Expert Systems With Applications 000 (2018) 1–16

Contents lists available at ScienceDirect

Expert Systems With Applications

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

An experimental comparison of classification techniques in debt recoveries scoring: Evidence from South Africa's unsecured lending market

Jonah Mushava*, Michael Murray

School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Westville Campus, Private Bag X54001, Durban 4000, South Africa

ARTICLE INFO

Article history:
Received 31 May 2017
Revised 3 January 2018
Accepted 19 February 2018
Available online xxx

Keywords:
Behavioural scoring
Classification techniques
South Africa
Recoveries
Credit risk

ABSTRACT

In South Africa, almost 50% of the people who take loans cannot afford it. Previously, lenders were able to make deductions from a borrower's payslip but this practice is no longer allowed. Consequently, lenders are now far more vulnerable to default particularly if these loans are no longer being backed by any form of meaningful collateral. The aim of this study is to investigate the predictive power of some of the more popular classification techniques currently in use with specific attention to predicting the propensity for a borrower who is 90 days or more in arrears on an unsecured loan to pay over a fixed window period at least 30% of the total amount due. Results show that these classification techniques perform best for predicting payment patterns over a future horizon period between 3 and 12 months. It is also found that generalized additive models (especially using a generalized extreme value link function), which have not been extensively explored within the credit scoring literature, outperformed all the other classifiers considered in this study.

1. Introduction

In South Africa (SA), debt collection has become a multi-billion-rand industry (NCR, 2016a). Per annum based statistics produced by the country's National Credit Regulator (NCR) reveal that 40–48% of all borrowers are at least three months in arrears on their repayments, have an adverse listing against them (where the debt has been written off, a repossession has taken place, or a credit card has been revoked) or have a judgement and/or administrative order issued against them.

When a debt obligation is unsecured, the effect of a default on a lending institution obviously becomes far more pronounced. Previously, lenders were able to make deductions from a borrower's payslip but this practice is now prohibited (Manuel, 2000). Unsecured lending in South Africa alone accounts for close to 20% of the total book value of all credit currently on issue (ZAR1.67 trillion at end of September 2016, NCR, 2016b).

This paper investigates the predictive power of some of the more popular classification techniques currently in use for score carding with specific attention being given to predicting the propensity of a borrower who is 90 days or more in arrears on an unsecured loan to pay over a fixed window period at 30% of the total amount due. An account is deemed to be 90 days in arrears if no repayments have been made for the past 3 months on that account. A classification technique (or simply a classifier) is a procedure used to assign an individual associated with a set of input variables to a pre-defined finite set of groups (Hastie, Tibshirani, & Friedman, 2009).

The practice associated with modelling the repayment behaviour of an existing customer, known as behavioural scoring, needs to be distinguished from that of application scoring which focuses on a new loan application (Thomas, Ho, & Scherer, 2001). This paper focuses on behavioural scoring noting that the literature in this field is still very limited (Kennedy, Mac Namee, Delany, O'Sullivan, & Watson, 2013). This is attributed to a number of factors, chief among them being the commercial sensitivities associated with such datasets. Moreover, lenders often privilege application scoring because it helps identify "credit worthy" customers, which is in line with the core business in credit lending – to push through large volumes of good sales.

Typically papers compare the performance of various classifiers on a single set of data arriving at a best performing classifier for that particular set of data by subdividing the data into a set that can be used for estimation (called the training/learning sample) and another that can be used for cross-validation (Abdou, 2009; Akkoç, 2012; Kruppa, Schwarz, Arminger, & Ziegler, 2013; Kültür & Çağlayan, 2017; Yap, Ong, & Husain, 2011). As the cases in the

* Corresponding author.
E-mail address: 208525323@stu.ukzn.ac.za (J. Mushava).

https://doi.org/10.1016/j.eswa.2018.02.030
0957-4174/

training sample and validation sample are likely to be very similar in pattern, Hand (2006) observes that this practice tend to overestimate the predictive power of sophisticated methods. Large scale internationally benchmarked studies such as those conducted by Lessmann, Baesens, Seow, and Thomas (2015) have all been able to rely on publicly available datasets to compare the performance of their classifiers using repeated cross-validation techniques. A large data set was obtained from a major South Africa unsecured money lending institution and the classification methods were applied. Therefore, in supplement to the standard practice of randomly subdividing a dataset into training and validation sets, an out-of-time dataset that has been collected in a future time period is used to help judge how the scorecards will perform when deployed in a real-life situation.

In credit scoring the idea is to assign to each borrower a particular score that can then be used to separate a 'bad' borrower from a 'good' borrower. In this paper a particular borrower who is 90 days or more in arrears on an unsecured loan will be classified as being good if they are able to repay at least 30% of the total amount due over that period (outcome period). In particular the study looks at outcome periods ranging over 1, 3, 6, 12 and 18 months. Furthermore a distinction between accounts that are being collected in-house and accounts that are being collected through an external debt collector is made. Given that there are five outcome window periods and two collection methods, as such a total of ten analyses will be done for a given classifier.

Various classification techniques have been considered in literature. Abdou and Pointon (2011), Lahsasna, Ainon, and Teh (2010), Lessmann et al. (2015) and Louzada, Ara, and Fernandes (2016) are some comprehensive review of classification techniques for credit scoring. This study is especially interested in comparing scoring techniques that make use of the industry standard logistic regression (LR) and the classical discriminant analysis (LDA) with binned covariates to handle nonlinear effects in the data to techniques that handle such nonlinear effects automatically. To achieve this, extensions of LR and LDA which have not been extensively explored within the credit scoring literature are considered. These are generalized additive models (GAM) and flexible discriminant analysis (FDA). A relatively novel flexible binary generalized extreme value (BGEVA) model designed to improve the predictive power of LR for imbalanced datasets and three popular machine learning algorithms, namely; support vector machines (SVM), boosting (AdaBoost) and random forests (RF) are also included.

Overall, a total of 9 classification methods are each fitted to the 10 data sets that have been created by considering 5 outcome periods over two collection methods. The predictive power of all the techniques are evaluated using the area under the receiver operating characteristic curve (AUC) and the H-measure. The AUC (or $Gini\ Index = 2AUC - 1$) is widely used in the credit industry to measure the discriminatory power of a binary classifier ignoring class distributions or misclassification costs. However, Hand (2009) showed that the AUC is generated by averaging the misclassification cost over a cost ratio distribution that depends on the probability distributions generated by the respective classifier. Therefore when comparing different classifiers using the AUC/Gini index, the comparison is no longer 'like-for-like' as the misclassification cost distribution may vary from classifier to classifier. The H-measure is a coherent alternative to the AUC that uses a classifier independent misclassification cost distribution that follows the beta distribution with two parameters α and β that need to be decided (Hand & Anagnostopoulos, 2014).

There is a general lack/improper use of statistical inferences to make the conclusion on whether the observed performance differences of the classifiers are statistically significant (Lessmann et al., 2015). Therefore recommendations in Santafe, Inza, and Lozano (2015) are followed: (1) determine if the observed performance differences are statistically significant and (2) if the outcome is affirmative perform pairwise comparisons to identify were the statistically significant differences are.

Having found a classification technique that performs best in a South African context, this score carding technique can be used worldwide to help reduce the collection costs that will arise from trying to recover money from a defaulting borrower. Conditions under which the winning classification technique(s) perform best will be noted to help researchers and practitioners with similar scoring problems. An illustration of how the best performing model can be used within a debt collections department to aid in the decision-making process will be provided.

Having outlined this, Section 2 provides a literature review of related work on credit behavioural scoring. The ensuing section gives a brief overview of the classification methods considered in this study. Section 4 provides a detailed account of the data preparation process and experimental set-up. Section 5 reports and discusses the results followed by Section 6 which serves as the summation of the study.

2. Literature review

This section reviews related literature on credit behavioural scoring using classification techniques. Fig. 1 illustrates the set-up needed to construct a classification model for behavioural scoring.

A sample of borrowers is chosen so that as at a chosen observation point, one has information on their repayment behaviour (Thomas et al., 2001). The period before the observation point is referred to as the performance period. Characteristics used for behavioural scoring are measured in the performance period as at the observation date. The period after the observation date is known as the outcome period. The outcome point denotes the end of the outcome period. A class label is assigned to each borrower based on their repayment behaviour during the outcome period.

Sarlija, Bensic, and Zekic-Susac (2006) used a Croatian revolving credit data that span over a 1 year in 2004 using June 30 as a single arbitrary observation point in the behavioural scoring set-up. The 6 months before June 30 were used as the performance period and the following 6-month used to assign a binary class label. Caution is made against the practice of using a single observation point when developing a behavioural scorecard as it fails to account for variations in behaviour that may occur due to the time of the year. Logistic regression, three neural network algorithms (back-propagation, radial basis function and the probabilistic networks) and a Cox regression survival model were fitted to 80% of the data and the performance tested on the remaining 20%. Sarlija et al. (2006) used receiver operating characteristic (ROC) curves to show that the radial basis function neural network performs best on the test sample compared to the other techniques considered in the study.

As highlighted before, Hand (2006) cautions how the aforementioned practice of comparing the performance of multiple classifiers based on a cross-validation sample created by randomly diving the data into two tend to overestimate the predictive power of sophisticated techniques. This type of pitfall when benchmarking multiple classifiers is also observed in Zhang (2007) who fit the models to 70% of a behavioural scoring data set and tested the models on the remaining 30%. The study concluded that using a genetic algorithm (GA) for feature selection then fitting a support vector machine (SVM) outperforms other techniques considered in that study such as a back-propagation neural network, genetic programming and logistic regression. The use of out-of-time data sets and statistical significant tests can help judge how the model will perform when deployed in a real-life situation.

Bijak and Thomas (2012) investigated the effect of segmenting the data on the performance of a scorecard. The behavioural data

Please cite this article as: J. Mushava, M. Murray, An experimental comparison of classification techniques in debt recoveries scoring: Evidence from South Africa's unsecured lending market, Expert Systems With Applications (2018), https://doi.org/10.1016/j.eswa.2018.02.030