid by debtors in each bin on average per month in the first payment sequence where the debtor paid nothing before default

Amount	Loan Term with £30 paid before default				
Loan Amount		2 years	3 years	4 years	5 years
	£1000	£34.86	£64.98	£42.41	£32.18
	£2000	£95.43	£48.35	£60.82	£72.75
	£3000	£105.79	£129.45	£51.71	£111.56
	£4000	£190.10	£106.70	£383.78	£183.48
	£5000	£157.02	£157.02	£157.02	£152.44
	£6000	£351.27	£351.27	£351.27	£368.68
	£7000	£217.15	£217.15	£217.15	£176.55
	£7,000+	£590.86	£590.86	£590.86	£546.09

Table 6.18b Mean amount paid by debtors in each bin on average per month in the first payment sequence where the debtor paid £30 before default

Amount	Loan Term with £60 paid before default				
Loan Amount		2 years	3 years	4 years	5 years
	£1000	£38.10	£39.73	£39.73	£39.73
	£2000	£71.44	£65.75	£86.99	£62.17
	£3000	£93.56	£106.31	£127.60	£83.13
	£4000	£143.07	£291.95	£170.98	£99.28
	£5000	£227.39	£227.39	£247.91	£230.54
	£6000	£194.13	£194.13	£194.13	£184.09
	£7000	£293.44	£293.44	£293.44	£300.68
	£7,000+	£260.19	£260.19	£260.19	£265.45

Table 6.18c Mean amount paid by debtors in each bin on average per month in the first payment sequence where the debtor paid £60 before default

Amount	Loan Term with £90 paid before default				
Loan Amount		2 years	3 years	4 years	5 years
	£1000	£74.87	£69.22	£69.22	£69.22
	£2000	£112.08	£88.44	£90.43	£90.43
	£3000	£112.36	£80.27	£112.36	£88.81
	£4000	£123.24	£76.56	£171.33	£130.36
	£5000	£107.27	£107.27	£138.25	£105.83
	£6000	£134.15	£134.15	£134.15	£143.17
	£7000	£215.00	£215.00	£215.00	£206.53
	£7,000+	£524.20	£524.20	£524.20	£428.20

Table 6.18d Mean amount paid by debtors in each bin on average per month in the first payment sequence where the debtor paid £90 before default

Amount	Loan Term with more than £90 paid before default				
Loan Amount		2 years	3 years	4 years	5 years
	£1000	£66.98	£74.83	£74.83	£74.83
	£2000	£135.26	£136.73	£131.73	£131.73
	£3000	£135.51	£306.28	£219.47	£168.27
	£4000	£220.76	£219.62	£240.32	£215.18
	£5000	£197.99	£197.71	£194.67	£203.55
	£6000	£280.22	£280.22	£356.69	£251.49
	£7000	£234.01	£234.01	£234.01	£264.38
	£7,000+	£432.73	£432.73	£160.99	£468.30

Table 6.18e Mean amount paid by debtors in each bin on average per month in the first payment sequence where the debtor paid more than £90 before default

The figure 6.12 shows the results of using the segmentation tree to predict the average amount paid each month during the first sequence. There is quite a large spread of payments made. The observed data shows that sometimes a debtor would just make one payment after default to pay off the whole debt. This meant that the payment amounts could vary from a few pounds up to thousands. This made predicting the payment amount very tricky. The large single payments were also quite rare meaning that a logistic regression to determine high and low payers would not be applicable.

The segmentation tree predicted payments from £32 up to £591. Using other prediction methods the predictions were all clustered around the mean of, £190 or the median of £66. Instead of segmenting into bins using loan amount, loan term and amount paid before default to create historic averages for the value of each bin, one could try to use non-linear regression. Using non-linear regression avoids the problem of clustering and is more adaptive than the segmentation tree. The next section suggests one non-linear regression approach, which gave good results for an individual model.

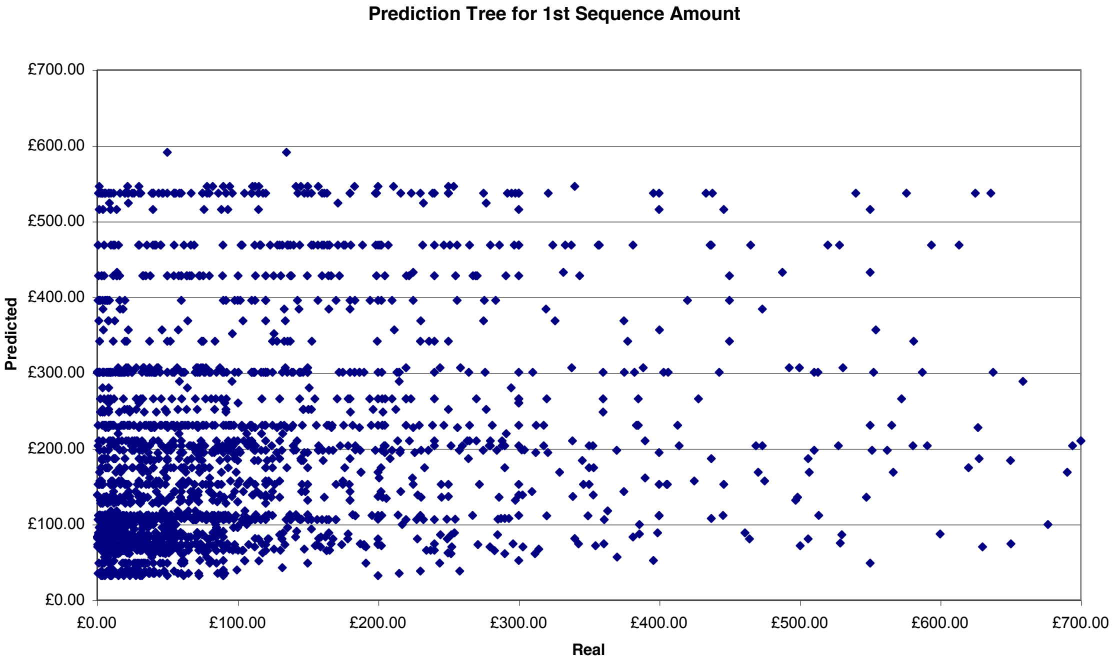


Figure 6.12 results of using the prediction tree for estimating the amount paid in the first sequence

6.9 Alternative Approach for Predicting the First Sequence Payments using Non Linear Regression

The average amount paid per month during the first sequence was modelled using linear regression with all available variables. Yet since the payment sequences are evidentially exponentially distributed and not normal (figure 6.5), linear regression gives very poor results (R^2~0.02), however the payments are lognormal. Therefore by converting the payments to $\log_{10}$ before using regression, the results are improved.

Equation 6.1

Predicted Log_{10} average payment per month in 1st payment sequence=1.45207 +average collected before default*0.0002483 +default amount*0.00015323 +loan amount*0.00000232 +no children*0.05573

+term*0.00312- defaulted in 1990*0.21361- defaulted in 1991*0.23166- defaulted in 1992*0.31639- defaulted in 1993*0.26216- defaulted in 1994*0.21751- defaulted in 1995*0.09943- defaulted in 1997*0.08927

Equation 6.1 gave an R^2=0.12 using the explained variance method on the holdout sample however the default year is not much use for future prediction therefore using the economic variables described in chapter 5.

Equation 6.2

Predicted Log_{10} average payment per month in 1st payment sequence = 1.20541 + average collected before default * 0.00025313 - default amount * 0.00015795 - loan amount*0.00000236 +no children * 0.05552 + term * 0.00292+ Halifax *0.00992 + net lending *0.00016674

This also gave an R^2=0.12 using the explained variance method on the holdout sample. Figure 6.12 shows the predicted results of this model compared to the real average monthly payment in the first sequence. As you can see the spread was not as good as in figure 6.11. On the holdout sample the correlation between the actual and the predicted was 0.05, which is far worse than the R^2 from the original regression model predicted.

Since the regression estimates the log of the payment amounts, the exponential of the estimates have to be taken to get the estimate of the actual payment amount.

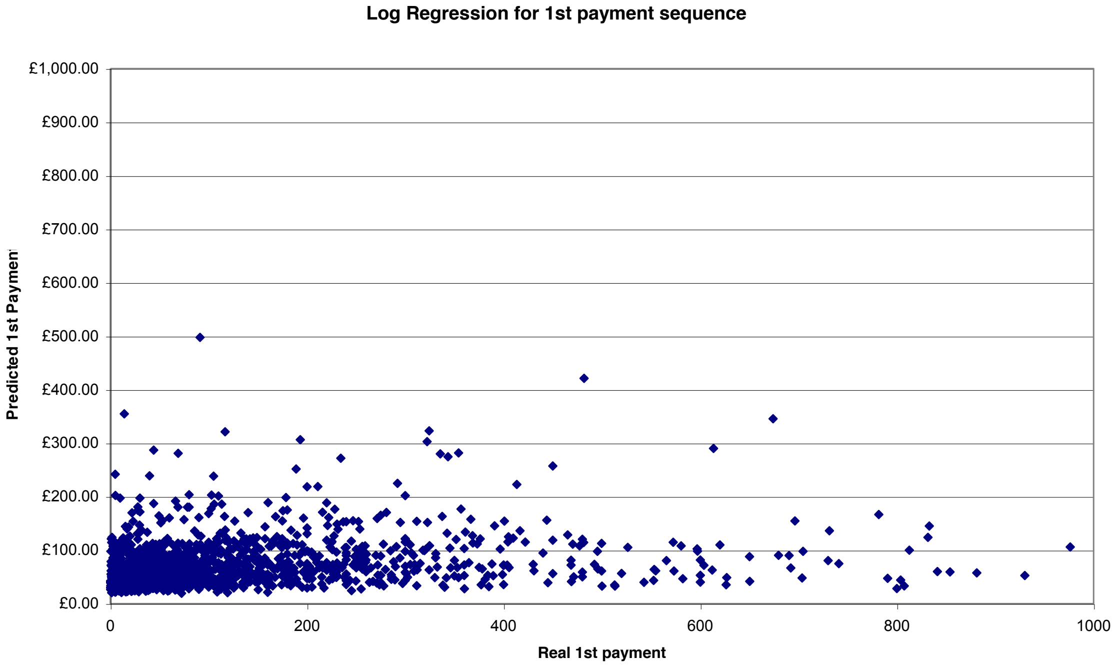


Figure 6.13 results of using lognormal regression for estimating the amount paid in the first sequence

6.10 Predicting Repayment Amounts in Future Sequences

When predicting the further payment sequence using regression the lognormal gave the best results but they were still poor in comparison to the first sequence. The model gave an R^2=0.06:

Equation 6.3

Predicted Log_{10} average payment per month in payment sequences after 1^{st} payment sequence= 1.7253 -loan amount* 0.000000171252 +term* 0.00748 + net lending at default* 0.00009003

An alternative is to use the average amount recovered in the first sequence to predict further payments. This means that the payment sequences have to have started, but the model is far better. The model no longer uses a lognormal but just a simple linear regression model. This is because of the linear relationship displayed between the first sequence payments and the second sequence payments. This achieved an R^2=0.28.

Equation 6.4

Predicted average payment per month in 2^{nd} payment sequences =-462.5+ payment in 1^{st} sequence *0.5384 + interest rate* 9.692 + net lending * 0.1583 + married * 9.31+ total recovered in first sequence * 0.047+unemployment * 3536.3 + amount left after first sequence*0.0114

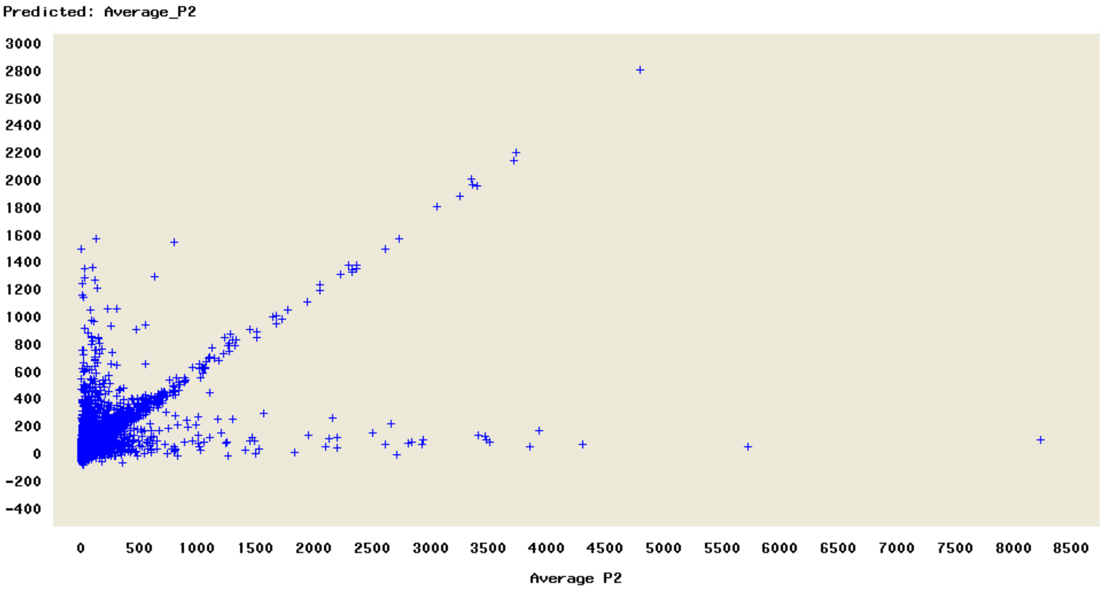


Figure 6.14 results of using lognormal regression for estimating the amount paid in the second sequence

Figure 6.14 shows the predicted results against the real average amount paid in the second sequence.

6.11 Expected Recovery Rate

The individual models discussed in this chapter can be used to predict the expected recovery rate. This is made up of the probability of having an i^{th} payment sequence for i=1,2… , multiplied by the length of the i^{th} payment sequence and the expected payment per month in the i^{th} payment sequence. Summing this for all i gives an estimate for LGD. This calculation is explained in more detail below.

The probability of starting to pay each sequence is summarised in table 6.1. Now once an individual has defaulted the probability of them starting to pay in the first sequence is 0.727 from table 6.1. The probability of starting the second sequence is the probability of starting the first payment sequence, multiplied by the probability of not paying off their debt in that first sequence, multiplied by the probability of starting to pay off the second sequence.

Equation 6.5

$$P(P^2) = P(P^1|NP^1) * P(NP^2|P^1) * P(P^2|NP^2)$$

Since figures 6.2 and 6.3 show that after the first payment sequence and first non payment sequence all of the subsequent sequences are very closely related and have been taken to be the same in all other models in this chapter, then the probabilities of starting the subsequent payment sequences should likewise be taken to be the same. Therefore the probability of stopping a non-payment sequence after NP^i, i>1 is 0.88 and the probability of stopping any payment sequence after P^i is 0.95. These probabilities are the averages from table 6.1 i.e. 0.88 is the average for (0.837, 0.862, … , 0.883) and 0.95 is the average for (0.957, 0.958, … , 0.935). Hence the probability of starting to pay off the second sequence is $P(P^2) = 0.727 * 0.95 * 0.88 = 0.61$.

The expected recovery amount from the first sequence $E(R^1)$ is the amount recovered per month (equation 6.2), multiplied by the number of months in the first sequence. The amount recovered per month is given below, let's call it M^1:

Equation 6.2 rewritten

Predicted amount recovered in the 1st payment sequence (M^1) = 10 ^ (1.20541 + average collected before default * 0.00025313 - default amount * 0.00015795 - loan amount * 0.00000236 + no children * 0.05552 + term * 0.00292 + Halifax * 0.00992 + net lending * 0.00016674)

The model for the conditional probability of stopping payment in P_1 is:

$P(NP_j|P_{j-1})$ = Probability of stopping payment in month j given that they have been paying and have not paid off by month j-1

$= -0.0128j + 0.2014$ for $j<7$

$= 0.11$ otherwise

Let's call the probability of paying in the i^{th} month of the first payment sequence q^1_i.

Equation 6.6

Therefore $E(R_k^1) = P(P^1)M_k^1\sum_{i=1}^{\infty} q_i^1 = 0.727 \times M_k^1 \times \sum_{i=1}^{\infty} q_i^1$

where *k* is for case *k*

So q^1_i = 1 –(-0.0128*(i -1) + 0.2014) for i<7 given that a payment was made in month i-1. Also $q^1_{i+1} = q^1_i$ (1 – 0.11) =0.89 q^1_i for i>6.

Therefore:

$q_1^1 = 1$, $q_2^1 = 1 \times (1 - (-0.0128 \times 1 + 0.2014)) = 1 - 0.2014 + 0.0128 = 0.811$,

$q_3^1 = 0.8114 \times (1 - (-0.0128 \times 2 + 0.2014)) = 0.8114 \times (1 - 0.2014 + 0.0256) = 0.8114 \times 0.8242$
= 0.669,

$q_4^1 = 0.669 \times (1 - (-0.0128 \times 3 + 0.2014)) = 0.669 \times (1 - 0.2014 + 0.0384)$
$= 0.669 \times 0.837 = 0.560$,

$q_5^1 = 0.560 \times (1 - (-0.0128 \times 4 + 0.2014)) = 0.560 \times (1 - 0.2014 + 0.0412)$
$= 0.560 \times 0.8498 = 0.476$
$q_6^1 = 0.476 \times (1 - (-0.0128 \times 5 + 0.2014)) = 0.476 \times (1 - 0.2014 + 0.054) = 0.476 \times 0.8626$
$= 0.410$
$q_7^1 = 0.410 \times (1 \text{-} (0.11)) = 0.410 \times 0.89 = 0.365$,

Hence,

$\sum_{i=1}^{6} q_i^1 =$ 1 + (1-0.2014+0.0128) (1+ (1-0.2014+0.0128*2) (1+ (1-0.2014+0.0128*3) (1+ (1-0.2014+0.0128*4) (1+ (1-0.2014+0.0128*5))))))

= 3.926

$\sum_{i=7}^{\infty} q_i^1 =$ 0.89 q^1_6 + 0.89^2 q^1_6 + … + 0.89$^\infty$ q^1_6 = $\frac{q_6^1}{1-0.89} - q_6^1$ =

= (1-0.2014+0.0128) (1-0.2014+0.0128*2) (1-0.2014+0.0128*3) (1-0.2014+0.0128*4) (1-0.2014+0.0128*5) (1/(1-0.89)-1) = 3.32

Thus $\sum_{i=1}^{\infty} q_i^1 = 7.25$

Therefore equation 6.6 becomes $E(R_k^1) = 0.727 \times M_k^1 \times \sum_{i=1}^{\infty} q_i^1 = 5.27 M_k^1$

Equation 6.7 (equation 6.2 substituted into equation 6.6)

$E(R^1_k)$ = 5.27 * 10 ^ (1.20541 + average collected before default$_k$ * 0.00025313 - default amount$_k$ * 0.00015795 - loan amount$_k$ * 0.00000236 + no children$_k$ * 0.05552 + term$_k$ * 0.00292 + Halifax$_t$ * 0.00992 + net lending$_t$ * 0.00016674)

Where *k* is for case *k* and *t* is for default date in case *k*

Moving on to the expected recovery amount from the second sequence $E(R^2)$

Equation 6.8

$$E\left(R^2_k\right) = 0.727 \times 0.95 \times 0.88 \times M^2_k \times \sum_{i=1}^{\infty} q^2_i$$

Where the amount recovered per month is M^2_k (equation 6.3 rewritten):

M^2_k = 10 ^(1.7253 -loan amount$_k$* 0.000000171252 +term$_k$* 0.00748 + net lending at default$_t$* 0.00009003)

And $P(NP_j|P_{j-1})$ = Probability of stopping payment in month j given that they have been paying and have not paid off by month j-1

= 0.6193 $j^{-1.32}$ for j<11

= 0.025 otherwise

So $q^2_i = 1 -(0.6193\ (i\text{-}1^{-1.32})$ for i<11 given that a payment was made in month i-1.

Therefore:

$q^2_1 = 1$, $q^2_2 = 1 \times \left(1 - \left(0.6193 \times 1^{-1.32}\right)\right) = 1 - 0.6193 = 0.3807$,

$q^2_3 = 0.3807 \times \left(1 - \left(0.6193 \times 2^{-1.32}\right)\right) = 0.3807 \times \left(1 - 0.248\right) = 0.2863$,

$q^2_{11} = q^2_{10}(1 - 0.025) - 1$=1 * (1-0.6193 *$1^{-1.32}$) (1-0.6193 *$2^{-1.32}$) (1-0.6193 *$3^{-1.32}$) (1-0.6193 *$4^{-1.32}$) (1-0.6193 *$5^{-1.32}$) (1-0.6193 *$6^{-1.32}$) (1-0.6193 *$7^{-1.32}$) (1-0.6193 *$8^{-1.32}$) (1-0.6193 *$9^{-1.32}$) (1-0.025) = 0.165555606

Hence:

$\sum_{i=1}^{10} q_i^2 =$ 1 + (1-0.6193 *1$^{-1.32}$) (1+ (1-0.6193 *2$^{-1.32}$) (1+ (1-0.6193 *3$^{-1.32}$) (1+ (1-0.6193 *4$^{-1.32}$) (1+ (1-0.6193 *5$^{-1.32}$) (1+ (1-0.6193 *6$^{-1.32}$) (1+ (1-0.6193 *7$^{-1.32}$) (1+ (1-0.6193 *8$^{-1.32}$) (1+ (1-0.6193 *9$^{-1.32}$)))))))))

= 3.057

$\sum_{i=11}^{\infty} q_i^2 =$ 0.975 q^2_{10} + 0.975^2 q^2_{10} + … + 0. 975$^{\infty}$ q^2_{10} = $\frac{q_{10}^2}{1-0.975} - q_{10}^2$

= (1-0.6193 *1$^{-1.32}$) (1-0.6193 *2$^{-1.32}$) (1-0.6193 *3$^{-1.32}$) (1-0.6193 *4$^{-1.32}$) (1-0.6193 *5$^{-1.32}$) (1-0.6193 *6$^{-1.32}$) (1-0.6193 *7$^{-1.32}$) (1-0.6193 *8$^{-1.32}$) (1-0.6193 *9$^{-1.32}$) (1/(1-0.975)-1)

= 6.622

Thus $\sum_{i=1}^{\infty} q_i^2 = 9.679$ = the expected number of months they would be paying

Therefore equation 6.8 becomes $E(R_k^2) = 0.61 \times M_k^2 \times \sum_{i=1}^{\infty} q_i^2 = 5.90 M_k^2$

The model to predict the recovery rate for all future sequences, where *j* is the sequence and *k* is case *k*, is as follows:

$$E(R_k^j) = \sum_{j=1}^{\infty} E(R_k^j) = \sum_{j=2}^{\infty} E(R_k^j) + E(R_k^j) = \sum_{j=1}^{\infty} 0.727 \times (0.95 \times 0.88)^{j-1} \times M_k^j \times \sum_{i=1}^{\infty} q_i^j$$

Hence $E(R_k) = 0.727 \times \sum_{j=2}^{\infty} (0.95 \times 0.88)^{j-1} \times M_k^j \times \sum_{i=1}^{\infty} q_i^j + E(R_k^1)$

$$= 0.727 \times \left(\frac{1}{1-(0.95 \times 0.88)} - 1 \right) \times M_k^j \times \sum_{i=1}^{\infty} q_i^j + E(R_k^1)$$

$$= 35.87 \times M_k^j + 5.27 * M_k^1$$

Equation 6.7 where equations 6.2 and 6.3 have been substituted in

E(R$_k$) =5.27*10 ^ (1.20541 + average collected before default$_k$* 0.00025313 - default amount$_k$ * 0.00015795 - loan amount$_k$ * 0.00000236 + no children$_k$ * 0.05552 + term$_k$ * 0.00292 + Halifax$_t$ * 0.00992 + net lending$_t$ * 0.00016674) +

35.87 * 10^ (1.7253 -loan amount$_k$ * 0.000000171252 +term$_k$ * 0.00748 + net lending at default$_t$ * 0.00009003)

Loan Amount in the equation 6.7 above is in the form of negative pence, since this was the format given in the data. In order to have all financial variables in the same form, E(R$_k$) can be rewritten as follows, where Loan Amount is in positive pounds.

E(R$_k$) =5.27*10 ^ (1.20541 + average collected before default$_k$ * 0.00025313 - default amount$_k$ * 0.00015795 + loan amount$_k$ * 0.000236 + no children$_k$ * 0.05552 + term$_k$ * 0.00292 + Halifax$_t$ * 0.00992 + net lending$_t$ * 0.00016674) + 35.87 * 10^ (1.7253 + loan amount$_k$* 0.0000171252 +term$_k$* 0.00748 + net lending at default$_t$* 0.00009003)

A real life example of calculating E(R) is given in table 6.19 below.

Variable	Debtor A	Debtor B	Debtor C	Debtor D
Average collected before default	£0.10	£2.26	£1,115.03	£0.94
Default amount	£2,498.38	£4,830.04	£8,730.06	£8,985.89
Loan amount	£2,500.00	£4,900.00	£6,500.00	£9,000.00
Loan term	48	60	36	36
No children	1	0	1	1
Halifax	0.2	-2.3	-1.7	5.4
Net lending	368	181	334	1190
Real amount recovered after default	£1,865.10	£4,104.68	£8,712.16	£7,529.42
Expected amount recovered	£6,555.72	£8,529.37	£6,337.46	£7,573.66

Table 6.19 real life examples from the hold out sample

Table 6.19 shows four real debtors from the holdout sample who all paid back part of their debt after default. The applicable variables to calculate their expected amount recovered after default for these debtors are shown above along with their real and predicted amounts recovered after default. As can be seen two out of the four where predicted to pay back over £3,000 more than they did in reality. This is partly due to the fact that the equation to calculate the expected amount recovered includes the assumption that there could be infinite payment sequences of infinite length. Therefore the expected recovery amount will be higher than the real recovery amount for most debtors, since the collectors will never allow this to happen.

Figure 6.15 shows the results of applying the expected recovery equation to the holdout sample. The vast majority of the debtors are estimated to pay back far more than they did. Looking at figure 6.15, on the other hand, which shows the results for estimating the amount recovered in the first sequence only, in the holdout sample. Here the reverse is true, the majority of debtors were under estimated. This is partly because there were quite a number of debtors who paid off the full amount in the first sequence and because within each sequence the estimated amount tends towards the mean.

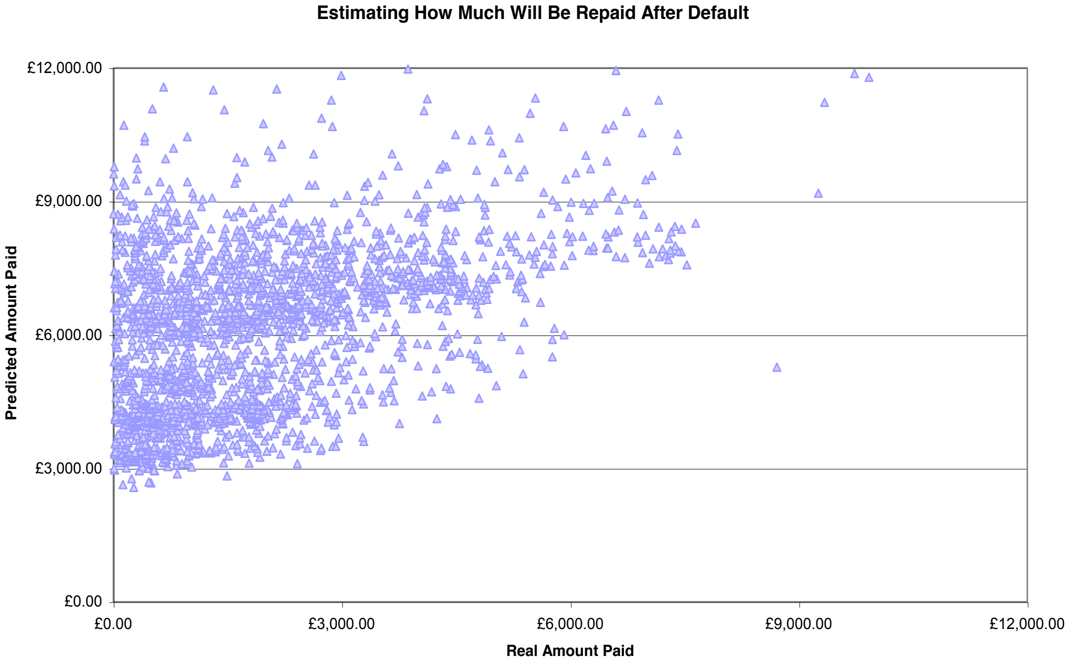


Figure 6.15, results of using the expected recovery amount for estimating the total recovery amount after default

Using Spearman's rank for expected recovery amount against the real recovery amount gives a 0.58. This is superior to the Spearman's rank for the third party predictions in chapter 4 which had a Spearman's rank between 0.4 and 0.3. The R-squared value for the holdout sample in this model is 0.21 which is nearly as good as the weight of evidence (WOE) approach discussed in chapter 3. In comparison this is a good result since the WOE approach was far less detailed, estimated the recovery rate and based on a limit two year period. The Mean Squared Error (MSE) for the model was 21,000,000. However all of the expected payments were far larger than the real payments. This is partly because the expected recovered amount assumes that there are infinite sequences, and each sequence can be of infinite length. Also there are no individual characteristics for determining the probability of starting a sequence or for the length of sequence. This is one aspect of the research, which could be looked into in future.

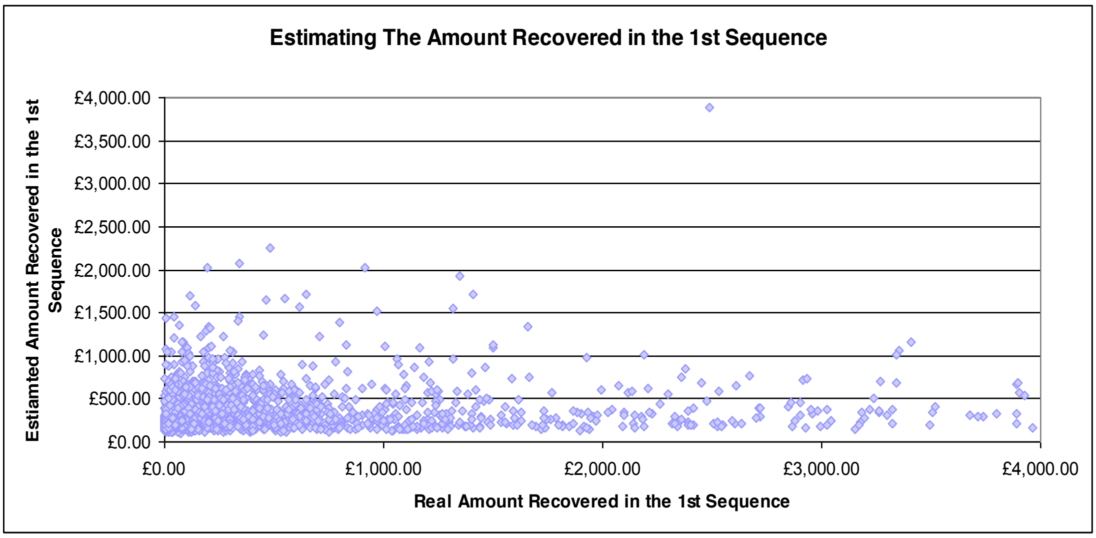


Figure 6.16, results of using the expected recovery amount for estimating the total recovery amount in the 1st payment sequence

This model resulted in R-squared = 0.008, and MSE=1,300,000. Since the R-squared = 0.05 for the model predicted the amount paid per month in the first sequence, this result while poor is not unexpected.

Therefore this model assumes that the average debtor will pay for five months in the first sequence and for a further 35 months for all following sequences. For most debtors this will be an over estimation. This model allows lenders to assess their write off policies by estimating the LGD for the different policies. The write off policy of the lender was unknown for this data set. However the lender didn't write off the debt until there had been no payments for over six months in all cases and often waited for years to be sure there would be no future payments. Therefore it can be assumed that if the debt was written off, the lender was fairly certain that they would not receive any further payments. The data can be used to test out the impact of different write off polices. These could include write offs after an agreed number of payments, non-payments or payment sequences. For example if the lender chose to write off the debt after the first payment sequence following default then the estimated LGD would be equal to:

$$\text{LGD after first sequence} = \frac{\text{Default Amount} - E(R_1)}{\text{Default Amount}}$$

All of these are practices by lenders but are not modelled here. This is one of the reasons for the over estimations. This is just one of a number of improvements which could be made to the model. Another expansion of this model could be to use the individual characteristics to estimate the length of the payment sequences and non-payment sequences. This could mean that the LGD could be estimated for any given length of time, e.g. the LGD 3 years after default.

Also as was previously demonstrated in figure 6.13, the more information collected after default like the first sequence payments leads to more accurate future predictions. However as would be expected the further into the future the predictions go the less accurate the prediction.

6.12 Summary

The payment patterns can be very useful for prediction models as they show how debtors pay back their debt after default. These models can be useful for not only predicting loss given default but also policy for collecting and predicting income from defaulted loans.

The amount recovered and the recovery rate for each sequence was dependent upon the length of the sequence as would be expected for any repayment plan. The length of the sequence was dependant upon the number of the sequence since the first non-payment and payment sequences were different from the others.

The expected recovery model assumes that all debtors have a 73% chance of starting to pay the first sequence. This has no individual characteristics, and neither do the probabilities for starting further sequences. This is one area that could be continued in future research.

The expected recovery model also assumes that all further payment sequences are the same, therefore they have the same probability of starting and the same amount will be recovered in each. This assumption is backed up by the data, but including some individual characteristic for determining the length of the payment sequence, might be applicable future research.

This model shows that modelling the payment sequences can in principle predict the expected recovery amount of the loan after default. This type of model may also be of more use to loan collectors than a simple regression to estimate LGD, because it allows the lender to estimate what would happen if different write off polices were implemented. Changing the write off policy would alter the predicted probability of starting each payment sequence. If the collector experimented in changing the write off policy with a few debtors and used this to estimate the probability of starting each payment sequence, then they could estimate the results of these changes within the model. This would mean they could assess the impact of the new policy after only a few sequences.

This model also shows that if a lender decided to write off the debtor after the first sequence, then there would be a potentially large loss of income from the debtor as table 6.14, and 6.15 demonstrates. These models could also be used to estimate the sale price for the debt no matter what sequence the debtor was in. And as figure 6.13 shows, once the debtor has started to pay back part of their debt the model can be improved by using the first sequence results to predict future payments more accurately.

Chapter 7 Conclusions

With debt on the increase, many consumers fail to pay back their debt. There are many factors contributing to customer delinquency. These include poor financial management skills, the economy and ease of access to loans and credit cards. When a debtor becomes delinquent for 180 days (FSA definition) then the loan is considered to be in default. The lender will try to collect the debt as soon as the debtor becomes delinquent. Then once the debtor defaults, the debt will get passed on to their in-house collections department who will try to collect the outstanding debt. However some companies use outside agents or will just sell off the debt. If the lender's collection department is unable to collect the debt, then they may also decide to use a collection's agency or just sell off the debt. The debt can be passed on several times, and can be collected up to six years after the last payment was made as stated in the Limitations Act of 1980 [40]. Debt collection agencies recovered $51 billion in 2005 [37].

The novelties of this research are that it looks at not only in-house collections but also compares them to third party recovery processes. Models for both in-house and third party LGD are calculated and discussed over similar time periods for real comparisons to be made. These models are also refined and improved; in the case of the in-house data set economic variables were included because the data was collected over different economic time periods. What is remarkably unique in this thesis is the use of payment patterns to predict the LGD of loans. This approach is far more flexible than other models because it can be used to not only calculate the final LGD but also the LGD at any given time.

Chapter 3 discusses the differences between debt that is collected in-house and debt that is collected by a third party. Although both analysed data sets are about debt recovery, the information available in each case is quite different and the average recovery rate varied from 5% to 46%. The two-stage model was appropriate for both, even though the spikes are at opposite ends of the LGD distribution. The in-house spike was at 0 indicating that a large proportion of debtors repaid everything whereas the third party spike was at 1 indicating that a large proportion of debtors repaid nothing. All of this is not surprising because third party debt will usually go through several collection processes, so by definition must be harder to collect.

What is remarkable about the models discussed in chapter 3 is that despite the in-house data set being more detailed, the goodness of fit of both was very similar. This is despite the third party model focusing on contact details and very few personal details. Whereas the in-house model focused more on loan characteristics; loan amount, time spent in arrears, lifetime of the loan.

Chapter 4 focused on predicting the recovery rate for third party collection over the 20-month time period. By splitting the debtors according to the amount of debt they owe the results of the models were far better than modelling the debtors as a whole. Only predicting the LGD for extra large debtors gave a poorer result than the linear regression model in chapter 3. The models for small and medium sized debt even managed to improve on the weight of evidence model.

The model created was a two-stage LGD predictor, using logistic regression to predict which debtors would have a LGD=1 and which would pay back part of their debt. To ensure the best classification of the debtors, those who achieved a result of less than 0.2 in the logistic regression model would have a predicted LGD of 1. The others would have their predicted LGD value estimated by using the linear regression model. Splitting at 0.2 meant that about 70% of debtors who paid were correctly classified and about 70% of debtors who did not pay were correctly classified.

Waiting until after the debtors had been in collections for at least 6-months gave better results for the logistic regression. That is not to say that the models should only be used after 6-months but rather that these models are for predicting the recovery rate after at least 6-months in collections.

For larger debts their predicted LGD was higher than for debts in smaller debt amount models. So for debts larger than £2000 none were predicted to pay back more than 35% of their debt.

Chapter 5 has looked at the effects of economic factors in debtors repaying their loans after they have defaulted. The in-house data set was ideal for testing economic variables on recovery rates, since it covers the loans' history during a recession, recovery and a period of stability.

When looking at the percentage of defaulters who pay back their loans each month, the economic variables were excellent at predicting how many will pay back. In particular, net lending was a very strong indicator. Net lending, GDP and house prices all had a strong positive linear relationship between them and, the percentage of debtors who are paying each month after default. Interest rates and unemployment had a negative relationship with the percentage of debtors who are paying each month after default.

When it came to predicting the recovery rates after the first 12 months, the debtors' behaviour during those first 12 months is an indicator for the following 12 and 24 months. On average it appeared that debtors were repaying around 5% of the default balance off each year. However the variability was very high where some debtors paid off everything in one payment and many failed to pay off anything.

Using economic variables for predicting for when debtors pay gave good results and for predicting how much debtors repay is also improved by using economic variables.

What is evident is that during the lifetime of a loan, economic conditions can vary wildly, especially as some loans can have debtors repaying even a decade after they have defaulted. This means that using the economic conditions at a certain point, e.g. at default, is not as useful as continuous monitoring within the models. Therefore in future research it would be useful to predict when debtors repay and use predicted economic conditions for each month to improve the predictions.

Both parties can use these models to determine the price at which to buy a debt if the lenders wish to sell. The third party model gives an indication of recovery rate so the third party can set an internal upper limit for the price of buying the debt. For the in-house collection; the question is how much more would they get by keeping the debt in their collection process for some further time? To get a feel for this one needs to estimate RR in the next year as covered in chapter 5.

All of these models are based on calculating the final LGD or the LGD after a predetermined time period. Chapter 6 discusses the advantages of a revolutionary LGD modelling approach. Once a debtor defaults on a loan they do not behave the same way as a non-defaulted debtor. Some payback all of their debt in one go, others never payback anything but the majority pay back what they can with instalments. These instalments are discussed with the collector, and often the lender describes these debtors as being “cured”. However these “cured” debtors do not stay “cured,” they stop paying again and again, causing the collector to renegotiate the instalments time and again. These instalments can potentially go on for years.

The payment patterns can be very useful for prediction models as they show how debtors pay back their debt after default. These models can be useful for not only predicting LGD but also policy for collecting and predicting income from defaulted loans.

The amount recovered and the recovery rate for each sequence was dependent upon the length of the sequence as would be expected for any repayment plan. The length of the sequence was dependant upon the number of the sequence (i.e. first, second, third sequence etc.). The first non-payment and payment sequences were different from the others.

The expected recovery model assumes that all debtors have a 73% chance of starting to pay the first sequence. This has no individual characteristics, and neither do the probabilities for starting further sequences. This is one area that could be continued in future research.

The expected recovery model also assumes that all further payment sequences are the same, therefore they have the same probability of starting and the same amount will be recovered in each. This assumption is backed up by the data, but including some individual characteristic for determining the length of the payment sequence, might be applicable future research.

The model in chapter 6 shows that modelling the payment sequences can in principle predict the expected recovery amount of the loan after default. This type of model may also be of more use to loan collectors than a simple regression to estimate LGD, because it allows the lender to estimate what would happen if different write off polices were implemented. Changing the write off policy would alter the predicted probability of starting each payment sequence. If the collector experimented in changing the write off policy with a few debtors and used this to estimate the probability of starting each payment sequence, then they could estimate the results of these changes within the model. This would mean they could assess the impact of the new policy after only a few sequences.

This model also shows that if a lender decided to write off the debtor after the first sequence, then there would be a potentially large loss of income from the debtor as tables 6.14, and 6.15 demonstrate. These models could also be used to estimate the sale price for the debt no matter what sequence the debtor was in. And as figure 6.13 shows, once the debtor has started to pay back part of their debt the model can be improved by using the first sequence results to predict future payments more accurately.

The main point of the payment pattern models was not to improve the current regression based models but to see if the approach is feasible. These finding show that the models are feasible and can be an improvement on the two-stage model also discussed in this thesis.

7.1 Further Research

As discussed research into consumer LGD is still in its infancy, therefore there is lots of potential further research. All of the models discussed in this thesis give results, which while on a par with other models in this industry are not terrific.

This is the first research to compare in-house and third party LGD results and recovery processes. This research could be expanded over a longer time period and if possible it would be very interesting to observe the complete history of some debtors after default. This would allow models to be created to calculate the LGD for the lender and the third party. Also assessing how the price at which the debt is sold is reflected in the LGD over time and when the third party begins to break even.

The third party models in this thesis are based on a snapshot of the debt. Further research could improve on these models by observing the debtors over a longer time period and at regular intervals to see how time affects the LGD.

The in-house models are more detailed, mainly because the data held by in-house collectors is far greater and the particular lender who donated the data set kept detailed records. The payment patterns model can be improved by more detailed research into the number of payment patterns, the length of the patterns and the length of the non-payment patterns. The probabilities for starting each of the sequences had no economic or individual characteristics. Neither did the length of the payment sequences. These could both be improved upon. The length of the non-payment sequences were not even included in the final model but only analysed in general. Creating a detailed model that includes the length of the non-payments as well as the payments means that the LGD at any time can be estimated. This would be very useful for collectors because they could estimate their potential income from defaulted loans of any given length; also all of the economic variables are from the default date. As the lifetime of some of these loans can be decades the economic situation can change radically. Knowing when the debtor will pay back means that the economic variables at these times could be predicted which may improve the models.

Appendix

Table A1, Logistic Regression Results (1st stage) for Third Party

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq	Standardised Estimate	Exp(Est)
Intercept	1	-1.8929	0.0932	412.34	<.0001		0.151
No Work Telephone	1	-0.1959	0.031	39.99	<.0001		0.822
No Mobile Telephone	1	-0.1582	0.0247	41	<.0001		0.854
Amount 100	1	0.6695	0.4723	2.01	0.1564		1.953
Amount 500	1	-0.00061	0.0951	0	0.9949		0.999
Amount 1000	1	-0.2258	0.0848	7.09	0.0077		0.798
Amount 1500	1	-0.0368	0.0886	0.17	0.678		0.964
Amount 2000	1	-0.0886	0.0959	0.85	0.3556		0.915
Amount 5000	1	-0.1469	0.0879	2.8	0.0945		0.863
Number of Telephones	1	0.6115	0.0247	615.19	<.0001	0.3196	1.843

Table A2, Logistic Regression Results (2nd stage) for Third Party

Label	*Coefficients*	*Standard Error*	*t Stat*	*P-value*
Intercept	0.23782	0.01928	12.33733	<.0001
Age 18-25	0.10969	0.01524	7.19650	<.0001
Age 25-35	0.06030	0.01324	4.55511	<.0001
Age 35-45	0.02765	0.01303	2.12251	0.0338
Age 45-55	0.00302	0.01376	0.21962	<.0001
Phone	0.13453	0.01826	7.36558	<.0001
Mobile	-0.05200	0.01037	-5.01253	<.0001
Default Amount	-0.00004	0.00000	-21.89901	<.0001
Owner	0.05380	0.01088	4.94404	<.0001

References

1. Allison, Paul.D., 2005, "Survival Analysis Using SAS A Practical Guide", SAS Institute Inc., Cary, NC, USA, pp. 248 ISBN 155544279X

2. Allred, Christopher, Hite, Kathryn,. Fonzone, Stephen,. Greenspan, Jennifer,. Larew, Josh,. Scherer, William,. Pomroy, Thomas,. Fuller, Douglas,. 2002, "Modelling And Data Analysis In The Credit Card Industry: Bankruptcy, Fraud, And Collections" IEEE Systems and Information Design Symposium, University of Virginia

3. Altman, Edward. I. and Kishore, Vellore. M. 1996, "Almost Everything You Wanted To Know About Recoveries On Defaulted Bonds", Financial Analysis Journal. Vol.52, No. 6, (November/December 1996), pp. 57-64.

4. Altman, Edward., Resti, Andrea and Sironi Andrea, Published London: Risk books 2005, “Recovery Risk : the next challenge in credit risk management”. ISBN: 9781904339502, 1904339506

5. Altman, Edward., Haldeman, Robert, and Narayanan, P. "ZETA Analysis: A New model to Identify Bankruptcy Risk of Corperations", journal of Banking and Finance Vol.1, Issue 1, June 1977, Page: 29-54.

6. Banasik, J L, and Crook, J N , 2005. “Explaining Aggregate Consumer Delinquency Behaviour over Time” Credit Research Centre, University of Edinburgh, Working Paper Series No 05/03

7. Basel Committee on Banking Supervision, June 2006, “International convergence of capital measurement and capital standards, A Revised Framework, Comprehensive Version”, ISBN print: 92-9131-720-9, ISBN web: 92-9197-720-9.

8. Basel Committee on Banking Supervision, June 2004, International Convergence of Capital Measurement and Capital Standards, A revised framework. ISBN print: 92-9131-669-5, ISBN web: 92-9197-669-5. Page 92-93, §Paragraph 452

9. Bastos, João, 2010, “Forecasting bank loans loss-given-default”, Journal of Banking & Finance, Volume 34, Issue 10, Pages 2510-2517

10. BBC "Debit card spending roars ahead", 3 July 2007, news.bbc.org.uk/1/hi/business/6265784.stm

11. BBC News 29th July 2004 quoting Bank of England figures for debt Retrieved on 3rd July 2007, from http://news.bbc.co.uk/1/hi/business/3935671.stm

12. Bellis, Mary, "Who invented Credit Cards?" www.inventors.about.com Retrieved on, 3rd July 2007, from http://inventors.about.com/od/cstartinventions/a/credit_cards.htm

13. Bellotti, T., and Crook, Jonathan., "Calculating LGD for credit cards”, QFRMC Conference on Risk Management in the Personal Financial Services Sector London 22-23 January 2009

14. Bellotti, Tony, and Crook, Jonathan., "Macroeconomic conditions in models of Loss Given Default for retail credit" Credit Scoring and Credit Control XI Conference, August 2009

15. Belyaev, Konstantin, and Aelita Belyaeva, 2009, "Application of survival analysis for LGD and EAD modelling", Czech National Bank, Credit Scoring and Credit Control XI Conference

16. Bennett, R.L., Catarineu, E., Moral, G., “Loss Given Default Validation”, Working Paper 14, Studies on the Validation of Internal Rating System. Basel Committee on Banking Supervision, Revised Version May 2005 p.60-93

17. Bermejo, Sergio,. Cabestany, Joan,. "Oriented principal component analysis for large margin classifiers", Neural Networks, Vol. 14, No. 10, December 2001. p.1447-1461

18. Bower, Ryan, Burkett, Matt, Jobs, Melaina, Marr, Brian, Scherer, William T., Pomroy, Thomas, Fuller, Douglas, Williamson, Jeff, 2001, "Automated Call Centre Analysis And Modelling".

19. Box, G.E., Cox, D.R., "An analysis of transformation", J. Royal Statistical Society. Series B, 1964, Vol.26, p.211-246

20. British Banking Association. Retrieved on 3rd July 2007 from www.bba.org.uk/bba/jsp/polopoly.jsp?d=149

21. Bruche, Max and González-Aguado, Carlos, “Recovery rates, default probabilities, and the credit cycle”, 2010, Journal of Banking & Finance, Volume 34, Issue 4, Pages 754-764

22. Caselli, S., Gatti, S., Querci, F., "The Sensitivity of the Loss Given Default Rate to Systemic Risk: New Empirical Evidence on Bank Loans" Journal of Financial Services Research, Vol. 34, No. 1, August 2008

23. Chin, A.G., Kotak, H., Improving the debt collection process using rule-based decision engines: a case study of Capital One, International Journal of Information Management, Vol.26, p.81-88 (2006)

24. “Civil Procedure Rules 1998”, No. 3132 (L.17) Supreme Court of England and Wales ISBN 0 11 080378 7.Retrieved on 3rd of July 2007 from http://www.opsi.gov.uk/si/si1998/19983132.htm

25. Consumer credit 1992-2002: Annual Abstract of Statistics Office for National Statistics. Retrieved on 3rd July 2007 from http://www.statistics.gov.uk/statbase/ssdataset.asp?vlnk=4925&More=Y

26. Credit Service Association frequently asked questions retrieved on the 3rd of July 2007 from www.csa-uk.com/csa/faq.php

27. De Servigny, Arnaud and Renault, Oliver, 2004., "Measuring and Managing Credit Risk" Published 2004 by The McGraw-Hill Companies, Inc. ISBN 0-07-141755-9

28. Dermine, Jean and NetoNeto Dde Carvalho,Cristina, 2006, "Bank loan losses-given-default: A case study," Journal of Banking & Finance, Vol.30, Issue 4, April 2006, p. 1219-1243

29. EconStats (VZRD), published 2010. Retrieved on 3rd July from http://www.econstats.com/uk/uk_md_____59m.htm

30. Engelmann, B., and Rauhmeier, R., 2006 "The Basel II Risk Parameters: Estimation, Validation, and Stress Testing" ISBN-10 3-

540-33085-2 Springer Berlin Heidelberg New YorkDe Servigny and Oliver, 2004, "Measuring and Managing Credit Risk"

31. Figlewski, Stephen, Liang, Weijian. and Frydman, Halina, 2008, "Modelling the effect of macroeconomic factors on corporate default and credit rating transitions" 5th September, 2006. NYU Stern Finance Working Paper No. FIN-06-007
32. Freeman, Jade, and Modarres, Reza. “Inverse Box-Cox: The Power-Normal Distribution” U.S. Environmental Protection Agency. Statistics & Probability Letters. Volume 76, Issue 8, 15th April 2006, p.764-772.
33. Grieb, T, Hegji, C, and Jones, S T., 2001, “Macroeconomic factors, consumer behaviour, and bankcard default rates”, Journal of Economics and Finance, 25(3), Fall, 316-327.
34. Gupton, G.M., Stein, R.M., "LossCalc v2; Dynamic Prediction of LGD", Moody KMV Company. modelling methodology. Jan 2005
35. Hillebrand, Martin, "Modeling and estimating dependent loss given default", September 2006 working paper version retrieved on the 3rd July 2007 from http://www-m4.ma.tum.de/pers/mhi/
36. Huang, Xinzheng, and Oosterlee, Cornelis. W., "Generalized Beta Regression Models for Random Loss-Given-Default" 9th September 2008, retrieved on the 8th of October 2009 from Citeseerx.ist.psu.edu
37. Hunt, Robert M., Collecting Consumer Debt in America. Business Review Q2 2007. Retrieved on 3rd July 2007 from www.philadelphiafed.org/econ/br/index.html
38. Lehmann, E. L., and Casella, George. "Theory of Point Estimation" (2nd ed.). New York: Springer. 1998. MR1639875. ISBN 0-387-98502-6
39. Leow, Mindy, Mues, Christophe & Thomas, Lyn, “Loss Given Default (LGD) Modelling For Mortgage Loans”, 2009, Credit Scoring Conference
40. Limitation Act 1980, The UK Statute Law Database, Part 1, paragraph 6

41. Lucas, A., "Basel II Problem Solving"; QFRMC Workshop and conference on Basel II & Credit Risk Modelling in Consumer Lending,. 6-8th September 2006. Retrieved on the 8th July 2007 from http://www3.imperial.ac.uk/mathsinstitute/programmes/research/bankfin/qfrmc/events/past/basel2

42. Makuch, W. M., J. L. Doge, J. G. Ecker, D. C. Granfors, G. J. Hahn, 1992, Managing consumer credit delinquency in the US economy: a multi-billion dollar management science application, Interfaces, 22 (1), 90-109

43. O'Sullivan, Arthur, and Sheffrin, Steven M,. "Economics: Principles in action", Upper Saddle River, New Jersey. Second edition Jan 2002. Pearson Prentice Hall p.339 ISBN 0-13-063085-3

44. Qi, Min and Yang, Xiaolong, "Loss Given Default of High Loan-to-Value Residential Mortgages", Journal of Banking & Finance, 2009, 33, 788-799.

45. Qi, Min and Zhaoa, Xinlei, "Comparison of modelling methods for Loss Given Default", Journal of Banking & Finance, 2011, Article in Press, Corrected Proof

46. Querci, F. Loss given default on a medium-sized Italian bank's loans: an empirical exercise. (2005). The European financial management association, Genoa, Genoa University. http://www.efmaefm.org/efma2005/papers/206-querci paper.pdf.

47. Somers, M., Whittaker. J., "Quantile Regression for Modelling Distributions of Profit and Loss" European journal of operational research, Vol.183, Issue 3, 16th December 2007, p.1477-1487

48. Steel, R. G. D., and Torrie, J. H., 1960, "Principles and Procedures of Statistics: with special reference to the biological sciences", New York: McGraw-Hill, p. 187, 287

49. Sullivan, Gary, 2009, "Forward-looking Odds: Incorporating economics into credit scoring"

50. Talbot, Richard, July 2010, Credit Action Debt Statistic, Retrieved on the 16th November 2010, http://www.creditaction.org.uk/debt-statistics/2010/july-2010.html

51. The Insolvency Act, 1986, retrieved on the 16th November 2010, http://www.governmentdebtassistance.co.uk/?gclid=CMa7mY3RraMCFQ6ElAodFgdM6A

52. Thomas, Lyn C., 2009, "Consumer Credit Models: Pricing, Profit and Portfolios", OUP Oxford, ISBN 978-0199232130

53. Thomas, Lyn, Christophe Mues and Anna Matuszyk, , "Modelling LGD for unsecured personal loans: Decision tree approach" JORS 2010 paper p.393-398

54. Thomas, Lyn, and Zhang, Jie, “Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD”, 2011, International Journal of Forecasting, Article in Press, Corrected Proof

55. Thomson, Sweet & Maxwell, 2007, Civil Procedure (The White Book) ISBN 1847031587

56. Whitley, J, Cox, P, and Windram, R, 2004, ‘An empirical model of household arrears’, Bank of England, Working Paper no. 214.

57. Wikipedia: credit card entry, retrieved on the 16th November 2010, en.wikipedia.org/wiki/credit_card

58. www.bankofengland.co.uk/publications/other/monetary/ConsumerCreditJanuary2010.xls retrieved on the 16th November 2010,