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

Trevor Fitzpatrick, Christophe Mues

PII: S0377-2217(15)00838-3
DOI: 10.1016/j.ejor.2015.09.014
Reference: EOR 13233

To appear in: *European Journal of Operational Research*

Received date: 18 January 2014
Revised date: 5 August 2015
Accepted date: 8 September 2015

Please cite this article as: Trevor Fitzpatrick, Christophe Mues, An empirical comparison of classification algorithms for mortgage default prediction: evidence from a distressed mortgage market, *European Journal of Operational Research* (2015), doi: 10.1016/j.ejor.2015.09.014

Highlights

- We evaluate default prediction performance of machine learning/regression models.
- Including boosted trees, random forests, penalised linear/semi-parametric logistic regression.
- Using data on over 300,000 residential mortgage loans.
- The results indicate varying degrees of predictive power.
- Statistical tests suggest boosted regression trees outperform penalised logistic regression.

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

Trevor Fitzpatrick[1,*]

Southampton Business School, University of Southampton, Highfield, Southampton, SO171BJ, UK

Central Bank of Ireland, PO Box 559, Dame Street, Dublin 2, Ireland

Christophe Mues

Southampton Business School, University of Southampton, Highfield, Southampton, SO171BJ, UK

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

1. Introduction

1.1. 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

*Corresponding author.

Email addresses: T.Fitzpatrick@soton.ac.uk (Trevor Fitzpatrick), C.Mues@soton.ac.uk (Christophe Mues)

[1]Ph+35312246000. The views expressed in the paper are those of the authors and do not represent the views of the Central Bank of Ireland or the European Central Bank.

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 technique 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 (Baesens et al., 2003; Kennedy et al., 2013b; Lessmann et al., 2015). 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).

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 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.

1.2. 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 Baesens et al. (2003), Crook et al. (2007), Crook and Bellotti (2009), Brown and Mues (2012), Kennedy et al. (2013b), and Lessmann et al. (2015). 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. To answer this question, 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, Deng et al. (2000) and Das and Meadows (2013) 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 (Foote et al., 2008; Haughwout et al., 2008; Kelly, 2011).[2] 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 (Caruana et al., 2012; Elith et al., 2008; Hastie et al., 2009; Liu et al., 2009). 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.[3] The algorithmic complexity involved in solving the general SVM

[2] 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.

[3] Sometimes, 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.

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.

1.3. 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), Thomas (2009), Hand (2009b),and Martin (2013). However, with the exceptions of Galindo and Tamayo (2000), or Feldman and Gross (2005), Kennedy et al. (2013a), 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.

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 Lo et al. (2010), Brown and Mues (2012), and Lessmann et al. (2015), 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 Burez and Van den Poel (2009), Brown and Mues (2012), and Kennedy et al. (2013b). 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 imbalanced real-world mortgage datasets 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. Statistical and classification models

The aim of each model is to produce a loan-level prediction for a binary variable; 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 greater than 90 days arrears.

2.1. Logistic regression

Logistic Regression (LR) is known as a classifier that performs reasonably well across many application settings and data types, including credit scoring (Brown and Mues, 2012; Kennedy et al., 2013b; Lessmann et al., 2015). 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 4.1).[4] 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 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.[5]

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

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 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 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).

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

[5] 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*.

2.2. 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 in this paper 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 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.

$$\text{logit}(P(y = 1|x)) = \beta_0 + \sum_{j=1}^{q} s_j(x_j) + \sum_{j=q+1}^{p} \beta_j x_j \tag{2}$$

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.[6] 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).[7]

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

[7] 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).

2.3. Decision tree-based methods

The tree-based models in this paper draw on Classification and Regression Trees (CART) (Brieman 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 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.

2.3.1. 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).[8]

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).[9] 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

[8] 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).

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

gradients.[10] 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.

2.3.2. Random Forests (RF)

Random Forests (RF) are another tree-ensemble classifier developed by Brieman (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 subsamples 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

[10] 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).

such as bioinformatics, image recognition, as well as in financial applications such as customer attrition and credit scoring (Kruppa et al., 2012; Lessmann et al., 2015; Malley et al., 2012).

3. Model building and data sets

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

3.1. 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 4.1). 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 Brieman (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.

3.2. 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.

When added together, the training, validation and test samples used in this paper amount to 322,915 cases across the four portfolios.[11] 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.[12]

3.2.1. 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

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

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

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.

3.2.2. Data description

The dataset variables are described in Table 1.[13] 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 more than 90 days past due (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.

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).[14]

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.[15] Due to data limitations, this variable is not available for Portfolio 3. Past arrears status (variable Recent 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

[13] 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).

[14] 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.

[15] 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.

Table 1: Description of variables

Variable	Description	Type
Default	Dependent variable: 1 if borrower greater than 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 greater than 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

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.[16] A value of 50 for the number of nearest neighbours (k) was chosen for the imputation.[17]

4. Performance measures and statistical comparison

4.1. 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.[18] An AUC value closer to 1 suggests better discrimination ability between defaults and non-defaults; a value of 0.5 implies that the classifier

[16]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.

[17]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.

[18]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.

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.[19]

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.

4.2. 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

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

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 (Garcia and Herrera, 2008; Holm, 1979).

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

5. Results and discussion

5.1. 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. 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), 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 3.

Next, the best-performing technique, BRT, is compared with the three other classifiers. As shown in Table 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: 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	1.25
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	1.25
RF	0.7781	0.7527	0.7701	0.7814	2.75

Table 3: Statistical comparison of classifiers using H-measures

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

Table 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) 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 5 and 6 in Appendix 1).

In addition to comparing the previous performance metrics, it is of interest to see 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.[20] Two sets of representative plots are shown for portfolios 1 and 4, in Figure 1 and Figure 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 having been rescaled to reflect the original class priors.[21] 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.

[20] 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).

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

5.2. 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 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 3 and 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.[22] 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 (Brieman, 2001), in which there may be a large number of variables that each can contribute to the model predictions. With a

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

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.

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 also 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 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 also 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).

6. 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 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 context (Baesens et al. (2003)).

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 cost information, if available, into a boosting algorithm in a manner similar to Berk and Kriegler (2010) may lead to improved performance.

Finally, exploring how population drift 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.

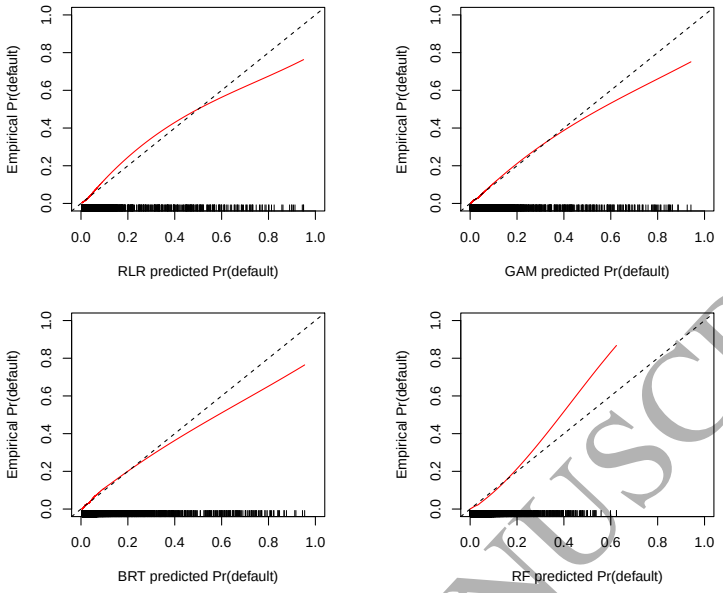

Figure 1: Calibration plots: portfolio 1

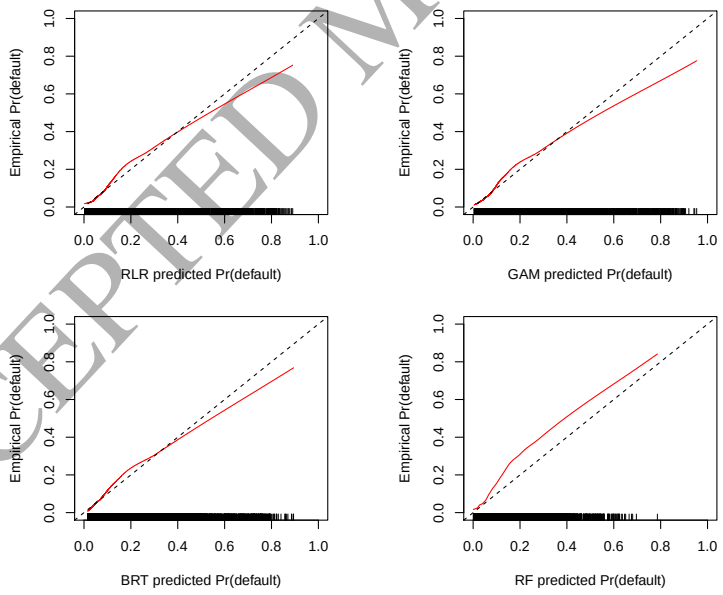

Figure 2: Calibration plots: portfolio 4

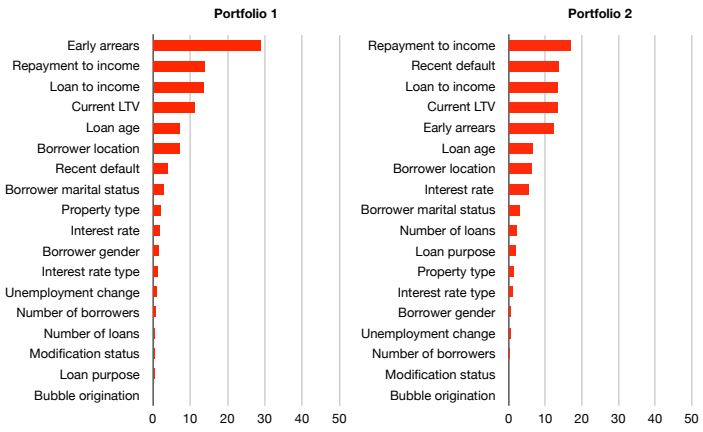

Figure 3: BRT variable importance plot: portfolios 1 and 2

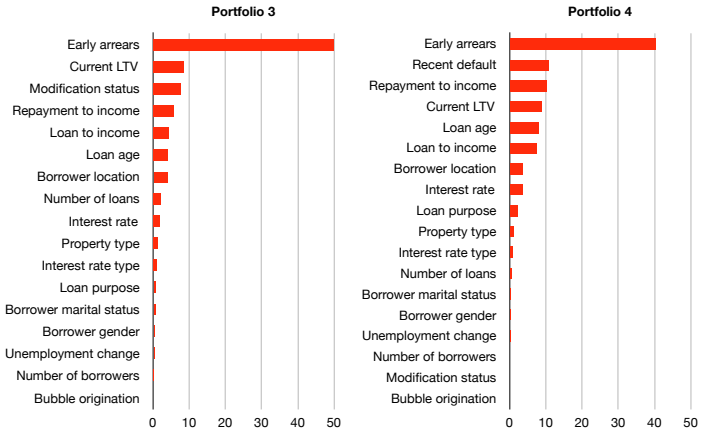

Figure 4: BRT variable importance plot: portfolios 3 and 4

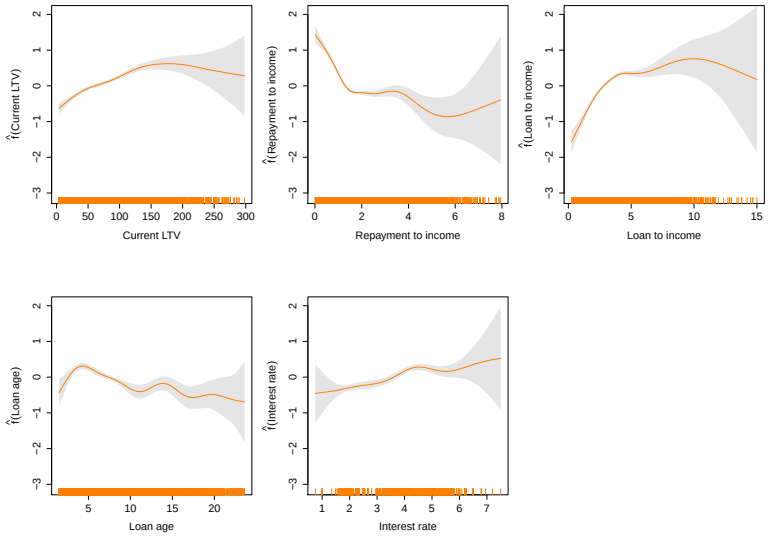

Figure 5: GAM estimated smooth functions for portfolio 4

References

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

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

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

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

Berg, D., 2007. Bankruptcy prediction by generalized additive models. Applied Stochastic Models in Business and Industry 23 (2), 129—143.

Berk, R., 2008. Statistical learning from a regression perspective, 1st Edition. Springer.

Berk, R., Kriegler, B., 2010. 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.

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

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

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

Brieman, L., Friedman, J., Stone, C. J., Olshen, R., 1984. Classification and Regression Trees, 1st Edition. Chapman and Hall/CRC.

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

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

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

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

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

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

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

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

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

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

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

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

Derksen, S., Keselman, H. J., 1992. 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.

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

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

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

Flach, P., Hernandez-Orallo, J., Ferri, C., 2011. 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.

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

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

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

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

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

Friedman, J. H., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33 (1), 1–22.
URL http://www.jstatsoft.org/v33/i01

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

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

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

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

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

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

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

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

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

Horton, N. J., Kleinman, K. P., 2007. 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.

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

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

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

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

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

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

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

Kruppa, J., Schwarz, A., Arminger, G., Ziegler, A., 2012. Consumer credit risk: individual probability estimates using machine learning. Expert Systems with Applications 40, 5125–5131.

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

Kuhn, M., Johnson, K., 2013. Applied Predictive Modelling, 1st Edition. Springer.

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

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

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

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

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

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

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

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

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

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

Ridgeway, G., 2013. gbm: Generalized Boosted Regression Models. R package version 2.1. URL http://CRAN.R-project.org/package=gbm

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

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

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

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

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

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

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

Wood, S., 2013. mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation. R package version 1.7-24.
URL http://CRAN.R-project.org/package=mgcv

Appendix 1: Classifier performance tests using AUC measure

This section contains results of testing for differences in performance across classifiers trained using the AUC and referred to in section 5.1 of the main text.

Table 5: Statistical comparison of classifiers using AUC measure

Test statistic	Value	p-value
Friedman	9.9	0.0194
Iman-Davenport	14.14	0.0000

Table 6: Holm's step down procedure for AUC; $\alpha = 0.05$ 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	3.0125	0.0026	0.0166
RF	1.6432	0.1003	0.025
GAM	0.8216	0.4113	0.05
10 % significance			
LR	3.0125	0.0026	0.0333
RF	1.6432	0.1003	0.05
GAM	0.8216	0.4113	0.1

Appendix 2: 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 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 7: 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 8: 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 9: 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