European Journal of Operational Research 249 (2016) 427–439

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

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

Trevor Fitzpatrick[a,b,1,*], Christophe Mues[a]

[a] Southampton Business School, University of Southampton, Highfield, Southampton SO171BJ, UK
[b] Central Bank of Ireland, PO Box 559, Dame Street, Dublin 2, Ireland

ARTICLE INFO

Article history:
Received 18 January 2014
Accepted 8 September 2015
Available online 16 September 2015

Keywords:
Boosting
Random forests
Semi-parametric models
Mortgages
Credit scoring

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.

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 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, Gestel, Viaene, M.Stepanova, Suykens, & Vanthienen, 2003; Kennedy, Nameea, & Delaney, 2013b; Lessmann, Seow, Baesens, & Thomas, 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.

* Corresponding author at: Southampton Business School, University of Southampton, Highfield, Southampton SO171BJ, UK. Tel.: +44 (0)23 8059 7677.
E-mail addresses: T.Fitzpatrick@soton.ac.uk (T. Fitzpatrick), C.Mues@soton.ac.uk (C. Mues).

[1] 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.

http://dx.doi.org/10.1016/j.ejor.2015.09.014

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, Edelman, and Thomas (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, Tibshirani, & Friedman, 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, Quigley, and Van Order (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 per cent. 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, Gerardi, & Willen, 2008; Haughwout, Peach, & Tracy, 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, Lou, & Gehrke, 2012; Elith, Leathwick, & Hastie, 2008; Hastie et al., 2009; Liu, Vu, & Cela, 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 & Crook, 2009) and bankruptcy prediction (Van Gestel, Baesens, & Martens, 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 quadratic programming problem is between $O(N^2)$ and $O(N^3)$, where N is the number of training observations (Bordes, Ertekin, Weston, & Bottou, 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, Vedaldi, Jawahar, & Zisserman, 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) and Kennedy, Namee, Delaney, O'Sullivan, and Watson (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, Khandani, and Kim (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.

[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, Baesens, Gestel, and Vanthienen (2007) demonstrate that it is possible to extract understandable rules that approximate an SVM classifier.