UNIVERSITY OF Southampton

Copyright © and Moral Rights for this thesis and, where applicable, any accompanying data are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis and the accompanying data cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content of the thesis and accompanying research data (where applicable) must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holder/s.

When referring to this thesis and any accompanying data, full bibliographic details must be given, e.g.

Thesis: Author (Year of Submission) "Full thesis title", University of Southampton, name of the University Faculty or School or Department, PhD Thesis, pagination.

Data: Author (Year) Title. URI [dataset]

University of Southampton
Faculty of Business and Law
Southampton Business School

Applications of machine learning in consumer credit risk modelling

by
Trevor Fitzpatrick

A thesis submitted for the degree of
Doctor of Philosophy

June 2020

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF BUSINESS AND LAW

SOUTHAMPTON BUSINESS SCHOOL

Thesis for the degree of Doctor of Philosophy

APPLICATIONS OF MACHINE LEARNING IN CONSUMER CREDIT RISK MODELLING

by Trevor Fitzpatrick

This thesis investigates three separate type of prediction problems, in differing contexts, with a common theme of experimental comparison of standard methods with more advanced machine learning methods. The objective is evaluation of the predictive power of machine learning methods through experiments on real world data.

The first paper is an application of machine learning classification methods to predict mortgage arrears. It finds that both machine learning and a flexible statistical model outperform standard approaches. This can help identification of important predictive factors for the management of loan arrears within banks and loan servicers.

The second paper applies both regression and classification methods to prediction of Peer to Peer (P2P) loan returns and default using different types of information.The main findings are that linear methods perform well on several (but not all) criteria; whether machine learning ensemble methods perform better than individual methods depends on the performance measure used to assess them. Use of alternative text-based information does not improve predictive outcomes. As a consequence, investors can be more informed about investments in this market.

The third uses survival analysis to predict time to sale of property collateral used for mortgage loans. When property sales occur, as separate set of statistical and machine-learning models are used to predict the haircut or discount between the indexed property valuation at the point of sale and the actual transaction price. Random survival forests worked well to predict the time to sale; while deep learning, random forests, and neural network regression methods performed best predicting the discount. Based on predictive models for these two parameters, a sensitivity analysis illustrated how predictive modelling of these parameters produces more conservative (i.e., higher) loss estimates than one current industry approach.

Table of Contents

List of Figures

List of Tables

Declaration of Authorship

I, Trevor Fitzpatrick, declare that this thesis entitled *Applications of machine learning in consumer credit risk modelling* and the work presented in it are my own and have been generated by me as the result of my own original research.

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:
    - Fitzpatrick, T., Mues, C., 2016. An empirical comparison of classification algorithms for mortgage default prediction: evidence from a distressed mortgage market. *European Journal of Operational Research*, 249, 427–439.

Signed:

Date:

Acknowledgements

I would like to express my appreciation and thanks to Professor Christophe Mues for his advice, encouragement, and support over the course of this work. I am grateful to other members of my supervision team Professor Mee Chi So and Dr. Kasia Bijak for their advice. I would like to thank Professor Cristián Bravo and Professor Taufiq Choudhry for being my internal readers. I would also like to thank Professor Stefan Lessmann for being my external examiner.

I am grateful to the team in the Central Bank of Ireland for the flexibility in allowing me to complete this work. A special thanks goes to the open source software community for providing such powerful tools to advance data science.

I would like to thank my family for their support, patience, and understanding.

Most of all, thanks to Mary for her love and invaluable support during "our Phd" studies and to our daughter Hazel for being her wonderful self.

Definitions and Abbreviations

ABS Asset-Backed Securities
AutoML Automated Machine Learning
AFT Accelerated-Failure Time
ANN Artificial Neural Networks
AUC Area Under the receiver operating characteristic Curve
BRT Boosted Regression Trees
CART Classification and Regression Trees
CV Cross Validation
ECL Expected Credit Loss
FSD Forced Sale Discount
GAM Generalised Additive Model
GCV Generalised Cross Validation
GFC Global Financial Crisis
IBS Integrated Brier Score
IPCW Inverse Probability of Censoring Weights
IRB Internal Ratings-Based
IRR Internal Rate of Return
KM Kaplan-Meier
kNN k-Nearest Neighbours
LC Lending Club
LGD Loss Given Default
LMM Linear Mixed Model
LR Logistic Regression
LTV Loan to Value
MAE Mean Absolute Error
MARS Multivariate Adaptive Regression Splines
MDP Markov Decision Process
MSE Mean Squared Error
NDCG Normalised Discounted Cumulative Gain
NFL No Free Lunch
NPL Non-Performing Loan
OLS Ordinary Least Squares
OR Operations Research
PD Probability of Default
PEC Prediction Error Curves
P2P Peer to Peer
RF Random Forests
RL Reinforcement Learning
RLR Regularised Logistic Regression
RSF Random Survival Forests
RLMM Robust Linear Mixed Model
RMBS Residential Mortgage-Backed Securities

ROC Receiver Operating Characteristic Curve
SME Small and Medium-sized Enterprises
SSM Single Supervisory Mechanism
SVM Support Vector Machine
TTS Time to Sale

Chapter 1

Introduction

Banks and providers of credit have been building credit risk models since the 1950s (Thomas, 2009). At first, the purpose of these models was to automate decisions on credit applications - hence their name, "application scorecards". After this, in the 1980s, behavioural scoring (i.e., using customer payment behaviour) was introduced to determine, among other things, whether credit limits could be extended or renewed. As noted by Thomas (2009), the mid-2000s saw a third wave of credit risk models related to optimising not just application or renewal decisions, but optimising several business criteria (acceptance, interest rate, limits) in profitability scoring, including a component related to credit risk (Finlay, 2010). Linear logistic regression or their variants have been the method of choice for these problems. The potential for improvements in predictive performance, allied with increasingly affordable and scalable computational power, availability of data of various types (structured, text, images), and the success of machine learning in computer vision and language translation, has led to a deepening interest in machine learning for credit risk modelling. This thesis, will therefore investigate its usefulness in three different settings – mortgage default prediction, P2P loan profit scoring, and recovery modelling.

Machine learning methods (in particular, supervised learning methods) will be more precisely defined in Section 1.1 but can be regarded as methods that learn prediction functions inductively from data. The potential improved predictive performance from machine learning has to be balanced with choosing the right algorithms for the task at hand, building the models appropriately, and comparing them to existing methods in a suitable manner. Other challenges include the computational complexity of the algorithm (how it scales with the number of observations, predictors, and the resulting run-time) and their ability to deal with imbalanced data (i.e., data where observations of the event of interest (default) are very small in comparison to observations with non-events (no default)).

Each of the papers in this three paper thesis investigates a separate type of credit risk prediction problem, the common theme being the experimental comparison of standard methods with more advanced machine learning methods. The objective of the research in this thesis is the evaluation of the predictive power of machine learning methods compared to standard statistical methods, through conducting experiments that replicate the conditions that modellers face in applied work in industry.

This introductory chapter of this thesis provides some background and context to the three papers contained in this thesis. Section 1.1 of this introduction provides an overview of the main types of machine

learning. Section 1.2 explains why the evaluation of machine learning methods is experimental. To provide context for the application domains of the papers, a brief introduction to consumer credit risk is provided in Section 1.3. This is followed by machine learning's relevance for credit risk in Section 1.4. This is followed by an overview of the three papers contained in the subsequent chapters. This final section describes the layout of the rest of the thesis.

1.1 Overview of machine learning concepts

Machine learning is a form of inductive learning from data. A precise definition of machine learning is given by Mitchell's book (Mitchell (1997), p.2)

> A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

These three components can be thought of as representing the class of tasks, evaluation of performance through using a performance measure, and optimising to improve based on experience. These are what Domingos (2012) terms the three components of learning: representation; evaluation; and optimisation.

There are a wide variety of machine learning methods and new ones are being created every day. There are three main types of machine learning. These are supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the goal of inductive learning is to learn a function or mapping from an input to an output or target. More formally, the aim of the learning process is to learn the relationship between a set of variables $\boldsymbol{x}$ that have an influence on the target y, given N instances of labelled input-output pairs from the training data $D = \sum_{i=1}^{N} (\boldsymbol{x_i}, y_i)$ (Murphy, 2012).

Unsupervised learning refers to finding patterns or groupings of the data for a specific purpose. The key difference compared with supervised learning is there are no labelled instances. Given $D = \sum_{i=1}^{N} \boldsymbol{x_i}$, the task is to find patterns or groupings for a specific purpose. These types of techniques include clustering, data summarisation, and density estimation (Rudin and Carlson, 2019).

The third type is reinforcement learning. The central idea of this method is that goal directed actions to maximise a particular reward measure. Unlike supervised learning, the method or learner is not instructed on the actions to take. There is an indirect link to operations research, where a specific version of this goal directed action can be framed as a Markov Decision Problem (MDP). In reinforcement learning, rather than specify the theoretical model for complex systems required in an MDP formulation, the learner discovers the actions leading to most reward by trying the actions and learning from what was successful or unsuccessful. In some types of reinforcement learning, actions may affect both the immediate reward, the next step including all subsequent rewards. These two characteristics - the trial-and-error search and delayed reward - are the main distinguishing features of reinforcement learning (Bishop, 2006; Barto and Sutton, 2018).

This thesis is concerned with supervised learning. Within this type of learning, there are two main types. These are classification - where is goal is predict a class/label or a 'yes/no' answer, that can output a probability. For example, predicting the probability of a disease or whether a loan will default. The second type is regression. These algorithms predict a real-valued output. Two examples are predicting the expected loss or profit on a loan or prediction of the sale price of a house. The methods used in

regression algorithms are similar to classification with differences in the loss function being minimised and performance measures used to evaluate the models.

A type of supervised learning regression is survival analysis. This is concerned with predicting a continuous variable where the outcome is the duration until the occurrence of an event. These time to event data differ from normal regression outcomes as the survival time for certain observations are incomplete. These are known as censored observations. For example, in a credit risk data set of 100 loans where outcomes for 100 loans are recorded over twenty-four months, and 1 loan defaults per month, then 66 loans are censored at the end of the study period.

The three application areas in this thesis determine the type of supervised learning method used to answer the research question. The first paper is an application of machine learning classification methods to predict mortgage arrears. The second paper applies both regression and classification methods to predict Peer to Peer (P2P) loan returns. The third uses survival analysis to predict time to sale of property collateral used for mortgage loans. When property sales occur, as separate set of statistical and machine-learning models are used to predict the discount between the indexed property valuation at the point of sale and the actual transaction price.

1.2 Why is machine learning evaluation experimental?

It was stated in Section 1 that machine learning is inductive learning from data. In a typical machine learning process, a model is applied to training data and then predictions are made on unseen or test data. The ability to categorise correctly or predict with a low degree of error based on new examples separate from those used for training is known as generalisation.

To use a model in practice, at least three steps can be taken. First, some algorithms or methods must be chosen from all of the possible methods. Second, once chosen, the methods performance can be optimised given the data. Finally, an evaluation step involves a performance comparison of the algorithms in the given domain on test or unseen data.

1.2.1 Choice of algorithm

Because learning is inductive, it is not clear how best to match algorithms to problems. One of the implications of the past three decades of research in machine learning and optimisation theory, implies that there is no one method that is universally the best given the context and problem at hand and the data used to investigate such problems (Alpaydin, 2016).

The detailed reasoning behind this was developed in a series of papers known as the No Free Lunch Theorems (NFL) by Wolpert (1996), Wolpert and Macready (1997), and Wolpert (2002). The Wolpert and Macready (1997) paper (p.2) set out to formally analyse how algorithms could be matched to data from various problem domains. The central point of these papers is that the average performance of algorithms across problem domains is the same. When one algorithm performs better than others in one domain, it may perform worse than comparators in other application domains. The implication is that no method could, therefore, be expected to be uniformly successful across all problem domains.

The papers generated substantial follow-up work and discussion in the machine learning community. Theorists and practitioners argued that the theorems were not relevant in practice as they did not for-

mulate them based on the *expected* generalisation performance, i.e., how well the methods would perform on functions not yet seen. Giraud-Carrier and Provost (2005) argue some functions are in reality more likely and there is some either explicit or implicit knowledge or assumptions that can be used to build learning algorithms. With those assumptions, researchers develop and apply general purpose algorithms that perform well.

However, it is a challenging problem to determine the properties of a dataset that makes one algorithm more appropriate than another. Kalousis et al. (2004) try to link performance between algorithms and datasets through the use of the dataset characteristics. They find that data size/dimensionality, class distribution, and information content of the predictors mattered for algorithm performance, given the data.

Indeed, in some important benchmark studies of classifiers like Fernández-Delgado et al. (2014) and cost-sensitive boosting Nikolaou et al. (2016), comparing across domains using the UCI datasets (Dua and Graff, 2019), a few algorithms (parallel random forest, cost-sensitive Ada-boosting) performed best given the various criteria considered. However, the authors note that many methods did not perform well across datasets from different domains. These studies are more focused versions of critiques by Rudin and Carlson (2019), Hand (2006) on the concentration on algorithm development, over-emphasis of specific methods, and whether the improvements are meaningful in practice. An extension of the meaningful improvement point is their relevance to real world impact for specific domain problems Wagstaff (2012).

Gómez and Rojas (2016) conducted a series of experiments to test the practical implications of the NFL theorems. They argue, based on their findings, that some methods appear to work better than others on a collection of real world datasets. They note the importance of how the learner forms its representation, the structure of the data (if there are a few important or many noise predictors), and preprocessing as all being important factors in having good average performance on the data they considered.

Instead of testing many different algorithms performance on datasets from several domains, another perspective is to try to determine the datasets that may match a particular algorithms. This is the approach developed by Eugster et al. (2014). They used statistical and information-theoretic measures, combined with recursive partitioning of preference ranking models (Bradley and Terry, 1952) to analyse which data sets 'prefer' certain algorithms. Traditional Bradley-Terry models assume that the preference of subjects (datasets) are the same across objects (classifiers). In this framework, if an algorithm performs well on a dataset, the data set 'prefers' that algorithm. Recursive partitioning can be used to group subjects with homogenous preference scalings in a consistent data-driven manner. A benefit of this approach is it provides some insight into the impact of dataset characteristics on performance of algorithms.

However, work by Montanez (2017a) and Montanez (2017b) suggest the favourable search space for algorithms to perform well is limited and that matching of problems to algorithms is provably difficult. Given algorithms can only perform well on a narrow subset of problems, novel algorithms are required for problem domains not covered yet or flexible algorithms that can adapt through parameterisation such as deep learning. Using external information (prior knowledge about the domain or the methods performance on similar domains) improves the chances of useful matches between algorithms and prob-

lems. Montanez (2017a) offers this as one explanation of the proliferation of algorithm development.

Because there is no universal method or master algorithm (Domingos, 2018), for a given domain data set, one specific method may perform best, but other methods could perform better on a similar but separate data sets from the same domain. As noted in Hastie et al. (2009) selecting the best approach can be one of the most important and challenging parts of fitting statistical and machine learning methods in practice.

To see why this is the case, consider a simple regression problem where the goal is to find a method approximating the function that minimises a loss function of the expected Mean Squared Error (MSE).[1] In this simple regression problem, $Y = f(X) + \varepsilon$, with $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$. Here, the Expected Prediction Error (EPE) using the MSE criterion is $EPE(x) = E[Y - f(x))]^2$ it can be shown that the prediction error on test data x_0 is given by

$$\left(y_0 - \hat{f}(x_0)\right)^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\varepsilon) \tag{1.1}$$

Variance is the amount by which $\hat{f}$ changes if estimated using differing training data. Bias is the error introduced by approximating the true unknown function by a simpler linear regression model. Equation 1.1 says that to minimise expected test error, one needs a low bias and low variance method (James et al., 2014). If the model is flexible and can become more complex - for example a spline rather than linear regression - it is more flexible in its use of the training data. This decreases bias but increases variance. The cost of additional flexibility is increased risk of *overfitting* the data where the model is reflecting random statistical noise rather than information from the predictors. This is known as the *bias-variance trade-off*.

1.2.2 Optimisation given data domain and algorithms

The discussion in this section has focused on the selection of an algorithm or method. Once a method or group of methods have been selected, the variables or parameters of the method are optimised with respect to the selected loss function Bennett and Parrado-Hernández (2006). These are learned during the training process.

Unless the model is a naïve benchmark, it will usually have tuning parameters. The role of these parameters is to limit overfitting. Depending on the model, this can be through influencing smoothness, adding some bias, limiting the number of logical conditions, covariates, or the type of models that the algorithms can use (Rudin and Carlson, 2019). Using the notation of Hastie et al. (2009) the model predictions depend on parameter(s) α with the best prediction $\hat{f}_\alpha(x)$ the one that maximises or minimises the loss function and associated performance measure.

The performance of the method on a given domain or across domains is determined by the domain data and any hyper-parameters (referred to above as α). Hyperparameters can be thought of as the range of settings specific to the method that can be tuned to optimise performance. These parameter values or range of values are fixed in advance.

[1] The expected MSE is mean test MSE that results from repeatedly estimated the function using large number of training sets and calculated on a test observations.

The various algorithms used in the applications are described in Chapters 2, 3, and 4. While an in-depth taxonomy of machine learning methods is outside the scope of this thesis, there are various ways to characterise types of learning algorithms. One taxonomy for various types of supervised learning methods is by Rudin and Carlson (2019). They outline:

- Models based on logical conditions or rules like decision trees or rule based models
- Linear combination (i.e., a sum) of decision trees (boosting, random forests)
- Case-based reasoning (k-Nearest Neighbours) and kernel-based methods (Support Vector Machines)
- Iterative summarisation (neural networks including deep learning)

Methods like decision trees are based on logical conditions or rules that partition data based on "if-then" conditions. This type of recursive partitioning of the predictors to predict the response results in predictions from the tree that are the average of the terminal nodes in the decision tree. Splits or partition points are determined by choosing the split point minimising some loss function condition such as the greatest reduction in the sum of the squared errors.

Decision trees form the basis for more complex machine learning methods such as random forests or boosting. In these methods, decision trees are used as the input or base learner, and many trees are grown and combined in different ways. These are known as ensemble methods. For random forests, many multi-level or deep trees are grown, on subsamples (i.e., rows) of the data and variables (columns), and the result is an average prediction over many trees. In a boosted regression tree model, trees with one or two splits are fit sequentially to the residuals of the previous tree, building up an additive model.

Case-based reasoning like k-Nearest Neighbours (kNN) predict the response by taking weighted average of the k-Nearest Neighbours in the data, where proximity is measured using a distance function like Euclidian distance. In a regression context, Support Vector Machines (SVMs) are based on minimising a loss function that includes a cost term for minimising large residuals. This method can flexibly specify how predictors enter the model either linearly or in a more general non-linear way using various types of kernels such as a Radial Basis Function.

Finally, in the simplest type of neural network models, the output variable is based on the input variables after they have been processed through a set of unobserved variables called hidden units or nodes. Each hidden unit is a linear combination of some or all of the input variables. These are transformed by a nonlinear activation function such as the logistic function. Several of these units can make up a layer, and neural network can have several layers, with the output from the first hidden layer being passed to the next hidden layer and so on. In the final step, these hidden layers are related to the outcome through another linear combination. In a regression setting, this is a non-linear regression model that is optimised by using specialised algorithms such as back-propogation to iteratively converge to the optimal fit.

Deep learning methods can be thought of as neural networks with much more elaborate architecture (the layers, activation functions) and optimisation methods applied to train the networks. The types of layers used depend on the data being represented. For example, Convolutional Neural Networks (CNNs) are a popular choice if the inputs are imaging data. More layers leads to a richer but a more complex representation of the data. Activation functions determine how the hidden layer transmits data

to the next layer; in deep learning models two popular methods are rectified linear or tan-h and similar to the types layers, the choice of activation function depends on the task at hand. The optimisation methods are adapted or developed specifically for these networks. Two of the most prevalent types of optimisation methods are Stochastic Gradient Descent based on small batches of training samples and Adaptive Moments Algorithm (ADAM) (Kingma and Ba, 2014).

Inspired by a figure from Rudin and Carlson (2019), Figure 1.1 is a visualisation of intended to provide an intuition of how these different types of methods work. This two-variable illustration of a non-linear relationship suggests each group of supervised learning methods produces different predictions in different ways. The data measure two aspects of air quality. Both the response variable (Temperature) and the predictor variable (Ozone) are scaled to lie between 0 and 1.[2] The top row of the figure shows a decision tree, random forests, and gradient boosted trees as two types of ensemble methods. The second row of the figure illustrates the two types of knn and a Radial Basis Function SVM. The third row includes a multi-layer perceptron with a logistic activation function and two deep learning methods with two hidden layers using the rectified linear and tanh activation functions.

The predictions for decision trees suggest they approximate the non-linear nature of the relationship by coarse steps where the predictions stay constant for a large range of the temperature variable. Random forests average the predictions of many individual decision trees to approximate the relationship and thus it is smoother than a single decision tree. The boosted regression trees produce predictions that are similar to both at different points on the range of the x-axis. The knn1 is a prediction based on one nearest neighbour which appears to overfit the data suggesting high variance; using 10 nearest neighbours produces a smoother curve. The SVM produces a fit that is somewhere in between each of the kNN predictions. Finally, the neural network and deep learning methods produce similar fits even though they have differing activation functions. Overall, this figure gives the impression that differing learners produce a range of different predictions depending the method. Some of these predictions are similar, and, in other cases like k-NN, one changed hyper-parameter setting leads to significant differences in fitting the data.

This leads to the question of how to tune the hyper-parameters to optimise performance for each method? The simplest approach to find the best set of hyper-parameters is to search over a grid of all possible combinations. The size of the search space depends on the number of parameters and their range, with each search taking place over every combination of values. With the development of methods with more complex adaptable structures, an active area of research is to find useful search strategies for optimising several hyper-parameters as the search space can be prohibitive or impossible given finite computing resources. This is one of the motivations for other search strategies such as random search as developed by Bergstra and Bengio (2012) and Bergstra et al. (2013). This has led to a variety of other approaches such as learning curves (van Rijn et al., 2015), and those summarised in table 2 of Luo (2016).

Performance measures of learned models are evaluated by one or more criteria. The criteria chosen to evaluate the model are informed by the purpose of the prediction and its type. For example, Area Under Curve (AUC) measures are useful for binary classification tasks. However, they implicitly assume that the costs of incorrect predictions (false positives and false negatives are the same). It is important

[2]The dataset is included in the R package *datasets*. It measures air quality in New York between May 1, 1973 and September 30, 1973. The data consist of six variables Ozone, Solar Radiation, Wind,Temperature, Day, Month. Ozone is measured in mean parts per million; Temperature is measured in Fahrenheit.

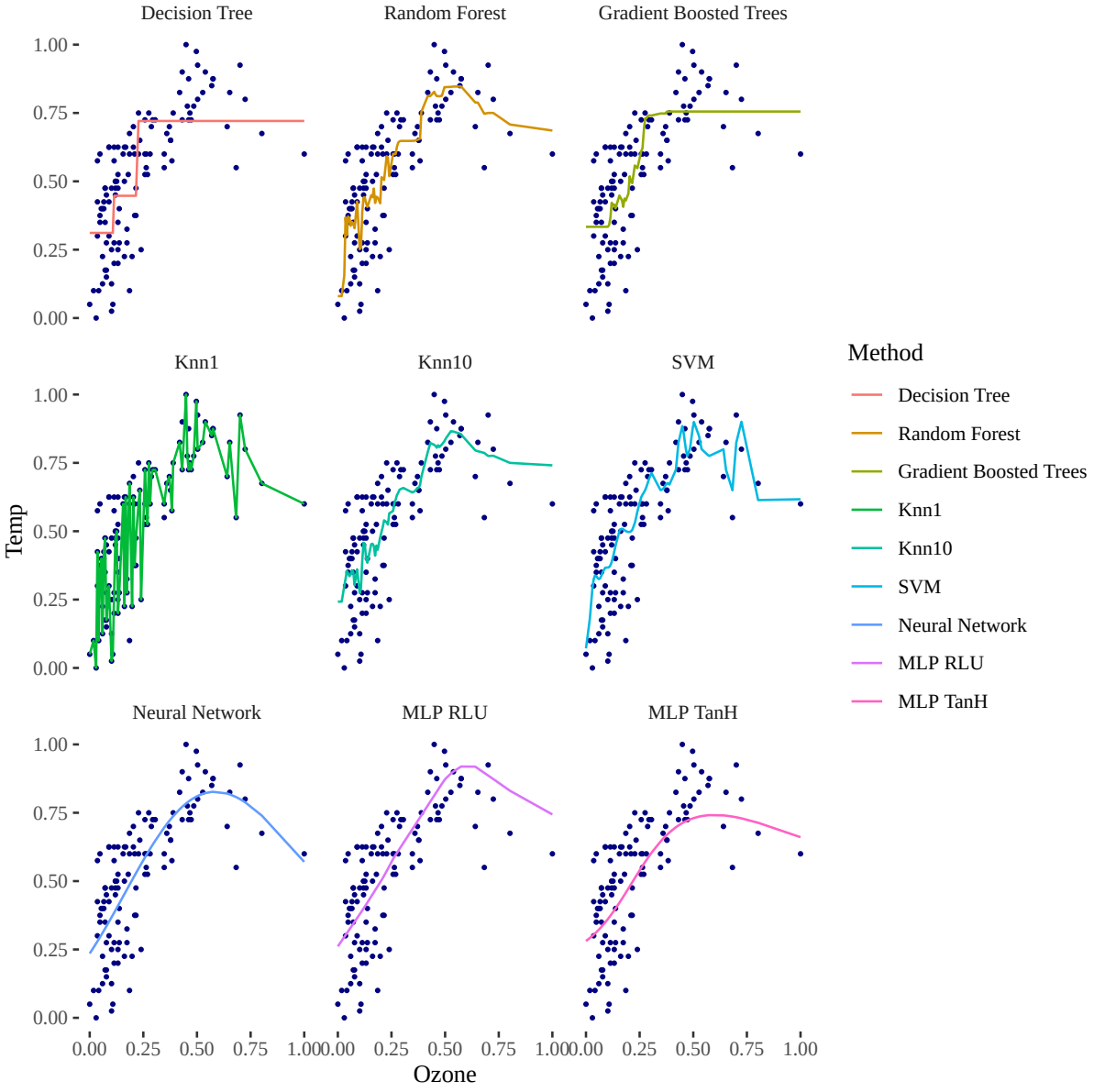

Figure 1.1: Illustration of machine learning method predictions on the Ozone dataset

to ensure that the performance measures is suitable for the overall experimental objective Flach (2019) and the type of prediction the model produces (a continuous variable, binary, survival, or multi-class). For example, in a regression context comparing predictions of continuous variable, Mean Squared Error (MSE) may be appropriate.

Up until now in this review, optimisation of performance implicitly assumed one performance measure. However, there are several measures that could be considered in applications. For example, optimising a performance measure subject to training time and the computational resources available. This is a type of multi-criteria optimisation and Dewancker et al. (2016) have suggested a method to combine several performance criteria and a ranking optimisation method to evaluate optimisation performance. Considering several criteria can be important as these often reflect real world constrains on computational resources, time, and possible trade-offs with accuracy. Caruana et al. (2008) note that some algorithms can efficiently explore the tuning parameter space more comprehensively than an algorithm that could be better but where fine-tuning would be computationally infeasible. Thus, the computational complexity of the algorithms and their ability to scale to the data available in problem domain is an important considering in algorithm selection.

This is one of the reasons why researchers such as Guyon et al. (2015) have invested effort in understanding if this process can be automated. One approach automating is an approach called Combined Algorithm Selection and Hyper-parameter optimisation (CASH) (Feurer et al., 2015). This is used to select the algorithm, then tune it. The results from this Automated Machine Learning (AutoML) research suggest that this is possible but there are still gaps between automatic and human-tuned models (Guyon et al., 2016). A recent book by Hutter et al. (2019) provides a useful review of the progress being made in this field.

1.2.3 Evaluation of predictive performance

Evaluation is the final step in comparison of algorithms. The purpose of the evaluation may be to test a new classifier or regression method against another group of methods on either domain data or benchmark datasets (Japkowicz and Shah, 2011). Other types of evaluation are testing multiple classifiers or regression methods on one domain data set or multiple benchmark datasets. Depending on the question being addressed and the experimental set-up to reflect that, appropriate use of statistical tests are set out in Demsar (2006), Garcia and Herrera (2008), Japkowicz and Shah (2011), Boulesteix et al. (2015). Examples of evaluation studies using some of these methods are contained in Section 1.4 and in the three papers in this thesis.

1.2.4 Summary

The implications of the issues discussed in this section are that in applied work, understanding the nature of the question to be answered, and how machine learning methods may help prediction is a first step. There is no uniformly best method per domain but there may be some good candidate methods that could be suitable if they have performed relatively well on similar data. However, this is not guaranteed and because of that, it is necessary to experimentally compare algorithms. In conducting this comparison, there are a variety of approaches to train and tune methods, and appropriate statistical tests to apply to the results to determine if differences in performance exist.

1.3 Consumer credit risk modelling: context

This section provides context for the consumer credit risk modelling applications contained in this thesis. Credit risk is the risk that a borrower will fail to meet their obligations in accordance with agreed terms (BCBS, 2000). For the purpose of this thesis, consumer credit risk refers to credit risk inherent in providing loans to individuals. This includes unsecured lending such as personal or Peer-to-Peer (P2P) loans and secured lending for mortgages. The next two subsections therefore provide an overview of the importance of consumer credit risk and context on developments in the Irish mortgage market.

1.3.1 Consumer credit risk: comparative context

The consumer credit market is a significant market and a source of risk to financial institutions (banks and non-bank lenders). In official statistics, loans to individuals is typically referred to as loans to households. In the US, household borrowing from banks and non-banks accounts for about $15,699 billion compared to total business borrowing of $15,579 billion. Of household borrowing, $2,806 billion was non-revolving consumer credit (i.e., non-mortgage) at Q1 2019.

In the euro area, the €5460 billion of loans to households was slightly larger than the amount of €5023 billion in loans to corporates, making mortgages one of the most important asset classes for banks. In Ireland, although the actual stock amount is orders of magnitude smaller, household borrowing accounts for proportionally more than corporate borrowing. In July 2019, household borrowing was €90 billion compared to corporate borrowing of €41 billion.[3]

In the aftermath of the Global Financial Crisis (GFC), non-bank origination of loans has grown as new business models such as Peer to Peer (P2P) models emerge. These are lending platforms intermediating between borrowers and lenders, with lenders directly funding individual borrowers. These types of firms use various types of application information as well as information generated by the applicants interaction with their platform/app to inform credit risk grading decisions. At present, the two largest P2P platforms in the US, Prosper and Lending Club, together lent over $ 63 billion by Q2 2019. The relative size of the platforms is still small but this market is growing quickly in the US and elsewhere. In the Asia-Pacific region including China, lending by alternative finance providers (including P2P lenders) amounted to €221 billion at the end of 2016; in Europe, the total amount lent was just under €7.6 billion by end 2016.[4] As noted by Claessens et al. (2018) China was the largest market. In this thesis data from the Lending Club (LC) platform is used in chapter 3, as it is one of the largest P2P lenders currently operating in the US. The continued growth in consumer lending from banks and the increased provision from non-banks illustrates the scale and continued importance of the consumer credit market to financial institutions and consumers.

1.3.2 Consumer credit risk: Irish specific developments

Chapter 2 and Chapter 4 focus on the Irish mortgage market. The scale of the Irish banking crisis from 2008-2012 and its aftermath are the motivation of two of the papers in this thesis. The applications demonstrate that improved predictive modelling can provide additional insight to manage mort-

[3]Euro area figures are from the table T02.03 of the SSM supervisory statistics. US figures are from page 7 of the Flow of Funds release, Q1 2019.The Irish figures are from table A1 Summary of Irish Private Sector Credit and Deposits.

[4]Based on Lending Club and Prosper website data, SEC filings, and Ziegler et al. (2018).

gage arrears and improve duration and loss severity modelling during workout of distressed mortgages. This section therefore provides some specific context for the scale of the mortgage arrears and repossession/sale process.

Before the onset of the Irish and Global Financial Crisis (GFC), during the credit-driven economic expansion in Ireland from the early 2000's, mortgage credit grew rapidly along with a dramatic increase in property prices. The residential property market peaked during Q2 2007. Between 1999Q1 and 2007Q2 nominal residential property prices grew by 264%; average annual mortgage credit grew by 24% on average per annum between 1999Q1 and 2007Q2. This was both a very pronounced and lengthy expansion of credit.

This was followed by a sharp economic contraction, a peak to trough fall of residential property prices of just under 54% in the five years from the peak of the residential market in Q2 2007 to the trough in Q2 2012. The unemployment rate more than tripled from just under 5% in July 2007 to just under 15% about four years in later in September 2010. During this period, mortgage arrears more than tripled from 3.9% in December 2009 to just under 13 % by the end of 2012 (figure 1.2). At a macroeconomic level, the financial costs of recapitalising some banks and resolving others amounted to €68 billion.[5] Valencia and Laeven (2018) estimate the fiscal cost to be 37 % of Irish GDP and overall lost output of 107% of GDP (comparing actual GDP growth to pre-crisis trend).

The scale of the long-term mortgages arrears meant that a Code of Conduct on Mortgage Arrears (CCMA) i.e., a defined process to engage with borrowers in a restructuring process, came into force in 2009. This evolved in subsequent years, providing additional protections for the borrower and combined with a supervisory focus to move banks away from short-term restructuring to longer term restructures (Donnery et al., 2018). At the same time, there were political decisions not to progress repossessions and legal issues surrounding change that had been made to conveyancing law (Phillips, 2013). The implications of the dramatic changes in the environment and the scale of the problem meant the time to repossession and sale of collateral became protracted. This has led to a two thirds of loss of ownership occurring through a voluntary surrender/sale; and one third from the repossession legal process (Figure 1.3).

[5] See Comptroller and Auditor General 2017 Report, Chapter 3.

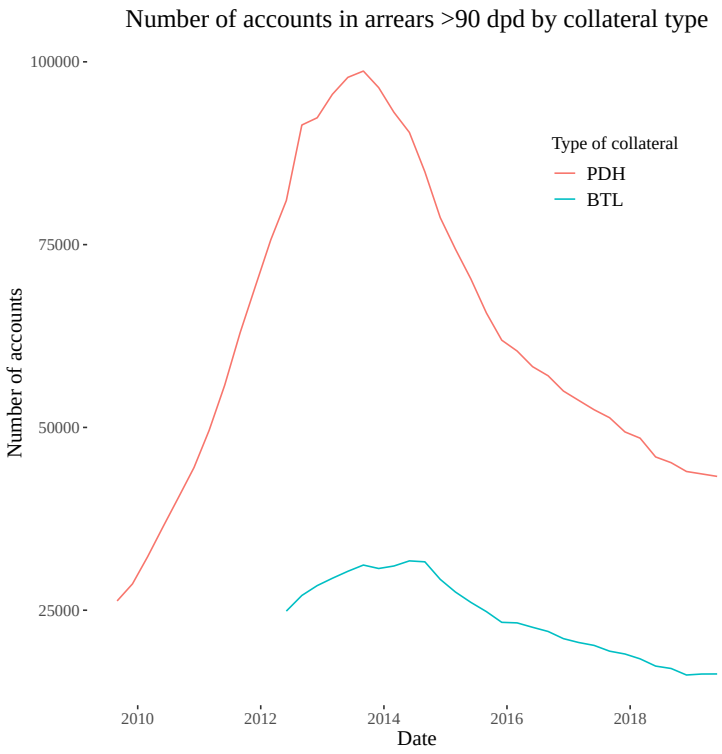

Figure 1.2: Irish mortgage arrears statistics: 2009-2019

Source: Central Bank of Ireland Mortgage Arrears Statistics

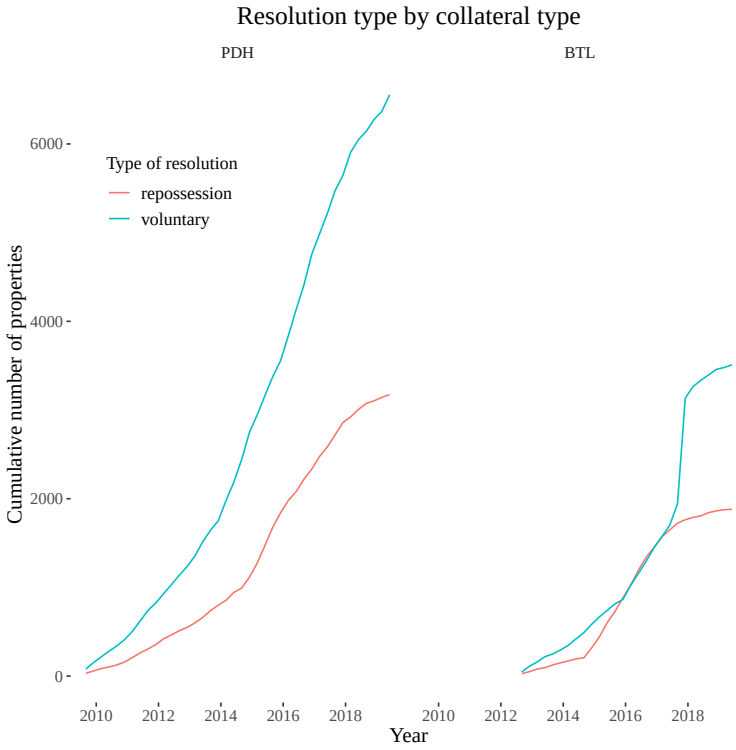

Figure 1.3: Irish loss of ownership statistics: 2009-2019

Source: Central Bank of Ireland Mortgage Arrears Statistics

1.4 Why is machine learning important for credit risk?

> "Consumer lending is the sleeping giant of the financial sector. ... Slowly over the years, with much more momentum since the millennium lenders have recognised that they need to extend the range of decisions that consumer credit risk assessment should be used in. This will mean changing the methodology used to build the assessment systems, changing the business measures used to assess the borrowers, and most of all, developing models to address the new decisions that credit scoring can be used for."

This prescient thought from Lyn Thomas (Thomas (2009), p.vii) encapsulates why more advanced methods such as machine learning are important for credit risk assessment. As noted earlier in Section 1.3, given the magnitude of the market improved tools for predicting consumer credit risk are of interest to both bank and non-bank loan originators, as well as borrowers and regulators.[6]

Because of the changes in the regulatory capital regime for banks with the Basel accords, banks can choose to develop their own credit risk models subject to certain conditions and regulatory approval. They can model Probability of Default (PD) which is the likelihood that a borrower cannot meet their debt obligation in defined time period. Banks typically use a variety of these types of credit risk models at origination (application scoring, PD models). Banks can also model the Exposure at Default (EAD) which is the gross amount of a facility owed by a borrower at the point of default. Models for these may be required as facilities can be granted up to a limit and then subsequently drawn on (i.e., overdrafts or credit lines). Credit risk models can also be used to make predictions of the Loss Given Default (LGD), that is the loss after default, collateral, and the time over which recovery cash-flows has been received by the bank.

While machine learning models are not yet in widespread use for this type of regulatory capital modelling, their predictive performance and ability to discover features, and in some cases, robustness to class imbalance is part of their appeal. Because of this, and the growing ability to interpret prediction and automation of tuning mean they could be used more widely in future.

1.4.1 Credit risk and benchmarking studies: a brief review

A central issue is their predictive performance. A comprehensive early benchmark study was by Baesens et al. (2003a). They compared 17 classifiers on 8 data sets and find that least squares-Support Vector Machines (SVM) and Neural Network (NN) classifiers performed best, but the margin of improvement was not much greater than logistic regression and linear discriminant analysis, both of which also performed well. One of the reasons that they say can explain this is their data sets are only weakly non-linear.

Lo et al. (2010) used Boosted Regression Trees (BRTs) to score credit card borrowers and comparison by Bastos (2008) found that BRT performed well compared to Neural Networks (multilayer perceptrons) and Support Vector Machines on two credit scoring tasks. Feature discovery as well as predictive performance of Support Vector Machines (SVMs) was assessed by Bellotti and Crook (2009) in predicting credit card default. They find SVMs competitive and useful in discovering features to predict default.

[6] Loan originator refers to the entity originating the loan. In the past, this was primarily banks or credit card companies. Since the Global Financial Crisis (GFC), non-bank originators such as Peer to Peer (P2P) platforms have become more prominent. In this thesis, the term refers to both types of entity unless specified otherwise.

Brown and Mues (2012) found that gradient boosted trees and random forests performed better than 8 other algorithms in the presence of pronounced class imbalance across 5 data sets. Kennedy et al. (2013b) had similar results; both papers did find that Logistic Regression (LR) performed relatively well, while not being the top classifier.

The most recent comprehensive benchmarking study was carried out by Lessmann et al. (2015). They compared single classifiers and ensemble classifiers that were all of the same base-learner type (homogenous ensemble) or composed of different types (heterogenous ensembles). In this study, the authors also considered how base classifiers in ensembles were selected and how the results were combined. The experiments were run on 8 separate datasets including some with large number of observations (c. 150,000) as well other data previously used in benchmarking. They used six performance measures, the AUC, H-measure, Brier score, the percentage correctly classified (PCC), a partial Gini index (PG), and the Kolmogorov–Smirnov statistic (KS).

There were several relevant findings. First, Logistic Regression (LR) is less accurate than other classifiers. Second, a selective ensemble developed by Caruana et al. (2006) (HCES-Bag) that was significantly better than any of the other classifiers. Finally, of the other top performing classifiers, Random Forests (RF) and Artificial Neural Networks (ANNs), there was no statistically significant difference.

As more transaction data becomes available due to changes in payment methods, this potentially is a useful source of information to manage credit risk. A recent paper by Tobback and Martens (2019) illustrate how payment transaction data, transformed using network analysis approaches, can be used to predict default and improve on an existing ratings system with a bank. They found that when the payments data are transformed using network methods, and a linear SVM is used to combined these outputs, the resulting model outperforms the rating system.

Apart from credit scoring or Probability of Default (PD) modelling, there are other uses for machine learning methods in credit risk measurement and management. Under the Internal Ratings Based (IRB) approach, banks may also model the Loss Given Default (LGD) to produce predictions for the loss given default for a specific account. To do so, they must comply with a range of regulatory requirements.[7] In one way, these requirements may limit the incentive to invest in more advanced modelling techniques that require more involved explanations of their risk drivers for a PD or LGD model. However, they could be used to improve predictions of more standard models. This is the so-called 'challenger-model' concept.[8] For example, this could be through using a different loss function or identifying important features or interactions that may not have been previously identified.

Work by Bastos (2010), Loterman et al. (2012), and Sheng Sun and Jin (2016) indicates that machine learning methods like SVMs and neural networks, and ensembles like Random Forests and Boosted Regression Trees (BRT) can perform better across a range of data sets. However, the work by Loterman et al. (2012) suggests that much of the variance in LGD is unexplained across their data, indicating further improvements in modelling and predictors are needed. Zhao et al. (2018) use a simulation study to benchmark linear regression models with more sophisticated parametric models (inverse gaussian, fractional response, gamma, beta and inflated beta regression. They find little difference between models when evaluated using mean predictions and squared error loss. When they extend their evaluation to

[7] See the SSM TRIM manual, page 4.

[8] Some of these ideas are explored in this blog post; other examples are included in Bellini (2019)

predicted distributions (i.e., conditional on the simulated predictor variables), they find large differences with some of the more sophisticated parametric models performing better.[9]

1.4.2 Summary

Later in the loan life cycle, banks may also use models to manage accounts or provide behavioural scores based on the history of the account. Should the account go into arrears or default, they can use models to provide insight into their arrears management process, and this could be used to help manage the process increasing efficiency of collections and reducing arrears. This is not yet as common in the literature, most likely because only a few European countries experienced the scale of distress experienced in Ireland. The relevance of the problem in the Irish case was one of the motivations behind Chapter 2

The growth of P2P lending and the potential change in how credit is intermediated between borrower and lenders means that performant predictive methods may help with managing default risk, but could provide new ways to measure return or profitability on P2P loans in a scaleable manner to different type of investors. Better loan selection could avoid defaults and increase returns. In large P2P markets, this could be significant. This is the motivation behind Chapter 3.

Another reason to consider machine learning methods for modelling collateral recovery is the legacy of the crisis is still being addressed at the same time as the introduction of the IFRS 9 expected credit loss standard. As the first generation of IFRS 9 models have been implemented, there is a lack of published research on methodological aspects of IFRS9 implementation. This is an appropriate time to assess whether machine learning methods can estimate more accurate and conservative impairment model parameters. This is one of the motivations behind Chapter 4.

1.5 Summary of papers and main contributions

This section summarises the main findings of the papers and outlines the main contributions of this research.

1.5.1 Summary of papers

One of the main themes in this research is whether machine learning algorithms can improve on more standard approaches when applied in different contexts. Three different aspects are considered.

The first paper focuses on mortgage credit risk management. The main research question is whether machine learning models have a better predictive performance than logistic regression. This paper evaluates the performance of a number of modelling approaches for future mortgage default status. Boosted regression trees, random forests, penalised linear and semi-parametric logistic regression models are applied to four portfolios of Irish owner-occupier mortgages. The main findings are that the selected approaches have varying degrees of predictive power and that boosted regression trees significantly outperform logistic regression. This suggests that boosted regression trees can be a useful addition to the current toolkit for mortgage credit risk assessment by banks and regulators. This paper has been published in the European Journal of Operational Research as Fitzpatrick and Mues (2016).

[9] They way they evaluate this is effectively using a two sample Kolmogorov-Smirnov (KS) comparing the conditional predictive distribution with that from the assumed data generating process, with the KS statistic measuring the largest divergence between the cumulative density functions of the two distributions.

The second paper topic is profit scoring of Peer-to-Peer (P2P) loans. Successful Peer-to-Peer (P2P) lending requires an evaluation of loan profitability from a large universe of loans. Predictions of loan profitability may be useful to rank potential investments. The paper investigates whether various types of prediction methods and the types of information contained in loan listing features matter for profitable investment. A range of methods and performance metrics are used to benchmark predictive performance, based on a large dataset of P2P loans issued on Lending Club. Robust linear mixed models are used to investigate performance differences between models, according to whether they assume linearity, whether they build ensembles, and which types of predictors they use. The main findings are that: linear methods perform surprisingly well on several (but not all) criteria; whether ensemble methods perform better than individual methods is measure dependent; the use of alternative text-based information does not improve profit scoring outcomes. This paper is under review at the European Journal of Operational Research.

The topic of the third paper is mortgage collateral recoveries. This paper focuses on collateral recovery vale for defaulted mortgages through modelling two important parameters determining this value: the time to sale (TTS) and Forced Sale Discount (FSD). Both of these parameters affect the loss on liquidation of collateral and are important for IFRS9 modelling of impairments as well as valuation of securitised loan portfolios. This paper is one of the first to assess the impact of estimation method for both the resolution times for mortgages that are in a loss of ownership process and the forced sale discount from credit risk perspective. Using a variety of survival modelling approaches to estimate time to sale, this paper evaluates their predictive performance and finds that Random Survival Forests and parametric survival models perform best. For the FSD, random forests, xgboost, and a deep learning neural network produce reasonable estimates. Using the predictions from these to steps, a sensitivity analysis of model outputs for estimated resolution time and FSD illustrating how predictive modelling of these parameters produce more conservative, ie, higher loss estimates than one current industry approach.

1.5.2 Main contributions

The main contributions of this thesis are briefly summarised below:

- Paper 3 (Chapter 4) in this thesis makes a contribution to the methodology of modelling collateral recoveries and comparing the impact of the methods chosen on two important parameters for that affect mortgage collateral recovery or loss severity: Time To Sale and the Forced Sale Discount. The research is one of the first to examine the impact of estimation method on both the TTS and the FSD as these are critical for determining the realised Loss Given Default (LGD). In particular, the predictive accuracy of three groups of methods for TTS are assessed as well as the important factors identified by those methods for time to resolution. In forced sale models, the type of forced sale (legal or bi-lateral agreement) is key determinants of the haircut. One of the implications of the results is modelling the parameters directly results in higher loss severities compared to assigning fixed average to a cohort, which is a common approach used within industry (Eder and Bank, 2019; Chawla et al., 2016).
- Within the application domain, a significant amount of research in the consumer credit risk literature deals solely with consumer credit (i.e., personal loans, overdrafts, credit cards) and not mortgages. This is surprising because retail mortgages are a significant share of the lending to

consumers. The work in this thesis extends the understanding of both prediction of mortgage arrears in Paper 1 (Chapter 2) and mortgage collateral recovery parameters in Paper 3 (Chapter 4). These findings may be relevant to financial institutions with mortgage lending businesses and in countries that have not recently experienced severe downturns or housing market crises and thus have limited data available to fit robust models under such scenarios.

- The results in this thesis provide a distinct contribution to the research literature comparing machine learning methods for classification, regression, and survival analysis methods. This includes comparison of classifiers such as BRTs and GAMs in Paper 1 (Chapter 2 of this work), various types linear and non-linear methods including ensembles in Paper 2 (Chapter 3), and survival and regression methods in Paper 3 (Chapter 4) using real world data.
- Chapter 3 contributes to the experimental evaluation literature in machine learning through using an robust linear mixed model to testing three experimental factors related to the various types of machine learning methods (linear/non-linear; individual/ensemble) methods and types of information included in the models. This contributes to understanding for investors of what may be useful modelling approaches in Peer to Peer (P2P) lending.

1.6 Outline of this thesis

The next three chapters contain the three papers making up the main body of this work. Chapter 2 focuses on predictive models of mortgage arrears. Chapter 3 contains a profit scoring application of Peer to Peer (P2P) lending. Chapter 4 analyses collateral recovery timing and loss severity. Chapter 5 concludes this thesis.

Chapter 2

An empirical comparison of classification algorithms for mortgage default prediction: evidence from a distressed mortgage market

2.1 Abstract

This paper evaluates the performance of a number of modelling approaches for future mortgage default status. Boosted regression trees, random forests, penalised linear and semi-parametric logistic regression models are applied to four portfolios of over 300,000 Irish owner-occupier mortgages. The main findings are that the selected approaches have varying degrees of predictive power and that boosted regression trees significantly outperform logistic regression. This suggests that boosted regression trees can be a useful addition to the current toolkit for mortgage credit risk assessment by banks and regulators.

Keywords: boosting, random forests, semi-parametric models, mortgages, credit scoring

2.2 Introduction

2.2.1 Background: mortgage default prediction and its applications

Credit default (i.e., failure to keep up with loan repayments) has cost implications for creditors in terms of losses or profits forgone and to other debtors in terms of higher prices (i.e., interest rates) and possible rationing of credit. Residential mortgages are one of the main types of lending and therefore a major potential source of credit risk for banks. Credit risk and credit scoring models to predict mortgage default are used by financial institutions and regulators to measure, assess, and inform decisions to mitigate various aspects of mortgage credit risk. A widely established techniques for this type of modelling is Logistic Regression (LR).

In recent years, there has been an increased research interest in a number of alternatives to LR and whether those could produce more accurate credit risk models. Particularly, with the development of new predictive modelling techniques in machine learning and the statistical literature, various studies have assessed how these newer approaches perform compared to more established methods with regards to scoring unsecured consumer loans such as personal loans and credit cards (Lessmann et al., 2015; Kennedy et al., 2013b; Baesens et al., 2003a). However, when it comes to secured lending, research findings regarding credit risk assessment of mortgage loans are much more scarce, despite the fact that they are among the largest class of assets on European banks' balance sheets. This paper attempts to assess, using real-world mortgage loan-level data, whether a selection of these newer methods can provide improved predictive performance over more established methods such as Logistic Regression (LR).

Mortgage credit differs from consumer credit as it is secured by property collateral and absorbs a significant amount of borrowers income-based repayment capacity. The fact that the loan is secured on property may reduce the probability of default and the loss given default. As property prices can fluctuate, this can reduce or increase the value of the collateral compared to the outstanding balance of the loan. Therefore, changes in the value of the collateral compared to the outstanding amount of the loan at a particular time may change the probability of default and loss in the event of default. Given the cost of buying a property means the loan obtained by a borrower may be a sizeable in relation to their annual income, and in turn, the monthly payments may be a substantial fraction of the monthly income. Both of these factors make the default rate on mortgage lending lower and sensitive to different macroeconomic variables like house prices and unemployment compared to some other types of consumer lending like credit card or unsecured personal loans.

Evaluating and comparing how various techniques perform with regards to mortgage default prediction serves a number of goals. First, for profitability and credit risk management purposes, financial institutions are interested in determining borrower creditworthiness through separation into good and bad categories. This is the central objective of credit scoring (Thomas, 2009). The outputs of these credit scoring methods can also contribute to the implementation of risk-adjusted loan pricing systems. Even a small improvement in the predictive power of such models could thus have a substantial impact on the quality of a bank's loan book and pricing strategy.

Second, adequate regulatory capital buffers are required so that banks would be able to cope with unforeseen losses in excess of expected loss. Accurate assessment of the risk or probability of mortgage loan default is critical for determining regulatory capital requirements. For retail credit risk classes such as mortgages, the Probability of Default (PD) models developed for this purpose are usually fixed in horizon (one year) and have so far been typically modelled using logistic regression; being to able to build more accurate models would enable more appropriate capital levels being set.

Third, the systemic banking crisis in Ireland and elsewhere in Europe has, in several of these countries, intensified the use of predictive models for operational management of credit arrears (Matthews, 2011). In this context, predictive models estimating the probability of a loan experiencing arrears in the near future are used to drive various decision-making strategies. This probability may depend on borrower attributes at application, borrower repayment behaviour such as past arrears or loan modifications, the presence of negative equity (i.e., the value of the property dropping below that of the loan), as well as regional economic conditions. Given that financial and operational resources are limited for financial

institutions and regulators, improvements to these models and their estimates could assist in better segmenting borrowers and targeting scarce resources to where they are needed most in early-prevention initiatives and active arrears management.

2.2.2 Research question; choice of techniques

Developments in statistical and machine learning approaches to classification (i.e., prediction problems where the target variable of interest is discrete, e.g. default or no default) have led to a variety of applications in credit risk. Previous reviews of various modelling approaches and empirical evaluations have been carried out by Lessmann et al. (2015), Kennedy et al. (2013b), Baesens et al. (2003a), Brown and Mues (2012), Crook and Bellotti (2009), and Crook et al. (2007). Some of their results suggest that newer approaches such as ensemble classifiers offer some improvement in predictive ability over logistic regression which could prove valuable for managing credit risk. However, the suggested performance boost is not guaranteed; on some datasets, newer techniques may not substantially improve predictive performance (Hand, 2006). This implies that empirical work is needed to determine if and where this is the case.

The main research question in this paper therefore is whether these alternative modelling approaches from the statistical/machine learning literature indeed offer improved predictive performance for mortgage credit risk compared to Logistic Regression (LR). LR is chosen as the baseline as it performs relatively well as a classifier in other credit scoring settings, and because of its relative ease of interpretation and widespread use in the financial services sector.

LR has been found to perform reasonably even in the case of imbalanced classes. While there is not a great deal of literature on this specific problem, some research suggests this is dependent on the data structure from the given domain. Owen (2007) found that LR has some drawbacks in the cases of extreme class imbalance. This can occur when the predictor values (X | y=1) are linearly separable from those for the cases when(X | y=0). In particular, Owen (2007) produced a result indicating when there is extreme class imbalance, the minority class (i.e., defaulters) only contribute to the logistic regression estimation through being collapsed to their sample mean vector. This result is built upon by Li et al. (2019) and they extend this result to penalised regression, which is not sufficient to address this problem. They suggest a clustering procedure to discover structure in minority class to improve the predictive performance of the model. Therefore, it is an open question whether LR is affected by class imbalance for the data considered in this paper.

To answer the question of whether alternative approaches can perform better, a number of alternative approaches were selected. The modelling approaches included in the empirical comparison are: semi-parametric Generalised Additive Models (GAMs), Boosted Regression Trees (BRT), and Random Forests (RF). These approaches each enable a flexible approach to modelling data with a complex structure (Hastie et al., 2009).

There are several reasons to choose these types of models among alternatives. First, there may be non-linear effects of predictors on the response variable. For example, using option pricing theory, Das and Meadows (2013) and Deng et al. (2000) argue that mortgage borrowers may hold an option to default if their home is in negative equity, i.e., the current loan to value is greater than 100 percent. Empirical work for various mortgage markets confirms that negative equity is an important predictor for default

and that loan to value does not have a simple linear relationship with the log odds of defaulting (Haughwout et al., 2008; Foote et al., 2008; Kelly, 2011).[1] Similarly, other variables such as loan vintage or borrower age are sometimes found to be non-linearly related to default risk. In contrast, one of the assumptions underpinning LR is that predictors are assumed to have a linear and monotonic effect. This may thus not hold in practice. Moreover, categorising or binning continuous variables, in an attempt to approximate this non-linearity, may result in mis-specification and loss of information. GAMs, BRT and RF on the other hand can all, to some extent, approximate non-linear functions of continuous predictors. This may allow identification of these effects and, if needed, the introduction of additional terms in a logistic regression model to approximate them.

Second, although arguably harder to interpret than LR, all three alternative approaches are not simply black-box models as they provide some degree of model explanation and insight into risk drivers. For example, GAMs can be assessed through statistical significance tests and spline plots. Variable importance measures and important interactions can be identified in BRT and RF (Hastie et al., 2009; Elith et al., 2008; Liu et al., 2009; Caruana et al., 2012). This may reduce the risk of model mis-specification and help make these models acceptable to practitioners. In addition, their use can potentially lead to improved predictive performance – i.e., the default predictions produced by these more recent techniques may be more accurate.

In the present application, a third justification for choosing LR, GAMs, BRT and RF is that their training algorithms tend to scale relatively well with the size of the data. All four techniques can cope with the large datasets analysed in the study within a reasonable amount of computation time. Although we experimented with Support Vector Machines (Vapnik, 1998), which have previously been found to be competitive for credit scoring (Bellotti and Crook, 2009) and bankruptcy prediction (Van Gestel et al., 2010), we did not include them in the final study due to the weaker scalability of available implementations.[2] The algorithmic complexity involved in solving the general SVM quadratic programming problem is between $O(N^2)$ and $O(N^3)$, where N is the number of training observations (Bordes et al., 2005). The complexity of Radial Basis Function SVMs may even be higher, i.e. between $O(dN^2)$ or $O(dN^3)$ (where d is the data dimensionality) (Sreekanth et al., 2010), which proved prohibitive for several of the training samples used in this study.

2.2.3 Related literature and main contributions

This paper extends the existing credit scoring literature in four main ways. First, it specifically focuses on mortgages. Detailed accounts of the various modelling approaches to credit scoring are included in Crook et al. (2007), Crook and Bellotti (2009), Hand (2009b), Martin (2013), and Thomas (2009). Lo et al. (2010) found Boosted Regression Trees (BRTs) were useful for scoring credit card borrowers. A comparison study by Bastos (2008) found that BRT performed well compared to Neural Networks (multilayer perceptrons) and Support Vector Machines on two credit scoring tasks. Feature discovery as well as predictive performance of Support Vector Machines (SVMs) was assessed by Bellotti and Crook (2009) in predicting credit card default. They find SVMs competitive and useful in discovering

[1] Negative equity is of course not the sole reason for default. As noted by Foote et al. (2008) and Van Order (2008), borrowers may default for a multitude of reasons which also include trigger events such as illness, unemployment, divorce, or a lack of financial resources to overcome the trigger event.

[2] Another partial reason for not considering SVMs (or Neural Networks) is that it is challenging to directly interpret the resulting model, which is considered a drawback in a highly regulated practical setting. However, in the case of SVMs, Martens et al. (2007) demonstrate that it is possible to extract understandable rules that approximate an SVM classifier.

features to predict default. Brown and Mues (2012) found that gradient boosted trees and random forests performed better than 8 other algorithms in the presence of pronounced class imbalance across 5 data sets. Kennedy et al. (2013b) had similar results; both papers did find that Logistic Regression (LR) performed relatively well, while not being the top classifier.

However, with the exceptions of Kennedy et al. (2013a), Galindo and Tamayo (2000), or Feldman and Gross (2005), most of the literature concentrates on credit card or personal lending only. This is somewhat surprising given the importance of mortgage lending as a business line to banks in advanced economies, but may be due to a lack of publicly available information from credit registers or third-party data providers in Europe, as well as commercial considerations by financial institutions (see Section 1.3).

Second, this paper adds to the findings on classifier comparison by making a focused comparison of four techniques on four portfolios of recently collected real-world data. Specifically, BRT, with the exceptions of Lessmann et al. (2015), Brown and Mues (2012), and Lo et al. (2010), have received relatively little attention to date in the credit scoring literature. Although Lo et al. (2010) used BRT to score credit card borrowers, they did not compare their performance to other classifiers. A comparison by Bastos (2008) found that BRT performed well compared to Neural Networks (multilayer perceptrons) and Support Vector Machines on two credit scoring tasks. GAMs were used by Berg (2007) to assess corporate credit risk, but they do not appear to be applied widely in mortgage credit risk modelling. In addition, several of the comparative studies of classifiers use datasets that may no longer be representative of the much larger scale of data available for predictive modelling within today's retail banks.

Third, the imbalanced nature of the portfolios considered in this paper, i.e., the large difference in the relative proportion of non-defaulters and defaulters, forms another topic of interest within the credit scoring literature. The impact that such imbalanced datasets have on the quality of the resulting models was studied by Kennedy et al. (2013b), Brown and Mues (2012), and Burez and Van den Poel (2009). Both Kennedy et al. (2013b) and Brown and Mues (2012) found that LR nonetheless holds up relatively well, along with other classifiers. However, the experiments set up in Brown and Mues (2012) indicated that BRT and RF started to outperform other classifiers when the level of class imbalance was further increased in their datasets – none of which were mortgage data. This paper thus contributes to these findings by applying the selected classifiers to four real-world mortgage datasets with a natural class imbalance so as to test whether BRT and RF offer a similar performance advantage in this setting.

Fourth and finally, the context for our study is a distressed European mortgage market within a recessionary economic environment, which sets it apart from other studies, as most of the published research is not informed by the current crisis or is based on the US mortgage market (Haughwout et al., 2008). Also, our findings may be relevant to financial institutions in other parts of the world that have not recently experienced severe downturns or housing market crises and thus have limited data available to fit robust models under such scenarios.

The remainder of this paper is structured in the following manner. The next section describes the specific modelling techniques or classification algorithms used in the paper. This is followed with a description of the parameter tuning and data. After that, the main results are presented and discussed; the final section concludes.

2.3 Statistical and classification models

The aim of each model is to produce a loan-level prediction for a binary variable where $Y = 1$ signifies default and Y=0 indicates no default. This prediction is made using n observations of training data with p predictor variables. Each observation $(x_i, y_i), i = 1, \ldots, n$, consists of a predictor vector (x_i) and an associated response (y_i = *0 or 1*). The predictor variables are a mix of continuous and categorical variables. We define default as 90 days arrears or greater.

This section describes the methods used. Based on the review of the literature contained in Section 2.2.3, BRT and GAMs do not appear to have been widely used in the credit risk literature nor for mortgage credit risk prediction. Random Forests are included because they performed well in various benmarking studies (Lessmann et al., 2015; Brown and Mues, 2012)

2.3.1 Logistic regression

Logistic Regression (LR) is known as a classifier that performs reasonably well across many application settings and data types, including credit scoring (Lessmann et al., 2015; Kennedy et al., 2013b; Brown and Mues, 2012). To avoid the problems associated with stepwise regression, and to make the model comparison as fair as possible, Regularised Logistic Regression (RLR) is used in this paper, with the final model chosen on the basis of the H-measure (see Section 2.5.1).[3] This type of logistic regression uses penalisation to improve the model fit. These penalties can include $\ell 1$ (the lasso), $\ell 2$ (ridge regression) or mixtures of the two (elastic-net) (Friedman et al., 2010). The best-fitting penalisation method is chosen by cross-validation.

The penalised negative binomial log-likelihood is given by equation 2.1. The β coefficients are chosen to minimise this objective function. The term on the left of the equation is the negative binomial log-likelihood. The additional term on the right (λ onwards) penalises the coefficients using two types of penalty terms, with $||\beta||_1$ and $||\beta||_2^2$ denoting the $\ell 1$ and the squared $\ell 2$ norms of the β coefficients.[4]

$$\min_{(\beta_0,\beta)} -\left[\frac{1}{n}\sum_{i=1}^{n} y_i \cdot (\beta_0 + x_i^T\beta) - \log(1 + e^{(\beta_0 + x_i^T\beta)})\right] + \lambda\big[(1-\alpha)\frac{1}{2}||\beta||_2^2 + \alpha||\beta||_1\big] \tag{2.1}$$

The effect of the $||\beta||_1$ term (also known as a lasso penalty) is to perform variable selection when λ is sufficiently large by setting their coefficients exactly equal to zero. The role of the $||\beta||_2^2$ term (also known as a ridge penalty) is to shrink coefficients towards zero as λ becomes larger. There are some drawbacks with the individual penalties. First, a model trained with a ridge penalty only will include all predictors, even if they are irrelevant, with the degree of coefficient shrinkage increasing with λ. Second, a lasso-based model may only select one predictor from a group of correlated variables and ignore the others. As it is usually difficult to determine before a model is estimated which predictors are truly important, a mixture of both penalties can be useful. The α parameter in equation 2.1 controls the degree of mixing between the lasso penalty (α=1) and ridge regression (α=0). Both λ and α are determined by cross-validation based on the training data. The advantages of this approach are that

[3] We are grateful to one of the reviewers of the published version of this paper for the suggested use of alternatives to stepwise regression. Note that stepwise regression was also tried, which produced similar performance ranks for LR.

[4] The coefficient β_0 is a scalar and is not typically penalised; β is a vector. This formulation is based on the implementation in the R package *glmnet*.

coefficient shrinkage and variable selection can be carried out simultaneously in a numerically stable manner through this penalty structure. This may improve predictive performance and avoid some of the problems with stepwise regression (Derksen and Keselmanl, 1992).

2.3.2 Generalised Additive Models (GAMs)

Generalised Additive Models (GAMs) retain many of the features of LR and are statistically interpretable. They are a useful alternative when the log-odds of default may be a non-linear function of some of the predictors as their output can be based on a sum of smoothed functions of predictor variables (Hastie et al., 2009). As the response data are binary, the logistic link function is used in the GAM. When linear terms and/or categorical variables are included alongside variables that are smooth terms, like in this application, the resulting model is termed as a semi-parametric GAM. Equation 2.2 shows the model that is estimated. The terms, $x_j, j = 1, \ldots, q$, represent variables from the training dataset that are smoothed, while $x_j, j = q+1, \ldots, p$, are variables assumed to have a linear effect on the log-odds of defaulting and are fit parametrically and are interpretable as in a LR. This approach retains a significant degree of interpretability and offers flexibility to incorporate potentially non-linear effects.

$$\text{logit}(P(y = 1|x)) = \beta_0 + \sum_{j=1}^{q} s_j(x_j) + \sum_{j=q+1}^{p} \beta_j x_j \tag{2.2}$$

The smooth functions in the GAM, $s_j(x_j)$, are estimated using penalised regression splines. An individual smooth term can use cubic splines as a building block.[5] This involves individual cubic polynomial regressions being run for different intervals of a given input variable, the results of which are combined at certain points (knots) to create a continuous curve or smooth function for that predictor. A penalty term for each smooth function of the covariates is included in the model. This is to ensure the smooth functions do not overfit the data. A parameter for each smoothed variable (λ) controls the trade-off between goodness of fit and smoothness.

Tuning of this smoothing parameter is critical: if the λ values are too high, the data will be over-smoothed; if they are too low, then the data will be under-smoothed (Wood, 2006). In both cases, the spline estimate will not closely approximate the true function, which will affect predictive performance. A technique called Generalised Cross Validation (GCV) is used to select the optimal smoothing parameter value given the data (Wood, 2006). This technique is similar to estimating prediction error based on a leave-one-out cross-validation estimation but using a more computationally efficient procedure (Wood, 2006).[6]

2.3.3 Decision tree-based methods

The tree-based models in this paper draw on Classification and Regression Trees (CART) (Breiman et al., 1984). This is a classification technique based on two central ideas: recursive partitioning and pruning. Recursive partitioning involves repeatedly splitting or dividing and then sub-dividing the predictor space into a series of smaller segments that are more homogeneous; i.e., each segment is ideally

[5] A cubic spline is a piecewise cubic function with continuous first and second derivatives.

[6] An alternative approach is to use a backfitting algorithm based on a scatterplot smoother or by other variants of penalised splines. The back-fitting algorithm is described in detail in Hastie et al. (2009).

composed of observations belonging to a single class. The resulting model assumes the structure of a tree. In CART, pruning is used to reduce the size of trees based on various measures of predictive error such as misclassification rate, Gini index, or deviance. This is necessary to avoid fitting every minor variation in the input data. The overall goal is to have a tree that explains relevant patterns and generalises well to unseen data. However, because CART is recursive, current splits depend on previous splits, making the resulting model outputs sensitive to small changes in the input data, such as when unseen data is applied to the model. Two subsequent algorithms – boosted regression trees and random forests – sought to improve upon CART.

Boosted Regression Trees (BRT)

Boosted regression trees combine tree-based recursive partitioning with the concept of boosting developed by Freund and Schaipre (1997) and extended with a statistical interpretation by Friedman et al. (2000), Friedman (2001), and Friedman (2002).[7]

Because the present application (mortgage default prediction) is a binary classification problem, the loss function used is binomial deviance. The algorithm used is called stochastic gradient boosting and is based on Friedman (2001) and Friedman (2002).[8] After initialisation, the algorithm minimises this loss function in each step by the stage-wise addition of a new tree that leads to the best reduction in the loss function, given the chosen tree size.

The procedure starts by choosing initial values such as the log odds of default based on the training data. A random sample of observations is drawn without replacement, and the difference between the response and the starting value is calculated. These are known as the vector of negative gradients.[9] Based on this data, a tree is constructed by choosing the variables and split points giving the maximum reduction in the loss function at this step. The algorithm updates by first calculating the predicted probability of defaulting based on the current tree and the random subset of data. These are then added to the existing fitted values up to that step and subtracted from the response to obtain a new set of negative gradients. A new random sample of observations is drawn from these and a new tree fit. This proceeds until the material improvement in the overall model fit is less than some small tolerance. Each time a tree is added to the model, its contribution is multiplied by a parameter termed the learning rate. The effect is to limit or shrink the contribution of any one tree to the overall model prediction. A final BRT model is the sum of several hundreds or thousands of trees multiplied by the learning rate.

Boosting has not been without its critics. In particular, Mease and Wyner (2008) have been critical regarding the reasons for the algorithm's resistance to overfitting and the way it has been interpreted in the statistical literature.

[7] These papers interpreted the algorithm in a likelihood framework and developed boosted logistic and other regression-based approaches. The papers also led to additions of shrinkage and bagging to the algorithm. Shrinkage refers to limiting the contribution of each sub-component of the model, through taking small increments in each forward stage-wise iteration. Bagging refers to only a random subset of data being used in each iteration. This random sampling is thought to reduce the variance, and thus improve predictive performance of the final model. A comprehensive overview of boosting is given in Hastie et al. (2009) and Bühlmann and Hothorn (2007).

[8] This section draws on the descriptions given in Elith et al. (2008), Berk (2008), Hastie et al. (2009), and Ridgeway (2013)

[9] The components of the negative gradient vector are sometimes referred to as pseudo-residuals, see Hastie et al. (2009), page 360-61, or Berk (2008), page 270. The use of a random subset of the data, known as the bag fraction, to construct a tree at each iteration in the algorithm has been found to improve predictive ability (Friedman, 2002).

Random Forests (RF)

Random Forests (RF) are another tree-ensemble classifier developed by Breiman (2001). There are three important differences between RF and the tree-based approaches outlined earlier. The first difference between RF and CART is that in a RF many trees are grown based on bootstrapped sub-samples of the training data. The second difference is that each time a split variable is chosen within an individual tree in a RF, the algorithm only chooses from a small random subset of predictors of size *mtry*. This is in contrast to CART or BRT where all of the predictors are evaluated to produce the best split. This process is repeated over many trees to create a 'forest' or ensemble of trees the predictions of which are averaged to produce an output. Randomly selecting a subset of predictors rather than trying all has the effect of reducing correlation among the trees in the random forest. Averaging predictions over all trees in the forest reduces variance, resulting in improved predictive ability compared to CART. A third difference is that random forests can be grown in parallel, as each tree can be grown independently, whereas the BRT algorithm proceeds sequentially depending on the output from the previous iteration. Random forests have been applied to a variety of domains such as bioinformatics, image recognition, as well as in financial applications such as customer attrition and credit scoring (Lessmann et al., 2015; Kruppa et al., 2013; Malley et al., 2012).

2.4 Model building and data sets

This section specifies how the various models were estimated and tuned, as well as describing the datasets.

2.4.1 Parameter settings and tuning

The penalised LR models include the main effects and pairwise interactions between predictors. The models are estimated using the R packages *glmnet* and *caret* (Friedman et al., 2010; Kuhn, 2008). The performance criterion for selecting the final model is the H-measure (to be further discussed in Section 2.5.1). The grid search considered a value range for the parameter α from 0 to 1, in 0.1 increments, and for λ, a sequence of 20 values from 0.005 to 1. The best combination was chosen using 10-fold cross-validation.

The semi-parametric GAM models are estimated using the R package *mgcv* (Wood, 2013). The degree of smoothing of the spline functions is chosen by Generalised Cross Validation (GCV).

Two parameters are key for BRT tuning. The learning rate (lr) or shrinkage parameter determines the contribution of each tree. A lower learning rate means that each tree has a lower weight in the final model. Tree complexity (tc) determines the degree of interaction between predictor variables. For example, a tc of 1 fits an additive model (each tree having a root and two leaves); a tc of 2 fits a model with up to two-way interactions. This paper uses the R package *gbm* and a modified version of the code from Elith et al. (2008). A grid search over these two parameters, i.e. learning rate [0.01, 0.005, 0.0025, 0.001], tree complexity [2-6], and a third parameter, bag fraction [0.5, 0.625, 0.75], was conducted to find the combination with the highest H-measure on the validation data. The number of trees (nt) is determined automatically by the function gbm.step using 10-fold cross-validation, for a given learning rate and tree complexity.

Finally, when tuning the RF, the number of predictors from which to select at each split (the *mtry*

parameter) was varied over the range [1-4, 6, 8]. The number of trees in the forest was fixed at 1000. The version of the algorithm used here is based on Breiman (2001) and implemented in the R package *randomforest* (Liaw and Wiener, 2002).

Initial results suggested that the class imbalance was affecting RF performance for some of the portfolios. Therefore, undersampling of non-arrears cases was carried out by taking balanced bootstrap samples from the original data. For example, if there were 1000 default cases in the training data, each time a tree is induced, this would be done on a different bootstrap sample containing all 1000 default cases and a random selection of 1000 non-default cases. This methodology is outlined in Breiman et al. (2004) and Kuhn and Johnson (2013). Compared to conventional undersampling, it has the advantage of making better use of all available training data, by not eliminating some majority class observations altogether but drawing a different sample at each step of the algorithm. The best parameter values are determined through 10-fold cross-validation using the R package *caret*; the optimal model is again selected based on the H-measure.

2.4.2 Data sets

This section describes the data collected by the Central Bank of Ireland on which our analysis was conducted. The data are composed of four separate portfolios of owner-occupier mortgage loans of Irish lenders. The sample represented 55 percent of the Republic of Ireland's mortgage market as of December 2010. For predictive modelling purposes, only those loans that were not yet in default at the observation point of December 2010 were retained; the target variable of interest is whether those loans moved to default status by December 2011. The predictor variables (i.e. the potential inputs to each model) are all measured either at December 2010 or prior to that.

The results presented in this paper are based on a combined training, validation, and test sample of 322,915 cases across the four portfolios.[10] The minimum training set size is over 31,000 and the maximum is just under 50,000 observations. The minimum test set size is approximately 18,000, the maximum just over 28,000. The proportion of default outcomes in the training data ranged from 3 to 9 percent.[11]

Split-sample setup

The data for each portfolio was divided randomly into training, validation, and test set, with a 50/20/30 split. The class distribution in the training, validation and test data was preserved to match the imbalance observed in each portfolio. The models are estimated or trained on the training data, where necessary tuned on the validation data, and performance is assessed using the test data. LR and GAMs are trained on a combined training plus validation sample as they do not require a separate validation sample for tuning. In the case of BRT and RF, only the training data are used for model fitting whilst the validation set is used to tune further the parameters and select the best performing model.

[10] Because of confidentiality restrictions, details for individual portfolios cannot be given.

[11] The training and test set sizes and class distribution is given for all portfolios and not for individual portfolios to preserve data confidentiality.

Data description

The dataset variables are described in Table 2.1.[12] The selected observations each relate to the main loan associated with a given property serving as collateral. The dependent variable is a binary variable defined as the equivalent of a borrower being 90 days past due or more (e.g. by missing three consecutive monthly payments) on their mortgage at some point over the outcome window. This is a standard measure of default used in capital requirement regulations in Ireland.

Table 2.1: Description of variables

Variable	Description	Type
Default	Dependent variable: 1 if borrower is reported at least 90 days past due on monthly instalments over the period Jan 2011 - Dec 2011; 0 otherwise	Categorical
Repayment to income	Monthly instalment amount in Dec 2010 over annual borrower income at origination in percent	Continuous
Loan to income	Ratio of origination loan balance over annual borrower income at origination	Continuous
Loan age	Time since origination in years (Dec 2010)	Continuous
Current LTV	Indexed loan-to-value (Dec 2010) in percent	Continuous
Number of loans	Number of loans (including current loan) registered against primary residence collateral	Continuous
Unemployment change	12-month change in NUTS 3 regional unemployment rates from Dec 2009 to Dec 2010	Continuous
Current interest rate	Mortgage interest rate in Dec 2010 in percent	Continuous
Interest rate type	Interest rate type: fixed, standard variable, or tracker	Categorical
Loan purpose	Mover, first-time buyer, or equity release switcher	Categorical
Property type	House type: detached, semi-detached, terraced, apartment/flat	Categorical
Borrower location	Borrower location at origination (8 NUTS 3 levels)	Categorical
Borrower gender	Borrower gender at origination	Categorical
Borrower marital status	Borrower marital status at origination: single, married, divorced/separated/widowed	Categorical
Number of borrowers	Number of borrowers servicing the mortgage: single or joint	Categorical
Modification status	Borrower received loan modification over Dec 2009 - Dec 2010: yes or no	Categorical
Recent default	Borrower was 90 days in arrears in Dec 2009 (i.e., one year prior to the observation point): yes or no	Categorical
Early arrears	Borrower has a material positive arrears balance of less than 30 days in Dec 2010: yes or no	Categorical
Bubble origination	Loan originated during 2004-2009: yes or no	Categorical

The predictor variables are a mix of continuous and categorical data and include a range of application and behavioural information. The updated loan-to-value ratio for December 2010 (variable Current LTV) is calculated by dividing the loan balance at that time by the indexed market value of the property (i.e., applying the December 2010 index to the original property value).[13]

Early arrears is a binary variable indicating whether the borrower had a non-zero arrears balance that was greater than 10 percent of but less than one month's full mortgage instalment in December 2010.[14] Due to data limitations, this variable is not available for Portfolio 3. Past arrears status (variable Re-

[12]For a more detailed description of a larger dataset from which these data were drawn, we refer the reader to Kennedy and McIndoe-Calder (2012).

[13]The house price index used to estimate market values in December 2010 is composed of Dublin and Non-Dublin property prices as well as house or apartment property types.

[14]The rationale for a floor of 10 percent of a one-month payment is to exclude borrowers that have a very small arrears amount, as this may be due to the loan nearly curing or technical reasons such as an incorrect standing order.

cent Default) may indicate that some borrowers could be at higher risk of defaulting in future. Finally, borrowers may have previously received a loan modification from their bank. This can occur while remaining current or after entering arrears and may be part of short-term forbearance.

There are some limitations to the data used in this study. First, some borrower-specific features are observed at origination (marital status, income) but not subsequently updated. Individual borrowers' personal and economic circumstances in December 2010 are likely to be important for prediction but remain unobserved after origination. Economic conditions such as the unemployment rate of the geographical region in which the borrower is located can only approximate the individual borrower's economic circumstances.

Second, additional unobserved features of borrower behaviour may also be relevant for default prediction. For example, borrowers could use the information advantage concerning their own economic and life circumstances that they have compared to their bank. They may be able to conceal their true ability to repay and default strategically (Das, 2012). These features are never observed and cannot be approximated using the data available for this study. Therefore, while the literature suggests several types of potential predictors of default, the predictors in this empirical study cannot be expected to explain all the idiosyncratic causes of default.

Third, after being checked for outliers and other errors, the data included missing values. Four categorical variables had missing values: property type, borrower's marital status and gender, and loan interest rate type. The percentage of cases with missing values for these variables ranged from 0-24% across the four portfolios. These were recoded as unknown rather than excluding the observation. The reason for this is that the alternative of imputation is a difficult problem which imposes a structure on the data, and if mis-specified, may itself lead to bias (Horton and Kleinman, 2007). Apart from these categorical variables, income at origination also contained some missing values with the percentage of cases with missing values for these variables ranged from 0-27% across the four portfolios. This is because of two reasons. A first cause were general data quality problems relating to banks inconsistently recording application information including income. Second, in some cases where a mortgage was topped up, extended, or refinanced, the institutions reported only the latest value for these income-related variables, as collected at the point of origination of those subsequent loans; the relevant values at the point of origination of the main mortgage were thus lost. Rather than proceeding by case-wise deletion or mean/median imputation, and thus potentially biasing the sample by excluding these cases, we imputed missing values using the k-Nearest Neighbour (kNN) algorithm.[15] A value of 50 for the number of nearest neighbours (k) was chosen for the imputation.[16]

[15] Replicating the same analysis on a smaller dataset following case-wise deletion gave results similar to those discussed in the remainder of this paper. The statistical performance tests showed BRT outperforming LR at a 5-percent significance level. The results for this robustness check are shown in Appendix 2.

[16] This was derived through empirical testing on two of the portfolios that either had no missing income or a very low number of missing income observations. After random deletion of a proportion of non-missing values in those datasets, using 50 nearest neighbours (k=50) in the imputation procedure led to the lowest estimation error for the income variable. Inclusion of a binary missing value indicator for income did not turn out to be a significant predictor of future default status.

2.5 Performance measures and statistical comparison

2.5.1 Model performance metrics

A commonly used measure for assessing the performance of a score-based classifier is the Area Under the Curve (AUC). This refers to the area under the Receiver Operating Characteristic (ROC) curve, which is a pairwise plot of the true positive rate against the false positive rate, as the classification threshold is varied over its entire range.[17] An AUC value closer to 1 suggests better discrimination ability between defaults and non-defaults; a value of 0.5 implies that the classifier performs no better than chance. Using the AUC as a performance measure is standard practice in credit scoring but not without its problems. Hand (2009a) argued that, when interpreted in terms of costs, the AUC measure treats the relative severities of misclassifications differently when multiple classifiers with different respective score distributions are compared, implying that the AUC is intrinsically incoherent.[18]

As a coherent alternative to the AUC, Hand (2009a) therefore proposed the *H-measure*. The advantage of using the H-measure as a classification performance measure is that it allows one to specify a distribution of likely misclassification costs (c) that is independent of the classifier; this choice is discussed in detail by Adams et al. (2012). Because of the class imbalance between defaulters and non-defaulters, this paper uses the default setting suggested there (corresponding to a Beta distribution with its mode set at $c = \pi_1$, i.e. the proportion of defaults in the dataset). This means that the reported H-measures put relatively greater weight on correctly classifying default cases than on incorrectly classifying non-default cases. As with the AUC, a higher H-measure is associated with better performance.

In this paper, unless otherwise stated, model comparisons are carried out using the H-measure. The AUC is nonetheless included as it is still widely used in practice. Where classifiers are compared based on the AUC, model selection/tuning for LR, BRT and RF has been done on the AUC instead.

2.5.2 Statistical comparison of performance differences

Statistical tests can indicate whether there is a significant difference between how well different classifiers perform over a set of available datasets. Friedman's test (Friedman, 1940) can be used to compare the various models based on their performance rankings for a chosen performance metric such as the H-measure (Demsar, 2006). The test statistic is χ^2 distributed with $k - 1$ degrees of freedom, where k is the number of classifiers. Its null hypothesis is that there is no difference between the classifiers' performance ranks. A less conservative variant of the Friedman statistic, also reported in this paper, is the Iman-Davenport test (Iman and Davenport, 1980).

In the event that there are significant differences according to either of these tests, various post-hoc tests can be used to compare pairs of individual classifiers. These tests adjust p-values to control for error propagation in multiple pairwise comparisons. Comparing the best-performing classifier with every other classifier requires the use of a particular approach which accounts for this family-wise error using what is known as Holm's procedure (Holm, 1979; Garcia and Herrera, 2008).

Holm's procedure starts by evaluating the performance rank differences between the best performer and

[17] In this application, the true positive rate, also known as the sensitivity, is the fraction of defaulters that are correctly classified using a given threshold value (i.e. having a score greater than the threshold). The false positive rate (1-specificity) is the fraction of non-defaulters classified incorrectly as defaulters, using the same threshold value.

[18] This point is debated by Flach et al. (2011).

each other model and, for each such pair, calculates the test statistic outlined in Garcia and Herrera (2008); each of these values is then compared against a normal distribution table to produce a significance value (p-value). Next, the procedure sorts these p-values in ascending order, comparing each p_i in the resulting sequence, $p_1, ..., p_{k-1}$, with an adjusted p-value, $\frac{\alpha}{k-i}$, where α is the required significance level. If p_i is less than the adjusted p-value, the relevant null hypothesis is rejected, in which case the corresponding model is considered significantly worse than the best performer. This proceeds until a null hypothesis cannot be rejected; any remaining performance differences can thus be ignored. The Java code by Garcia and Herrera (2008) is used to calculate the Friedman, Iman-Davenport statistics, and Holm's post-hoc tests.

2.6 Results and discussion

2.6.1 Results

The model performance results for the H-measure and AUC (both of which measured on an independent test set) are shown in Table 2.2. The results vary across portfolios and by classifier. In the upper-half of the table, the four classifiers can be ranked from 1 (best) to 4 (worst) on each portfolio, based on their H-measures; the resulting average ranks over the four portfolios are put in the rightmost column. BRT thus receive the highest average performance ranking of 1.25 (underlined in Table 2.2), followed by GAMs (2.25), RF (2.75), and, ranked lowest, LR (3.75). The null hypothesis that there are no differences in average rank between classifiers is rejected by both the Friedman (at the 10 % level) and Iman-Davenport tests (5% level) reported in Table 2.3.

Table 2.2: Performance summary of classifiers

Technique	Port 1	Port 2	Port 3	Port 4	Avg. Rank
H-measure					
LR	0.2302	0.2344	0.2825	0.2776	3.75
GAM	0.2579	0.2591	0.2928	0.2824	2.25
BRT	0.2776	0.2626	0.2909	0.2948	<u>1.25</u>
RF	0.2719	0.2411	0.2800	0.2854	2.75
AUC					
LR	0.7448	0.7466	0.7700	0.7737	4.0
GAM	0.7653	0.7617	0.7768	0.7816	2.0
BRT	0.7806	0.7630	0.7759	0.7878	<u>1.25</u>
RF	0.7781	0.7527	0.7701	0.7814	2.75

Table 2.3: Statistical comparison of classifiers using H-measures

Test statistic	Value	p-value
Friedman	7.8	0.0503
Iman-Davenport	5.6	0.0194

Next, the best-performing technique, BRT, is compared with the three other classifiers. As shown in Table 2.4, the results from the post-hoc procedure indicate that, only BRT and LR differ significantly (at the 5% level), whereas the other null hypotheses cannot be rejected, at either the 5% or 10% level. On the basis of these results, it can be concluded that BRT perform significantly better than LR, but

that no statistically significant difference in performance is evident between BRT and the other two classifiers, GAMs and RF.

Table 2.4: Holm's step down procedure for H-measure ranks; $\alpha = 0.05$ and $\alpha = 0.1$ (BRT is control classifier)

Classifier	$z = (R_0 - R_i)/SE$	p_i	Holm's adjusted p-value
5 % significance			
LR	2.7386	0.0062	0.0166
RF	1.6432	0.1003	0.025
GAM	1.0954	0.2733	0.05
10 % significance			
LR	2.7386	0.0062	0.0333
RF	1.6432	0.1003	0.05
GAM	1.0954	0.2733	0.1

The results are generally unchanged if the models/algorithms are tuned and compared using the AUC. The performance ranks according to the AUC (displayed in the lower-half of Table 2.2) are very similar to those observed for the H-measure. The results of the corresponding statistical tests show that BRT are again significantly better than LR, whereas no significant difference between BRT and GAMs or RF is found (see Tables A.2 and A.3 in Appendix A).

The classifiers used in this chapter are scoring classifiers that produce a rank ordering of cases based on their likelihood of being greater than 90 day arrears in 12 months. If these predictive models were to be used in implementing various actions such as differing types of borrower contact (letters, calls, meetings) to changing the type of restructure or taking legal action, then how well their probabilities are calibrated becomes important. This is because the cost of potential actions and the decision to carry them out are different across actions (So et al., 2020). While a detailed analysis of this topic is outside the scope of this thesis, it is relevant to consider how well the estimated class probabilities match the empirical default rates in the test sets.

One intuitive method to do so is through a calibration plot. These plots have been used in bioinformatics and in credit risk (Malley et al., 2012; Medema et al., 2009). They plot the class probability produced by the model (x-axis) against a non-parametric regression of the empirical proportion of defaulters with the same predicted probability (y-axis). The intuition is that if the smoothed curve runs along the 45-degree axis, a model is perfectly calibrated; either side of this and it is either under- or over-predicting default rates.

To construct the plots, a non-parametric loess regression of actual outcomes against predicted values was used.[19] Two sets of representative plots are shown for portfolios 1 and 4, in Figure 2.1 and Figure 2.2, respectively. For each of these portfolios, the figures show that the models are, for the most part, reasonably well calibrated, except at the less densely populated highest-risk segments on the right-hand side of each figure. Elsewhere the fitted loess curve (solid line) generally does not depart much from the 45-degree reference line (dashed line), for most of the models. The plots for the RF models however suggest that they are not as well calibrated as some of the other models, despite the class probabilities

[19]The optimal bandwidth for the smoothing window was chosen using the AIC and the polynomial is of degree 1. This is based on the AIC method outlined in Hurvich et al. (1998).

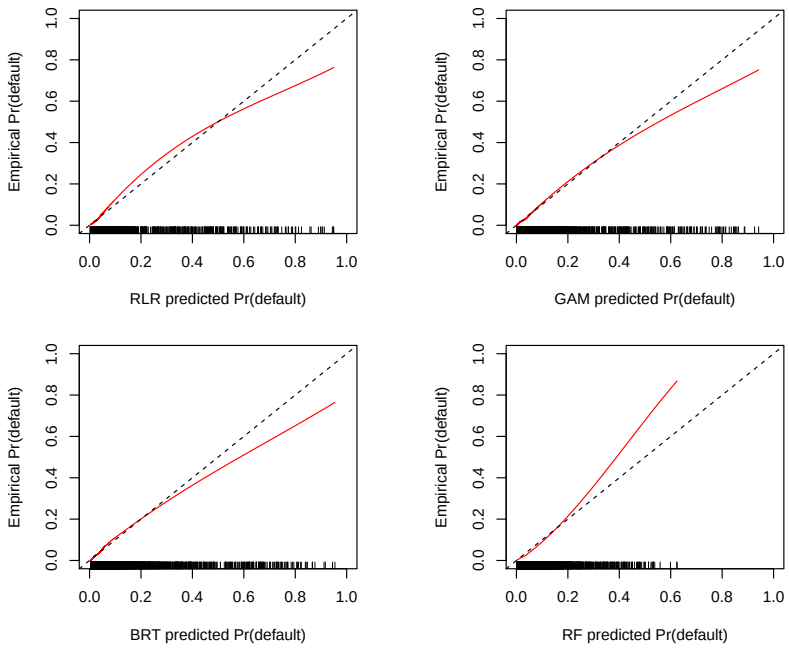

Figure 2.1: Calibration plots: portfolio 1

having been rescaled to reflect the original class priors.[20] In both portfolios 1 and 4, RF appear to underestimate default outcomes over a wider prediction range than the other models. The other three models also exhibit some minor divergences from the reference diagonal at lower levels, but the larger divergences are for predicted probabilities of default from 0.4 upwards: for those, in contrast to RF, predictions over-estimate rather than under-estimate the actual default risk.

In summary, this visual inspection suggests that, for the most part, the majority of the approaches produce reasonable class probability estimates, but that further work on calibration for high predicted class probabilities would be beneficial before these models could be used in practice.

2.6.2 Discussion

Overall, the results indicate that BRT significantly outperformed the conventional method, LR. That said, there was no uniform winner amongst the newer approaches, BRT, GAMs, and RF. While there appears to be particular promise in the BRT and GAM approaches based on our results, the extent of the performance improvement varies across portfolios.

When trying to relate these findings to the existing credit scoring literature, a direct comparison is less straightforward as that literature has tended to concentrate more on unsecured consumer credit (credit cards, personal loans) than on secured lending products such as residential mortgage loans. However, we can make several observations. First, the reasonably good predictive performance of the BRT algorithm, even with a very pronounced class imbalance, is in line with the findings of Brown and Mues (2012), Burez and Van den Poel (2009), and Bastos (2008). Second, unlike in Brown and Mues (2012), Lessmann

[20] Note that the probabilities are rescaled using a method outlined in Elkan (2001) as they were produced using an undersampled RF.

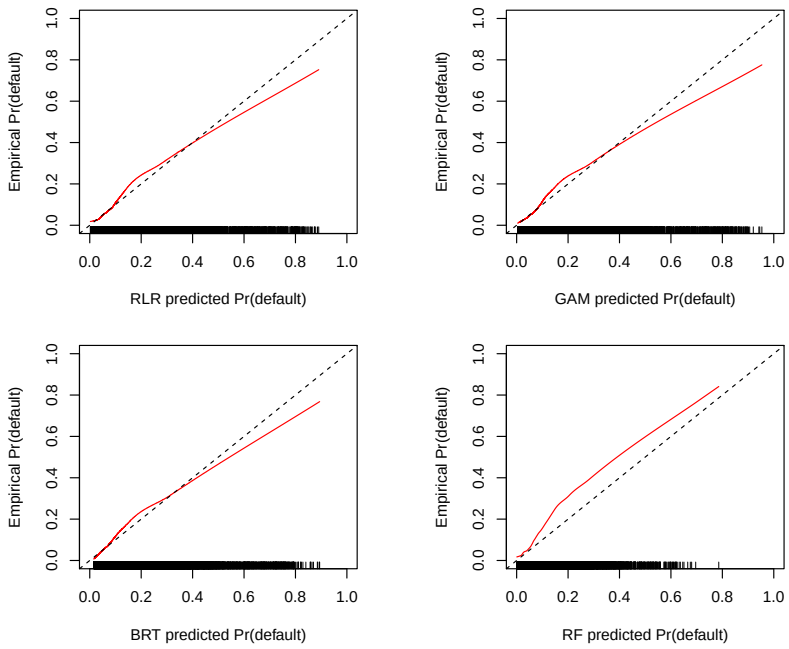

Figure 2.2: Calibration plots: portfolio 4

et al. (2015), and Burez and Van den Poel (2009), RF have a lower average ranking compared to BRT over the four loan portfolios (although the difference is not statistically significant).

We suggest that BRT performed very well in this context thanks to their ability to select important predictors and model higher-order interactions through the tree complexity parameter. BRT identified a small group of important predictors alongside a larger group of relatively less importance. This can be seen in Figures 2.3 and 2.4, where 4-5 features (early arrears, repayment to income, loan to income, current LTV, and, in portfolio 2, recent default) account for a substantial portion of the variable importance in the BRT for portfolios 1 and 2.[21] In portfolios 3 and 4, a single predictor (early arrears) provides most of the predictive power. Second, higher tree complexity can be thought of as modelling higher-order interaction effects than the two-way terms included in our penalised logistic regression models (Hastie et al., 2009); this may also partially explain the observed predictive performance difference between BRT and LR.

The observation that much of the predictive power of the models is down to a relatively small subset of dominant predictors could partially explain why RF did not perform better. They have been shown to perform especially well on high-dimensional data (Breiman, 2001), in which there may be a large number of variables that each can contribute to the model predictions. With a small number of strong predictors, there is the risk that those may often end up being overlooked by the random selection of *mtry* variables considered at each tree split, particularly if *mtry* is set to a small value. Furthermore, because of the imbalanced nature of the data, RF also required the introduction of undersampling into the algorithm (Breiman et al., 2004), which may have been a further factor.

[21] For BRT, this measure is based on the number of times a variable is selected for splitting, weighted by the squared reduction in deviance averaged over all the trees in the model.

As past/recent delinquency is usually a powerful predictor in any behavioural scoring system, the fact that this variable has a strong but varying influence in all of the portfolios is not surprising. It is interesting to see that, while current LTV ratios, repayment ratios, and loan to income multiples at origination are important in BRT, their relative importance ranking differs across the portfolios. This suggests that, even with a relatively homogenous mortgage product in the same geographical market, each of the portfolios still benefits from a custom-built default prediction model that makes different use of available characteristics.

Semi-parametric GAMs performed almost as well as BRT in terms of H-measure performance. Unlike BRT, they required minimal tuning. Another attractive feature of GAMs, which has likely contributed to their performance, is their ability to handle situations where some of the continuous predictors may have a non-linear effect on the response. For example, a series of plots showing how smooth terms vary with a selection of predictors are included in Figure 2.5, for portfolio 4. They indicate that, keeping all other predictors fixed, higher current LTV or loan to income, and lower loan age, tend to increase the log odds of default, but not linearly. Also, near the lower end of its value range, a smaller repayment-to-income ratio could actually be associated with higher log odds of default; this may be due to modification/forbearance policies which reduce monthly repayments for borrowers in difficulty. Clearly, with a linear classifier, one would struggle correctly specifying such non-linearities.

Note that in the results presented here, no interactions have been included in the GAM specification. Extending the GAM-based approaches to include interactions identified by BRT could help reduce the search space for important interactions. It is possible to go one step further and use GAMs as the base classifier in ensembles, combined with various ways of augmenting the input data such as bagging (DeBock et al., 2010) and boosting (Caruana et al., 2012).

The portfolios all exhibited class imbalance. In the empirical experiments, this did not affect most classifiers. However, as noted earlier in sub-section 2.4.1, the class imbalance was affecting RF performance for one portfolio (portfolio 3). This required under-sampling for this portfolio. Note this would not have altered the conclusions of the statistical tests as this was the lowest ranked under the H-measure; for the AUC, this would change the slightly the t-statistic and p-values by the not its significance.

2.7 Conclusions and future research

This paper compared four techniques for the purpose of predicting mortgage defaults. Two of these techniques have their roots in the machine learning: Boosted Regression Trees (BRT) and Random Forests (RF). The other two are statistical models: penalised Logistic Regression (LR) and semi-parametric Generalised Additive Models (GAMs). The predictive performance of these approaches was assessed using the H-measure and performance differences on four large real-life datasets were evaluated using an appropriate statistical testing procedure.

The results of the empirical study showed that BRT performed significantly better than LR. Although BRT and GAMs were first and second in the overall ranking, there were no statistically significant differences between BRT and GAMs or RF. The ability of BRT and RF to capture variable interactions and the handling of non-linear effects in a GAM may have contributed to their performance in this setting. The study thus suggests that the tree-based methods and semi-parametric GAMs could be

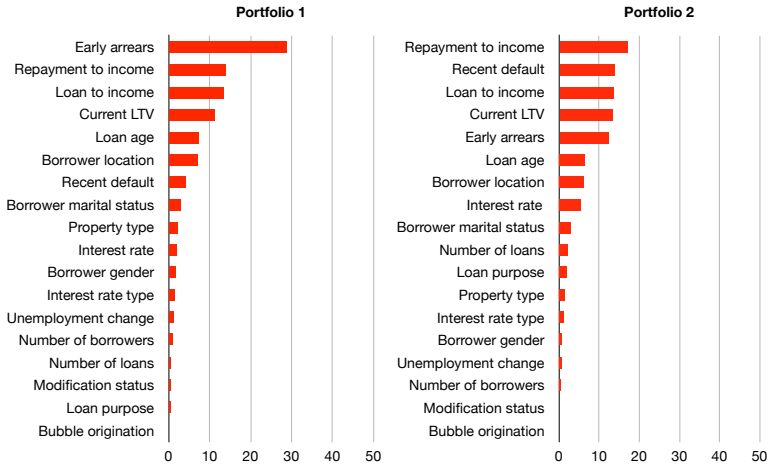

Figure 2.3: BRT variable importance plot: portfolios 1 and 2

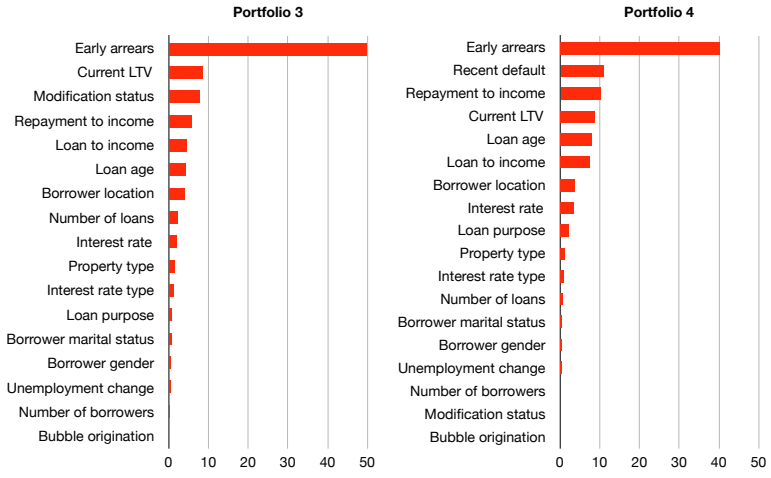

Figure 2.4: BRT variable importance plot: portfolios 3 and 4

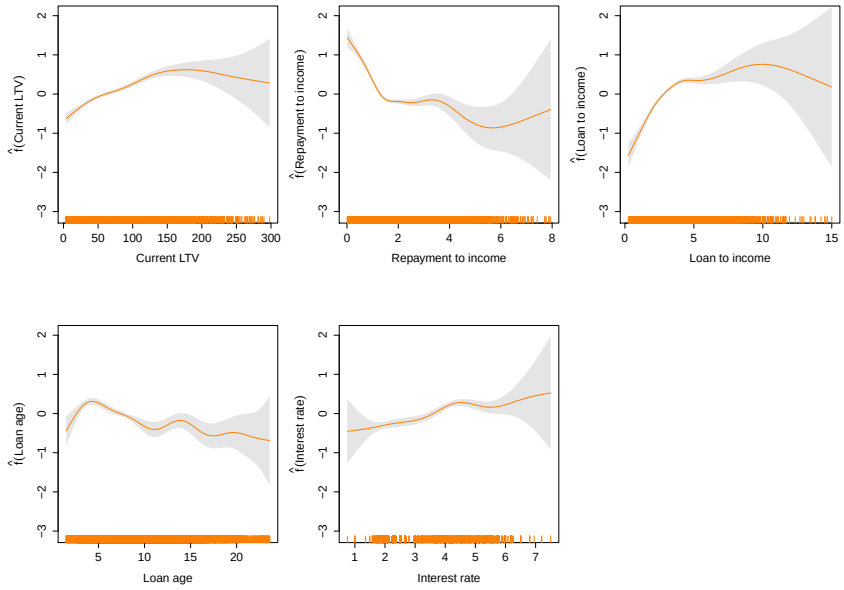

Figure 2.5: GAM estimated smooth functions for portfolio 4

more widely used in credit risk applications, particularly in exploratory modelling where it is not known ex-ante which predictors are important. Even if the end product is not a BRT model or GAM, these models may help to identify suitable interaction or non-linear terms to add to more conventional logistic regression models. This may be particularly relevant if linear classifiers such as logistic regression are still preferred for business or regulatory reasons. While the overall differences in performance between some of the methods may appear small, even small improvements may mean significant revenue savings depending on the application Baesens et al. (2003a).

The practical relevance of these results for a bank are GAMS or BRTs could be incorporated into a collections system such as that outlined in So et al. (2020); for a regulator they can help identify arrears cases with a high likelihood of being in arrears and the factors associated with them. This could be useful in off-site and on-site supervision.

Care should be taken when generalising these findings to other jurisdictions or other types of (unsecured) lending, as the context and drivers of arrears and default are likely to be different. Furthermore, the models in this paper are based on data observed during a time of severe economic distress, during which the distribution of good and bad borrowers may have shifted (Hand, 2006). It is also unclear, due to data limitations, whether changes in borrower behaviour and financial sector policies such as forbearance have had an impact on arrears incidence. Therefore, it is up to practitioners to test empirically whether these techniques produce similar results for their particular portfolios.

Several directions for future research could be considered. First, boosting could be carried out on the semi-parametric GAM to see if this produces further performance gains (Bühlmann and Hothorn, 2007; Tutz and Binder, 2008). Second, using a different type of GAM may offer alternative ways to handle class imbalance (Calabrese and Osmetti, 2013).

A third extension could be to consider the use of misclassification costs for ensemble-based approaches. This may be important in applications where the costs of misclassifying arrears cases vary between the two types of errors, i.e. false positives and false negatives. For example, arrears management teams or regulatory authorities may view the costs of incorrectly classifying an arrears case as a non-arrears case as higher than the converse. Incorporating this into a boosting algorithm in a manner similar to Berk and Kriegler (2010) may lead to improved performance.

Finally, exploring how to model population drift and how that may affect model performance would also be an interesting area of research (Krempl and Hofer, 2011). More practically, testing over various prediction horizons (18, 24 months) and perhaps fitting models to a longer time span than the one used in this study would be beneficial before deployment either within financial institutions or by regulatory authorities.

Chapter 3

Do lenders prosper? Assessing returns in peer-to-peer (P2P) lending using machine learning

3.1 Abstract

Successful Peer-to-Peer (P2P) lending requires an evaluation of loan profitability from a large universe of loans. Predictions of loan profitability may be useful to rank potential investments. In this paper, we investigate whether various types of prediction methods and the types of information contained in loan listing features matter for profitable investment. A range of methods and performance metrics are used to benchmark predictive performance, based on a large dataset of P2P loans issued on Lending Club. Robust linear mixed models are used to investigate performance differences between models, according to whether they assume linearity, whether they build ensembles, and which types of predictors they use. The main findings are that: linear methods perform surprisingly well on several (but not all) criteria; whether ensemble methods perform better than individual methods is measure dependent; the use of alternative text-based information does not improve profit scoring outcomes.

Keywords: risk analysis, investment analysis,P2P lending, predictive modelling, ensemble learning

3.2 Introduction

Peer-to-Peer (P2P) lending is a type of crowdfunding in which an online platform enables borrowers to obtain credit from a large number of individual lenders. Unlike other types of crowdfunding, which may be for altruistic motives, in P2P lending the lender has a financial return motive. The growth in this type of lending has been spurred by technological advances, changing consumer habits, higher costs of and lower access to bank finance for borrowers, and lower returns for investors from traditional investments (Vallee and Zeng, 2018). At present, the two largest P2P platforms in the US, Prosper and Lending Club, together lent over $42 billion by the end of 2017. In the Asia-Pacific region including

China, lending by alternative finance providers (including P2P lenders) amounted to €221 billion at the end of 2016; in Europe, the total amount lent was just under €7.6 billion by end 2016.[1] In this paper, data from the Lending Club (LC) platform is used, as it is one of the largest P2P lenders currently operating in the US.

Similarly to traditional retail credit scoring, P2P loan platforms screen potential borrowers against their own acceptance criteria. For example, borrower identity verification, a minimum credit bureau score, and other criteria may need to be met. After acceptance, borrowers are scored and allocated to a certain grade, based on their characteristics, the requested loan amount, and their credit history. The loan is then listed on the platform. At this point, the decision whether to lend lies with the investors, as do the associated return and credit risk – if the borrower defaults on their payment obligations, the investor takes a loss. This is in contrast to bank lending, where once a borrower is accepted, credit is advanced by the bank and it is the bank itself that bears the risk and makes the return. To make this investment decision, P2P investors must weigh the importance of various attributes in determining whether a loan may present a profitable investment. However, it is not feasible for an investor to manually assess the large volume of listings. Nonetheless, the potential gains of a systematic assessment could be significant as, in recent years, realised returns for this type of investment are comparable to those earned on high-yield bond portfolios.

This prospect has attracted various types of investors. In the early years of P2P investments, they mostly consisted of retail investors funding individual loans. In recent years, institutional investors have become important in this market as well.[2] For some platforms, recent research has suggested that active or "loan-picking" strategies may yield more than passive institutional strategies (Balyuk and Davydenko, 2018). Therefore, an algorithmic approach that can produce loan-level predictions of (risk-adjusted) loan returns could be useful to rank potential investments. A comprehensive assessment is both timely and relevant because there are a wide range of prediction models and algorithms to choose from, various types of predictors, and different experimental settings to judge the effectiveness of such methods. The main goal of this paper is to provide this assessment.

In so doing, the paper makes three main contributions. First, we contribute to the emerging P2P literature (Vallee and Zeng, 2018; Jagtiani and Lemieux, 2018) and profit-scoring literature (Garrido et al., 2018; Verbraken et al., 2014), by assessing whether a profit-scoring approach is more useful to investors than one solely focused on avoiding default. We examine three differing alternative performance metrics from classification, ranking, and regression. This may help investors assess a suitable approach for loan selection.

Second, we contribute to the empirical assessment literature of machine learning models through using a variety of performance measures and a specific experimental framework to compare profit scoring methods. Given the relative success of non-linear and ensemble prediction methods in other application settings, we augment the standard testing framework to test the importance of these factors for performance. This broadens the literature to include factors associated with the variability of performance across methods, rather than solely identifying differences using the standard methods of omnibus tests for differences across methods.

[1] Based on Lending Club and Prosper website data, SEC filings, and Ziegler et al. (2018).

[2] On the Prosper platform, over 90% of loans were provided through institutional channels (Balyuk and Davydenko, 2018)

Third, we investigate whether alternative text-based information provided along with the loan listing has predictive value in this setting. This adds to the emerging research area of the use of alternative data for scoring in this alternative form of financial intermediation. If additional sources of information have predictive content, then it may provide more profitable investment opportunities.

The paper is organised as follows. The next section reviews related work and formulates the research questions. Sections 3 and 4 describe the data and methods, respectively. Section 5 then outlines the experimental design. The results of the experiments are reported in Section 6. Section 7 provides further discussion and elaborates on some of the robustness checks carried out. Section 8 concludes.

3.3 Related work and research questions

Against the backdrop of an evolving P2P lending market, a body of literature on P2P loan profit scoring is emerging. This work cuts across two different research communities – the Operations Research (OR) community, which tends to focus on P2P loan scoring methods, and finance, which studies the specific properties of this new form of financial intermediation and its implications for risk and profit.

A first perspective in the OR literature on credit scoring for P2P lending is provided by Malekipirbazari and Aksakalli (2015); Emekter et al. (2014). Using the Random Forests algorithm, Malekipirbazari and Aksakalli (2015) find that credit history variables and score/grade application information are the most important determinants for Lending Club (LC) loans that default. The paper by Emekter et al. (2014) uses a logistic and a Cox proportional hazards model to investigate determinants of default. They find that credit grade, the borrower's debt-to-income ratio, FICO credit score band, and revolving credit utilisation rate are significant predictors of default.

Although default risk is indeed a concern for investors, they are primarily interested in identifying high-return loans, i.e. those loans that present a good trade-off between default risk and interest returns. Hence, a profit scoring perspective may be more appealing to them. In the P2P context, loan selection based on estimated profitability is particularly important, since a P2P investor, unlike a traditional bank, cannot benefit from the risk diversification of taking on large portfolios of loans and, on most platforms, they cannot set risk-adjusted prices.

The current P2P literature on profit scoring methods, however, is limited. Both Serrano-Cinca and Gutiérrez-Nieto (2016) and Guo et al. (2016) find that various profit scoring approaches are useful to generate returns for investors. They considered a limited selection of regression and non-parametric methods such as CART, logistic and kernel-based regressions, respectively. These are valid approaches. However, other methods such as deep learning (Kim et al., 2019; Sirignano et al., 2016) and ensemble methods such as random forests (which build not just one model but combine multiple estimates) have been found to be competitive in various related tasks. These include profit scoring applications (Verbraken et al., 2014), credit scoring (Lessmann et al., 2015), and in other related applications (Fuster et al., 2018; Lessmann and Voß, 2017; Kim et al., 2019). This suggests a need for a more systematic comparison, in particular one that comprises both non-linear and ensemble methods and assesses their ability to improve predictive performance in the P2P profit scoring setting.

A second perspective on P2P lending is provided by the finance community. Their research considers various aspects of P2P financial intermediation. These include how investors adapt to specific changes

in platform operation and available information (Miller, 2015), as well as broader considerations of how this type of lending could change financial intermediation mechanisms (Balyuk and Davydenko, 2018; Vallee and Zeng, 2018; Jagtiani and Lemieux, 2017). As well as assessing the impact of more traditional factors linked to creditworthiness (e.g. credit score, grades, debt ratios), this body of literature has also focused on alternative or "soft" information available in the P2P context, such as appearance and text descriptions (Duarte et al., 2012; Hertzberg et al., 2016; Jagtiani and Lemieux, 2018).

Whereas "hard" information can be easily compressed to numerical values or attributes (Liberti and Petersen, 2017), alternative information may be unverifiable/costly to verify, or based on some non-standard format like images or free text. Emerging research points towards some role for alternative or soft information, once processed appropriately, in predicting the probability of attracting P2P funding and subsequent credit risk performance (Duarte et al., 2012; Lin, 2016; Dorfleitner et al., 2016; Jagtiani and Lemieux, 2018). However, it is unclear whether this finding is platform-specific as most of this research, with the exception of Dorfleitner et al. (2016), is based on the Prosper platform. Analysing data collected from German P2P platforms, Dorfleitner et al. (2016) instead find that the list text may influence the funding probability but does not appear to be informative for default prediction. In any event, further work is needed to establish whether soft information has any added value for profitability scoring.

Based on these gaps in the literature, the main goal of this paper is to empirically investigate which types of methods and sources of data are able to provide investors with more accurate predictions of P2P loan profitability. This is a broad objective; hence, it is useful to distinguish three specific research questions.

The first question relates to how different methods characterise the relationship between the profitability measure and the predictors. Linear methods, such as penalised linear regression approaches, may suffice if the underlying relationship between loan profitability and its predictors is linear. However, if the underlying relationship is non-linear, then methods originating from the machine learning community could provide a significant edge over linear regression based methods. Given that there is little theory to guide the selection of either approaches in this application setting, this forms the basis of the first research question: *Are non-linear models better at predicting P2P loan profitability than linear models?*

Second, having seen some evidence in the credit scoring and related literature that ensemble methods tend to perform better than single models (Lessmann et al., 2015; Lessmann and Voß, 2017), it is natural to ask whether this finding extends to P2P profit scoring as well. Hence, the second research question is: *Do ensemble methods predict P2P loan profitability better than individual models?*

Third, and finally, while certain forms of soft (e.g. free text based) information appear to matter for the likelihood of being funded or for default prediction, the predictive power of alternative information remains to be assessed for profit scoring. Given that this source of information is becoming more prevalent as platforms grow, understanding its relevance for investment decisions is also becoming more important. The third research question therefore is: *Does including alternative information into predictive models lead to more accurate and more profitable P2P investments than solely using hard information?*

3.4 Data

The data are from Lending Club's statistical information on application and subsequent payment data for loans originated from its platform. The application data all relate to loans with a 36-month maturity, originated between October 2008 and January 2014. The payment data for these loans start in October 2008 and end in March 2017. All of the loans are closed – they have either been paid off early (i.e., prepaid), paid off at maturity, or the borrower defaulted. The loan-level predictors are a combination of loan, borrower, credit risk and text-derived characteristics; we further added macroeconomic variables to this dataset.

The loan characteristics include loan amount and purpose. Credit risk attributes include the sub-grade assigned by Lending Club at issuance and the FICO credit score band. Borrower characteristics include previous inquiries in the past six months, adverse public records, and delinquencies within the past two years. They also include months of credit history, total open accounts, revolving balance on other credit lines, utilisation of revolving lines, monthly loan instalment to total income, annual income after borrower incomes below/above the 0.01% and 99.99% quantiles are given these quantile value, and overall debt-to-income for the borrower. The categorical variables indicate whether: the borrower's length of employment is unknown; their employment title is missing in the listing; their income is verified; and the borrower is a home owner.

The listing text for each loan is included as a series of features. The text is a concatenation of two free text fields: the listing title and the description provided by the borrower. The text per listing is relatively short with two sentences on average and an average sentence length of 6 words. While there are a variety of possible approaches to including the text as features including word-embeddings, this may not be productive here as these short-texts suffer from sparsity, i.e., limited word co-occurrence in each listing. We use a method adapted to short-texts called a Bit-Term Topic Model (BTTM) Yan et al. (2013) utilising the word co-occurences in all of the training listing texts rather than individual listings. Because this is based on terms in all of the training listing texts, it can overcome the sparsity problem in a single listing document. The BTTM is fit with a total of 18 topics. These topic probabilities were then included as features in the models and the resulting performance differences tested between including this set of features and not. Additional detail on the feature construction and topic model are included in Appendix B.

Two controls for prevailing macroeconomic conditions are the state-wide unemployment rate, lagged two quarters before issuance of the loan and the year-on-year change in the OFHEO house price index, lagged two quarters prior to the issuance quarter of the loan.

The profitability measure is the Internal Rate of Return (IRR). This is the discount rate that equates the present value of a loan's monthly cash inflow to the face value of the loan. Formally, the IRR is defined as the value δ for which:

$$Amount_{t0} = \sum_{t=1}^{36} \frac{CF_t}{(1+\delta/12)^t} \tag{3.1}$$

The cash flows, CF_t, are positive as a borrower pays back the loan. If the borrower fails to pay back a loan for four periods, the loan is charged-off/defaulted, and the cash flows are terminated at that point.

The IRR is chosen as a dependent variable because loan-level cash flow data are readily available and, as the IRR incorporates the actual payments made by borrowers, it is a direct measure of return for investors. This helps with comparisons to the literature, where IRR has been one of the main ways of measuring returns. P2P IRRs can be easily benchmarked against returns on alternative investment assets such as consumer credit card Asset-Backed Securities (ABS). To solve for the IRR, a root-finding algorithm is used. Note that, for this problem, the solution of this numerical procedure is unique as there are no irregular repayment cash-flows.[3]

3.5 Methods

Based on the literature, a representative set of regression methods of varying complexity were selected to predict profitability. They can be grouped into two main classes: individual and ensemble. Individual methods or models produce IRR estimates based on a single model. Ensemble methods use multiple instances of a base estimator, e.g. regression trees, combined in different ways.

Using Figure 3.1 as a guide, there are six individual methods specifying a linear relationship between the response variable and predictors. These individual methods are an implementation of a regularised glm based on the elastic net (Zou and Hastie, 2005; Friedman et al., 2010), lasso regression (Tibshirani, 1996); ridge regression (Hoerl and Kennard, 1970), partial least squares (Mevik and Wehrens, 2007), and linear Support Vector Machines (SVM). The L2 linear regression is from (Fan et al., 2008; Helleputte, 2017).

The glmnet is a generalisation of lasso and ridge regression, combining regularisation via the ridge penalty and feature selection via the Lasso penalty. The relative weighting between the two penalties is determined adaptively from the data. Lasso and ridge are special cases of this. Partial least squares forms a linear combination of predictors, chosen in a way to summarise the variation in the predictors themselves and correlated with the response. The linear SVM was chosen to reduce computational complexity (Karatzoglou et al., 2004).

In Figure 3.1, the schematic indicates that a second group of individual non-linear methods including Multivariate Adaptive Regression Splines (MARS) (Friedman, 1991), a simple neural network, and a deep learning model (Candel et al., 2020). The neural network is a very simple single layer feed-forward neural network, with weight regularisation. The deep learning method is a multi-layer feed-forward neural network with two hidden layers of 40 units each with drop out, input drop out, and regularisation. The effect of the hidden layer/input drop out and regularisation is help constrain overfitting. Deep learning has been successfully applied in several finance applications, such as bond return forecasting (Bianchi et al., 2018).[4] MARS is a non-linear regression method that uses an additive piecewise linear representation of the original predictors to approximate a non-linear relationship with the dependent variable (Friedman, 1991).

The ensemble methods selected for the experiments are: random forests, bagged trees, gradient boosted trees, and five stacked models, as illustrated in the lower right-hand side of Figure 3.1. Random Forests

[3] We removed 504 loans that were repaid over a period of more than 36 months, that defaulted but were not charged off, or that were recorded as in default but were actually fully paid. A further 184 loans with zero payments were set to a -100% IRR. This means that cash flows after origination are always positive or zero.

[4] We are grateful to a reviewer of the submitted paper based on this chapter who suggested the inclusion of additional methods.

(RF) are an ensemble method developed by Breiman (2001) which uses the Classification and Regression Trees (CART) recursive partitioning algorithm as a base learner. Many such trees are grown from bootstrapped training samples of the data, the predictions of which are averaged. Each time a split variable is chosen for an individual tree node, the RF algorithm only chooses from a small subset of *mtry* predictors instead of trying all available predictors. This process is repeated over many trees to create a forest. This has the effect of reducing correlation among the trees in the RF, thus reducing variance when averaging the trees; this typically results in improved predictive ability compared to CART or bagged trees. The latter can be thought of as a special case of random forests where the number of predictors is set equal to *mtry*. RFs have been applied successfully in a variety of domains including credit scoring (Lessmann et al., 2015).

Gradient boosted trees use a sequence of base learners that minimise a chosen loss function by the stage-wise addition of a new tree that leads to the largest reduction in loss, given the tree size. With a squared loss function, the focus at each of these steps is on the residuals, i.e. the variation in the response not yet explained by the terms in the ensemble up to that step.

Finally, stacked ensembles use a library or set of first-level models to make a combined prediction. The first-level base models are meant to be a reasonably diverse group. A second-level model, referred to as a "metalearner", learns the optimal combination of these base learners. In this paper, the meta-learners that were tried were a simple average of the first level models, linear (stacked ridge, stacked L2liblinear) or non-linear (stacked gbm, stacked mars).

All of the methods have tuning parameters to optimise their predictive performance. The range of settings considered for each of the methods are summarised in Table 4.2

The software used for all experiments is R. The following packages were used to implement the methods: *mlr* (Bischl et al., 2016); an implementation of MARS in a package called *Earth* (Millborrow, 2018); random forests/bagged trees using the *Ranger* package (Wright and Ziegler, 2017); partial least squares using the *pls* package (Mevik and Wehrens, 2007); *h2o* (Aiello et al., 2019) for the regularised glm, neural network, and deep learning; *glmnet* for ridge and lasso (Friedman et al., 2010); *LiblineaR* for the L2 linear regression (Fan et al., 2008; Helleputte, 2017) *kernlab* linear SVMs (Karatzoglou et al., 2004); *XGBoost* for the gradient boosted trees (Chen and Guestrin, 2016), and *gbm* (Ridgeway, 2012). Finally, the robust linear mixed models were estimated using *rlmer* (Koller, 2016).

3.6 Experimental design

This section describes the overall process flow for the experiments, outlining the choices made at each step of the setup. The prediction problem is to estimate a chosen profitability measure, y_i, for each P2P loan, i, from a vector of selected predictors, $\boldsymbol{x_i}^\top$. A range of individual models/algorithms and ensembles are trained to produce these estimates. As the form of this regression function is unknown, model tuning/selection is guided by optimising a suitable performance measure on the training data.

The various steps and choices in the experimental design are summarised in Figure 3.1. Details are described in the following subsections.

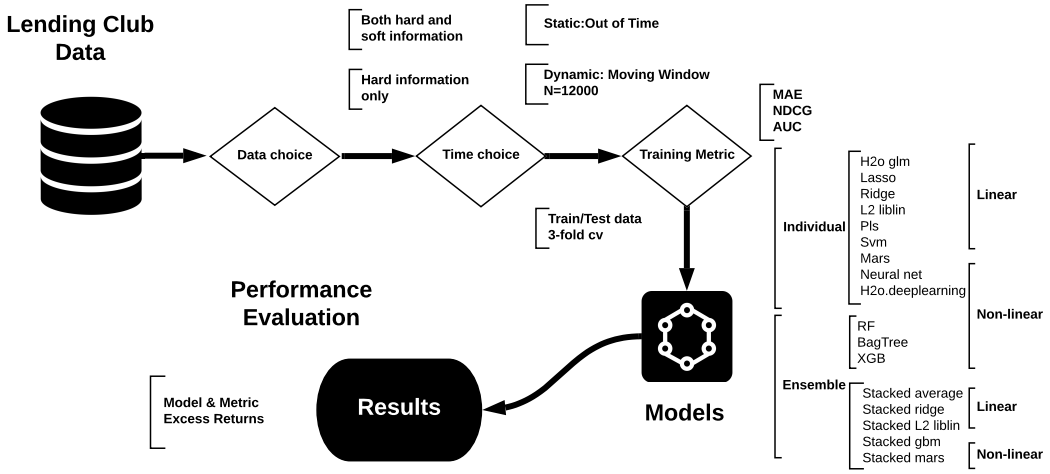

Figure 3.1: Experiment workflow

3.6.1 Predictor selection

The first step involves making a selection from the two groups of predictors outlined in Section 3.4 – i.e., hard and soft. Either all data (including alternative/soft information, such as text-based predictors) or only hard information (excluding the text-based predictors) are selected. The added value of soft predictors can be tested later on in the workflow.

3.6.2 Moving-window and out-of-time tests

In the next step, a series of training/test samples are taken, either using a moving window or out-of-time test framework. Moving window experiments can be useful to investigate how changes in time periods/sample size may affect performance and allow more robust answers to the research questions. Note that in previous work (Serrano-Cinca and Gutiérrez-Nieto, 2016), calendar periods were used instead, which has the disadvantage that results could be specific to one period or may not generalise to other periods, even if careful selection of calendar periods is carried out (Butaru et al., 2016).

Second, for advanced prediction methods to be useful to investors, and as part of a comprehensive empirical approach, an out-of-time test design is added. Using only data on completed loans available at the time of the investment decision, this can provide a more realistic assessment of the performance of various methods.

Both approaches are illustrated in Figure 3.2. For the moving window approach, a window size, n, is selected. The first n observations (according to origination time) are then used as training data; the next n are test data. In the subsequent step, the previous test data now become the training data and a new test set is selected. This continues until the full data set is exhausted. The same window size for train and test is chosen for simplicity and to not introduce another experimental variable. The window size is set to 12,000. In earlier versions of this chapter, window sizes of 6,000-30,000 were used and the results did not differ markedly. For the out-of-time test, the training data used consist of loans with an origination date from October 2008 to November 2010. Given that a 36-month gap is required to observe the returns for the most recent of these training loans, the test data are loans that originated

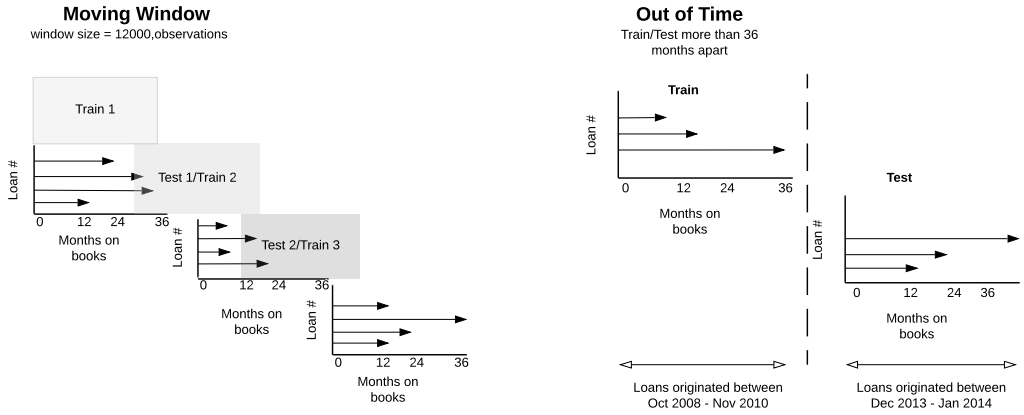

Figure 3.2: Experiment data structure schematic

Table 3.1: Types of models and respective evaluation metrics

Dependent variable	Performance metric
IRR	MAE
Rank-transformed IRR	NDCG
Default $(y\|n)$	AUC

between December 2013 to January 2014. In the out-of-time framework, the training set has 12,799 observations; the test set has 10,658 observations.

3.6.3 Choice of dependent variable and performance evaluation metrics

The next step is to choose the type of dependent variable and appropriate performance measure for model tuning and evaluation. Previous work on P2P profit scoring focused on just one error metric and a limited range of models. However, in most profit scoring settings, the loans are ranked according to predicted profitability and a decision is made to invest in some proportion of the top ranked loans. There are particular challenges in such an application setting to pick one single metric. Therefore, three different evaluation measures are used – the Mean Absolute Error (MAE), the Normalised Discounted Cumulative Gain (NDCG), and the Area Under the ROC Curve (AUC) (see Table 3.1).

The MAE is the absolute residual between the predicted IRR and the actual IRR of each loan, averaged over all n observations.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - f(x_i)| \tag{3.2}$$

Although MAE has an intuitive interpretation and is a robust measure of prediction accuracy, its use may not necessarily lead to higher returns. Investors have limited budgets and may only be interested in the top-k loans. Hence, an alternative approach is to focus on the relative loan return, i.e., using a rank-transformed IRR as the dependent variable. An additional rationale for transforming IRRs is given by the non-normality of the distribution of IRR which may lead to non-normal residuals in a standard regression model. In this setting, it is natural to turn to an Information Retrieval (IR) metric.

One such metric, which is widely used in the learning to rank literature (Liu, 2011), is the Normalised Discounted Cumulative Gain (NDCG). This metric is useful when there are non-binary relevance scores in a complete ranked order, such as ranked IRRs. In this paper, loans are evaluated over the top k results; e.g. with k=100, the accuracy of each method is evaluated by how well the predicted rank of the top 100 loans compares to their actual ranking. The value chosen for k= 100 reflects one way of comparing to the profit scoring literature.

NDCG is calculated by using the predicted relevance score $R(m)$ for item m – here the predicted rank (equation 3.3). In this paper, a loan with a higher IRR receives a lower numerical rank. This is divided by a discount factor to reward better predictions of the rank of items at the top of the list compared to further down the list. In the literature, the discount typically is $log(1+m)$ or $log_2(1+m)$. This results in the Discounted Cumulative Gain (DCG). The term Z_k is a normalisation term to scale the ranking from 0 to 1.[5] A higher value indicates a better ranking; i.e., a value of 90% or 0.9 deviates by 10% from the ideal ranking.

$$NDCG(k) = Z_k \sum_{m=1}^{k} \frac{R(m)}{log_2(1+m)} \tag{3.3}$$

Since borrower default would be a key event turning any potential profit into a loss, and given that default prediction is the standard scoring practice in consumer lending, we also build models to predict default. In this case, the dependent variable is binary with 1 indicating a default event, and 0 indicating no default. To measure the predictive ability of this third series of models, we use a widely used metric – the AUC, which is short for the Area Under the Receiver Operating Characteristic Curve (ROC). The ROC curve is created by evaluating event probabilities produced by the model across a range of cut-off or threshold values. For each threshold value, the true positive rate (sensitivity) and the false positive rate (1 - specificity) are plotted against each other.[6] The AUC is the area underneath this curve; the higher the AUC, the better the model is able to discriminate between default and non-default.

Other measures such as Expected Maximum Profit (Verbraken et al., 2014) or the H-measure (Anagnostopoulos and Hand, 2012) could be used. EMP is a useful measure but would likely need further adaption to the present setting, and we have left this for further research. We used the H-measure but the results were very similar to the AUC results and are therefore not included.

In summary, we built models for three different choices of dependent variable – IRR, rank-transformed IRR, and a binary variable representing whether the loan defaulted. To train and evaluate those models, we used the following three performance metrics – MAE, NDCG, and AUC, respectively.

3.6.4 Model training

The meta parameters are shown in Table 4.2. Each method or model is trained using random search and three-fold Cross Validation (CV). For the moving window test, three-fold CV is carried out for each

[5] It is calculated by assuming that R_{perf} is the perfect relevance or ranking order score, and discounting by the same discount term. Dividing the calculated DCG by the ideal DCG leads to the Normalised Discounted Cumulative Gain (NDCG).

[6] Sensitivity or the true positive rate is the proportion of all events of interest (i.e. defaults) that are correctly predicted by classifying all instances with an event probability greater than the threshold as events. Specificity is the proportion of actual non-events correctly predicted as non-events.

Table 3.2: Regression method training parameters

Name	Meta parameters	Values
h2o glm	alpha	alpha =(0.0001,. . . ,0.5)
lasso	alpha, lamba	lamba =(0,. . . ,1); alpha=1
ridge	alpha, lamba	lamba =(0.0625,. . . ,4,); alpha=0
l2liblin	cost	cost = (0.0001,. . . ,10)
pls	num principal components	number=(1,. . . ,10)
svm	cost	cost=$2^{(-5,\ldots,2.2)}$
mars	degree, nprune, nk	degree=(1,2); nprune=(15,. . . ,40); nk=(10,. . . ,30)
bagged trees	ntrees	ntrees =(100,500,1000); min node size=5
rf	mtry	ntrees=1000; min node size=3; mtry=(3,. . . ,9)
xgboost	eta, max depth, sub-sample, lambda	nrounds=1000; min.child.weight=3; eta=(0.0075,0.01); max depth=(3,4,5,6); sub-sample=(0.5,0.632,0.75); lambda = $2^{(-10,\ldots,-1)}$
neural net	size, l2	size =3; l2= (0.0001,. . . ,0.5)
h2o deep learning	l1, l2, epochs	epochs =(10,20,30); l1=l2= (0.00001,. . . ,0.001); input dropout =0.05; hidden layers =(40,40,40); hidden drop out =(0.5,0.5,0.5)
sl.avg	none	none
sl.ridge	lambda	lambda=0.0625
sl.liblin	cost	cost=0.1
sl.mars	degree, nprune, nk	degree=2; nprune=5; nk=10
sl.gbm	ntrees, shrinkage, train fraction	ntree=500; shrinkage=0.01; train fraction=0.75

window. For the out-of-time test, we take five bootstrap samples (0.632 fraction) of the training data to train the methods also using three-fold CV. The same tuning parameter ranges are applied when the methods are trained and evaluated using the three performance measures.

3.6.5 Statistical testing framework

To answer the three research questions outlined in Section 3.3, a suitable statistical framework must be chosen. This is a different type of exercise than conventional model benchmarking, where the goal is to identify which methods significantly outperform which others; there, a common methodology is to apply a Friedman test to the observed differences in rank performance, followed by post-hoc tests controlling for multiple comparisons (see e.g. Lessmann et al. (2015)). The Friedman test, however, is single factor and tests if there are differences between methods; in this paper, we want to determine if there are differences between predictive performance of the methods and what role the three specific factors play associated with the research questions:

1. whether a linear or nonlinear method is used;
2. whether an individual model or ensemble is used;
3. whether soft information is added to the predictors used in the model.

The experimental factors are approximately balanced for linear vs non-linear (9 vs 8 models) and ensemble vs individual (9 vs 8 models) and balanced for including or excluding soft information. One option to address the questions above is a repeated measures ANOVA in which the model performance metric is the dependent variable and the three experimental factors are the between (linear/non-linear, individual/ensemble) and within (no soft/with soft information) variables.

However, in the current application, some of the assumptions required for ANOVA may not hold. These include that the performance measures be drawn from a normally distributed population and that the variances in performance across methods are assumed to be equal (sphericity assumption). A second challenge to non-normality lies in the nature of the performance measures themselves. The MAE is left-bounded at zero; both NDCG and AUC are bounded between zero and one, with the AUC typically between 0.5 and 1. The first challenge is likely to be more relevant than the second, as models rarely produce an AUC/NDCG of 0 or 1, or an MAE of 0 or a large positive real number.

To deal with these challenges, a Robust Linear Mixed Model (RLMM) is used to produce the results

presented in the main text (Koller, 2016). This approach has two advantages in this experimental design. First, the method can cope with non-normality and outlier observations, allows for differences in error variance and incorporates random effects to account for repeated measures. Second, it allows testing of the experimental factors of interest.

There are some downsides: inference using RLMMs is not yet well developed. Therefore, t-statistics are referred to in the text only.

The results for the first two research questions are presented in one set of regression tables. For the third research question addressing the effect of adding soft information predictors, the result on the information coefficient from these regressions are contained in a separate set of tables. This question requires treating both the model and information type as within factors; i.e., each model experiences both levels of the information factor excluding/including the soft information.[7]

3.7 Results

Because of the two types of experiments conducted, the moving window and out-of-time results are discussed in separate sub-sections. Each sub-section presents the results in three ways. First, results are presented in a table summarising the performance of each method averaged over all model runs. Second, a graph in which the methods are ranked according to their mean performance on each individual metric (note that ranks are used here as the performance metrics are on different scales). Third, each set of results is then subjected to the statistical procedure outlined in section 3.6 to determine whether there are significant differences in performance related to the research questions.

3.7.1 Moving window

To help compare their predictive performance on the moving window experiments, the slopegraph in Figure 3.3 shows the performance ranking of the different methods, across each of the metrics used. A lower numerical rank (lower on the y-axis) reflects better performance. In other words, a lower numerical rank for AUC and NDCG corresponds to a larger AUC or NDCG (see bottom-left and bottom-right sections of the figure, respectively); a lower numerical rank for MAE means a lower absolute error value (see bottom-middle section). The performance values used to produce these rankings are listed in Table 3.3. Each value represents the average test sample performance over the moving window of 12,000 observations.

Overall, three stacked ensembles (stacked ridge, stacked average, stacked liblinear) have the best average rank across the three performance metrics. These are followed by stacked mars and ridge regression. For the (binary) default prediction models evaluated using AUC, stacked methods (ridge, average, liblinear) are top three best performing methods; using NDCG, stacked ridge, linear individual methods such as lasso, ridge are the top three. For both AUC and NDCG, there is very little difference between the top three models.[8] For the MAE measure, svm, l2liblinear, and stacked l2liblinear are the top three methods, followed by average stacked ensemble and h2o.glm.

[7] The linear mixed model fit is a two factor within-subjects repeated-measures model. The first factor is *info* with two levels (hard information only, both types); the second factor is *modname* - the seventeen different model types.

[8] The AUC values are in the range reported by Malekipirbazari and Aksakalli (2015) but lower than the best performing random forest found by those authors. This is likely because the results in Table 3.3 are averages over moving windows, and unlike Malekipirbazari and Aksakalli (2015) are not based on static samples.

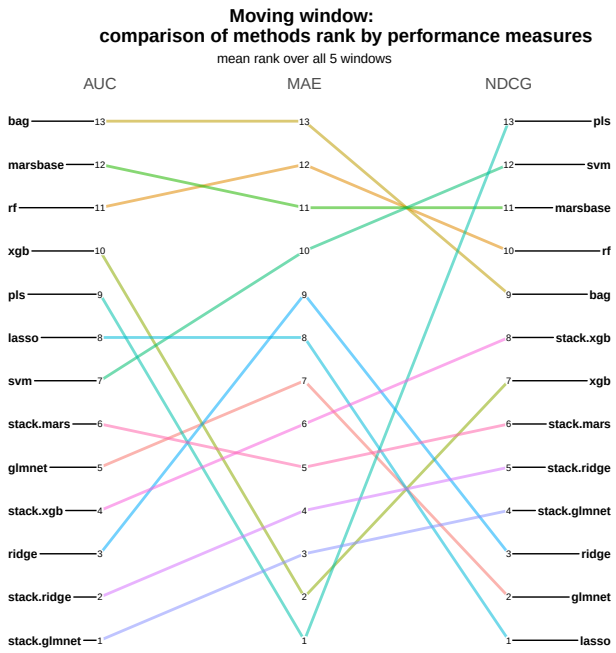

Figure 3.3: Performance ranked over metrics: moving window

A closer look at the results reveals that the performance ranking of some of the methods varies extensively depending on the performance measure. For example, svm is ranked as the top performer when using the MAE criterion, but ranked 17th for AUC and NDCG, respectively. Stacked ridge is the best method using the AUC and NDCG criteria, yet has a much lower ranking when trained using the MAE. This variability could potentially be linked to the chosen tuning strategy, as the parameter range used to optimise the performance metric is not varied across MAE, NDCG, or AUC. However, we believe that the added complexity of further varying the tuning strategies for each performance measure is not required to answer the chosen research questions.

Robust LMM: moving window

Table 3.4 contains the test results for the moving window approach. The variables representing the research questions are categorical. Following the discussion in sub-section 3.6.5, the reference category for the two-level factor *lin.nonlin (non-linear/linear)* is non-linear; the reference category for the two-level factor *ensemble (ensemble/individual)* is ensemble. Dividing the lin.nonlin variable coefficient estimates by their standard error gives t-statistics of -8.34, 0.96,-6.57, respectively, for the MAE, NDCG, and AUC criteria; For the ensemble variable, there is a small t-statistic for MAE (-0.2) and larger values for NDCG (-4.75), and AUC (-5.51); i.e., the effect of using individual to ensemble methods is apparent for two of the three criteria.

In the MAE column, the negative coefficient for the *lin.nonlin* variable means that, surprisingly, linear methods tend to have a lower MAE than non-linear methods; similarly, in the AUC column, the positive coefficients indicate that linear methods are associated with increased performance compared to non-linear methods. For NDCG, the effect is negative with a low t-statistic. On the other hand, compared to ensembles, individual methods are negatively associated with performance when using the

Table 3.3: Rolling window: mean performance by metric

method	MAE	NDCG	AUC
h2o.glm	14.45	0.74	0.66
ridge	14.48	0.83	0.66
lasso	14.55	0.84	0.64
svm	9.10	0.69	0.54
pls	14.47	0.82	0.66
l2liblin	9.11	0.71	0.66
mars	14.50	0.76	0.64
nnet	14.54	0.82	0.64
h2o.dl	14.55	0.80	0.64
rf	14.75	0.81	0.66
bag	15.59	0.81	0.63
xgb	14.50	0.81	0.65
sl.avg	13.59	0.82	0.67
sl.liblin	9.12	0.82	0.67
sl.ridge	14.45	0.84	0.67
sl.mars	14.46	0.82	0.66
sl.gbm	14.54	0.82	0.67

Table 3.4: Robust linear mixed effect model: rolling window

	MAE	NDCG	AUC
Intercept	14.6955	0.8214	0.6533
	(0.2049)	(0.0055)	(0.0015)
lin.non = linear	−1.9465	−0.0060	0.0111
	(0.2332)	(0.0063)	(0.0017)
ensemble = individual	−0.0474	−0.0300	−0.0095
	(0.2365)	(0.0063)	(0.0017)
Num. obs.	442	442	442

Standard errors in parentheses

NDCG and AUC measures, indicating that ensemble methods produce better performance than individual methods with regards to those two criteria. Finally, to test the role of information on model performance, Table 3.5 summarises the relevant coefficient for the information variable from the within subjects regression of model and information. The magnitude of the coefficient on the *info: both* variable has a t-statistic for MAE of 0.24 and 0.17 for NDCG, i.e. the effect of adding soft information on MAE and NDCG is likely insignificant. For AUC, the coefficient is -0.0069; the t-statistic is 3.45 suggesting a small negative effect to adding text-based information predictors on AUC compared to excluding it.

Table 3.5: Rolling: coefficients for information variable in within-subjects regression

metric	Estimate	Std. Error	t value
MAE	0.0207	0.0875	0.2365
NDCG	0.0025	0.0146	0.1684
AUC	-0.0069	0.0020	-3.4598

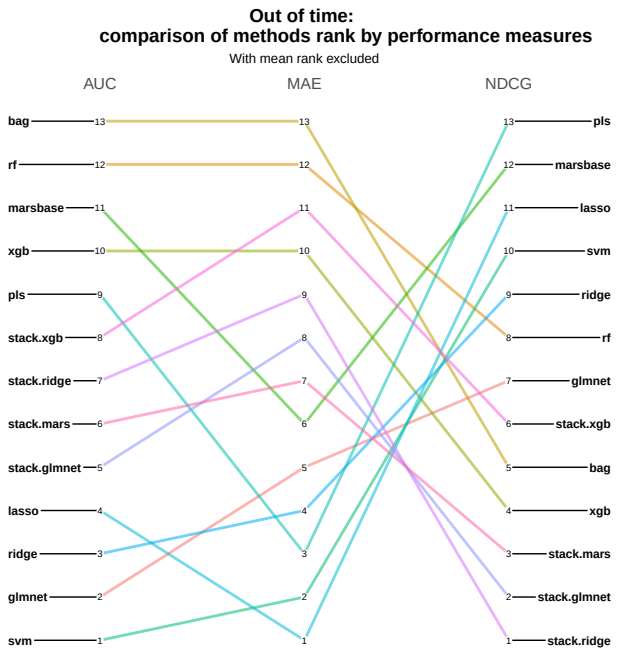

Figure 3.4: Performance ranked over metrics: out of time test

3.7.2 Out of time

The out-of-time setting is a sterner test of each method's predictive ability, in which we expect some deterioration in predictive performance. This is because, at a minimum, at least 37 months have elapsed between the origination dates in the training and test samples (see Figure 3.2). This setting may be more informative to investors who would only use data on closed loans to build their predictive models.

Figure 3.4 and Table 3.6 indicate that performance is more variable compared to the rolling window. In the out-of-time setting, individual models (l2liblin, h2o.glm) and stacked liblin are the best performers averaged over the three measures. For the MAE performance measure, svm, l2liblin, and stacked liblin are the top three; for AUC, l2liblin, h2o.glm, and average of stacked models perform best. Finally for NDCG, the stacked ridge, bagged trees, and stacked liblin are the top three, although there is little difference between mars and the next best: xgboost. However, in some instances, there is a substantial variability in performance; e.g., svm performs well on one of the three measures (MAE) but is one of the worst performers on the NDCG and AUC criteria.

Robust LMM: out of time

This sub-section reports the results of the robust LMM applied in the out-of-time setting. The results in Table 3.7 indicate differences across performance measures between linear and non-linear methods for MAE, NDCG, and AUC. For the coefficient on the factor *lin.nonlin*, the t-statistics are -7.36, 2.49, -4.74 respectively. When the MAE is used as a performance measure, the average reduction in MAE from using linear methods instead of non-linear methods is -2.94, other factors unchanged. When NDCG and AUC are used, linear methods are associated with a somewhat increased performance.

The t-statistics suggest differences between ensemble and individual methods for the MAE and NDCG

Table 3.6: Out of time: mean performance by metric

method	MAE	NDCG	AUC
h2o.glm	13.92	0.77	0.64
ridge	13.73	0.75	0.63
lasso	14.09	0.76	0.58
svm	9.15	0.70	0.56
pls	13.76	0.75	0.62
l2liblin	9.16	0.77	0.64
mars	22.28	0.73	0.59
nnet	14.82	0.73	0.63
h2o.dl	14.46	0.73	0.63
rf	17.26	0.74	0.60
bag	22.49	0.78	0.58
xgb	16.66	0.69	0.61
sl.avg	14.81	0.74	0.64
sl.liblin	9.30	0.78	0.63
sl.ridge	15.70	0.79	0.63
sl.mars	15.52	0.76	0.62
sl.gbm	15.29	0.74	0.63

performance measures (with t-statistics of -4.35 and -4.29, respectively), with ensemble methods outperforming individual methods on NDCG and the opposite for MAE. There are no detectable differences for the AUC. Finally, as shown in Table 3.8 and in the appendix, including soft information has an adverse effect on MAE (with t-statistics of -1.41) on $info : both$, respectively), with some difference for NDCG (t-statistic = 2.79), and none apparent for AUC.

Table 3.7: Robust linear mixed effect model : out of time

	MAE	NDCG	AUC
Intercept	16.7943	0.7622	0.6123
	(0.3513)	(0.0158)	(0.0052)
lin.non = linear	−2.9468	0.0137	0.0139
	(0.4000)	(0.0055)	(0.0029)
ensemble = individual	−1.7671	−0.0240	0.0014
	(0.4057)	(0.0056)	(0.0030)
Num. obs.	170	170	170

Standard errors in parentheses

Table 3.8: Out of time: coefficients for information variable in within-subjects regression

metric	Estimate	Std. Error	t value
MAE	-0.5561	0.3951	-1.4074
NDCG	0.0452	0.0162	2.7977
AUC	0.0021	0.0054	0.3941

3.8 Robustness checks and discussion

3.8.1 Robustness checks

Several robustness checks have been carried out. The first is a consistency check on the moving window and out-of-time results by rank-transforming the dependent variable in the robust LMM to check that any non-normality in the residuals does not lead to invalid inference. The results for this alternative test are shown in Appendix C (see Table C.5 and Table C.6, for the moving window and out-of-time setting, respectively). Comparing these results with Table 3.4 and Table 3.7 leads to similar conclusions.

A second analysis considers the extent to which, in the out-of-time set, superior performance with regards to an evaluation metric is also linked to greater returns. Each method's excess returns is calculated by selecting the top 100 most attractive loans based on that method's predictions and comparing their average IRR return against the mean return rate in the whole test set. Higher such excess returns suggest that an active loan picking strategy using that method produces greater returns than random loan selection. We then rank the methods from largest excess returns (rank 1) to smallest excess returns (rank 17). This IRR rank can now be compared against the same method's performance rank according to MAE, NDCG or AUC (lower rank numbers again indicating better performance).

For each method, Figure 3.5 plots the mean rank based on the performance measure against the return-based measure (IRR rank), for each performance measure and the mean averaged over the three performance measures. The most appealing methods are those that perform consistently well on both criteria (both have a low numerical rank) and thus appear on the lower left-hand side of each panel. The bottom-right corner is where good performance on the metric does not correspond to good performance on IRR rank.

A large difference between performance measure and IRR rank suggests inconsistent performance; i.e., in those cases, better/worse performance on the evaluation metrics may not translate to larger/smaller excess returns, relative to the other methods. For example, in the figure, linear regularised methods (h2o.glm, stacked liblin) have a reasonably good ranking compared to xgb for the mean ranks across performance measures. The figure illustrates the point that a method that minimises MAE or maximises AUC/NDCG for all loans does not necessarily correspond to an investment strategy that deliver excess returns (over the test mean) of the top 100 loans.

Real-world investors may be interested in those methods and strategies linked to greater excess returns, we examine the relationship between excess returns and two sets of variables – the experimental factors (i.e. *lin.nonlin* and *ensemble*) and the choice of dependent variable and tuning strategy (i.e. whether we build models to predict IRR, rank-transformed IRR or default Y/N, using MAE, NDCG and AUC as respective training metrics). To do so, we again estimated a robust linear mixed model with a random effect for inclusion of soft information.[9] The results are shown in the first column ('ALL') of Table 3.9. The respective reference categories for the factors *lin.nonlin* and *ensemble* are non-linear and ensemble; the reference category for metric is MAE. Next to these pooled results, the other three columns in the table assess the impact of the experimental factors separately for each choice of dependent variable and corresponding tuning metric.

The results for the column ALL show a t-statistic for the categorical variable linear and non-linear

[9]Information is treated as a within factor with two levels: *both* and *hard.only*.

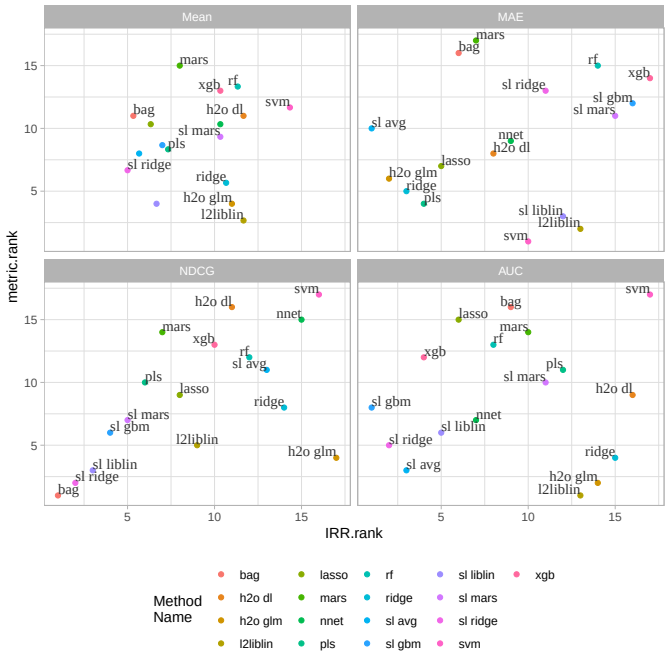

Figure 3.5: Out-of-time difference in rank performance (metric vs. excess returns)

methods (t-statistic = 0.76), suggesting little evidence that either leads to meaningful differences in excess returns. However, looking at the breakdown for the individual performance metrics (columns 2-4), this result is largely down to linear methods performing worse for AUC and NDCG; for the IRR models trained with MAE, the coefficient is 0.84, with a t-statistic for 2.5, whilst for AUC and NDCG, the signs are reversed. This means linear methods are associated with larger excess returns when loans are ranked using the methods trained using the MAE as the performance criterion. Overall, ensemble methods do not have a large t-statistic compared to individual methods (t-statistic =-1.42). Another important finding relates to the choice of dependent variable and tuning strategy: relative to MAE, NDCG and AUC are associated with significantly reduced excess returns (t-statistics = -6.95; -8.99). In other words, the best modelling strategy from an average profit perspective is to predict IRR directly, with MAE as the tuning metric. Controlling for the other factors, this strategy produces larger excess returns than focusing on the IRR ranking (NDCG) or the traditional scoring approach of picking loans based on the risk of default (AUC). In terms of relative magnitude, this effect outweighs that of the other two factors.

3.8.2 Discussion

The variable nature of performance across the three evaluation measures and the test setting used, and its non-trivial relationship with returns suggest that findings recommending specific methods in the existing profit scoring literature may not be generalised easily. The results may be dependent on these factors, in addition to the usual considerations such as the application domain and data used.

In this paper, a successful profit scoring approach is associated with positive excess returns (Table 3.9). The findings suggest that while it pays to model using profitability directly (i.e. using IRR as the dependent variable), performance depends on the methods adopted, the performance measure itself, and the

Table 3.9: Robust linear mixed effect model: excess returns

	ALL	MAE	NDCG	AUC
Intercept	1.6897	0.8540	0.2228	0.1525
	(0.2415)	(0.2940)	(0.2699)	(0.4128)
lin.non = linear	0.1508	0.8390	−0.0797	−0.3067
	(0.1964)	(0.3347)	(0.3073)	(0.3643)
ensemble = individual	−0.2832	0.4978	−0.3883	−0.9453
	(0.1991)	(0.3395)	(0.3116)	(0.3695)
metric = NDCG	−1.6450			
	(0.2366)			
metric = AUC	−2.1283			
	(0.2366)			
Num. obs.	510	170	170	170

Standard errors in parentheses

type of information used. This first finding is in line with Serrano-Cinca and Gutiérrez-Nieto (2016) for P2P lending and other profit scoring literature (Garrido et al. (2018), Verbraken et al. (2014)).

Given the range of methods, and prediction problems, it is not straightforward to identify one set of reasons or features that is associated with better predictive performance. Each of the individual models represent the data in different ways, depending on the type of prediction (continuous, ranking, or binary) and performance measure. For MAE, some models like ridge regression identify credit risk focused variables like utilisation, balance, unemployment and debt to income ratios, as well as categorical information like fico score and LC sub grade; linear SVMs for MAE identify grades as some of the most important factors, whereas using AUC as the performance measure, they identify a range of very different factors. Ensemble methods outperformance of individual methods is metric dependent (e.g. NDCG vs MAE). This suggests that while ensembles can demonstrate good predictive performance as found in Lessmann et al. (2015) credit scoring benchmarking study, in our profit scoring setting, this finding is dependent on the measure used.

Our ability to produce positive excess returns suggests that the predictors may contain information not directly incorporated into Lending Club's grading system during the sample period in this paper. Studying a different research question, Jagtiani and Lemieux (2018) come to similar conclusions for a sample period covering much of the same period as in this paper.

The implications for platform pricing and investing are more nuanced. The information and methods in this study are public and the returns are ex-post, based on closed loans. Therefore, one cannot be overly optimistic about excess returns in future. Platforms like Lending Club do not bear the credit risk; their main income comes from receiving a small fraction of the monthly repayments on all loans. Adjustments to pricing models are one of several considerations for this type of business model, in addition to platform growth due to the supply of new listings.

The negligible to negative impact of soft information may give pause for thought. There may be limitations in this study in the sense that text data has been represented through using a type of topic model adapted for short text. In an earlier version of this paper, we represented the text as certain features such as the fraction complex words and measures of lexical diversity, and obtained similar results. It is likely that the result is negative because the listing text was sparse. The finding that these features

do not help improve predictive performance for profit scoring contrasts with other studies that were based on the Prosper platform, where the text is richer. Instead, our results concur with Dorfleitner et al. (2016) who modelled default in German P2P loans. This provides further indication that results in the literature could be platform-dependent.

3.9 Conclusions

This study explored three research questions motivated by a P2P investment setting. First, we compared whether non-linear methods could provide improved profitability predictions compared with linear methods. Second, drawing on findings in Lessmann et al. (2015), we investigated whether ensemble methods gave better performance than individual methods. Third, as new types of data including soft information in the form of text become available through these platforms, we also assessed their relevance for P2P investment.

In our experiments, we find empirical evidence supporting a profit-scoring instead of modelling default risk. Specifically, we find that linear methods were actually often associated with improved predictive performance, although the magnitude of the effect varied with the performance measure and, in a robustness test, they did not produce greater excess returns on the top-100 loans than non-linear methods. Ensemble methods outperformed individual methods on some metrics (e.g. NDCG) but not all (e.g. MAE). In general, we did not find significantly better performance by including soft information in the predictor set.

The results add to the findings on P2P lending in general, and specifically contribute to the empirical assessment of P2P profit scoring. Considering the research findings, the results suggest that relatively straightforward approaches such as MAE and linear models provide good performance as well as potentially positive out-of-time excess returns, at least for this sample period. A binary classification approach that models default and uses a performance criterion such as AUC results in some excess returns out-of-time, though not as much as using the MAE. Using a ranking performance measure such as NDCG is a reasonable approach on paper but results in lower excess returns on average than using the MAE or AUC.

A relatively consistent result regardless of the performance criterion is that the inclusion of soft information either makes little difference or makes the model perform slightly worse than when trained with only hard information. However, incorporating unstructured data from text and other sources and its utilisation in predictive modelling contexts is an evolving area of research and other representations could provide better predictive ability. Specifically, soft information from P2P platforms with more abundant sources of text-based information could be incorporated using other methods than those considered in this paper. Finally, alternative sources of information such as digital footprint information could be explored further for predictive modelling of P2P loans (Berg et al., 2018).

Chapter 4

Modelling mortgage collateral recoveries

4.1 Abstract

This chapter addresses the problem of predicting the collateral recovery value of defaulted mortgages, by modelling two important parameters determining this value: time to sale (i.e. the length of time before the default is resolved through sale) and forced sale discount (i.e., the percentage loss in sales proceeds relative to the indexed valuation). The predictive performance of a variety of survival analysis approaches to estimate time to sale and regression approaches for the forced sale discount are empirically evaluated. For time to sale prediction, random survival forests and parametric survival models perform best. For forced sale discount prediction, random forests, xgboost, and deep learning methods produce the lowest errors. A sensitivity analysis illustrates how predictive modelling of these parameters produces higher loss estimates than a current industry approach.

Keywords: survival analysis, resolution time, mortgages, non-performing loans, forced sale discount

4.2 Introduction

Mortgage lending is one of the most important types of lending for retail banks in the euro area and Ireland. This chapter focuses on the estimation of resolution times for defaulted mortgages, the value of the collateral upon resolution, and the implications that resolution duration and sale value have for loss severity estimates.

Recoveries on mortgage loans are typically worked-out with cash flows coming from either repayment by the borrower or repossession/sale of the property collateral discounted over the Time To Sale (TTS) of the collateral. The Time to Sale (TTS) is an estimated future time point at which sale of collateral would occur. In distressed mortgage and housing markets, this can mean that loss severities can be higher due to longer sale times. This is a particular problem in some European banking sectors recovering from the Global Financial Crisis (GFC) including Ireland (see Chapter 1, sub-section 1.3.2). This topic is a central focus in Europe (Baudino et al., 2018; European Commission, 2019) and underpins the

introduction of new regulatory requirements such as the prudential backstop for NPLs by the European Commission.[1]

Loss severities at the time of sale can be higher post-GFC due to decreased property prices following a property bubble, illiquidity in property market segments, buyer expectations, and various other effects (Campbell et al., 2011; Donner et al., 2016). For these reasons, and the fact that property sales can be forced through repossessions/surrender of property, there can be a divergence in the sales value compared to the indexed value at the point of sale. This is known as the Forced Sale Discount (FSD). It is the discount on the sale price achieved for distressed assets compared to the indexed valuation at the sale date.

In this chapter, the FSD refers to residential property mortgage sales as a result of repossessions or other bilateral agreements with the bank to sell the collateral (i.e., property) following substantial mortgage arrears and with no other restructuring solution being found. The type of sales considered in this chapter arise from borrower's inability to service their mortgage resulting in a substantial time in default with significant arrears. The borrower will lose ownership either through legal proceedings or a borrower/bank agreed sale or surrender of the property. Understanding these aspects of the work-out process is important for Non Performing Loan (NPL) resolution and related polices. This is a particularly important consideration in the EU and euro area because of the high concentration of NPLs within certain banking sectors and banks.

It is critical for banks, investors, and regulators that loans are correctly valued and credit risk is appropriately understood and managed. In the aftermath of the financial crisis and following criticism of the slow pace of loss recognition for credit risk, a new accounting standard - IFRS 9 - defined Expected Credit Losses (ECL) for a financial instrument as the difference between contractual cash-flows due and the expected cash flows to be received. Credit losses are the present value of all cash shortfalls, i.e., the difference between contractual cash-flows due and the cash flows expected to be received. Expected credit losses are an estimate of credit losses over the lifetime of the financial instrument, i.e., loan. According to the standard, ECLs should reflect an unbiased estimate based on several possible outcomes weighted by their probabilities of occurrence. The critical change in IFRS 9 compared to the previous backward-looking standard (IAS 39) is that now there is a requirement to incorporate forward-looking information to estimate losses.

The three main components of an ECL calculation are: Probability of Default (PD), Exposure at Default (EAD), and Loss Given Default (LGD). In this chapter, we are concerned with the LGD component of the ECL calculation. Within this, the TTS and FSD parameters are important for calculating mortgage ECLs. These parameters link the recovery period to house price index projections that form part of the forward-looking information for lifetime-loss estimation of impairments.

This accounting standard is a principles-based standard with only a high-level definition of ECL and does not require a specific methodology to calculate it in practice. Regardless of the chosen methodology, one reason for modelling these parameters is that part of their variation may be predictable by covariates available at the time of default in the case of TTS and at the time of sale for the FSD. It is an open question, however, whether estimating these parameters through various types of modelling provides a more appropriate approach to credit risk management than assigning a simple average to a

[1] See the EU Commission statement here.

cohort, which is a common approach used within industry (Eder and Bank, 2019; Chawla et al., 2016). Adoption of differing approaches may lead to different loss severity and Expected Credit Loss (ECL) estimates across banks with similar business models.

There are four particular challenges in modelling these parameters. For TTS, recovery data are incomplete in the sense that some loans do not have a completed workout process (sale or full cash repayment by the borrower). Therefore, the possible workout time is unknown. This type of right censoring can occur at the end of the follow-up period. Even though some of the loans do not have an end point, the information on both the closed and censored loans can be combined to produce an estimate of time to sale for all loans in a portfolio. This has the advantage of not downward biasing estimates of TTS by using only loans with a completed sales process.

The second challenge relates to the choice of approach to estimate TTS. Because of the censored nature of the data, a range of methods from survival analysis are considered. It is not apparent which method is to be preferred ex-ante. The methods can range from a basic Kaplan-Meier (KM) (Kaplan et al., 1958) estimator to parametric, semi-parametric, and fully non-parametric machine learning approaches. The implications for the estimation of TTS means that impairments for the same cohort of loans or individual loans themselves could be very different depending on the method chosen to estimate the TTS. Therefore, an understanding of the estimation of TTS based on a variety of model/method selections is an important consideration.

The third challenge is that Forced Sale Discounts (FSD) or haircuts on sale are only available for collateral that has been sold through a bilateral agreement (between borrower and bank) or at the conclusion of a legal process.[2] One approach to estimating the FSD is calculating an average for certain groups of disposals similar to the segmentation approach for TTS as mentioned earlier in this section. Whether those groups or cohorts are the most appropriate could be based on what are the important predictive factors for the FSD.

The fourth and final challenge for estimating TTS and the FSD is as a result of the GFC in several euro area and EU countries, there were significant macroeconomic volatility and policy changes occurring throughout the recovery period (Dendramis et al., 2018). Combined with depressed collateral values during a volatile economic environment, this may affect resolution times and loss severity. Understanding the effects of both the policy and macroeconomic developments on the estimation of these parameters is important to ensure models are sufficiently robust.

4.3 Related Literature

4.3.1 Default resolution time

Because of the focus of the research questions, this paper draws on two strands of the LGD modelling literature. The first strand of the literature adapts survival analysis to investigating the work-out LGD of SME loans in Dermine and de Carvalho (2006) or mortgage loans in Chen (2018), and recovery times of SME loans in Betz et al. (2016, 2017).

In this literature, rather than directly regressing LGD against its relevant drivers, there is a focus on

[2]The term forced sale is used in this chapter includes both legal repossession/receivership and properties sold as a result of a bilateral voluntary sale or surrender agreement.

Default Resolution Time (DRT) for two reasons. First, Dermine and de Carvalho (2006) and Betz et al. (2016, 2017) suggest that for SME loans, longer resolution times are associated with higher LGDs and increased costs. Betz et al. (2016) provide an interesting illustration of the impact of different specifications of a Cox survival model on inference for SME/Corporate loans and the resulting portfolio's unexpected losses. An improved understanding of TTS drivers may lead to a better understanding of these issues. Second, adaptation of survival analysis methods has practical advantages for prediction of estimated work-out time for incomplete workouts. Incomplete workouts are a feature of the data considered in this paper, in common with many other recoveries data sets.

Various proposals for the estimation of resolution times have been made in Rapisarda and Echeverry (2013) and Betz et al. (2016, 2017), along with an analysis of how factors included in these models can affect recovery time. The latter find that loan and collateral-specific features are important for explaining default resolution times of SME loans in the US, UK, Canada, and Germany. Based on this useful cross-country perspective, they find mixed evidence for macroeconomic effects in recovery time modelling, as this tends to be country-specific.

The possibility of using survival analysis for mortgages was mentioned in Leow and Mues (2011), and applied in Leow et al. (2011) in the context of a two-stage LGD model. Elsewhere in the literature, however, there is limited consideration of the impact of resolution time modelling approaches on estimating mortgage losses.

The second strand of the literature deals with various drivers of LGD (Loterman et al., 2012; Gurtler and Hibbeln, 2013; Tong et al., 2014; Krüger and Rösch, 2017). One related challenge in LGD modelling is that having only short model development samples (c.5 years) available compared to the longer work-out time necessary for completed recoveries means incomplete observation for some loans. While the type of credit portfolio being modelled is varied, the quality of LGD predictions depend on the modelling approach followed (Bade et al., 2011; Loterman et al., 2012; Krüger and Rösch, 2017) and on the groups of covariates used. These include transaction-related factors such as Loan to Value (LTV), seniority, collateral types, product features, and obligor characteristics (Qi and Yang, 2009; Zhang and Thomas, 2014; Bellotti and Crook, 2013; Tobback et al., 2014; Andersson and Mayock, 2014). For US mortgage LGDs, higher LTVs, and borrower liquidity constraints have been found to be important predictors of non-zero LGDs by Xuan et al. (2019). Betz et al. (2016) found that both macroeconomic variables and the inclusion of frailty terms improved inference for SME resolution times.

4.3.2 Forced sale discount

The Forced Sale Discount (FSD) parameter has received somewhat less attention in the credit risk modelling literature. In the economics literature Campbell et al. (2011) and Donner et al. (2016), explored a range of factors explaining the existence and magnitude of the this discount. These included the nature of the sale itself being forced as well as the illiquidity of property markets in downturn or distressed conditions, seller incentives, the holding cost of the property, and legal frameworks. Two relevant findings are that the process of foreclosure itself is associated with substantial discounts compared with normal sales transactions; its magnitude varies by geographical area and the nature of the property being sold (Lee, 2010). This literature does not consider a wide range of models to predict the discount.

In the credit risk literature, Park and Bang (2014) find for Korean mortgages that current LTV matters

the most for loss severity estimation, and the foreclosure auction process is an important determinant of LGD variability. This suggests a role for institutional factors that may influence the recovery process (Park and Bang, 2014; Zhao et al., 2019).

There are two particularly relevant studies in the credit risk literature for prediction of mortgage default haircuts. The first of these, Somers and Whittaker (2007) use quantile regression to model the haircut on a European mortgage data set for LGD modelling. They use this technique to understand variation around the median discount as well as its dispersion. Using data on defaulted loans from a UK bank, Leow and Mues (2011) modelled a post-default process where the outcome could be no repossession or repossession and sale. Leow and Mues (2011) find that inclusion of variables such as previous defaults and the type of security improved LGD predictions for a two-stage mortgage model, compared to a single-stage approach.

The latter paper is related to the present work but there are important differences. First, in this chapter, the data are from the final stage of the collateral recovery process where loss of ownership has occurred or will occur via repossession or borrower agreed sale/surrender. Second, in this research, the distinction between legal (forced) and borrower agreed (voluntary) sale/surrender is particularly important. Because of the scale of the crisis in Ireland, public policy, and legal issues surrounding repossession, approximately two thirds of loss of ownership take place through bilateral voluntary sales/surrenders and one-third through court ordered repossession. As noted in Section 4.2, this type of forced sale is an important feature of collateral recovery in Ireland and some other euro area countries that has not yet been explored in the post-GFC recoveries modelling literature.

4.4 Research objectives and contributions

The two previous sections suggest two specific gaps in the research literature. First, there is a limited body of work concerning mortgages and no evaluation of the performance of a range of methods to predict mortgage resolution times. This is somewhat surprising given that models are used to produce mortgage loss estimates within industry. This gap is particularly relevant when there are substantial amounts of NPLs yet to be worked out, post-GFC, and just after the introduction of IFRS 9 in Europe where lifetime ECL losses have to calculated.

Second, there is a relatively sparse literature on modelling the forced sale discount from an IFRS 9 credit risk perspective. This is an important parameter for estimation of loss severity and understanding its driving factors is important for both accurate impairment forecasting and prudent estimation of capital requirements through appropriate collateral haircuts in LGD models.

This leads to the three main questions addressed by this research:

- Which modelling approaches for defaulted mortgages produce the most accurate TTS and FSD predictions?
- What are the important predictive factors identified for TTS and FSD in this context?
- What are the real world impacts of assuming fixed parameter values versus those based on predictive models in loss severity calculations?

This research makes three contributions. To the author's knowledge, it is one of the first papers to

examine the impact of estimation method on both the TTS and the FSD as these are critical for determining the realised LGD. In particular, the predictive accuracy of three groups of methods for TTS are assessed as well as the important factors identified by those methods for time to resolution. This an area of active policy concern in the EU to understand the time to recovery and recovery rates on Non-Performing Loans.[3] It is relevant also in Ireland and other countries where mortgage arrears are still significant a decade after the GFC (Figure 1.2).

This work contributes to understanding the drivers of the FSD by considering three modelling approaches and the importance of the type of sale (i.e., forced or voluntary) for the estimation of the FSD. This is an aspect that has not yet been focused on in the FSD literature. This is an important aspect of Irish mortgage loss experience where two-thirds of loss of ownership occurring through a voluntary surrender/sale; and one third from the repossession legal process (Figure 1.3). Analysing both TTS and FSD together provides a more complete understanding of NPLs resolution process.

This paper makes a second contribution to understanding how the roles of two key individual model parameters matter for calculating life-time losses for an important asset class under IFRS 9 standards used in the EU and Asia. For practitioners and regulators concerned with IFRS 9 impairment models, estimation of life-time losses for collective impairment calculations requires appropriate estimation of these parameters.

This accounting standard is still relatively new compared to Basel Internal Rating Based (IRB) approaches to LGD modelling. Therefore, this work improves understanding the suitability of various approaches (cohort-based averaging or modelling), as well as the performance of different types of predictive models. Furthermore, estimation of these parameters and their validation through using multiple approaches provides a sensitivity analysis of model risk.

This work may be of interest for LGD estimation by practitioners and regulators as banks may group collateral into homogenous groups based on recovery processes and duration, taking into account the potential biases arising from incomplete recoveries.[4] The findings are relevant for the repossession stage in multi-stage modelling of Loss Given Default (LGD) (Leow and Mues, 2011; Tanoue et al., 2017; Xuan et al., 2019) through improved understanding of factors associated with the FSD.

The paper's third contribution broadens the scope of retail asset classes considered in the default resolution time modelling literature. Thus far, a significant amount of the emerging literature on resolution time focuses on Small and Medium-sized Enterprises (SME) and corporate credit, in part, because of availability of data from various sources including the Global Credit Data consortium.[5] As discussed in Section 4.2, mortgages are one of the most substantial asset classes for European banks and are the subject of this paper.

Both the first and second questions address the gaps in the research literature. The third question assesses if the insights gained from improved TTS and FSD predictive modelling are meaningfully different from one current industry approach of using fixed parameter values.

[3] See European Commission request for advice to the European Banking Authority on the efficacy of judicial enforcement frameworks in the EU.

[4] See EBA 2017, paragraphs 128 and 159(c)

[5] This is voluntary private sector initiative to collect/pool historical loss information.

4.5 Data

The data used in this study are confidential mortgage recoveries data collected by the Central Bank of Ireland as part of their supervisory analysis of mortgage models. They are residential mortgage recoveries from two banks. The data are for loans in long-term mortgage arrears where legal proceedings have been undertaken by the banks to enforce their security on the collateral or from a bilateral borrower and bank agreed sale/surrender to settle the mortgage debt including arrears. Because of limitations of the data, there is no information on cash recoveries (i.e., payments made by the borrower prior to the sale of the property). As these are loans already within a loss of ownership process (i.e., bilateral sale or legal process), there are no cures. The data cover loans originated between 1998 and 2015 and recovery period between 2009 and 2017.

Time To Sale (TTS) is defined as the minimum of the time to sale date or (censored at) the end of follow up time minus the date of default. The unit of measurement is months. A loan has a resolution event if a sale has concluded with a sale date. An unresolved loan is one that has not been sold (i.e., resolved) by the end of the follow-up date. As an example, a TTS of 60 and a resolution event means it took 60 months, i.e., five years, between the date of default and sale. Because of confidentiality restrictions, the exact proportion of events and censoring for each bank cannot be disclosed.

The FSD is the haircut on sale of the property: $FSD = (1 - \frac{\text{sale proceeds}}{\text{indexed valuation at sale}}) \times 100$. The ratio in parentheses is the sales to index ratio (also known as the sales ratio).

A table of summary statistics is shown in Table 4.1. The available categorical information includes the type of loan collateral, i.e. Primary Dwelling Home (PDH) (owner-occupier) vs. Buy to Let (BTL), and the type of property (house, apartment or other). Dublin flag is a location indicator for the collateral location in Dublin or non-Dublin. For the FSD model, additional location data by county is used for the 26 counties in the Republic of Ireland. This replaces the collateral location variable Dublin flag for the FSD modelling.[6]

The default LTV is the ratio of the facility balance at the point of default divided by the indexed valuation of the collateral at the date of default. The facility balance is the unpaid principal balance on the facility. The continuous variable loan age at default is the time (in months) since origination to the date of default.

Unemployment was chosen on the basis of previous research (Kelly, 2011; Kelly and McCann, 2016) which found that the unemployment rate and changes in employment status have been found to be the most important economic predictor for arrears and transition into later arrears states, such as the long-term arrears cases in this paper. The unemployment rate used is the year-on-year change in the unemployment rate lagged one month prior to the date of default. Note that house price developments are taken into account through their effect on the indexed valuation at default as the denominator in the default LTV variable.

Due to data limitations regarding the exact date of initiation of the legal or bilateral process, a categorical variable indicating the type of process (forced via legal proceeding or bilateral) cannot be used as

[6]This categorical variable (26 categories representing the county-level regions in Ireland) could be used in the TTS for most models, except the Aalen model where it is not possible to estimate the model using these data beyond a fixed time point that is shorter than the end of the sample period. This is discussed later in more detail later in this chapter. In summary, it is because the least-squares type-estimator of the cumulative regression functions become singular if there are fewer than number of predictors+1 subjects remaining at risk.

Table 4.1: Data Summary

Concept	Characteristic	Value
N	Number of observations	18079
Time	Median (Range)	64.63 (0.23, 175.2)
Event	Resolved	5861 (32.42%)
	Non Resolved	12218 (67.58%)
Bank	Bank name	18079
Property	House	14652 (81.04%)
	Apt or other	3427 (18.96%)
Collateral	PDH	9914 (54.84%)
	BTL	8165 (45.16%)
Dublin flag	Dublin	4237 (23.44%)
	Non Dublin	13842 (76.56%)
Default LTV	Median (Range)	94.3 (5.18, 350.54)
Loan age (at default)	Median (Range)	61.28 (0.1, 214.6)
Resolution type (FSD)	Voluntary	4745 (81%)
	Forced	1116 (19%)
Unemployment rate (y/y change, 1 month lag)	Median (Range)	0.8 (-2.3, 6.7)

a predictor for TTS. For the FSD model, this feature can be used and may be important because legal proceedings may take much longer to conclude than a consensual bilateral agreement and may thus be associated with a larger discount.

The distribution of the haircut on sale by resolution time bucket are shown in Figure 4.1.[7] This figure shows that voluntary haircuts tend to be smaller than forced sale haircuts. Longer default resolution durations are associated with increased haircut magnitudes for both forced and voluntary processes.

4.6 Methods

This section describes the methods and the empirical approach used to model both Time to Sale (TTS) and the Forced Sale Discount (FSD).

4.6.1 Time to sale modelling

Three types of methods are considered for modelling TTS. The first are Accelerated Failure Time (AFT) parametric survival models. In these models, the explanatory variables can accelerate, decelerate, or have no effect on the survival process compared to the baseline survival function. Three types of distributions are chosen to parameterise these models - Lognormal, Weibull, and Log-logistic. They have the advantages of straightforward interpretation and estimation. Their main disadvantage is that the distributional assumptions underpinning the models may not be sufficient descriptions of the data.

The second group of methods are semi-parametric survival models. These are the Cox proportional hazards model (Cox, 1972) and the Aalen semi-parametric model (Aalen, 1989). The Cox model is

[7] The box plots show for each time bucket that shows the median (black line), the first (Q1) and third quartiles (Q3) (the box). The length of the box corresponds to the Inter Quartile Range (IQR) of a variable x and the whiskers are defined as: upper whisker = min(max(x), Q3 + 1.5 * IQR); lower whisker = max(min(x), Q1 – 1.5 * IQR).

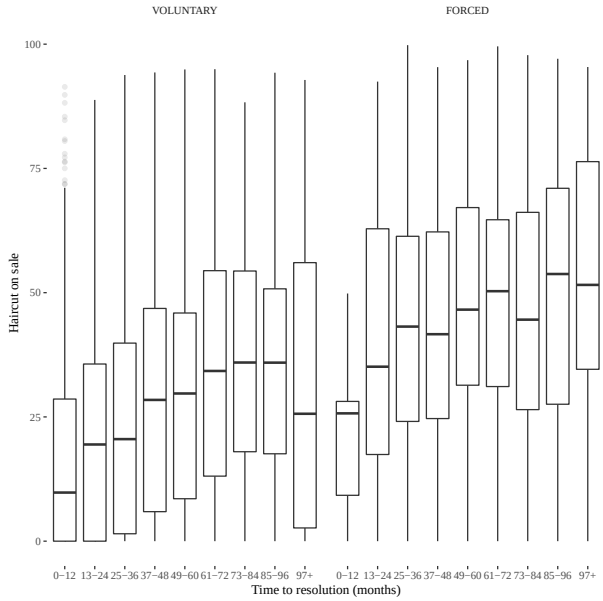

Figure 4.1: Haircut distribution and time to resolution

a standard model in the credit risk and resolution time literature (Betz et al., 2016, 2017). Its main advantages are it is computationally quick to estimate and interpretation is straightforward. The main disadvantage is that the proportional hazard assumption may not be met. The R package *rms* (Harrell Jr, 2019) is used to estimate both the parametric and Cox survival models.

The Aalen model can be thought of as generalising the Nelson-Aalen estimator (Nelson, 1972; Aalen, 1978). This model is not as widespread in credit risk modelling, though Lando et al. (2013) used this to model corporate defaults data. The appeal of this method is that it does not assume proportional hazards, and can treat some of the predictor effects as changing through time and others as time constant. Part of the price of this flexibility is that as the approach is additive, the hazard is not constrained to be positive. This can result in a cumulative hazard that is non-monotonic and survival functions that are outside of $[0, 1]$ when covariates take extreme values. The R package *timereg* (Scheike and Zhang, 2011) is used to fit the Aalen model.

The third group are non-parametric estimators from the machine learning literature. These include a k-Nearest Neighbour (kNN) survival method, (Lowsky et al., 2013) and Random Survival Forests (Wright and Ziegler, 2017). The kNN survival method generates survival functions by using a weighted average Kaplan-Meier estimator based on the k most similar observations from the training data. This can be combined with data augmentation methods such as bootstrap sampling of training data and a random selection of fixed number of predictors to produce an average ensemble estimate over n base learners.

Random Survival Forests (RSF) were introduced by Ishwaran et al. (2008) to deal with right-censored survival data. The underlying principles are similar to the original algorithm of Brieman (2001). Their adaptation to survival analysis requires changes to the split criterion and tree prediction. In a survival forest setting, an individual survival tree is grown by selecting predictors split points that maximise the log-rank statistic in each node. This is repeated for all trees in the forest. Ensemble predictions are produced by averaging the cumulative hazard estimates in the terminal nodes of each tree. The

cumulative hazards are estimated using the Nelson-Aalen estimator.

The advantages of both of these approaches is that they are non-parametric and this may suit certain types of data. The disadvantages are interpretability and for kNN, that the nearest neighbours parameter and number of base learners can substantially affect computational time given the training dataset size. The R packages *bnnSurvival, Ranger* (Wright, 2017; Wright and Ziegler, 2017) are used to fit these algorithms. The parameters for these methods were as follows:

- A random forest (*RF*) with *mtry*=3 and 750 trees.
- A kNN survival method with the number of neighbours (k) = 40; number of base learners =20; the number of features randomly selected in each base learner = 2.

Finally, in addition to benchmarking the methods against each other, a null model is included. This is a an estimate with no covariates produced by the Kaplan Meier estimator (Kaplan et al., 1958).

Empirical approach and evaluation for TTS

The purpose of the empirical work in this section is twofold: to understand the important drivers of resolution time across methods and to assess how these approaches perform on unseen data. The data are divided (80%/20%) into a training and test sample. In industry, the TTS parameter is typically an average per cohort group, usually calculated from completed sales only, and updated infrequently (i.e., every two or three years) and are not forecast. Therefore, the testing approach used here is out of sample, not out of time.

Several measures are available for assessing the predictive power of survival models at various time points (Gneiting and Raftery, 2007; Harrell et al., 1996; Gerds and Schumacher, 2006; Gerds et al., 2008). In this paper, a measure based on the Brier score is used (Brier, 1950). For a given observation of a subject at a certain time point, the Brier score is defined as the squared error between the observed survival status (alive/dead) and a model based prediction of survival at a chosen timepoint t. The observed event time $y_i = min(t_i, c_i)$ is the minimum of the event time (t_i) for the ith loan and the censoring time for that observation (c_i). Because of this, the status variable for observations will be undefined when y_i is less than t. Therefore, they need to be weighted to account for censoring bias. This weight can be calculated using the Inverse Probability of Censoring Weights (IPCW) method in Gerds and Schumacher (2006).

The TTS predictive models should be evaluated over the time horizon for a workout process, and as the overall prediction error and the error at specific time points are of interest, the time-dependent expected Brier score is used. On an independent test data set D with n observations, the expected Brier score can be estimated by:

$$\hat{BS}(t, \hat{S}) = \frac{1}{n} \sum_{i \in D} \hat{W}_i(t) \{\tilde{Y}_i - \hat{S}(t|X_i)\}^2 \tag{4.1}$$

The status variable for subject i in the test data is given by $\tilde{Y}_i$. The predicted probability of survival by time t, for subject i, based on the training data predictor variables X_i, is given by $\hat{S}(t|X_i)$. The IPCW are $\hat{W}_i(t)$.

These can be summarised graphically in Prediction Error Curves (PEC) (Mogensen et al., 2012). Prediction Error (PE) is the prediction error from any method generating predicted survival times. $\tau > 0$ is a value smaller than the maximum time for which prediction errors for each method are estimated. The curve plots the PE versus the time point for that prediction. The PE curve for each method can be summarised with the Integrated Brier Score (IBS) also known as the Integrated Prediction Error Curve (IPEC, Lowsky et al. (2013)).The IBS is the integral over the relevant evaluation time period as per equation 4.2.

$$IBS(PE, \tau) = \frac{1}{\tau} \int_0^\tau PE(u, S)du \tag{4.2}$$

In this study, 127 months or approximately 10.5 years is the maximum time for evaluation.[8] A lower value for the Brier score at defined time points and a lower overall IBS indicate better predictive performance. The R package *pec* and additional code is used to calculate the prediction error curves for the relevant prediction horizon.

4.6.2 Forced sale discount modelling

This section describes the approach to the forced sale discount predictive modelling. The FSD reflects the difference between the indexed valuation at the point of sale and the actual sale price. The main objective is to produce a prediction of the FSD given that a sale has occurred, conditioning on the covariates included in the model.

Eight types of methods are used for the prediction of the FSD. To start, a simple OLS model is used with the haircut as the untransformed dependent variable. Next, a more flexible beta-regression modelling is used. This approach is described in Ferrari and Cribari-Neto (2004); Smithson and Verkuilen (2006) and Cribari-Neto and Zeileis (2010). This beta regression approach can model bounded and skewed distributions through modelling the mean and the variance (precision).[9] A linear Quantile Regression (QR) model is used to estimate median quantile ($\tau = 0.5$) (Koenker and Hallock, 2001) as a robust estimator of the response. This type of approach has been applied to a similar haircut estimation for LGD modelling in Somers and Whittaker (2007).

Several machine learning methods are chosen as non-parametric alternatives based, in part, on previous benchmarking studies such as Loterman et al. (2012). The methods are:

- A single layer Neural Network (NN)
- A linear Support Vector Machine (SVM)
- A Random Forest (RF)
- Gradient boosted trees (XGB)

[8] This is chosen because it is the maximum time over which the semi-parametric Aalen model can be estimated. This is because the estimator of the cumulative regression functions in the model is a least-squares type estimator. This requires $X'X$ to be invertible. One condition for singularity in the Aalen model is if there are fewer than number of predictors+1 subjects remaining at risk. In the data used in this study, this corresponds to a survival time of greater than 127 months.

[9] The logit link is used to map the linear predictor to the sample space of the observations. The logit link is used to map the open unit interval, i.e., (0,1) of a beta distributed dependent variable to a real line. A log link is used for the precision parameter.

Table 4.2: Regression method training parameters

Name	Meta parameters	Values
OLS	none	none
Beta	none	none
Quantile	quantile	$\tau = 0.5$
Neural Net	size, decay	size =(3,...,10); decay= (0.0,...,0.3)
SVM	cost	cost=(0.01,...,10)
RF	mtry, ntrees, sample fraction	ntrees=(250,...,1000); mtry=(1,...,20); sample fraction =(0.5,0.632)
XGB	eta, max depth, min child weight, sub sample, lambda, nrounds	nrounds=(250,500,...,1000); min.child.weight=(1,3,...,7,9); eta=(0.0075,...,0.3); max depth=(1,2,...,9); sub-sample=(0.5,0.632,0.75); lambda = $2^{(-10,\ldots,-1)}$
DL1	epochs	epochs =(1,...,30); hidden layers =(200,200)
DL2	epochs,hidden	epochs =(1,...,30); hidden layers =(280,100),(320,120);(280,120);(300,120)

- A deep learning feed-forward network with two hidden layers of fixed size (*DL1*)
- A deep learning feed-forward network with two hidden layers of varying size (*DL2*)

The regression methods are not optimised or tuned as there is no regularisation or tuning of parametric models. The machine learning method tuned over 5-fold cross-validation using the parameter settings in Table 4.2. [10]

All of the modelling was carried out in R. The R function *lm* is used to fit the OLS model. The R package *Betareg* (Cribari-Neto and Zeileis, 2010) is used to fit the Beta regression. Quantile regressions were estimated using *quantreg* (Koenker, 2019). All of the machine learning methods were implemented in the R package *mlr* by Bischl et al. (2016).

Empirical approach and evaluation for FSD

There are 5861 loans with a resolution resulting in the sale of collateral (a resolution event). To understand the generalisation performance of the methods, these data are divided into a 80% training sample and a 20% test sample. Similar to the TTS parameter, for FSD this is typically an average per cohort group from completed sales only is updated infrequently and not forecast. Therefore, the testing approach used here is also out of sample, not out of time.

The data are range normalised to lie between zero and one as this is needed for the SVM and deep learning methods. The response variable, the haircut, is transformed so that the endpoints lie strictly between (0,1) so the beta-regression method can be used.[11]

Two performance measures are used: the Mean Absolute Error (MAE) and Mean Squared Error (MSE). These are used chosen as we are interested in the assessment of the mean predictive ability of the methods, as well as being simple to understand and used in previous work (Leow et al., 2011; Loterman et al., 2012).

4.7 Results

The results of the TTS and FSD modelling and their predictive performance are summarised in subsections 4.7.1 and 4.7.2. Detailed results for TTS and FSD regression models are contained in Appendix D.

[10] To check whether this would otherwise bias the results towards the machine learning methods, the same analysis was carried out using the default settings with no hyper-parameter optimisation. The results were similar to those presented in this section.

[11] The transformation is the one used Smithson and Verkuilen (2006) in $y = [y(N-1)+s]/N$, where y is the haircut, N is the sample size, and s=0.5 is a constant added to restrict the range to the (0,1) interval.

4.7.1 TTS results summary

Overall, the variable effects summarised in Table 4.3 are broadly consistent across the parametric and semi-parametric models.[12] For TTS, the most important predictive variables are the bank indicator, collateral/property types, the unemployment rate, and location.

Table 4.3: Summary of variable effects on resolution time: parametric/semi-parametric regression models

Variable	Log Normal	Weibull	Log Logistic	Cox	Aalen	Explanation
Bank: B (A)	-	-	-	-	-	Bank B resolves collateral quicker than A.
Default LTV	-	-	-	-	-	Loans with higher default LTV resolved faster.
Loan age default	-	-	-	-	-	Older loans at default resolved faster.
Property type: Apt.other (House)	+	+	+	+	+	Apartments have longer resolution times compared to Houses.
Collateral type: BTL (PDH)	-	-	-	-	-	BTL loans have longer resolution times compared to PDH.
Dublin Flag: Non Dublin (Dublin)	+	+	+	+	+	Non Dublin collateral has a longer resolution time compared to Dublin location.
Unemployment rate	+	+	+	+	+	Positive changes in unemployment rate level increase resolution time.

[1]The effects for the Cox model are based on reversing the coefficient signs in Table D.1.

The detailed parametric survival model results are shown in Table D.1. All of the variables are statistically significant for the three parametric models. It is important to note the difference in coefficient signs between AFT and Cox models. In an AFT model, the covariates act multiplicatively on time. The coefficients are logarithms of ratios of survival times, a positive coefficient meaning that the covariate's effect is to increase resolution time, a negative coefficient means the covariate effect is to reduce resolution time. In the Cox model, the covariates act multiplicatively on the hazard. A negative coefficient refers to decreased risk of resolution and a longer resolution time; conversely, a positive coefficient indicates a higher resolution intensity and a shorter resolution time. Therefore, for our purposes, a positive coefficient in an AFT model has a similar effect as a negative coefficient in the Cox model.

Across all three AFT models in Table D.1, bank-specific differences play an important role in resolution times. For example, according to the lognormal model, bank B accelerates resolution times (log TTS) by 0.287 ($exp(-1.25)$) compared to bank A. Increased default LTV, loan age at default, and collateral

[12]The Aalen model time varying cumulative regression function plot for loan age suggests that loan age additively increases the hazard of sale up to about 60 months, after which is remains relatively constant.

type BTL all reduce log TTS. Properties that are apartments (versus houses) and collateral being in a non-Dublin location (versus Dublin) act to increase log TTS, compared to their reference categories. An increase in the lagged change in the unemployment rate increases log TTS. Overall, the coefficient estimates are relatively similar despite the different distributional assumptions in the three parametric models.

The results from the Cox PH model in Table D.1 indicate that all of the predictor variables are significant. Bank B compared to bank A (the reference level) has a much higher resolution intensity, and therefore a shorter TTS. Default LTV, loan age at default, and collateral type BTL are associated with an increased hazard of resolution and a shorter resolution time. Properties that are apartments (versus houses) and collateral in a non-Dublin location (versus Dublin) have a decreased hazard of resolution compared to their reference categories. An increase in the lagged change in the unemployment rate decreases the hazard of resolution, increasing TTS. The results are similar to the AFT models in Table D.1.

While default LTV and loan age are significant, they do not have a large effect size on resolution time. This is illustrated in plots of the exponentiated coefficients from the AFTs and Cox model are shown in Figures D.2 and D.3.[13]

Unlike the AFT or Cox models, for the Aalen model, there are two sets of hypotheses tested. The first one is whether a covariate can be modelled as time-invariant. The null hypothesis is that the covariate effect is time-invariant. The second is whether a coefficient is significant; the null hypothesis is that the effect not different from zero.

The time-invariance test statistics are described in detail in Martinussen and Scheike (2006), Lando et al. (2013). This hypothesis can be tested using two test statistics - the first is a Kolmogorov-Smirnov (KS) test statistic, sensitive to large deviations from the null hypothesis of time constancy; the second is a Cramer-von Mises (CvM) statistic sensitive to small but persistent deviations from the null. The second set of tests are coefficient significance tests. These involve computing the maximal deviation of the estimated cumulative regression coefficient $\beta_j(t)$ from zero divided by a robust estimate of its variance.

Based on the aforementioned tests, the intercept and loan age at default were found to be time-varying, whilst the remainder of the covariates could be treated as time-constant effects and modelled as parametric terms (see Table D.2). All of the variables are significant at the 5% level (last column; Table D.3). Similarly to the parametric survival models, in this model, default LTV is relatively less important than the other factors.

What are the important factors for TTS?

One of the most important factors for resolution time prediction is the bank specific indicator. Bank-specific differences may reflect differing resolution strategies for individual banks and therefore resolution time. The results suggest that whereas previous studies have pooled data from several institutions, explicitly taking these factors into account may be important where it is feasible to do so. The importance of loan age at default suggests loan seasoning is a statistically relevant factor to consider in reso-

[13]It is important to note that these are exponentiated coefficients and the continuous variables have been normalised to have a mean of zero to facilitate interpretation of the Aalen model.

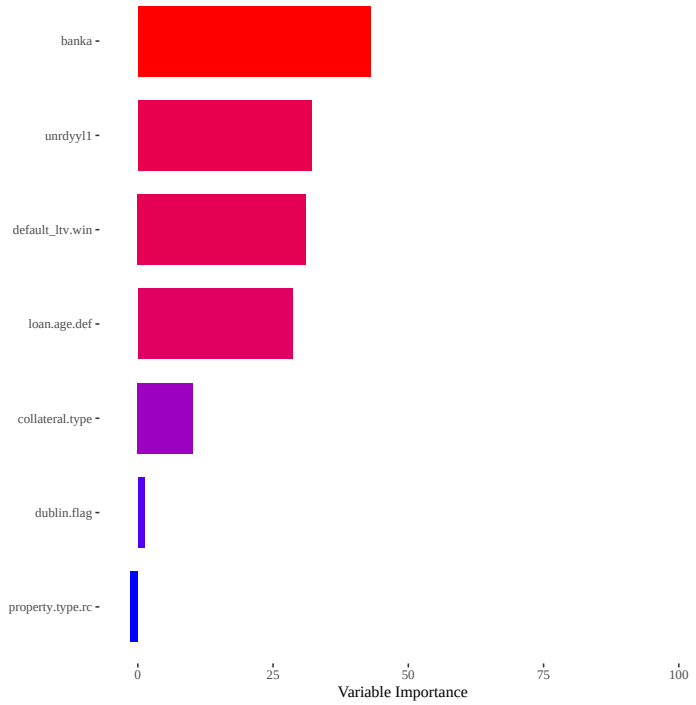

Figure 4.2: Survival forest impurity corrected variable importance

lution time models based on this data. However, the effect size is limited in the regression models (see Figures D.2 and D.3). Conditioning mortgage resolution time estimates on variables like unemployment appears to be important in most models.

As survival forests are a non-parametric technique, they do not have coefficients that can be compared to the regression methods. However, it is possible to produce a relative variable importance ranking for them. This variable importance measure shown in Figure 4.2 is the impurity importance measure. For a variable x_i, impurity importance is calculated by summing the decrease in the impurity measure for all the nodes in the forest where x_i has been split and divided by the number of trees. In a survival forest context, the specific impurity measure is the maximum of the log rank statistic over all the possible split points for the covariates consider for splitting. In this paper, the unbiased split variable selection method outlined in Wright et al. (2017) is used.

The most important variables according to this measure are the bank factor, the lagged change in the unemployment rate, default LTV, loan age at default, and collateral type. It is interesting that the survival forest indicates that unemployment is more important than some of the other variables identified by the parametric and semi-parametric survival models.

Default LTV, while significant in the parametric survival models, did not have a large effect on resolution time. By contrast, it was identified as important in the survival forest. This may suggest that this covariate may have non-linear effects. However, inclusion of interactions for default LTV in the AFT/Cox models did not markedly improve predictive performance.

4.7.2 FSD results summary

The results for the FSD regression models are contained in Table D.4. The variables bank, forced vs. voluntary, time in default, property/collateral types, and the unemployment rate are significant in the three models. Default LTV is only marginally significant in the OLS regression, and loan age at default is not significant in any regression.

The factor variable for 26 counties in the Republic of Ireland, (i.e., region-specific differences) is omitted from the table due to space requirements. The intercept in the OLS method estimates a mean haircut of 17% unadjusted for any covariates. Bank B has a 7% smaller discount compared to bank A. The impact of forced sale through the legal process is approximately 12% higher compared to voluntary. BTL collateral has about a 5% higher forced sale discount than owner occupier. Apartments have a 4% lower FSD than houses based on this data. An extra 12 months resolution time increases the FSD by 1.8% with other covariates remaining unchanged.

The beta regression has broadly similar results with much smaller standard errors. Because the response is on the logit scale, the coefficient interpretation is not as intuitive as for OLS or QR. For resolution type, the coefficient 0.57 is the log of the ratio between the chance of a predicted haircut for a forced sale compared to the reference category of voluntary sale. The QR model intercept is smaller than the OLS (0.09 vs 0.17), but has a slightly larger value for the resolution type (0.136 vs 0.126). The other coefficients are similar in magnitude to the OLS estimates.

What are the important factors for FSD?

Bank-specific differences, the mechanism that leads to the sale, the type of collateral/property, and the duration of the work-out are all important determinants of the FSD for mortgages. This adds to the findings of Rapisarda and Echeverry (2013) and Betz et al. (2018) who included TTS and collateral types in their LGD modelling approaches for SME/Corporate loans. In terms of how these compare to other results available for FSD magnitude, they are larger than the range reported in Donner et al. (2016) and Lee (2010), with a mean of about 18%. However, given the scale of the property bubble and subsequent decline in Ireland, this is not surprising.

For the ML methods, the feature importance can be approximated using permutation importance. The relative importance of a feature can be approximated by the change in prediction error after permuting the feature. A feature is important if permuting its values increases the error as the model relies on the feature for prediction. A feature is not important if, after it is permuted, its values leave the error unchanged. A negative permutation importance for a feature indicates predictive value worse than random noise. Figure 4.3 suggests that a few variables are important in terms of the MAE for forced sale discount prediction. These are the default LTV, time in default, and bank-specific differences as well as the voluntary versus forced mechanism. It is interesting that the ML methods identify default LTV as being an important predictors when regression models suggest a more limited role in terms of importance. This is similar to what is found for TTS.

4.7.3 Predictive performance of TTS and FSD models

The predictive performance for TTS were evaluated using the measures outlined in Section 4.6.1. This can be seen in Figure 4.4 which shows the prediction error for each of the methods over 120 months

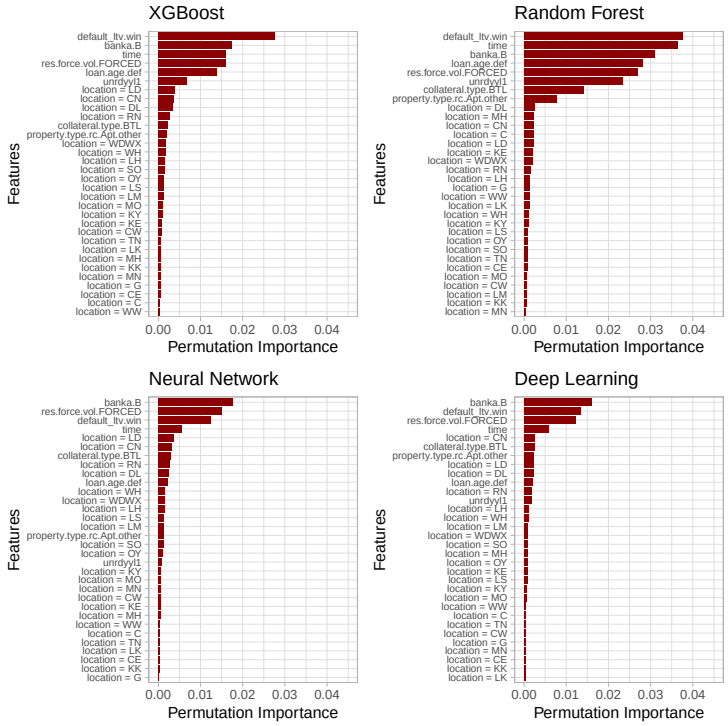

Figure 4.3: Machine learning predictive models of FSD: MAE permutation importance

and the Integrated Brier Score (IBS). Overall, the RSF is the most accurate evaluated on the test data. Parametric AFT models like the log-logistic and lognormal model are second and third respectively as measured by the IBS. The Cox proportional hazards and Weibull model have little between them. The log-logistic and other parametric models perform relatively similarly until after about 90 months, when the log-logistic performs slightly better. Finally, the bagged kNN survival method and the Aalen semi-parametric method have a higher prediction error than the other methods, but lower than the KM null method. Table 4.4 shows the Brier scores for selected time points and methods.

Table 4.4: Brier score for select times (months)

	modname	12	36	60	72	120
1	aalensemi	0.03	0.10	0.18	0.20	0.23
2	bnn	0.03	0.10	0.17	0.19	0.24
3	cox	0.03	0.09	0.15	0.16	0.21
4	km	0.03	0.11	0.19	0.22	0.25
5	llog	0.03	0.09	0.15	0.16	0.21
6	ln	0.03	0.09	0.15	0.16	0.21
7	ranger	0.03	0.08	0.15	0.16	0.20
8	wei	0.03	0.09	0.15	0.16	0.21

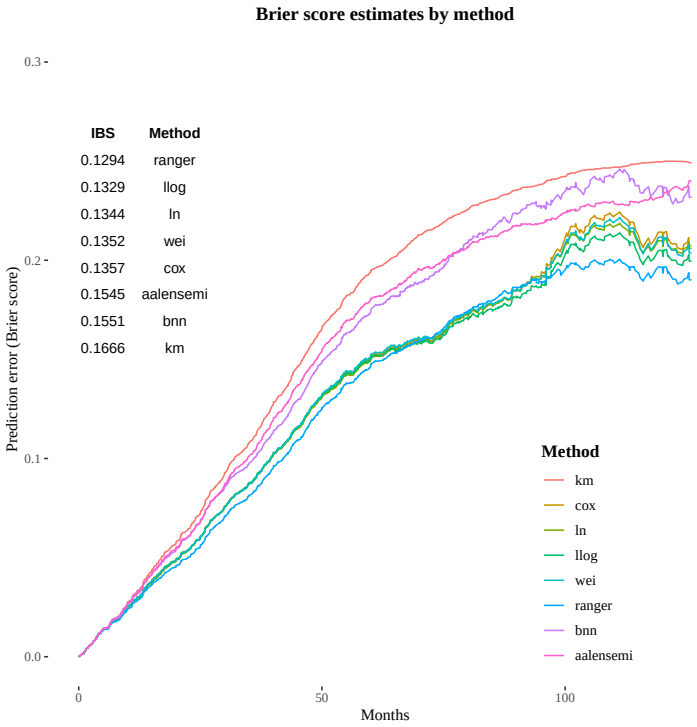

Figure 4.4: Prediction error comparison

For FSD prediction, all of the methods perform better than the median or mean as the no-information benchmark. Table 4.5 shows the predictive performance measures. The performance of the parametric methods is similar, with QR performing slightly better on the MAE, and OLS on the MSE criterion respectively, compared to Beta-regression. The parametric models are mostly out-performed to a small extent by some of the machine learning methods. SVM perform worse than QR on the MAE criterion, and only slightly better on the MSE criterion. The remainder of the machine learning methods have similar results across the two performance measures. Deep Learning methods, Neural Networks, and Random Forests and perform the best, with only slight differences between them.

Given the modest differences between most methods and that a single data set is used for the ten methods (nine models plus a no-information benchmark), typical post-hoc tests of differences adjusted for multiple comparisons suggest that there were no statistically significant differences between methods. In a practical setting, further investigation of practically significant performance differences on a larger samples (Benavoli et al., 2017), and if available, differing datasets would be required before it could be concluded that the predictive performance of advanced machine learning methods offer a significant improvement in FSD modelling.

Table 4.5: Forced sale discount model performance

Metric	Mean	OLS	QR	Beta	NN	SVM	RF	XGB	DL1	DL2
MAE	0.2106	0.1793	0.1772	0.1809	0.1716	0.1782	0.1725	0.1730	0.1712	0.1728
MSE	0.0622	0.0473	0.0486	0.0495	0.0445	0.0481	0.0445	0.0452	0.0443	0.0444

4.8 Implications for loss severity estimation

This section combines the results for TTS/FSD predictive models through illustrating the implications of model-predictions of these parameters compared to assuming fixed values for TTS/FSD. In some industry implementations, the TTS may be estimated from completed sales only and not from sales that were progressing but not yet completed, i.e., censored. As noted in Section 4.2, this implementation of a loss severity models use TTS and FSD parameters that are average values for certain groups of loans. For example, different fixed values of TTS and FSD could be looked-up based on whether the collateral is an owner-occupied property (PDH) or Buy To Let (BTL); property type is house or apartment; regional location and so on. Model predictions for individual property collaterals may be both lower and/or higher than these constant TTS/FSD values per cohort as they do not account for other factors.

To illustrate the implications of the differences of this industry approach compared to that informed by the approach in this paper on loss severity, a simple simulation is carried out. There are four main steps:

- Assume 5000 loans with the same origination balance (€450,000) and an initial valuation of €500,000 representing an 88% origination LTV. Each loan has an annual interest rate of 3% and a 25 year term with monthly repayment frequency. The loans default with the timing of default governed by a Weibull distribution with a shape parameter of 1.2 and a scale parameter of 35. The EAD at default is the outstanding balance at the default time. A lognormal distribution is assumed for simulating HPI index changes with a drift of -2.5% and a volatility of 13% per year.
- Three groups of TTS and FSD values are calculated. The first are random samples based on the range of average TTS excluding non-closed loans and average for the FSD based on the grouping or cohorts typically used in an IFRS 9 model such as collateral type and location. The second group has two differences. The TTS are now based on the mean predicted TTS including censored (non-closed) properties; the FSD are calculated in a similar fashion. The third are similar to the second (i.e., mean of model predictions), but for FSD, the grouping is based on an additional factor identified as important - time in resolution.
  - assumed TTS: these are generated by draws from a uniform distribution between 15 and 47 months. The assumed FSD is drawn from a uniform distribution between 17-37%. The TTS is the naïve estimate excluding non-closed loans; the FSD is based on the range of averages by cohort groups (resolution type (voluntary vs forced); collateral type (PDH vs BTL); location (Dublin/Non-Dublin)).
  - predicted TTS and FSD: TTS is based on the range (36 to 120 months) of median TTS across methods with continuous covariates at their median values and for each categorical variable, each category is represented. The FSD range (12 to 55%) is based on mean model predicted FSD for the same cohort groups as above but including resolution type (voluntary vs forced); collateral type (PDH vs BTL); Dublin/Non-Dublin location.
  - modelled stratified FSD: FSD are cohort groups based on the top three models based on the MAE/MSE criteria.[14] For TTS the range is the same as above - 36 to 120 months. The FSD

[14]The purpose of the exercise is to make a general observation about the impact of modelling the TTS and FSD parameters, not specifically to discuss individual bank loss severity estimates.

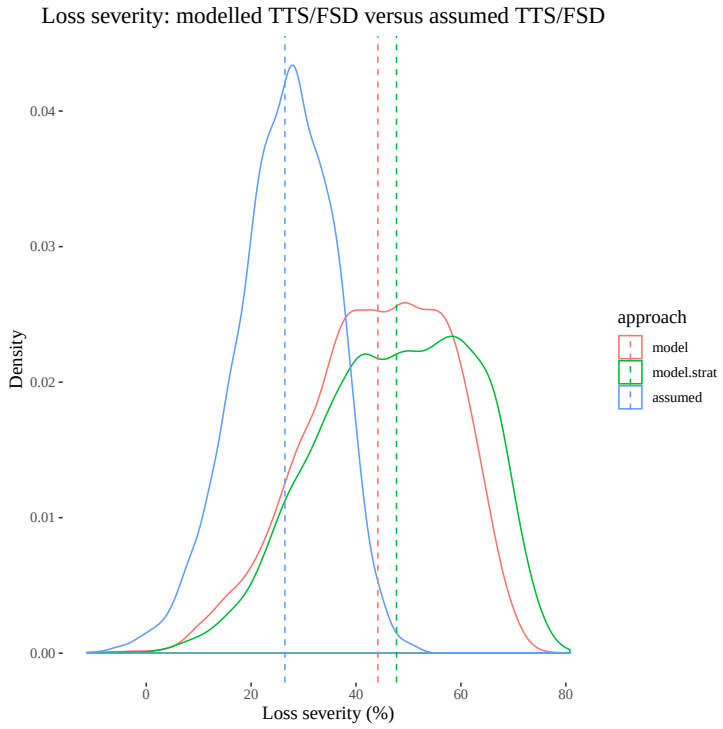

Figure 4.5: Time to sale, forced sale discounts, and loss severity.

range (13 to 63 %) is the mean model predicted FSD for the same cohort groups as above and each quartile of the time in resolution.

- For each loan, its default time and TTS are added to give the sales time. This combined with the property price index paths and FSD determines the collateral value at sale. A maximum time period for the simulation is set at 175 months or nearly 15 years to limit implausibly long sale periods. The Discounted Collateral Recovery (DCR) is calculated using equation 4.3.
- Assuming no repayments since default, $Loss\,Amount = EAD - DCR$; $Loss\,Severity = Loss\,Amount/EAD$. This is repeated for each loan and the resulting density is plotted in figure 4.5.

$$DCR = \frac{1}{(1+R)^{TTS}}(FSD - Disposal\,Costs) \tag{4.3}$$

Two points are apparent from Figure 4.5 in this stylised example. The first is assuming constant parameters results in a lower average loss severity of 26% versus 44%-48% modelling these parameters; a relative increase in loss severity of 69%. Second, using a modelled approach results in more variable loss severity with increased mass at the right tail. Stratifying the groups by the model factors does not substantially change the mean loss severity (44 vs 48%) but it increases loss severities in the right tail the green-coloured density somewhat compared to the red-coloured. While this is quite a simple stylised example, one implication is that assuming fixed parameter values for TTS and FSD may result in less conservative estimates of loss severities.

4.9 Conclusions

This chapter has compared various approaches to estimating two important parameters determining mortgage collateral recoveries: the Time to Sale (TTS) and Forced Sale Discount (FSD). The first research question in this chapter focused on what approaches can produce accurate predictions of each of the two parameters. For TTS, Random Survival Forests (RSF) were found to be the most accurate, but parametric survival models performed reasonably. For the FSD modelling, from the range of the methods considered, machine learning methods performed better than parametric methods, with Deep Learning, Random Forests, and Neural Networks performing best, with minor differences in performance.

Regarding the second research question, the important factors identified for TTS were bank-specific differences, loan seasoning, and the change in the unemployment rate prior to default. RSF identified default LTV as being important, but this was not identified by other models as being equally important. This may reflect some non-linear effects and as well as a feature of the Irish crisis when negative equity was pronounced among all defaulted loans, and not only properties in a collateral liquidation process. The finding that unemployment is an important predictive feature may mean that conditioning TTS estimates on macroeconomic factors should be considered further by model builders.

In terms of important factors for FSD prediction, default LTV, bank-specific differences, and the nature of the resolution process itself were the most important features identified by machine learning methods. Default LTV importance concurs with previous empirical results in this area. Bank-specific differences may reflect varying resolution strategies that in turn affect the type and timing of resolutions. The results suggest that previous work using pooled data from several institutions, may need to take this factor into account. Sales concluded through bilateral agreement between bank and borrower reduced the forced sale discount significantly compared to sales as a result of a legal process to enforce security. This could be an important consideration for jurisdictions where work-out through a legal process takes much longer, and bilateral agreements become an important method of Non-Performing Loan (NPL) resolution. Time in resolution is also important for FSD prediction. This empirical finding supports recent LGD research that includes time to resolution as a covariate.

Finally, in Section 4.8, a simple simulation, informed by the predictive models built in Sections 4.6.1 and 4.6.2, revealed underestimation of loss severity arising from two sources. First, ignoring censored loans and apply a TTS calculated based on completed sales only. Incorporating censored loans using survival analysis results in longer TTS durations in this study. Second, from assuming parameters as fixed group means within certain grouping variables such as collateral type (i.e., PDH vs BTL). This grouping could be carried out based on the most important factors for TTS or FSD predictive models. Compared to a limited number of group means - one approach currently used within industry - the simulation suggests there are practically significant differences in loss severities between both approaches. This suggests that more conservative, i.e., higher loss severities are produced as a result of modelling these parameters directly.

As the study is country-specific, it would be interesting to compare the results for other markets and approaches to estimating both the TTS and FSD parameters. The FSD model approach in this study is somewhat limited by the available data, and having a richer set of predictors could be exploited by more advanced statistical or machine learning approaches.

Chapter 5

Conclusions

5.1 Introduction

This chapter outlines the general conclusions of this thesis, some of the limitations of the work, and potential for further work on some related aspects of the three papers that make up this thesis.

5.2 Mortgage arrears prediction

5.2.1 Main findings and conclusions

Chapter 2 was a study comparing whether machine learning methods could outperform standard methods like logistic regression to predict mortgage arrears. The main findings are that some methods like Boosted Regression Trees (BRTs) can. However, there are other non-machine learning methods, such as Generalised Additive Models (GAMs), that can be usefully applied in this context. A frequent criticism of machine learning prediction methods is that they are "black boxes" regarding the variables that are important for their prediction. In both GAMs and BRTs allow for (at least partial) explanations through variable importance plots and partial dependence or plots of the smoothing terms (which is important in this context). They could also be useful as an input to the building of white-box models through identifying where there are non-linearities or interactions.

5.2.2 Limitations and further research

The research reflected the financial conditions at a point in time in 2010-2012, and like all results from inductive learning these findings may not readily be generalised to other times and places as the crisis context and drivers of arrears and default are likely to be different. It is also unclear, due to data limitations, whether changes in borrower behaviour and financial sector policies such as forbearance have had an impact on arrears incidence, as well as the severe economic distress, during which the distribution of good and bad borrowers may have shifted (Hand, 2006).

This topic could be tackled by using approaches to model concept drift (i.e., the change in joint distribution of the data) (Krempl and Hofer, 2011), and drawing on some of the advancements in on-line

learning since the paper was published could be an interesting area of research. This could be investigated along with the issue of class imbalance, as this was a feature of the data in this paper, and the distribution of classes changed significantly over time. This is a challenging problem no doubt, but there are examples of arguably more acute versions of this problem in different domains. For example, in cybersecurity applications such as spam prevention, intrusion detection, vulnerabilities exist and are exploited/found or are found and patched before being used. This can lead to a time changing distribution as adversaries try new techniques to defeat the security system (Sethi and Kantardzic, 2017).

5.3 P2P loan return prediction

5.3.1 Main findings and conclusions

Chapter 3 investigated whether various types of prediction methods and the types of information contained in loan listing features matter for profitable investment in Peer to Peer (P2P) loans. The main findings are that: linear methods perform surprisingly well on several (but not all) criteria; whether ensemble methods perform better than individual methods is measure dependent; the use of alternative text-based information does not improve profit scoring outcomes.

5.3.2 Limitations and further research

A key finding is that it may pay (in terms of returns) to model profitability directly; however, the performance depends on the methods adopted and the type of information used. Similar to Chapter 2, there are data limitations related to the specific time window when the P2P platform published text information. The findings regarding the efficacy of more advanced machine learning techniques may be a result of the methods chosen and/or the separability of the data. Focusing on more recent periods, i.e., not using text data, and assessing the information content of the Lending Club grade versus the other covariates (Fico score, application information) could be interesting to explore. It is likely that the platform have been through many iterations of their grading model since they first launched, and it may provide investors with further insights into the rating process.

A second interesting area is the portfolio allocation of investors given their investment budget and risk tolerance. Over time, the P2P platform investor base has changed from being small investors to a mix of smaller investors and large professional investors. An interesting research question is given two sizes of investment budgets and risk preferences reflecting a small investor and a professional investor, how big does a portfolio have to be and of what grade to minimise the value of large tail losses? This is a variant of a mixed integer linear programming problem, and these methods, subject to the constraints being formulated appropriately, may a useful avenue for further work from a portfolio perspective.

5.4 Mortgage collateral recovery prediction

5.4.1 Main findings and conclusions

Chapter 4 addresses the problem of predicting the collateral recovery value of defaulted mortgages, by modelling two important parameters determining this value: time to sale (i.e., the length of time before the default is resolved through sale) and forced sale discount (i.e., the percentage loss in sales proceeds

relative to the indexed valuation). Using data from two Irish banks, the predictive performance of a variety of survival analysis approaches to estimate time to sale is evaluated. The main finding of this part of the work is that Random Survival Forests and parametric survival models perform best. For forced sale discount, Deep Learning, Random Forests, Neural Networks, and XGBoost methods produced the lowest errors. Using these two parameters (TTS and FSD), a sensitivity analysis illustrated how predictive modelling of these parameters produces higher (i.e., more conservative) loss estimates than a current industry approach consisting of average values per cohort of loans.

5.4.2 Limitations and further research

Similar to the research in Chapter 2, one limitation of the study is that it may not be directly generalised to some other European countries because of the extreme nature of the crisis in Ireland. However, while the context may be different, there are parallels to other crisis hit countries. This is an area of continuing policy focus in Europe for several other euro area and two EU countries. A related limitation is the data available for this study were from two banks only. A larger sample of banks and a richer predictor dataset could provide additional insight. As the Forced Sale Discount (FSD) literature in credit risk modelling is somewhat limited, future research in this area could focus on using alternative data sources that including additional property level features.

5.5 General conclusions of this thesis

In this research, three types of prediction problems in three application domains have been explored. The following conclusions can be made. Learning theory and evidence of the widespread adoption of machine learning into our daily lives as users of technologies based on these methods tells us that machine learning works. Learning theory and empirical research, including some of the results in this thesis, also tells us that no one algorithm will perform best in many contexts. It follows that applying more advanced methods can be useful, given a good understanding of the problem domain, the context, and awareness of the limitations of these methods.

In a credit risk context, some of the performance improvements were significant. However, even where there were smaller improvements, scaled to large portfolios or more accurate predictions of defaults/impairment parameters can matter substantially in a business context.

Given the results in this thesis, improved predictive performance combined with appropriate use and model risk management, suggests that machine learning methods could be more widely used in both banks and non-banks for consumer credit risk management as well as within the supervisory community assessing these types of risks within regulated firms. Three developments make this more likely. The first is the growing awareness of policy makers of how this type of modelling could be beneficial, as well as how the associated risks need to be managed as various industries including financial services adopt this technology (Brainard, 2018). The second is the emerging research topic of explainable machine learning (Rudin and Shaposhnik, 2019). This field is related to rule extraction methods (de Fortuny and Martens, 2012; Martens et al., 2007; Baesens et al., 2003b) to create summary explanations of predictions that are consistent with the underlying model. This could reduce the perception of a 'black-box' barrier to explanation of model predictions.

In this thesis, the methods chosen, their performance optimisation, and experimental evaluation were undertaken with domain experience, understanding of some of the prior research literature, and based on current standards for optimising performance in the contexts they were applied. The third factor in making adoption of machine learning more probable is the emerging research on what is termed "AutoML". This automates the search for a best performing algorithm given the problem domain. These methods may produce better predictions for the problems of today - considered in this thesis - and the problems of tomorrow.

Appendix A

Additional statistical testing results for chapter 2

A.1 Classifier performance using only complete observations for income-based variables

These results are based on a smaller sample than those used in the main part of the paper. After excluding cases with missing income, a sample size of approximately 280,000 observations remained. The model training, validation and testing was carried out as in the main part of the paper. For portfolio 3, the values are the same as in Table 2.2 as this portfolio was not missing any income data. Overall, the results indicate that the performance ranking remains similar regardless of our treatment of missing income variable values.

Table A.1: Summary performance of classifiers: complete cases income variables

Technique	Port 1	Port 2	Port 3	Port 4	Avg. Rank
H-measure					
LR	0.2256	0.2354	0.2900	0.2578	3.75
GAM	0.2467	0.2619	0.2928	0.2607	1.875
BRT	0.2599	0.2647	0.2909	0.2711	1.5
RF	0.2586	0.2475	0.2814	0.2607	2.875

Table A.2: Complete cases income: statistical comparison of classifiers using H-measures

Test Statistic	Calculated	Calculated p value
Friedman	7.425	0.0595
Iman-Davenport	4.869	0.028

Table A.3: Complete case income: Holm's step down procedure for H-measure ranks

($\alpha = 0.05$ and $\alpha = 0.1$; BRT is control classifier)

Classifier	$z = (R_0 - R_i)/SE$	p_i	Holm's adjusted p-value
LR	2.4648	0.0137	0.0166
RF	1.5062	0.1320	0.025
GAM	0.4108	0.6812	0.05
LR	2.4648	0.0137	0.0333
RF	1.5062	0.1320	0.05
GAM	0.4108	0.6812	0.1

Appendix B

Text mining feature construction for chapter 3

B.1 Overview of text features

This appendix summarises the text based features and the approach used to fit the topic model.

The text comprises of the title of the listing text and the listing text itself. The text is a concatenation of two free text fields. There are no specific requirements to follow for the listing title or text. The provision of this text is voluntary, not mandatory. This text description was discontinued on Wednesday, March 19th, 2014. Compared to other P2P platforms like Prosper, the text on Lending Club is relatively short.

B.1.1 Preprocessing and summary statistics

The main steps in pre-processing removing whitespace, non-ASCII characters, removing HTML tags, and other artifacts related to the platform such as "Borrower added on <Date>". This is applied to the title and description text of the listing. The text is concatenated. This is because some borrowers provide short listing titles like card or move, and longer listing text text. Other borrowers do the opposite providing long titles with details that other borrowers have provided in the listing text, and short or no listing text.

- Following merged with the payment and application information, selection of loans issued from October 2008 - March 2014.
- Town/city and state fields are concatenated to a string and geo-coded to longitude and latitude.
- Listing with title or description texts with less than 4 characters in length were removed (474 loans)
- Convert numbers to words, remove punctuation, alphanumeric characters, trims strings, encode strings as UTF-8-MAC, remove any remaining non-ASCII characters, convert to lower case, remove stop words.

The summary of the text information indicates that it is short - on average two sentences, and each sentence is on average just over six words long.

Table B.1: Summary statistics for text features

var	min	median	mean	max	sd
number of words	1.00	4.00	20.55	819.00	38.98
number of sentences	1.00	1.00	2.18	97.00	2.26
sentence length	1.00	3.00	6.13	141.00	5.82
number frequency	0.00	0.00	0.45	11.00	1.44
complex.words	0.00	1.00	2.34	148.00	4.23

B.1.2 Bit-term topic model

The short listing texts presented a challenge to construct representations of the text as feature vectors as there are a limited number of words per listing. One way to deal with this is to use topic modelling. However, topic models designed for standard length text (i.e., full web-pages, multiple page lengths of text) still face a problem of the sparsity of text within individual listings in this data.

One solution to this is a bit-term topic model. A bit-term is an unordered pair of words from a text string. A bit-term topic model is a short-text topic model that is based on global word co-occurrence (i.e., across texts) to overcome text sparsity within individual documents (Yan et al., 2013). Bit-terms can be extracted using local word co-occurrence so that words that are within a window size are used, and words that occur outside of this window (i.e., too far apart) are not.

The main steps in fitting the topic model are to prepare the text input into tokens or one word per row, per listing. The key parameters are the window size (2 words), the default priors alpha and beta for the Bayesian estimation of the model (beta=0.01; alpha=50/k where k is the number of topics), and 1000 iterations of the Gibbs sampling procedure. The number of topics k was chosen by searching over 1-20 topics, recording the resulting log-likelihood as well as assessing the top 5 words within each topic for each set of iterations to ensure there were distinct topics. This resulted in 18 topics being chosen. This means we now have 18 additional feature vectors reflecting probabilities that a given listing has a certain topic.

For the out of time setting, feature generation involved using the text from the training data and scoring both the train and test data with the resulting model. For the rolling window, it is necessary to carry out the tokenisation process for each iteration of the moving window of 12000 observations to ensure that only bit-terms present in those texts were used. This is to prevent data leakage among windows and involved fitting the bit-term model separately for each slice of the moving window data. For simplicity, we kept k=18 topics as searching for different numbers of topics within each window slice would be computationally expensive and introduce additional variability within this part of the experiments.

Appendix C

Additional statistical testing results for chapter 3

Testing information type as a within factor

This section includes the detailed results for testing the role of differing types of information referred to in the main text. This is a robust linear mixed model with two within-subject factors. The first is a two level variable information type *info* (hard only; both) and *model* (17 levels, bagged trees is the reference category). The coefficients of interest are those for *info:both* (underlined in the tables). In Table C.1 the coefficient on info is negative for AUC in the rolling experimental set-up, with a t-statistic of 3.45 indicating negative predictive value. There are no large t-statistics for the other performance measures. In the out of time setting in Table C.2, the effect on MAE is negative, with a t-statistic of 1.40, and NDCG is positive (t-statistic -2.8), indicating a degree of predictive value for NDCG. There are no large t-statistics for AUC.

Alternative Testing Approaches for Research Questions

This section contains an alternative approach considered in exploring the research questions. The response was rank transformed and then used as a dependent variable in a linear mixed model. The results broadly confirm those of the main text. The exception is the inclusion of hard and soft information. This now has no detectable effect on the rank performance.

In Table C.3 across the three criteria (MAE, AUC, and NDCG) there are statistically significant differences for the factors lin.nonlin. For ensemble, there are significant differences MAE, NDCG, and AUC. This is similar to Table 3.4 in the main text. The results contained in Table C.5 are similar to those in Table 3.5 except the t-statistic for information is no longer large for AUC.

For the out of time setting, the results in Table C.4 are similar to those in the main text in Table 3.7. The results in Table C.6 as similar to Table 3.8 except the t-statistic on information NDCG is now much lower, and similar to MAE and AUC indicating additional text information is not important for performance.

Table C.1: Robust linear mixed effect model: rolling window within-subjects

	MAE	NDCG	AUC
(Intercept)	15.5639	0.8118	0.6313
	(0.3370)	(0.0144)	(0.0029)
infoboth	0.0207	0.0025	−0.0069
	(0.0875)	(0.0146)	(0.0020)
modnameh2o.dl	−1.1015	−0.0204	0.0104
	(0.0875)	(0.0146)	(0.0020)
modnameh2o.glm	−1.1796	−0.0747	0.0320
	(0.0875)	(0.0146)	(0.0020)
modnamel2liblin	−6.5217	−0.1026	0.0311
	(0.0875)	(0.0146)	(0.0020)
modnamelasso	−1.0745	0.0212	0.0162
	(0.0875)	(0.0146)	(0.0020)
modnamemars	−1.1194	−0.0527	0.0071
	(0.0875)	(0.0146)	(0.0020)
modnamennet	−1.2490	0.0160	0.0083
	(0.0875)	(0.0146)	(0.0020)
modnamepls	−1.1645	0.0068	0.0288
	(0.0875)	(0.0146)	(0.0020)
modnamerf	−0.9048	−0.0088	0.0310
	(0.0875)	(0.0146)	(0.0020)
modnameridge	−1.1438	0.0248	0.0308
	(0.0875)	(0.0146)	(0.0020)
modnamesl.avg	−2.0293	0.0068	0.0355
	(0.0875)	(0.0146)	(0.0020)
modnamesl.gbm	−1.0589	0.0046	0.0341
	(0.0875)	(0.0146)	(0.0020)
modnamesl.liblin	−6.5399	0.0032	0.0354
	(0.0875)	(0.0146)	(0.0020)
modnamesl.mars	−1.1696	0.0106	0.0344
	(0.0875)	(0.0146)	(0.0020)
modnamesl.ridge	−1.2161	0.0252	0.0356
	(0.0875)	(0.0146)	(0.0020)
modnamesvm	−6.5511	−0.1228	−0.0911
	(0.0875)	(0.0146)	(0.0020)
modnamexgb	−1.1690	−0.0037	0.0219
	(0.0875)	(0.0146)	(0.0020)
infoboth:modnameh2o.dl	−0.0177	0.0105	0.0080
	(0.1237)	(0.0206)	(0.0028)
infoboth:modnameh2o.glm	0.0110	0.0002	0.0066
	(0.1237)	(0.0206)	(0.0028)
infoboth:modnamel2liblin	−0.0528	0.0033	0.0068
	(0.1237)	(0.0206)	(0.0028)
infoboth:modnamelasso	−0.0148	0.0043	0.0051
	(0.1237)	(0.0206)	(0.0028)
infoboth:modnamemars	0.0029	−0.0073	0.0076
	(0.1237)	(0.0206)	(0.0028)
infoboth:modnamennet	0.3444	−0.0215	0.0070
	(0.1237)	(0.0206)	(0.0028)
infoboth:modnamepls	0.0230	0.0020	0.0061
	(0.1237)	(0.0206)	(0.0028)
infoboth:modnamerf	0.0434	0.0130	−0.0006
	(0.1237)	(0.0206)	(0.0028)
infoboth:modnameridge	0.0021	−0.0052	0.0061
	(0.1237)	(0.0206)	(0.0028)
infoboth:modnamesl.avg	0.0071	−0.0028	0.0062
	(0.1237)	(0.0206)	(0.0028)
infoboth:modnamesl.gbm	−0.1907	0.0067	0.0067
	(0.1237)	(0.0206)	(0.0028)
infoboth:modnamesl.liblin	−0.0362	0.0166	0.0060
	(0.1237)	(0.0206)	(0.0028)
infoboth:modnamesl.mars	0.0195	0.0041	0.0047
	(0.1237)	(0.0206)	(0.0028)
infoboth:modnamesl.ridge	0.0229	−0.0021	0.0061
	(0.1237)	(0.0206)	(0.0028)
infoboth:modnamesvm	−0.0219	−0.0097	0.0144
	(0.1237)	(0.0206)	(0.0028)
infoboth:modnamexgb	0.0345	−0.0057	0.0035
	(0.1237)	(0.0206)	(0.0028)
Num. obs.	442	442	442

Standard errors in parentheses

Table C.2: Robust linear mixed effect model: out of time within-subjects

Coefficient	MAE	NDCG	AUC
(Intercept)	21.2311	0.7741	0.5818
	(0.3335)	(0.0136)	(0.0042)
infoboth	−0.5561	0.0452	0.0021
	(0.3951)	(0.0162)	(0.0054)
modnameh2o.dl	−7.0175	−0.0599	0.0469
	(0.3951)	(0.0162)	(0.0054)
modnameh2o.glm	−7.3669	0.0042	0.0573
	(0.3951)	(0.0162)	(0.0054)
modnamel2liblin	−12.1443	−0.0139	0.0592
	(0.3951)	(0.0162)	(0.0054)
modnamelasso	−7.5554	−0.0236	−0.0727
	(0.3951)	(0.0162)	(0.0054)
modnamemars	−6.4677	−0.0401	0.0144
	(0.3951)	(0.0162)	(0.0054)
modnamennet	−6.2857	−0.0449	0.0473
	(0.3951)	(0.0162)	(0.0054)
modnamepls	−7.5055	−0.0338	0.0416
	(0.3951)	(0.0162)	(0.0054)
modnamerf	−4.0386	−0.0258	0.0322
	(0.3951)	(0.0162)	(0.0054)
modnameridge	−7.5182	−0.0291	0.0520
	(0.3951)	(0.0162)	(0.0054)
modnamesl.avg	−6.5127	−0.0418	0.0575
	(0.3951)	(0.0162)	(0.0054)
modnamesl.gbm	−5.9853	−0.0361	0.0502
	(0.3951)	(0.0162)	(0.0054)
modnamesl.liblin	−12.0191	−0.0150	0.0505
	(0.3951)	(0.0162)	(0.0054)
modnamesl.mars	−5.5899	−0.0336	0.0486
	(0.3951)	(0.0162)	(0.0054)
modnamesl.ridge	−5.7505	−0.0071	0.0531
	(0.3951)	(0.0162)	(0.0054)
modnamesvm	−12.1779	−0.0746	−0.0159
	(0.3951)	(0.0162)	(0.0054)
modnamexgb	−4.7460	−0.0509	0.0416
	(0.3951)	(0.0162)	(0.0054)
infoboth:modnameh2o.dl	0.6875	−0.0384	−0.0098
	(0.5587)	(0.0229)	(0.0076)
infoboth:modnameh2o.glm	0.4040	−0.0492	−0.0090
	(0.5587)	(0.0229)	(0.0076)
infoboth:modnamel2liblin	0.6526	−0.0277	−0.0088
	(0.5587)	(0.0229)	(0.0076)
infoboth:modnamelasso	1.4568	−0.0303	0.1056
	(0.5587)	(0.0229)	(0.0076)
infoboth:modnamemars	−0.1388	−0.0545	−0.0022
	(0.5587)	(0.0229)	(0.0076)
infoboth:modnamennet	−0.3583	−0.0476	0.0024
	(0.5587)	(0.0229)	(0.0076)
infoboth:modnamepls	0.4351	−0.0197	−0.0087
	(0.5587)	(0.0229)	(0.0076)
infoboth:modnamerf	0.1013	−0.0414	−0.0300
	(0.5587)	(0.0229)	(0.0076)
infoboth:modnameridge	0.5192	−0.0242	−0.0046
	(0.5587)	(0.0229)	(0.0076)
infoboth:modnamesl.avg	−0.0749	−0.0078	−0.0094
	(0.5587)	(0.0229)	(0.0076)
infoboth:modnamesl.gbm	0.3532	0.0061	−0.0136
	(0.5587)	(0.0229)	(0.0076)
infoboth:modnamesl.liblin	0.5894	−0.0028	−0.0086
	(0.5587)	(0.0229)	(0.0076)
infoboth:modnamesl.mars	−0.4500	−0.0080	−0.0126
	(0.5587)	(0.0229)	(0.0076)
infoboth:modnamesl.ridge	−0.4233	−0.0129	−0.0082
	(0.5587)	(0.0229)	(0.0076)
infoboth:modnamesvm	0.6704	−0.0155	−0.0167
	(0.5587)	(0.0229)	(0.0076)
infoboth:modnamexgb	−0.6312	0.0044	−0.0266
	(0.5587)	(0.0229)	(0.0076)
Num. obs.	170	170	170

Standard errors in parentheses

Table C.3: Robust linear mixed effect model: rolling window (rank transformation)

	MAE	NDCG	AUC
(Intercept)	11.6013	7.7179	8.6139
	(0.4244)	(0.4725)	(0.3355)
lin.nonlinlinear	−4.9632	−0.2587	−4.4475
	(0.4832)	(0.5379)	(0.3819)
ensembleindividual	−0.0397	2.5336	4.5629
	(0.4901)	(0.5456)	(0.3874)
Num. obs.	442	442	442

Standard errors in parentheses

Table C.4: Robust linear mixed effect model: out of time (rank transformation)

	MAE	NDCG	AUC
(Intercept)	14.5563	9.6189	11.4084
	(0.4407)	(0.7065)	(0.6355)
lin.nonlinlinear	−5.9760	−3.4869	−4.4505
	(0.5017)	(0.8044)	(0.7236)
ensembleindividual	−4.0198	2.0930	−0.3984
	(0.5088)	(0.8158)	(0.7339)
Num. obs.	170	170	170

Standard errors in parentheses

Table C.5: Robust linear mixed effect model: rolling window within-subjects (rank transformation)

	MAE	NDCG	AUC
(Intercept)	16.8462	8.9219	15.1618
	(0.6928)	(1.0571)	(0.4355)
infoboth	−0.0769	−0.3515	0.3766
	(0.9798)	(1.4950)	(0.6158)
modnameh2o.dl	−6.7794	2.4948	−1.6666
	(0.9798)	(1.4950)	(0.6158)
modnameh2o.glm	−8.3470	4.8358	−8.8334
	(0.9798)	(1.4950)	(0.6158)
modnamel2liblin	−14.3077	5.9949	−7.4920
	(0.9798)	(1.4950)	(0.6158)
modnamelasso	−5.3797	−3.5054	−2.4444
	(0.9798)	(1.4950)	(0.6158)
modnamemars	−5.5393	4.1813	−1.1711
	(0.9798)	(1.4950)	(0.6158)
modnamennet	−9.7977	−2.8885	−1.0867
	(0.9798)	(1.4950)	(0.6158)
modnamepls	−6.6166	−1.4175	−5.7501
	(0.9798)	(1.4950)	(0.6158)
modnamerf	−2.0769	1.2538	−7.5669
	(0.9798)	(1.4950)	(0.6158)
modnameridge	−6.0037	−4.8298	−7.1764
	(0.9798)	(1.4950)	(0.6158)
modnamesl.avg	−12.6154	−1.4591	−12.4265
	(0.9798)	(1.4950)	(0.6158)
modnamesl.gbm	−4.7580	−1.2219	−10.9358
	(0.9798)	(1.4950)	(0.6158)
modnamesl.liblin	−15.0000	−1.3740	−11.8867
	(0.9798)	(1.4950)	(0.6158)
modnamesl.mars	−7.8170	−2.4169	−11.6682
	(0.9798)	(1.4950)	(0.6158)
modnamesl.ridge	−8.0057	−4.3834	−12.6081
	(0.9798)	(1.4950)	(0.6158)
modnamesvm	−15.2308	6.6296	1.8382
	(0.9798)	(1.4950)	(0.6158)
modnamexgb	−7.4166	−0.2031	−3.7772
	(0.9798)	(1.4950)	(0.6158)
modnameh2o.dl:infoboth	0.8962	−1.4275	−1.0731
	(1.3857)	(2.1142)	(0.8709)
modnameh2o.glm:infoboth	1.1162	0.7607	−0.8906
	(1.3857)	(2.1142)	(0.8709)
modnamel2liblin:infoboth	−0.3846	0.7424	−1.2280
	(1.3857)	(2.1142)	(0.8709)
modnamelasso:infoboth	−0.0360	−0.7028	0.0436
	(1.3857)	(2.1142)	(0.8709)
modnamemars:infoboth	0.1236	0.3527	−0.6839
	(1.3857)	(2.1142)	(0.8709)
modnamennet:infoboth	6.6917	4.1189	−0.7890
	(1.3857)	(2.1142)	(0.8709)
modnamepls:infoboth	0.6638	0.0278	−1.0335
	(1.3857)	(2.1142)	(0.8709)
modnamerf:infoboth	0.2308	−2.3063	2.2336
	(1.3857)	(2.1142)	(0.8709)
modnameridge:infoboth	−0.6547	1.8429	−1.0705
	(1.3857)	(2.1142)	(0.8709)
modnamesl.avg:infoboth	−0.0769	0.8754	−0.7274
	(1.3857)	(2.1142)	(0.8709)
modnamesl.gbm:infoboth	−4.8221	0.3734	−0.3735
	(1.3857)	(2.1142)	(0.8709)
modnamesl.liblin:infoboth	0.2308	−1.6581	−0.8382
	(1.3857)	(2.1142)	(0.8709)
modnamesl.mars:infoboth	0.2490	0.2180	1.6166
	(1.3857)	(2.1142)	(0.8709)
modnamesl.ridge:infoboth	−0.0323	1.3515	−0.3422
	(1.3857)	(2.1142)	(0.8709)
modnamesvm:infoboth	0.3846	0.8026	−0.3766
	(1.3857)	(2.1142)	(0.8709)
modnamexgb:infoboth	0.3639	0.8641	−0.6653
	(1.3857)	(2.1142)	(0.8709)
Num. obs.	442	442	442

Standard errors in parentheses

Table C.6: Robust linear mixed effect model: out of time within-subjects (rank transformation)

	MAE	NDCG	AUC
(Intercept)	16.4000	4.4228	15.4000
	(0.8803)	(1.6003)	(0.8071)
infoboth	0.4000	−2.0228	−0.2000
	(1.2449)	(2.2632)	(1.1414)
modnameh2o.dl	−8.7945	9.9062	−6.2000
	(1.2449)	(2.2632)	(1.1414)
modnameh2o.glm	−10.0041	−1.4996	−12.0000
	(1.2449)	(2.2632)	(1.1414)
modnamel2liblin	−14.4000	2.9062	−12.8089
	(1.2449)	(2.2632)	(1.1414)
modnamelasso	−11.2000	1.7503	0.0223
	(1.2449)	(2.2632)	(1.1414)
modnamemars	−5.1094	7.9041	−1.0000
	(1.2449)	(2.2632)	(1.1414)
modnamennet	−5.7547	7.7772	−8.4000
	(1.2449)	(2.2632)	(1.1414)
modnamepls	−10.2000	5.5541	−3.9916
	(1.2449)	(2.2632)	(1.1414)
modnamerf	−1.4000	3.7456	−2.4000
	(1.2449)	(2.2632)	(1.1414)
modnameridge	−10.4000	3.2503	−8.5584
	(1.2449)	(2.2632)	(1.1414)
modnamesl.avg	−5.5998	7.2019	−13.0000
	(1.2449)	(2.2632)	(1.1414)
modnamesl.gbm	−4.8000	6.1555	−7.4000
	(1.2449)	(2.2632)	(1.1414)
modnamesl.liblin	−14.0000	0.9772	−7.7536
	(1.2449)	(2.2632)	(1.1414)
modnamesl.mars	−4.2000	4.9560	−6.9911
	(1.2449)	(2.2632)	(1.1414)
modnamesl.ridge	−3.4063	−0.8228	−10.0000
	(1.2449)	(2.2632)	(1.1414)
modnamesvm	−14.8000	10.2019	0.6000
	(1.2449)	(2.2632)	(1.1414)
modnamexgb	−2.6000	6.3877	−4.4000
	(1.2449)	(2.2632)	(1.1414)
modnameh2o.dl:infoboth	2.0008	1.7709	−0.0000
	(1.7606)	(3.2006)	(1.6141)
modnameh2o.glm:infoboth	0.9119	6.0196	0.4648
	(1.7606)	(3.2006)	(1.6141)
modnamel2liblin:infoboth	−0.2000	1.7890	−0.5911
	(1.7606)	(3.2006)	(1.6141)
modnamelasso:infoboth	6.1066	5.0365	−5.3807
	(1.7606)	(3.2006)	(1.6141)
modnamemars:infoboth	−3.4009	4.6959	−0.2000
	(1.7606)	(3.2006)	(1.6141)
modnamennet:infoboth	−3.7519	5.2228	−4.4584
	(1.7606)	(3.2006)	(1.6141)
modnamepls:infoboth	−0.8000	2.0459	−0.2973
	(1.7606)	(3.2006)	(1.6141)
modnamerf:infoboth	−0.0436	6.2297	2.4012
	(1.7606)	(3.2006)	(1.6141)
modnameridge:infoboth	−0.0235	3.9339	−2.4036
	(1.7606)	(3.2006)	(1.6141)
modnamesl.avg:infoboth	−3.5549	−2.9275	1.8000
	(1.7606)	(3.2006)	(1.6141)
modnamesl.gbm:infoboth	0.8000	−3.1555	1.6427
	(1.7606)	(3.2006)	(1.6141)
modnamesl.liblin:infoboth	−1.0000	0.4228	0.1288
	(1.7606)	(3.2006)	(1.6141)
modnamesl.mars:infoboth	−1.2235	−0.2560	0.9669
	(1.7606)	(3.2006)	(1.6141)
modnamesl.ridge:infoboth	−3.2565	3.0844	0.4000
	(1.7606)	(3.2006)	(1.6141)
modnamesvm:infoboth	0.0000	1.9981	0.8000
	(1.7606)	(3.2006)	(1.6141)
modnamexgb:infoboth	−0.3372	0.0401	2.4000
	(1.7606)	(3.2006)	(1.6141)
Num. obs.	170	170	170

Standard errors in parentheses

Appendix D

Regression results and supplementary graphs for chapter 4

D.1 TTS regression model results

D.1.1 Parametric and Cox model results

Table D.1: Survival model results estimating TTS

	Model			
	lognormal	Weibull	log-logistic	cox
	(1)	(2)	(3)	(4)
bank =B	−1.2501***	−0.9739***	−1.0732***	1.4679***
	(0.0278)	(0.0241)	(0.0255)	(0.0316)
default ltv	−0.0013***	−0.0007***	−0.0014***	0.0009***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)
loan.age.def	−0.0082***	−0.0074***	−0.0074***	0.0108***
	(0.0004)	(0.0003)	(0.0003)	(0.0005)
property.type=Apt.other	0.2862***	0.2627***	0.2825***	−0.3952***
	(0.0326)	(0.0294)	(0.0304)	(0.0436)
collateral.type=BTL	−0.2395***	−0.1928***	−0.2568***	0.2721***
	(0.0242)	(0.0209)	(0.0221)	(0.0311)
dublin.flag=Non.Dublin	0.0868***	0.0773***	0.0674***	−0.1185***
	(0.0273)	(0.0236)	(0.0249)	(0.0352)
unrdyyl1	0.1230***	0.0875***	0.1009***	−0.1448***
	(0.0060)	(0.0055)	(0.0056)	(0.0085)
Constant	4.9111***	5.0411***	4.8246***	
	(0.0310)	(0.0271)	(0.0278)	
Observations	14,463	14,463	14,463	14,463
R^2	0.2138	0.1928	0.2024	0.1906
χ^2 (df = 7)	3,413.1550***	3,037.1620***	3,207.0500***	3,048.8390***
Note:				*p<0.1; **p<0.05; ***p<0.01

D.1.2 Aalen semi-parametric survival model results

Table D.2: Aalen semi-parametric: cumulative regression functions

	Supremum-test	p-value	KS test	p-value	CvM test	p-value
(Intercept)	14.9311	0.0000	0.0763	0.0260	0.1925	0.0330
loan.age.def	13.4026	0.0000	0.0020	0.0000	0.0001	0.0010

Table D.3: Aalen semi-parametric: time constant variables

	Coef.	SE	Robust SE	z	P-val
bank = B	0.0106	0.0003	0.0006	17.5000	0.0000
default ltv	0.0000	0.0000	0.0000	2.4400	0.0145
property.type = Apt.other	-0.0020	0.0002	0.0002	-8.5000	0.0000
collateral.type = BTL	0.0014	0.0002	0.0002	6.3000	0.0000
dublin.flag = Non.Dublin	-0.0006	0.0002	0.0002	-3.2000	0.0014
unrdyyl1	-0.0006	0.0000	0.0001	-7.4000	0.0000

The panels in Figure D.1 thus illustrate the cumulative baseline hazard and the estimated cumulative coefficients for loan age at default. The baseline cumulative hazard, $\beta_0(t)$, is the estimated cumulative hazard for with the categorical variables at their reference levels (bank = A, property type = house, collateral type = PDH, Dublin flag = Dublin) and the continuous variables at their median (Default LTV (99.31 %), Loan age at default (64.5 months), unemployment rate changes lagged (1.08 %)). The figure on the right suggests that the effect of loan age at default initially increases strongly over time, then then at approximately 60 months remains relatively flat, then decreases slightly.

The plots for the time-changing cumulative regression function are contained in Figure D.1. These types of plots illustrate the change in hazard at time t, from the baseline hazard function, $\beta_0(t)$, for a one-unit change in the particular covariate, holding all other covariates constant.

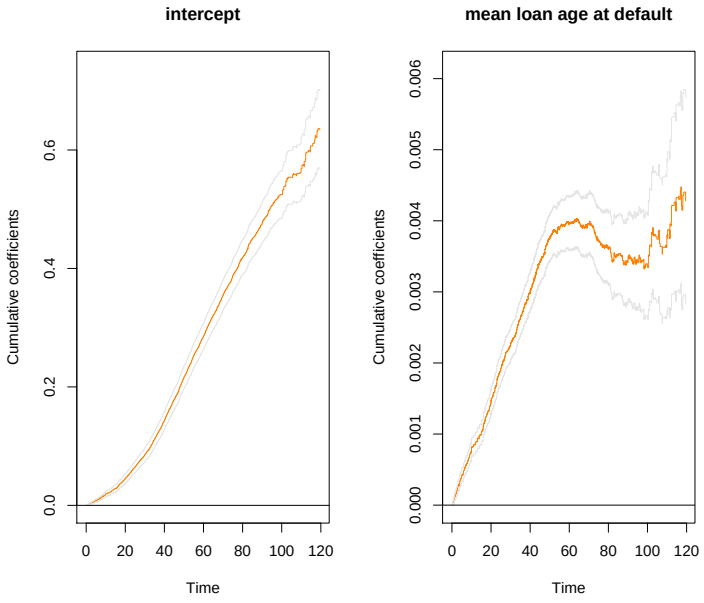

Figure D.1: Semi-parametric Aalen model

D.1.3 Effect sizes for survival regression models

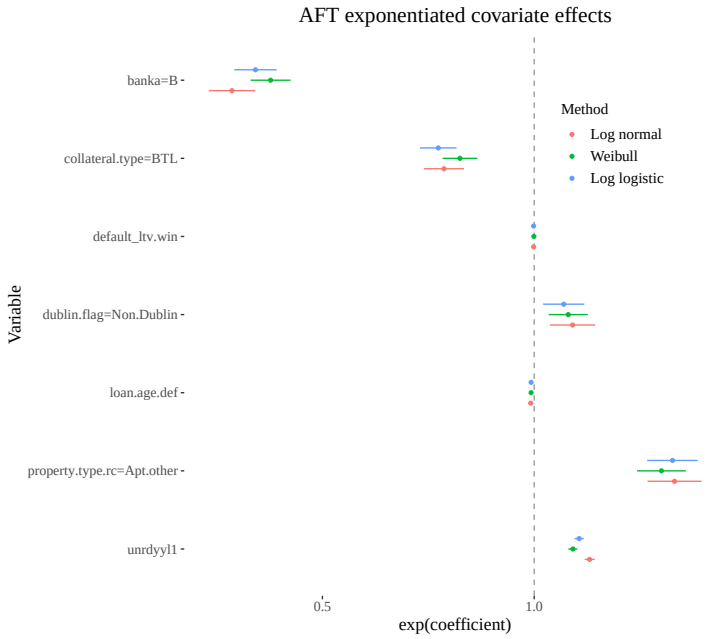

Figure D.2: TTS AFT models effect size

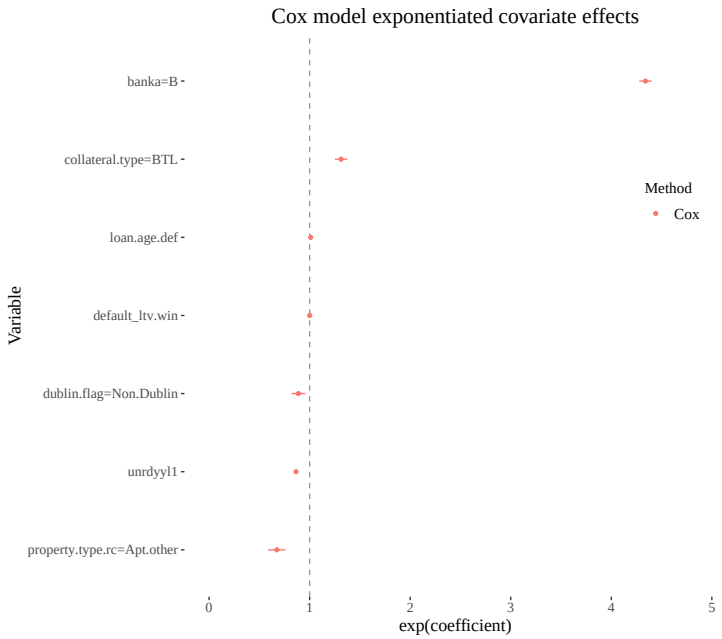

Figure D.3: TTS Cox model effect size

D.2 FSD regression model results

Table D.4: FSD parametric methods results

	OLS (untransformed)	Beta regression	Quantile regression (tau=0.5)
(Intercept)	0.1720***	−1.6418***	0.0933***
	(0.0179)	(0.0920)	(0.0213)
bank=B	−0.0662***	−0.3248***	−0.0679***
	(0.0074)	(0.0388)	(0.0097)
res.force.vol=FORCED	0.1223***	0.5741***	0.1364***
	(0.0091)	(0.0476)	(0.0120)
time	0.0015***	0.0063***	0.0019***
	(0.0002)	(0.0008)	(0.0002)
default ltv	−0.0001	0.0005	0.0002**
	(0.0001)	(0.0003)	(0.0001)
loan.age.def	−0.0000	−0.0012*	−0.0001
	(0.0001)	(0.0006)	(0.0002)
property.type=Apt.other	−0.0382***	−0.2699***	−0.0506***
	(0.0096)	(0.0495)	(0.0119)
collateral.type=BTL	0.0471***	0.2440***	0.0514***
	(0.0073)	(0.0420)	(0.0093)
unrdyyl1	0.0069***	0.0299***	0.0085***
	(0.0017)	(0.0086)	(0.0023)
Precision: (Intercept)		0.5991***	
		(0.0290)	
Precision: res.force.vol=FORCED		0.2561***	
		(0.0477)	
Precision: collateral.type=BTL		−0.1528***	
		(0.0397)	
R^2	0.2298		
Adj. R^2	0.2245		
Num. obs.	4689	4689	4689
RMSE	0.2188		
Pseudo R^2		0.1635	

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

References

O. Aalen. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6(4): 701–726, July 1978.

O. O. Aalen. A linear regression model for the analysis of life times. *Statistics in Medicine*, 8(8):907–925, Aug. 1989.

N. M. Adams, C. Anagnostopoulos, and D. Hand. Measuring classification performance: the hmeasure package. Technical report, Imperial College, London, 2012.

S. Aiello, E. Eckstrand, A. A. Fu, M. Landry, and P. Aboyoun. *h2o: R interface for H2O*, 2019. URL `http://www.h2o.ai`. R package version 3.26.0.11.

E. Alpaydin. *Machine Learning: The New AI*. MIT Press, 1 edition, 2016.

C. Anagnostopoulos and D. Hand. *hmeasure: The H-measure and other scalar classification performance metrics*, 2012. URL `http://CRAN.R-project.org/package=hmeasure`. R package version 1.0.

F. Andersson and T. Mayock. Loss severities on residential real estate debt during the Great Recession. *Journal of Banking and Finance*, 46(C):266–284, Sept. 2014.

B. Bade, D. Rösch, and H. Scheule. Empirical performance of loss given default prediction models. *The Journal of Risk Model Validation*, 5(2):25–44, June 2011.

B. Baesens, T. V. Gestel, S. Viaene, M.Stepanova, J. Suykens, and J. Vanthienen. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6):627–635, 2003a.

B. Baesens, R. Setiono, C. Mues, and J. Vanthienen. Using neural network rule extraction and decision tables for credit-risk evaluation. *Management Science*, 49(3):312–329, Mar. 2003b.

T. Balyuk and S. A. Davydenko. Reintermediation in fintech: evidence from online lending. *SSRN*, pages 1–54, June 2018.

A. G. Barto and R. S. Sutton. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

J. Bastos. Credit scoring with boosted decision trees. Working paper, CEMAPRE, School of Economics and Management, Lisbon, 2008.

J. Bastos. Forecasting bank loans loss given default. *Journal of Banking and Finance*, 34(10):2510–2517, Oct. 2010.

P. Baudino, J. Orlandi, and R. Zamil. The identification and measurement of non-performing assets: a cross country comparison. Fsi insights, Financial Stability Institute, BIS, Basel, 2018.

BCBS. Principles for the management of credit risk. Technical report, BIS, Basel, Sept. 2000.

T. Bellini. *IFRS 9 and CECL Credit Risk Modelling and Validation.* Academic Press, 1 edition, Jan. 2019.

T. Bellotti and J. Crook. Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications*, 36:3302–3308, 2009.

T. Bellotti and J. Crook. Forecasting and stress testing credit card default using dynamic models. *International Journal of Forecasting*, 29(4):563–574, Sept. 2013.

A. Benavoli, G. Corani, J. Demsar, and M. Zaffalon. Time for a Change: a Tutorial for Comparing Multiple Classifiers Through Bayesian Analysis. *Journal of Machine Learning Research*, 18(77):1–36, 2017.

K. P. Bennett and E. Parrado-Hernández. The interplay of optimization and machine learning research. *Journal of Machine Learning Research*, 7(Jul):1265–1281, 2006.

D. Berg. Bankruptcy prediction by generalised additive models. *Applied Stochastic Models in Business and Industry*, 23(2):129–143, 2007.

T. Berg, V. Burg, A. Gombović, and M. Puri. On the rise of FinTechs – credit scoring using digital footprints. *SSRN*, pages 1–50, July 2018.

J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

J. Bergstra, D. Yamins, and D. R. Cox. Making a science of model search: hyper-parameter optimization in hundreds of dimensions for vision architectures. In *Workshop on Automatic Machine Learning*, pages 115–123. Springer International Publishing, Cham, Feb. 2013.

R. Berk. *Statistical learning from a regression perspective.* Springer, 1 edition, 2008.

R. Berk and B. Kriegler. Small area estimation of the homeless in los angeles: an application of cost-sensitive stochastic gradient boosting. *The Annals of Applied Statistics*, 4(3):1234–1255, 2010.

J. Betz, R. Kellner, and D. Rösch. What drives the time to resolution of defaulted bank loans? *Finance Research Letters*, 18(C):7–31, Aug. 2016.

J. Betz, S. Krüger, R. Kellner, and D. Rösch. Macroeconomic effects and frailties in the resolution of non-performing loans. *Journal of Banking and Finance*, pages 1–26, Oct. 2017.

J. Betz, R. Kellner, and D. Rösch. Systematic effects among loss given defaults and their implications on downturn estimation. *European Journal of Operational Research*, pages 1–32, July 2018.

D. Bianchi, M. Büchner, and A. Tamoni. Bond risk premia with machine learning. *SSRN*, pages 1–84, 2018.

B. Bischl, M. Lang, L. Kotthoff, J. Schiffner, J. Richter, E. Studerus, G. Casalicchio, and Z. M. Jones.

mlr: machine learning in R. *Journal of Machine Learning Research*, 17(170):1–5, 2016. URL `http://jmlr.org/papers/v17/15-066.html`.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2 edition, 2006.

A. Bordes, S. Ertekin, J. Weston, and L. Bottou. Fast kernel classifiers with online and active learning. *Journal of Machine Learning Research*, 6:1579–1619, 2005.

A.-L. Boulesteix, R. Hable, S. Lauer, and M. J. A. Eugster. A statistical framework for hypothesis testing in real data comparison studies. *The American Statistician*, 69(3):201–212, Aug. 2015.

R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–23, Dec. 1952.

L. Brainard. What are we learning about artificial intelligence in financial services?, Nov. 2018. URL `https://www.federalreserve.gov/newsevents/speech/files/brainard20181113a.pdf`.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, Aug. 2001.

L. Breiman, J. Friedman, C. J. Stone, and R. Olshen. *Classification and Regression Trees*. Chapman and Hall/CRC, 1 edition, 1984.

L. Breiman, C. Chen, and A. Liaw. Using random forest to learn imbalanced data. Technical Report 666, Statistics Department, University of California at Berkeley, 2004.

L. Brieman. Random forests. *Machine Learning*, 45:5–32, 2001.

G. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, pages 1–3, 1950.

I. Brown and C. Mues. An experimental comparison of classification algorithms for imbalanced credit scoring datasets. *Expert Systems with Applications*, 39:3446–3453, 2012.

P. Bühlmann and T. Hothorn. Boosting algorithms: regularisation, prediction, and model fitting. *Statistical Science*, 22:477–505, 2007.

J. Burez and D. Van den Poel. Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36:4626–4636, 2009.

F. Butaru, Q. Chen, B. Clark, S. Das, A. W. Lo, and A. Siddique. Risk and risk management in the credit card industry. *Journal of Banking and Finance*, 72(C):218–239, Nov. 2016.

R. Calabrese and S. A. Osmetti. Generalized extreme value regression for binary rare events data: an application to credit defaults. *Journal of Applied Statistics*, 40(6):1172–1188, 2013.

J. Y. Campbell, S. Giglio, and P. Pathak. Forced sales and house prices. *American Economic Review*, 101 (5):2108–2131, Aug. 2011.

A. Candel, E. LeDell, A. Arora, and V. Parmar. *Deep learning with H2O*, January 2020. URL `http://h2o.ai/resources`.

R. Caruana, A. Munson, and A. Niculescu-Mizil. Getting the most out of ensemble selection. *Sixth International Conference on Data Mining (ICDM'06)*, pages 828–833, Sept. 2006.

R. Caruana, N. Karampatziakis, and A. Yessenalina. *An empirical evaluation of supervised learning in high dimensions*. ACM, New York, New York, USA, July 2008.

R. Caruana, Y. Lou, and J. Gehrke. Intelligible models for classification and regression. In *Proc. 23rd ACM SIGKDD Conference, Beijing, China, August 2012*. SIGKDD, 2012.

G. Chawla, L. R. Forest, and S. D. Aguais. Point-In-Time (PIT) LGD and EAD models for IFRS9/CECL and stress testing. *Journal of Risk Management in Financial Institutions*, 9(3):249–263(, Mar. 2016.

H. Z. Chen. A new model for bank loan loss given default by leveraging time to recovery. *The Journal of Credit Risk*, 14(3):1–29, 2018.

T. Chen and C. Guestrin. XGBoost. In *The 22nd ACM SIGKDD International Conference*, pages 785–794, New York, New York, USA, 2016. ACM Press.

S. Claessens, J. Frost, G. Turner, and F. Zhu. Fintech credit markets around the world: size, drivers and policy issues. *BIS Quarterly Review*, (3):29–49, Sept. 2018.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972. ISSN 00359246. URL `http://www.jstor.org/stable/2985181`.

F. Cribari-Neto and A. Zeileis. Beta regression in R. *Journal of Statistical Software*, 34(2):1–24, 2010.

J. Crook and T. Bellotti. Credit scoring with macroeconomic variables using survival analysis. *The Journal of the Operational Research Society*, 60(12):1699–1707, 2009.

J. Crook, D. Edelman, and L. C. Thomas. Recent developments in consumer credit risk assessment. *The Journal of the Operational Research Society*, 183:1447–1465, 2007.

S. R. Das. The principal principle. *Journal of Financial and Quantitative Analysis*, 47(6):1215–1246, 2012.

S. R. Das and R. Meadows. Strategic loan modification: an options based response to strategic default. *Journal of Banking and Finance*, 37:636–647, 2013.

E. J. de Fortuny and D. Martens. Active learning based rule extraction for regression. In *2012 IEEE 12th International Conference on Data Mining Workshops*, pages 926–933. IEEE, Nov. 2012.

K. W. DeBock, K. Coussement, and D. V. den Pol. Ensemble classification based on generalised additive models. *Computational Statistics and Data Analysis*, 54:1535–1546, 2010.

J. Demsar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Y. Dendramis, E. Tzavalis, and G. Adraktas. Credit risk modelling under recessionary and financially distressed conditions. *Journal of Banking and Finance*, 91:160–175, June 2018.

Y. Deng, J. Quigley, and R. Van Order. Mortgage terminations, heterogeneity and the exercise of mortgage options. *Econometrica*, 68(2):275–307, 2000.

S. Derksen and H. J. Keselmanl. Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, 45:265–282, 1992.

J. Dermine and C. N. de Carvalho. Bank loan losses-given-default: a case study. *Journal of Banking & Finance*, 30(4):1219–1243, Apr. 2006.

I. Dewancker, M. McCourt, S. Clark, P. Hayes, A. Johnson, and G. Ke. A strategy for ranking optimization methods using multiple criteria. *JMLR Workshop and Conference Proceedings*, 64:11–20, Dec. 2016.

P. Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10): 78–87, 2012.

P. Domingos. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Basic Books, Inc., New York, NY, USA, 2018. ISBN 0465094279, 9780465094271.

H. Donner, H.-S. Song, and M. Wilhelmsson. Forced sales and their impact on real estate prices. *Journal of Housing Economics*, 34(C):60–68, Dec. 2016.

S. Donnery, T. Fitzpatrick, D. Greaney, F. McCann, and M. OKeeffe. Resolving non-performing loans in Ireland:2010-2018. *Central Bank of Ireland Quarterly Bulletin*, (1):1–17, Apr. 2018.

G. Dorfleitner, C. Priberny, S. Schuster, J. Stoiber, M. Weber, I. de Castro, and J. Kammler. Description-text related soft information in peer-to-peer lending: evidence from two leading European platforms. *Journal of Banking and Finance*, 64(C):169–187, Mar. 2016.

D. Dua and C. Graff. UCI machine learning repository, 2019. URL `http://archive.ics.uci.edu/ml`.

J. Duarte, S. Siegel, and L. Young. Trust and credit: the role of appearance in peer-to-peer lending. *Review of Financial Studies*, 25(8):2455–2484, July 2012.

B. Eder and M. Bank. A survey on the estimation of expected credit losses for IFRS 9. In *Proceedings of the Credit Scoring and Credit Control Conference XVI*, pages 1–67, Edinburgh, Aug. 2019. University of Edinburgh.

J. Elith, J. Leathwick, and T. Hastie. A working guide to boosted regression trees. *Journal of Animal Ecology*, 3(1):802–813, 2008.

C. Elkan. The foundations of cost sensitive learning. In *Proc. of the Seventeenth International Joint Conference on Artificial Intelligence, Seattle, Washington, USA*. IJCAI, 2001.

R. Emekter, Y. Tu, B. Jirasakuldech, and M. Lu. Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics*, 47(1):54–70, Oct. 2014.

M. J. A. Eugster, F. Leisch, and C. Strobl. (Psycho-)analysis of benchmark experiments: A formal framework for investigating the relationship between data sets and learning algorithms. *Computational Statistics & Data Analysis*, 71:986–1000, Mar. 2014.

European Commission. Call for advice to the EBA for the purposes of a benchmarking of national loan enforcement frameworks (including insolvency frameworks) from a bank creditor perspective. Commission communication, European Commission, 2019. URL `https://eba.europa.eu/documents/10180/2556373/CfA+EBA+ins+bmkg+draft+7424809.pdf/1593a7ff-69e1-487f-9e8f-adbfe294a496`.

R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: a library for large linear classification. *Journal of Machine Learning Research*, 9(Aug):1871–1874, 2008.

D. Feldman and S. Gross. Mortgage default: classification tree analysis. *The Journal of Real Estate Finance and Economics*, 30(4):369–396, 2005.

M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15:3133–3181, 2014.

S. Ferrari and F. Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815, Aug. 2004.

M. Feurer, A. Klein, K. Eggensperger, J. T. Springenberg, M. Blum, and F. Hutter. Efficient and robust automated machine learning. In *NIPS Proceedings of the 28th International Conference on Neural Information Processing Systems*, pages 2755–2763. MIT Press, Dec. 2015.

S. Finlay. Credit scoring for profitability objectives. *European Journal of Operational Research*, 202(2): 528–537, Apr. 2010.

T. Fitzpatrick and C. Mues. An empirical comparison of classification algorithms for mortgage default prediction: evidence from a distressed mortgage market. *European Journal of Operational Research*, 249(2):427–439, Mar. 2016.

P. Flach. Performance evaluation in machine learning: the good, the bad, the ugly, and the way forward. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:9808–9814, July 2019.

P. Flach, J. Hernandez-Orallo, and C. Ferri. A coherent interpretation of AUC as a measure of aggregated classification performance. In *Proc. 28rd International Conference on Machine Learning, Bellevue,WA, USA, June 2011*. ICML, 2011.

C. Foote, K. Gerardi, and P. Willen. Negative equity and foreclosure: theory and evidence. *Journal of Urban Economics*, 2(64):234–245, 2008.

Y. Freund and R. Schaipre. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

J. H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–141, Jan. 1991.

J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5): 1189–1232, 2001.

J. H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38(4):367–378, 2002.

J. H. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28(2):337–374, 2000.

J. H. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, Feb. 2010.

M. Friedman. A comparison of alternative tests of significance for the problem of m rankings. *Annals of Mathematical Statistics*, 11(1):82–92, 1940.

A. Fuster, P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther. Predictably unequal? The effects of machine learning on credit markets. *SSRN*, pages 1–73, Nov. 2018.

J. Galindo and P. Tamayo. Credit risk assessment using statistical and machine learning: basic methodology and risk modelling applications. *Computational Economics*, pages 1–37, 2000.

S. Garcia and F. Herrera. An extension on statistical comparison of classifiers over multiple data sets for all pairwise comparisons. *Journal of Machine Learning Research*, 9:2667–2694, 2008.

F. Garrido, W. Verbeke, and C. Bravo. A robust profit measure for binary classification model evaluation. *Expert Systems With Applications*, 92:154–160, Feb. 2018.

T. A. Gerds and M. Schumacher. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040, Dec. 2006.

T. A. Gerds, T. Cai, and M. Schumacher. The performance of risk prediction models. *Biometrical Journal*, 50(4):457–479, Aug. 2008.

C. Giraud-Carrier and F. Provost. Toward a justification of meta-learning: is the no free lunch theorem a show-stopper. In *In ICML workshop on meta-learning*, pages 9–16, 2005.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, Mar. 2007.

D. Gómez and A. Rojas. An empirical overview of the no free lunch theorem and its effect on real-world machine learning classification. *Neural Computation*, 28(1):216–228, Jan. 2016.

Y. Guo, W. Zhou, C. Luo, C. Liu, and H. Xiong. Instance-based credit risk assessment for investment decisions in P2P lending. *European Journal of Operational Research*, 249(2):417–426, Mar. 2016.

M. Gurtler and M. Hibbeln. Improvements in loss given default forecasts for bank loans. *Journal of Banking and Finance*, 37(7):2354–2366, July 2013.

I. Guyon, K. Bennett, G. Cawley, H. J. Escalante, S. Escalera, T. K. Ho, N. Macià, B. Ray, M. Saeed, A. Statnikov, and E. Viegas. Design of the 2015 ChaLearn AutoML challenge. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, Feb. 2015.

I. Guyon, I. Chaabane, H. J. Escalante, S. Escalera, D. Jajetic, J. R. Lloyd, N. Macià, B. Ray, L. Romaszko, M. Sebag, A. Statnikov, S. Treguer, and E. Viegas. A brief review of the ChaLearn AutoML challenge: any-time any-dataset learning without human intervention. In *Workshop on Automatic Machine Learning*, pages 21–30. Dec. 2016.

D. Hand. Classifier technology and the illusion of progress. *Statistical Science*, 21(1):1–14, 2006.

D. Hand. Measuring classifier performance: a coherent alternative to area under the ROC curve. *Machine Learning*, 77:103–123, 2009a.

D. Hand. Mining the past to determine the future: problems and possibilities. *International Journal of Forecasting*, 25:441–451, 2009b.

F. E. Harrell, K. l. Lee, and D. B. Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15 (4):361–387, Feb. 1996.

F. E. Harrell Jr. *rms: regression modelling strategies*, 2019. URL `https://CRAN.R-project.org/package=rms`. R package version 5.1-3.1.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Addison-Wesley, 2nd edition, 2009.

A. Haughwout, R. Peach, and J. Tracy. Juvenile delinquent mortgages: bad credit or bad economy. *Journal of Urban Economics*, (64):246–257, 2008.

T. Helleputte. *LiblineaR: linear predictive models based on the LIBLINEAR C/C++ library*, 2017. R package version 2.10-8.

A. Hertzberg, A. Liberman, and D. Paravisini. Adverse selection on maturity: evidence from online consumer credit. In *Financial Innovation Online Lending to Households and Small Businesses*, pages 1–66, Dec. 2016.

A. E. Hoerl and R. W. Kennard. Ridge regression: biased estimation for non-orthogonal problems. *Technometrics*, 12(1):55–14, Feb. 1970.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6: 65–70, 1979.

N. J. Horton and K. P. Kleinman. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, 61(1):71–90, 2007.

C. M. Hurvich, J. S. Simonoff, and C.-L. Tsai. Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society B*, 60: 271–293, 1998.

F. Hutter, L. Kotthoff, and J. Vanschoren, editors. *Automated Machine Learning*. Methods, Systems, and Challenges. Springer, Cham, Aug. 2019.

R. L. Iman and J. M. Davenport. Approximations of the critical region of the friedman statistic. *Communications in Statistics*, 9(6):571–595, 1980.

H. Ishwaran, U. Kogalur, E. Blackstone, and M. Lauer. Random survival forests. *Annals of Applied Statistics*, 2(3):841–860, 2008.

J. Jagtiani and C. Lemieux. Fintech lending: financial inclusion, risk pricing, and alternative information . Technical report, Research Department, Federal Reserve Bank of Philadelphia, Philadelphia, PA, July 2017.

J. Jagtiani and C. Lemieux. The roles of alternative data and machine learning in fintech lending: evidence from the Lending Club consumer platform. Technical report, Research Department, Federal Reserve Bank of Philadelphia, Philadelphia, PA, Mar. 2018.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer, 2014. ISBN 9781461471370.

N. Japkowicz and M. Shah. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, 1st edition, 2011. ISBN 9780511921803.

A. Kalousis, J. Gama, and M. Hilario. On data and algorithms: understanding inductive performance. *Machine Learning*, 54(3):275–312, Feb. 2004.

E. Kaplan, , and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis. kernlab: an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20, 2004.

R. Kelly. A model of irish mortgage default. Technical Report 2011/04, Central Bank of Ireland, 2011.

R. Kelly and F. McCann. Some defaults are deeper than others: understanding long-term mortgage arrears. *The Journal of Banking and Finance*, 72:15–27, 2016.

G. Kennedy and T. McIndoe-Calder. The irish mortgage market: stylised facts, negative equity, and arrears. Technical Report 2011/04, Central Bank of Ireland, Dublin, 2012.

K. Kennedy, B. M. Namee, S. J. Delaney, M. O'Sullivan, and N. Watson. A window of opportunity: assessing behavioural scoring. *Expert Systems with Applications*, 64(4):1372–1380, 2013a.

K. Kennedy, B. M. Nameea, and S. J. Delaney. Using semi-supervised classifiers for credit scoring. *The Journal of the Operational Research Society*, 64:513–529, 2013b.

A. Kim, Y. Yang, S. Lessmann, T. Ma, M. C. Sung, and J. E. V. Johnson. Can deep learning predict risky retail investors? A case study in financial risk behaviour forecasting. *European Journal of Operational Research*, pages 1–18, Dec. 2019.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2014. URL `https://arxiv.org/abs/1412.6980`.

R. Koenker. *quantreg: quantile regression*, 2019. URL `https://CRAN.R-project.org/package=quantreg`. R package version 5.51.

R. Koenker and K. F. Hallock. Quantile regression. *Journal of Economic Perspectives*, 15(4):143–156, Nov. 2001.

M. Koller. robustlmm: an R package for robust estimation of linear mixed-effects models. *Journal of Statistical Software*, 75(1):1–24, Dec. 2016.

G. Krempl and V. Hofer. Classification in presence of drift and latency. In *2011 11th IEEE International Conference on Data Mining Workshops*. ICDM, 2011.

S. Krüger and D. Rösch. Downturn LGD modelling using quantile regression. *Journal of Banking and Finance*, 79:42–56, June 2017.

J. Kruppa, A. Schwarz, G. Arminger, and A. Ziegler. Consumer credit risk: individual probability estimates using machine learning. *Expert Systems With Applications*, 40(13):5125–5131, Oct. 2013.

M. Kuhn. Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5): 1–26, 2008. URL `http://www.jstatsoft.org/v28/i05`.

M. Kuhn and K. Johnson. *Applied Predictive Modelling*. Springer, 1st edition, 2013.

D. Lando, M. Medhat, M. S. Nielsen, and S. F. Nielsen. Additive intensity regression models in corporate default analysis. *Journal of Financial Econometrics*, 11(3):443–485, June 2013.

K.-y. Lee. Examining REO sales and price discounts in Massachusetts. In P. Chakrabarti, M. Lambert, and M. H. Petrus, editors, *REO Vacant Properties Strategies for Neighbourhood Stabilisation*, pages 55–65. Boston, Aug. 2010.

M. Leow and C. Mues. Predicting loss given default (LGD) for residential mortgage loans: a two-stage model and empirical evidence for UK bank data. *International Journal of Forecasting*, 28(1):183–195, Dec. 2011.

M. Leow, C. Mues, and L. C. Thomas. Competing risks survival model for mortgage loans with simulated loss distributions. In *Proceedings of the Credit Scoring and Credit Control XII conference, United Kingdom*, Aug. 2011.

S. Lessmann and S. Voß. Car resale price forecasting: the impact of regression method, private information, and heterogeneity on forecast accuracy. *International Journal of Forecasting*, 33(4): 864–877, Sept. 2017.

S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas. Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research. *European Journal of Operational Research*, 247 (1):124–136, Nov. 2015.

Y. Li, T. Bellotti, and N. Adams. Issues using logistic regression with class imbalance, with a case study from credit risk modelling. *Foundations of Data Science*, 1(4):389–417, 2019.

A. Liaw and M. Wiener. Classification and regression by random forest. *R News*, 2(3):18–22, 2002. URL `http://CRAN.R-project.org/doc/Rnews/`.

J. M. Liberti and M. A. Petersen. Information: hard and soft. Working paper, Kellogg School, Northwestern University, 2017.

M. Lin. Economic value of texts: evidence from online debt crowdfunding. In *Financial Innovation Online Lending to Households and Small Businesses*, pages 1–37, Nov. 2016.

T.-Y. Liu. *Learning to Rank for Information Retrieval and Natural Language Processing, Second Edition*, volume 7. Springer, Berlin, Heidelberg, Oct. 2011.

W. Liu, C. Vu, and J. Cela. Generalisations of generalised additive models (gam): a case of credit risk modelling. Conference paper 113-2009, SAS Forum, Washington, 2009.

A. K. Lo, A. E. Khandani, and A. J. Kim. Consumer credit risk models via machine-learning algorithms. *Journal of Banking and Finance*, 34(11):2767–2787, 2010.

G. Loterman, I. Brown, D. Martens, C. Mues, and B. Baesens. Benchmarking regression algorithms for loss given default modelling. *International Journal of Forecasting*, 28(1):161–170, Dec. 2012.

D. J. Lowsky, Y. Ding, D. K. K. Lee, C. E. McCulloch, L. F. Ross, J. R. Thistlethwaite, and S. A. Zenios. A k-nearest neighbours survival probability prediction method. *Statistics in Medicine*, 32(12): 2062–2069, May 2013.

G. Luo. A review of automatic selection methods for machine learning algorithms and hyper-parameter values. *Network Modelling Analysis in Health Informatics and Bioinformatics*, 5(1):18–15, May 2016.

M. Malekipirbazari and V. Aksakalli. Risk assessment in social lending via random forests. *Expert Systems With Applications*, 42(10):4621–4631, June 2015.

J. Malley, J. Kruppa, A. Dasgupta, G. Malley, and A. Ziegler. Probability machines. *Methods of Information in Medicine*, 1:74–81, 2012.

D. Martens, B. Baesens, T. V. Gestel, and J. Vanthienen. Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research*, 183: 1466–1476, 2007.

N. Martin. Assessing scorecard performance: a literature review and classification. *Expert Systems and Applications*, 40:6340–6350, 2013.

T. Martinussen and T. H. Scheike. *Dynamic regression models for survival data*. Springer, 1 edition, 2006.

P. Matthews. Effectively deploying analytics to support collections. Presentation at the Credit Scoring and Credit Control XII Conference, Edinburgh, 2011.

D. Mease and A. Wyner. Evidence to the contrary of statistical boosting. *Journal of Machine Learning Research*, 9:131–156, 2008.

L. Medema, R. Koning, and R. Lensink. A practical approach to validating a PD model. *Journal of Banking and Finance*, 31:701–706, 2009.

B.-H. Mevik and R. Wehrens. The pls package: principal component and partial least squares regression in R. *Journal of Statistical Software*, 18(1):1–23, Jan. 2007.

S. Millborrow. *earth: multivariate adaptive regression splines*, 2018. R package version 4.6.0.

S. Miller. Information and default in consumer credit markets: evidence from a natural experiment. *Journal of Financial Intermediation*, 24(1):45–70, Jan. 2015.

T. Mitchell. *Machine Learning*. McGraw-Hill Science, London, 1st edition, 1997. ISBN 0070428077.

U. B. Mogensen, H. Ishwaran, and T. A. Gerds. Evaluating random forests for survival analysis using prediction error curves. *Journal of Statistical Software*, 50(1):1–23, Sept. 2012.

G. D. Montanez. The famine of forte: few search problems greatly favour your algorithm. In *2017 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 477–482. IEEE, Apr. 2017a.

G. D. Montanez. *Why machine learning works*. PhD thesis, Department of Machine Learning,Carnegie Mellon University, 2017b.

K. P. Murphy. *Machine learning: a probabilistic perspective*. The MIT Press, 1 edition, 2012.

W. Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4): 945–966, 1972. ISSN 00401706. URL `http://www.jstor.org/stable/1267144`.

N. Nikolaou, N. Edakunni, M. Kull, P. Flach, and G. Brown. Cost-sensitive boosting algorithms: do we really need them? *Machine Learning*, 104(2):359–384, July 2016.

A. B. Owen. Infinitely imbalanced logistic regression. *Journal of Machine Learning Research*, 8(Apr): 761–773, 2007.

Y. W. Park and D. W. Bang. Loss given default of residential mortgages in a low LTV regime: role of foreclosure auction process and housing market cycles. *Journal of Banking and Finance*, 39(C): 192–210, Feb. 2014.

M. Phillips. Welcome to ireland, where mortgage payments are apparently optional, 2013. URL `https://qz.com/50615/`.

M. Qi and X. Yang. Loss given default of high loan-to-value residential mortgages. *Journal of Banking and Finance*, 33(5):788–799, May 2009.

G. Rapisarda and D. Echeverry. A nonparametric approach to incorporating incomplete workouts into loss given default estimates. *The Journal of Credit Risk*, pages 1–16, June 2013.

G. Ridgeway. Generalized boosted models: a guide to the gbm package. Technical report, 2012. URL `https://github.com/harrysouthworth/gbm/blob/master/inst/doc/gbm.pdf`. [Online; accessed 10-July-2014].

G. Ridgeway. *gbm: generalized boosted regression models*, 2013. URL `http://CRAN.R-project.org/package=gbm`. R package version 2.1.

C. Rudin and D. Carlson. The secrets of machine learning: ten things you wish you had known earlier to be more effective at data analysis. *Tutorials in Operations Research*, pages 1–29, June 2019.

C. Rudin and Y. Shaposhnik. Globally-consistent rule-based summary-explanations for machine learning models: application to credit-risk evaluation. *SSRN*, pages 1–19, 2019.

T. H. Scheike and M.-J. Zhang. Analyzing competing risk data using the R timereg package. *Journal of Statistical Software*, 38(2):1–15, 2011. URL `http://www.jstatsoft.org/v38/i02/`.

C. Serrano-Cinca and B. Gutiérrez-Nieto. The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. *Decision Support Systems*, 89(C):113–122, Sept. 2016.

T. S. Sethi and M. Kantardzic. On the reliable detection of concept drift from streaming unlabeled data. *Expert Systems With Applications*, 82(1):77–99, Apr. 2017.

H. Sheng Sun and Z. Jin. Estimating credit risk parameters using ensemble learning methods: an empirical study on loss given default. *The Journal of Credit Risk*, 12(3):1–28, Aug. 2016.

J. Sirignano, A. Sadhwani, and K. Giesecke. Deep learning for mortgage risk. *arxiv*, pages 1–83, July 2016.

M. Smithson and J. Verkuilen. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, 11(1):54–71, 2006.

C. M. So, C. Mues, A. T. de Almeida Filho, and L. C. Thomas. Debtor level collection operations using Bayesian dynamic programming. *Journal of the Operational Research Society*, 70(8):1332–1348, Feb. 2020.

M. Somers and J. Whittaker. Quantile regression for modelling distributions of profit and loss. *European Journal of Operational Research*, 183(3):1477–1487, Dec. 2007.

V. Sreekanth, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Generalized RBF feature maps for efficient detection. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2010.

Y. Tanoue, A. Kawada, and S. Yamashita. Forecasting loss given default of bank loans with multi-stage model. *International Journal of Forecasting*, 33(2):513–522, Feb. 2017.

L. C. Thomas. *Consumer Credit Models*. Springer, New York, 2nd edition, 2009.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, Apr. 1996.

E. Tobback and D. Martens. Retail credit scoring using fine-grained payment data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 54:627–20, May 2019.

E. Tobback, D. Martens, T. Van Gestel, and B. Baesens. Forecasting loss given default models: impact of account characteristics and the macroeconomic state. *Journal of the Operational Research Society*, pages 1–17, Jan. 2014.

E. N. C. Tong, C. Mues, and L. Thomas. A zero-adjusted gamma model for mortgage loan loss given default. *International Journal of Forecasting*, 29(4):548–562, Oct. 2014.

G. Tutz and H. Binder. A comparison of methods for the fitting of generalized additive models. *Statistics and Computing*, 18:87–99, 2008.

F. Valencia and L. Laeven. Systemic banking crises revisited. Technical Report 206, International Monetary Fund, Washington DC, Sept. 2018.

B. Vallee and Y. Zeng. Marketplace lending: a new banking paradigm? pages 1–60, jan 2018. URL `https://www.hbs.edu/faculty/Pages/item.aspx?num=53870`.

T. Van Gestel, B. Baesens, and D. Martens. From linear to non-linear kernel based classifiers for bankruptcy prediction. *Neurocomputing*, 73(2):2955–2970, 2010.

R. Van Order. Modelling and evaluating the credit risk of mortgage loans: a primer. *Journal of Risk Model Validation*, 2(2):63–82, 2008.

J. N. van Rijn, S. M. Abdulrahman, P. Brazdil, and J. Vanschoren. Fast algorithm selection using learning curves. In E. Fromont, T. De Bie, and M. van Leeuwen, editors, *Advances in Intelligent Data Analysis XIV. IDA*, pages 1–12. Springer International Publishing, 2015.

V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

T. Verbraken, C. Bravo, R. Weber, and B. Baesens. Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research*, 238(2):505–513, Oct. 2014.

K. L. Wagstaff. Machine learning that matters. In *ICML*, pages 1–6, May 2012.

D. H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7): 1341–1390, Oct. 1996.

D. H. Wolpert. The supervised learning no-free-lunch theorems. In *Soft Computing and Industry*, pages 25–42. Springer London, London, 2002.

D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, Apr. 1997.

S. Wood. *Generalised additive models: an introduction with R*. CRC, London, 1 edition, 2006.

S. Wood. *mgcv: mixed GAM computation vehicle with GCV/AIC/REML smoothness estimation*, 2013. URL `http://CRAN.R-project.org/package=mgcv`. R package version 1.7-24.

M. N. Wright. *bnnSurvival: bagged k-nearest neighbours survival prediction*, 2017. URL `https://CRAN.R-project.org/package=bnnSurvival`. R package version 0.1.5.

M. N. Wright and A. Ziegler. ranger: a fast implementation of random forests for high dimensional data in C++and R. *Journal of Statistical Software*, 77(1):1–17, 2017.

M. N. Wright, T. Dankowski, and A. Ziegler. Unbiased split variable selection for random survival forests using maximally selected rank statistics. *Statistics in Medicine*, 36(8):1272–1284, Jan. 2017.

D. H. Xuan, D. Rösch, and H. Scheule. Liquidity constraints, home equity and residential mortgage losses. *The Journal of Real Estate Economics and Finance*, pages 1–39, June 2019.

X. Yan, J. Guo, Y. Lan, and X. Cheng. A bit-term topic model for short texts. In *WWW '13 Publication WWW Proceedings of the nd international conference on World Wide Web*, pages 1445–1456, May 2013.

J. Zhang and L. C. Thomas. Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD. *International Journal of Forecasting*, 28(1): 204–215, Oct. 2014.

D. Zhao, Y. Wang, and T. F. Sing. Impact of foreclosure laws on mortgage loan supply and performance. *Journal of Real Estate Economics and Finance*, pages 1–42, Jan. 2019.

X. Zhao, P. Li, and X. Zhang. Modelling loss given default. Technical Report 3, FDIC, July 2018.

T. Ziegler, R. Shneor, K. Garvey, K. Wenzlaff, N. Yerolemou, R. Hao, and B. Zhang. Expanding horizons - 3rd european alternative finance report. Technical report, Cambridge Center for Alternative Finance, Judge Business School, University of Cambridge, Cambridge, 2018.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, Apr. 2005.