Journal of Banking & Finance 56 (2015) 72–85

Contents lists available at ScienceDirect

Journal of Banking & Finance

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

An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes

Stewart Jones [a,*], David Johnstone [a], Roy Wilson [b]

[a] The University of Sydney Business School, The University of Sydney, Building H69 Cnr Codrington and Rose St, Darlington, NSW 2006, Australia
[b] Principal Consultant – Data Science, HP Enterprise Services, Australia

ARTICLE INFO

Article history:
Received 9 May 2014
Accepted 4 February 2015
Available online 14 March 2015

JEL classification:
C1
M4

Keywords:
Credit ratings changes
Prediction
Binary classifiers
Statistical learning

ABSTRACT

In this study, we examine the predictive performance of a wide class of binary classifiers using a large sample of international credit ratings changes from the period 1983–2013. Using a number of financial, market, corporate governance, macro-economic and other indicators as explanatory variables, we compare classifiers ranging from conventional techniques (such as logit/probit and LDA) to fully nonlinear classifiers, including neural networks, support vector machines and more recent statistical learning techniques such as generalised boosting, AdaBoost and random forests. We find that the newer classifiers significantly outperform all other classifiers on both the cross sectional and longitudinal test samples; and prove remarkably robust to different data structures and assumptions. Simple linear classifiers such as logit/probit and LDA are found nonetheless to predict quite accurately on the test samples, in some cases performing comparably well to more flexible model structures. We conclude that simpler classifiers can be viable alternatives to more sophisticated approaches, particularly if interpretability is an important objective of the modelling exercise. We also suggest effective ways to enhance the predictive performance of many of the binary classifiers examined in this study.

1. Introduction and prior literature

There is an extensive literature in credit risk and corporate bankruptcy prediction (Duffie and Singleton, 2003; Jones and Hensher, 2008). Credit ratings research attempts to explain and predict how credit ratings are assigned by the issuer at a given time, based on observable covariates that affect the credit quality of firms (Duffie and Singleton, 2003). While techniques have evolved, much of the literature relies on conventional classifiers such as standard logit/probit models and linear discriminant analysis (LDA)[1] (see e.g. Altman et al., 1977; Blume et al., 1998; Duffie and Singleton, 2003; Altman and Rijken, 2004; Amato and Furfine, 2004; Nickell et al., 2000; Jorion et al., 2009).[2] To a lesser extent, statistical learning techniques including neural networks, support vector machines (SVMs) and tree structure classifiers, including recursive partitioning have been utilised (see Duffie and Singleton, 2003).[3]

Despite some developments,[4] the literature in credit risk and bankruptcy has not kept pace with developments in the theoretical

* Corresponding author. Tel.: +61 2 93517755.
E-mail addresses: stewart.jones@sydney.edu.au (S. Jones), david.johnstone@sydney.edu.au (D. Johnstone).

[1] The logit model appears to be the dominant classifier in credit ratings and related literatures. Our review of over 150 empirical studies indicates that the logit model appears (either as the primary classifier or as a comparator model) in 27% of studies, followed by (in percentage of studies): LDA (15.1%), neural networks (14.6%), SVMs (6.8%), probit models (6.3%), recursive partitioning (3.6%); with the remainder of studies providing an assortment of different approaches, including rough sets, hazard/duration models, genetic algorithms, ensemble techniques, unsupervised learning models and other methods.

[2] In contrast, studies which have modeled ratings-transition probabilities have relied more on cohort and duration models (see e.g. Lando and Skødeberg, 2002).

[3] This study is limited to binary classifiers. We acknowledge other literature that has examined corporate failure and credit rating models in multinomial settings (see e.g. Jones and Hensher, 2008 for a review of relevant literature). We test similar model structures, but only in a binary classification setting. Other studies have also used hazard model functions to predict corporate failure and other risk events (see e.g. Shumway, 2001). However, hazard models are not binary classifiers because they predict a *single* event/outcome as a function of time and other explanatory variables (see Kleinbaum and Klein, 2012).

[4] A small group of studies has attempted to examine a broader range of modelling approaches. For instance, Doumpos and Zopounidis (2007) use a sample of Greek credit defaults to compare the performance of a stacked generalization methodology with individual models such as LDA, logit, neural networks, classification trees and other techniques (see also Hu, 2008). Baesens et al. (2003) compared several statistical learning techniques and conventional models based on European consumer credit data.

http://dx.doi.org/10.1016/j.jbankfin.2015.02.006
0378-4266/

statistics literature. Empirical evidence from other fields of application suggests that more recent classifiers (such as generalised boosting models, AdaBoost and random forests) can clearly outperform conventional classifiers such as logit/probit, LDA and also neural networks (see Hastie et al., 2009; Schapire and Freund, 2012). Similarly, there has been relatively little attention in the field to evaluating the empirical performance and theoretical merits of alternative classification models. Studies have compared the performance of conventional classifiers, particularly logit, probit and LDA (see e.g. Jones and Hensher, 2004; Greene, 2008) but there has not been equal scrutiny of more recent classifiers, apart from some evaluation of neural networks and SVMs.[5] While neural networks and SVMs remain established statistical learning techniques in the literature, these classifiers have been superseded by arguably more powerful techniques, particularly generalised boosting and random forests[6] (Hastie et al., 2009; Schapire and Freund, 2012).

This study adds to the existing empirical research in at least two important ways. First, we identify and test a wide class of binary classifiers against a large sample of credit ratings changes. While not an exhaustive list, the 20 classifiers selected for this study are broadly representative of the most widely used and cited classifiers in the literature (see Hastie et al., 2009 for a review of modern applications). On one side of the spectrum we have relatively simple linear classifiers such as standard form logit, probit and LDA. These classifiers are common in the literature, but have limited capacity to model nonlinearity and unobserved heterogeneity in the dataset. However, they are generally more interpretable in terms of understanding the functional relationship between predictor variables and the response outcome. In the middle of the spectrum we have classifiers which are better equipped to handle nonlinearity and unobserved heterogeneity, including mixed model approaches (such as mixed logit), multivariate adaptive regression splines (MARS) and generalized additive models (GAM). The greater flexibility of these models usually translates into better fit and enhanced predictive performance, at the cost of lower interpretability. At the end of the spectrum we have fully general, nonlinear models that are designed to capture all nonlinear relationships and interactions in the dataset. These classifiers include neural networks, SVMs, generalised boosting models, AdaBoost, random forests and oblique random forests. While the complex algorithms underpinning many of these classifiers are designed to enhance classification accuracy they too pose major hurdles for interpretation, since the relationship between the predictor variables and response variable is largely hidden in the internal mathematics of the model system.

The benefit from using a more complex nonlinear classifier (such as neural networks or generalised boosting) should come from improved out-of-sample predictive success. Following Occam's Razor, if two classifiers have comparable predictive performance, a simpler and more interpretable classifier is preferred to a less interpretable classifier, particularly if statistical inference and not just prediction is a major goal of the modelling exercise (James et al., 2013). An important objective of this paper is to assess whether more complex classifiers do in fact lead to better out-of-sample prediction success, particularly compared to simpler more interpretable classifiers.

Second, we assess to what extent the predictive performance of different classifiers is impacted by the underlying shape and structure of input variables, and whether predictive performance can be enhanced by modifying these conditions. This issue appears to be an important but much neglected issue in the literature. We examine two common problems which can potentially impact the performance of classification models: non-normality of the input variables and missing values (see Hastie et al., 2009). The predictive performances of all classifiers are tested both with and without variable transformation (using the Box–Cox power transformation procedure), and missing value imputation (using the SVD procedure), including the combined effects of both techniques. Arguably, Occam's Razor holds when it comes to data intervention issues. If two classifiers predict well, and the interpretability level of each classifier is comparable, we prefer a classifier that performs well without requiring significant data intervention by the researcher (e.g. extensive variable transformation and/or missing value imputation). As many of the binary classifiers examined in this study are quite new to accounting and finance research, and credit research in particular, an empirical assessment of their characteristics and forecasting potential, particularly under different data assumptions and restrictions, provides an important motivation for this study.

The remainder of the paper is organised as follows. Section two discusses the sample, methodology and empirical models to be examined in this study. Section three presents the results, which is followed by concluding remarks and directions for future research.

2. Empirical context and methodology

2.1. Sample

We utilize the credit ratings data from Standard & Poor's *RatingsDirect* (accessed through the Standard's and Poor *Capital IQ* service). Corresponding annual financial, market and other input variables are extracted from a customized software application developed with the researchers by *Capital IQ* technical staff. Our international sample is based on 5053 long-term issuer credit ratings changes which occurred over a 30 year period (1983–2013). Of this sample, 2681 rating changes relate to public companies incorporated in the US, while 2372 ratings changes relate to public companies incorporated outside of the US.[7] We focus on rating changes (rather than initial ratings) to avoid potential staleness in the ratings data which can bias performance and limit the usefulness of empirical results (Amato and Furfine, 2004).[8] Moreover, ratings changes represent the dynamics of the ratings process that most influence capital markets.[9]

2.2. Input variables

We test the predictive performance of the binary classifiers using a set of financial indicators, market variables, corporate gov-

[5] While there are mixed findings in this literature, many studies indicate that neural networks do tend to outperform LDA on both training sets and test samples (see e.g. Tam and Kiang 1992; Zhang et al. 1999). However, the incremental improvement in predictive performance from neural networks is not always evident. Some empirical studies find that simple classifiers such as LDA are preferable to neural networks, particularly given interpretability issues associated with neural networks (see Altman et al. 1994).

[6] The potential of these classifiers has not been extensively explored in the credit risk and related literatures. Of the few studies that have examined these classifiers, early results seem promising. For example, based on a failure sample of 1365 private firms, Cortés et al. (2007) find that the generalised boosting model significantly improved test sample predictive accuracy by up to 28 percent (see also Kim and Kang 2012).

[7] The international sample includes public companies from: (1) Europe (n = 1038); (2) the Asia Pacific (n = 687); (3) Latin America (n = 299); (4) Canada (n = 199); and (5) the African and Middle East region (n = 149).

[8] For instance, it is unlikely that ratings agencies would monitor all rated firms on an on-going basis due to cost factors and resource constraints. With ratings changes we can be more confident that the ratings decision was based on a recent assessment of a company's performance and credit worthiness.

[9] There is substantial research showing that ratings changes convey value relevant information to bond and equity markets (see e.g. Hand et al. 1992; Dichev and Piotroski, 2001; Amato and Furfine 2004).