AN EMPIRICAL EVALUATION OF THE PERFORMANCE OF BINARY CLASSIFIERS IN THE PREDICTION OF CREDIT RATINGS CHANGES

Stewart Jones, David Johnstone and Haran Segram

Abstract

In this paper, we examine the predictive performance of a range of binary classifiers on US credit ratings over the period 1981-2006. We compare classifiers ranging from conventional logit/probit and linear discriminant analysis classifiers (LDA) to fully non-linear classifiers, including neural networks, support vector machines (SVMs) and 'new age' techniques such as generalised boosting and random forests. The out-of-sample predictive performance of all classifiers is tested on both randomized cross-sectional and longitudinal validation samples. Out-of-sample predictive performance is also tested using different variable transformation and missing value imputation assumptions. The results contribute to the literature in several ways: (1) the generalised boosting method (and related techniques such as AdaBoost and random forests) significantly outperformed all other classifiers on the cross-sectional and longitudinal validation samples; and proved remarkably robust to different data structures and assumptions; (2) variance stabilizing transformations significantly improved the out-of-sample predictive performance for all binary classifiers; while missing value imputation techniques made little difference to performance; (3) while simple classifiers such as logit, probit and LDA underperformed some of the sophisticated 'new age' classifiers such as generalised boosting, they performed as well if not slightly better than more traditional data mining techniques such as neural networks and SVMs; (4) conventional classifiers performed relatively strongly on the validation samples; suggesting that they still have an important role to play in this literature, particularly if *interpretability* is a major goal of the modelling exercise. Overall, this study identifies a range of new classifiers and their performance characteristics which may have significant potential for future research; as well as suggesting effective ways to enhance the predictive performance of many binary classifiers examined in this study.

Introduction

Over the past four decades, a sizeable literature has developed in the field of credit risk and corporate bankruptcy prediction (see Jones and Hensher, 2008 for a recent review). The most common binary classifiers utilised in this research are discrete choice models, such as logit, probit and linear discriminant models; and (to a lesser extent) statistical learning techniques such as neural networks, support vector machines and tree structure classifiers, such as recursive partitioning (see Duffie and Singleton, 2003).[1] Despite some innovative modelling developments, two observations are worth making about the current literature. First, the literature has not explored or kept abreast of many important developments in the statistical modelling literature which can potentially provide fruitful avenues for future research. For instance, empirical evidence from other discipline fields suggests that more recent 'new age' classification models (such as generalised boosting, AdaBoost and random forests) can significantly outperform more conventional classifiers, such as logit, probit, LDA and neural networks. Second, very little research has been devoted to evaluating the empirical performance, theoretical merits and characteristics of alternative classification models, even among the relatively narrow range of classifiers utilised in the credit risk and related literatures.

The literature evidences a variety of conventional modelling approaches and techniques, often applied to different sample sizes drawn from different jurisdictions, and using a range of explanatory variables frequently measured differently across studies. Model selection and evaluation can also vary widely across studies, as do model performance (Jones and Hensher, 2008). This makes it difficult to evaluate and compare the relative performance and contribution of alternative classifiers in a comprehensive, uniform and controlled manner. This study adds to the existing empirical research in several important ways. First, we explain the purpose, strength and limitations of a wide range of binary classifiers, ranging in their sophistication and

[1]This study is limited to binary classifiers. We acknowledge other literature which has examined corporate failure models in multinomial settings (see Jones and Hensher, 2004 for a review of this research). We test similar models, but only in a binary classification setting. Other studies have also use hazard model functions to predict corporate failure (see e.g., Shumway, 2001). However, hazard models are not binary classifiers (rather they predict a single event or outcome as a function time and other explanatory variables).

complexity; and highlighting major points of similarity and difference. A useful way to conceptualise the role and function of different binary classifiers is in terms of a trade-off between model flexibility and interpretability (see Section 3 and Appendix 2). On one side of the spectrum we have relatively simple or inflexible classifiers such as standard form logit, probit and LDA. These classifiers have limited capacity to model nonlinearity and unobserved heterogeneity in the dataset. While these models are quite inflexible, they are more interpretable in terms of understanding the functional relationship between predictor variables and the response outcome. Towards the middle of the spectrum we have classifiers which are better equipped to handle nonlinearity and unobserved heterogeneity in the data. Important examples of such models are mixed model approaches (such as mixed logit); multivariate adaptive regression splines (MARS) and general additive models (GAMs). The greater flexibility of these models usually translates into better model fits and enhanced predictive performance, however the interpretability of these models becomes ever more challenging. These classifiers are only partially nonlinear because their functional form is constrained by the additivity condition (see Appendix 2). Towards 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; support vector machines (SVMs); generalised boosting models (including variants such as AdaBoost); and random forests (see Section 3 and Appendix 2 for a detailed discussion). While the complex algorithms underpinning many of these classifiers are designed to enhance classification accuracy they can pose major hurdles for interpretation. For instance, neural networks are often described as the penultimate 'black box' (James et al., 2013), as the relationship between the predictor variables and response variable is largely indecipherable, being hidden in the internal mathematics of the model system.

The benefit from using a more complex nonlinear classifier (such as a neural network) should come from improved out-of-sample predictive success. It is taken as a given in the statistical modelling literature (following the Occam's razor principle) that if two classifiers have comparable predictive performance, a simpler more interpretable classifier should be preferred to a less interpretable classifier, particularly if statistical inference is an objective of the modelling exercise (see James et al., 2013). We believe the same principle holds when it comes to data handling issues. If two classifiers predict well, and the interpretability level of each classifier is

comparable, we would prefer a classifier that performs well without requiring significant data intervention by the analyst, such as extensive variable transformations and missing value imputation. 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. We also evaluate to what extent the predictive performance of different classifiers is affected by the underlying shape and structure of data, and whether predictive performance can be enhanced by modifying these conditions.

We empirically examine the classifiers discussed above in a controlled setting, using a large dataset of US credit ratings data covering a 25 year period (1981-2006). Many of the binary classifiers examined in this study have not been extensively tested or applied in previous research in this field, hence an empirical assessment of their characteristics and forecasting potential provides important motivation for this study.

Second, in order to compare the empirical performance of alternative binary classifiers, we provide a framework for model comparison. We use a common set of predictor variables used in previous research, and identical processes for data handling, as well as model selection and evaluation. The predictive performance of all classifiers is tested on eight permutations of the dataset, including: (1) out-of-sample predictive accuracy, using both randomized cross-sectional and longitudinal validation samples; (2) cross-sectional and longitudinal predictive performance both with and without variance stabilizing data transformations (ie Box Cox power transformations); (3) cross-sectional and longitudinal predictive performance both with and without missing value imputation (using single value decomposition or SVD). The predictive performance of all classifiers is compared with receiver operating characteristic (ROC) curves using the area under the curve (AUC) as the basis for statistical comparisons. Each classifier is trained on the full datasets using its own inbuilt optimal variable selection procedure to limit any intended or unintended biases introduced by the analyst, such as 'data snooping' (White, 2000). This is tantamount to the analyst trying to optimise the performance of different classifiers by manipulating variable selection with the benefit of hindsight. Furthermore, for all classifiers tested in this study we introduce some form of subset selection procedure, penalisation and hyper-parameter selection to limit possible model over fitting.

Third, it is expected that the empirical analysis provided in this study will lead to a number of new insights. For instance, we identify and explain the characteristics and theoretical merits of some promising new classifiers (such as generalised boosting and random forests) which have received little attention in previous research. Generalised boosting, AdaBoost and random forests appear to offer significantly better predictive performance relative to many conventional classifiers. Consistent with findings in other literatures (see Schapire and Freund, 2012), we also find that these classifiers not only predict very well, but appear resilient to model over-fitting and are relatively insensitive to the shape and structure of data (such as non-normalness and missing values).

This study also offers insight into how the out-of-sample predictive performance of many conventional classifiers (such as standard form logit, probit and LDA models) can be enhanced without compromising the attractive feature of interpretability germane to these classifiers. The results of this study may also point to more optimal modelling strategies in this field. For instance, it may not be optimal for researchers to use complex, less interpretable classifiers (such as neural networks and SVMs) if the out-of-sample predictive success is no better (and in some cases inferior) to simpler, more interpretable classifiers. The results of this study can also assist researchers find more optimal modelling strategies appropriate to the particular features of their datasets. For instance, our results indicate that the performance of some classifiers deteriorate in the presence of missing values and non-normalness in the data; whereas predictive performance for other classifiers actually improves. Our results could suggest modelling strategies that work best for different types of data structures and assumptions, and provides insight into how classification performance can potentially be improved as a result.

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

2. Prior Literature

There is now an extensive literature in credit risk (which embraces credit ratings research) and corporate bankruptcy prediction (Duffie and Singleton, 2003; Jones and Hensher, 2008). The key issue in credit ratings research is to explain and predict how credit ratings are assigned by the issuer at a given time, based on observable covariates that determine the credit quality of firms (Duffie and Singleton, 2003). While a variety of techniques have evolved over the years, much of the formal modelling literature in credit ratings prediction relies on conventional classifiers such as standard form probit/logit models and LDA[2] (examples include Altman, Haldeman, and Narayanan, 1977; Kaplan and Urwitz, 1979; Ederington Yawitz, 1987; Iskandar-Datta and Emery, 1994; Blume, Lim and MacKinlay, 1998; Duffie and Singleton, 2003; Altman and Rijken, 2004; Amato and Furfine, 2004; Nickell, Perraudin and Varotto, 2000; Jorion, Shi and Zhang, 2009).[3]

Comparatively little attention has been devoted to evaluating the performance of alternative classifiers. However, in recent years there have been significant advances in the broader statistical modelling literature, particularly with respect to statistical learning classifiers such as generalised boosting models and random forests; and in the discrete choice literature, most notably mixed models, such as mixed logit (see Hastie et al., 2009; Greene, 2007, 2008; Jones and Hensher, 2008). Even within the relatively narrow range of classifiers utilised in the credit ratings and related literatures, there have been comparatively few studies that have directly compared the performance characteristics and theoretical merits of alternative classifiers.

Some early empirical studies have compared LDA with systems such as recursive partitioning and linear probability modelling (see e.g., Collins, 1980; Frydman, Altman and Kao, 1985). There are also a variety of studies which have compared the performance of conventional classifiers, particularly logit, probit and LDA and to lesser extent quadratic discriminant analysis

[2]The logit model appears to be overwhelming the dominant classifier in credit ratings and related literatures. Our review of over 150 empirical studies indicates that the logit model appeared (either as the primary model or as a comparator model) in 27% of cases, followed by LDA (14.6% of cases), neural networks (14.6% of cases), SVMs (6.8% of cases), probit models (6.8% of cases), recursive partitioning (3.65%) of cases, with the remainder an assortment of models, including rough sets, hazard models, genetic algorithms, ensemble approaches, unsupervised learning models and other approaches.

[3]In contrast, studies which have modelled ratings-transition probabilities have relied more on cohort and duration models (see e.g., Carty and Fons, 1994; Wilson, 1997a, b; Behar and Nagpal, 2001; Kavvathas, 2001; Lando and Skodeberg, 2002).

(QDA) (see e.g., Hopwood, McKeown and Mutchler 1988; Lawrence and Bear; Lau, 1987; Lennox, 1999; Jones and Hensher, 2004; Barniv and McDonald, 1999; Greene, 2008). Some of these studies find evidence for the superior performance of the logit model, possibly because this classifier relies on less rigid statistical assumptions. However, more recent research has concluded that classifiers such as LDA are surprisingly robust to violation of the multivariate normality and IID assumptions; and that logit, probit and LDA classifiers often yield similar empirical results (Greene, 2008).

Several other studies have compared the performance of neural networks with conventional classifiers such as LDA and logistic regression. No doubt this reflected the early interest in LDA popularised by Altman (1968); as well as the advent of powerful new statistical learning algorithms, such as neural networks, which came into prominence during the 1980s and 1990s. The comparison between LDA and neural networks is particularly interesting given the strong theoretical links between the two classifiers.[4] Naturally, the empirical focus of these studies is to ascertain whether a fully flexible nonlinear classifier such as a neural network can outperform a simpler, more restrictive (but highly interpretable) classifier such as LDA. While there are some mixed findings in this literature, many of these studies indicate that neural networks do tend to outperform LDA on both training sets and validation samples (see e.g., Tam and Kiang; 1992; Coats and Fant, 1993; Wilson and Sharda, 1994; Jo, Han and Lee, 1997; Olmeda, 1997; Argawal, 1999; Zhang, Patuwo, Indro, 1999; Charitou, Neophytou, Charalambous, 2004; Wu et al., 2007; Sun and Li, 2008; Rafiei, Manzari, and Bostanian, 2011). However, the improvement in predictive performance is not always evident and some empirical studies find that simple classifiers such as LDA are preferable to neural networks for this reason, particularly given serious interpretability concerns surrounding neural networks (see Altman, Marco, and Varetto, 1994). These issues may have contributed to a declining interest in neural networks in recent years.

While neural networks remain an established statistical learning technique in the literature, to some extent this classifier has been superseded by newer and arguably more powerful techniques in recent years, such as generalised boosting, AdaBoost and random forests (Schapire and

[4]Neural networks have similar properties to a nonlinear discriminant analysis (see Hastie et al., 2009).

Freund, 2012). The potential of these 'new age' classifiers has not been explored extensively in the credit ratings and related literatures. Of the few studies that have examined these classifiers, the early results seem promising. For example, based on a failure sample of 1365 private firms, Martynez and Rubio (2007) find that the generalised boosting model improved validation sample predictive accuracy by up to 28%, which includes a significant reduction in type 1 errors (see also Kim and Kan, 2012).

A small group of studies have 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. Huysmans et al., (2006) examined the performance of self organizing maps (SOMs), multi-layer perception (MLP) and support vector machines (SVMs). Dimitras et al., (1999) compared rough sets with LDA on a small sample of Greek firms (a 40 firm matched sample); while Hu (2008) compared a multi-layer model, LDA, logit, probit and perceptron multi-layer (MLP) models on 65 failed firms sampled from the Moody's Industrial Manuals. Based on a matched pair of 50 failed and non-failed UK firms, Neophytou and Molinero (2004) find nonlinear techniques such as multidimensional scaling (MDS) tend to outperform LDA and logit classifers. Based on a sample of 1133 firms listed on the LSE, Lin and McClean, (2001) find that hybrid models (that combine and weight a number of models) tend to outperform LDA, logit, decision trees and neural networks but only when these models are used in isolation. Baesens et al., (2003) compared several statistical learning techniques and conventional models based on European consumer credit data and found that both the neural network and SVM classifiers demonstrated strong classification performance, but simpler classifiers such as LDA and logit also performed well.[5]

Against this background, there appears to be significant scope to develop and extend on the current literature, in particular: (1) most empirical studies are limited to a narrow range of classifiers, and comparative studies investigating the predictive performance of alternative

[5]Most of the classifiers explored in the above literature are included in our empirical analysis in some form. For instance, self-organising maps and multilayer perception models are types of artificial neural networks, which is a classifier tested in this study. Other techniques are forms of ensemble learning or tree structure models which share conceptual similarities with generalized boosting and random forests.

classifiers is sparse; (2) many empirical studies are based on small localised samples, and tend to adopt different approaches to data handling, variable measurement, model selection and evaluation. This can severely limit the generalizability of empirical findings across studies; (3) research has not kept abreast of a number of recent and potentially important developments in the statistical modelling literature, such as generalised boosting models, AdaBoost and random forests. Using a large sample of US credit ratings data from 1981-2006, this study evaluates the performance of a wide range of alternative classifiers using a consistent approach for model selection and evaluation.

3. Empirical Context and Methodology

This section describes the sample selection, methodology and classification models to be examined in this study.

Sample

The sample is based on 3813 firm years covering the period 1981 to 2006. Corporate credit ratings data (issuer rating) and financial statement variables are obtained from Standard and Poor's RatingsXpress and Center for Research in Security Prices (CRSP)/Compustat Merged databases respectively. These databases are accessed via Wharton Research Data Services (WRDS). The sample comprises all non-financial[6] public firms in the United States that have initial ratings and ratings changes. This approach was adopted to avoid possible staleness in the ratings data (see Amato and Furfine, 2004). For instance, it is highly unlikely that ratings agencies would monitor all rated firms on an on-going basis due to cost factors and availability of resources. With ratings changes (and initial ratings) we can be reasonably confident that the ratings decision was based on a recent assessment of a company's performance and credit worthiness. The notion that dynamic ratings (i.e. initial ratings and ratings changes) convey value

[6]Consistent with Carey and Hrycay (2001) and Altman and Rijken (2004), this study is limited to non-financial US firms. Non-financial firms have Standard Industrial Classification (SIC) codes other than 6000 to 6999.

relevant information (on bond prices) has been documented in prior studies (see e.g., Hand, Holthausen and Leftwich, 1992; Amato and Furfine, 2004).

Corporate ratings data are obtained from S&P's RatingXpress. First, we extracted credit ratings data with Compustat identifiers.[7] Next we eliminated duplicate entries, blanks and ratings prior to 1981. To simplify the empirical context for model comparisons, ratings changes are classified as a binary outcome dependent variable, where a ratings upgrade is coded '1'; and a ratings downgrade is coded '0'. Rating changes for a company is defined as a rating change from one major rating category to another major rating category (for example, if a company's rating changes from AA to AAA, or from BBB to BB). Using the Compustat identifiers the financial statement data is extracted for the corresponding initial rating and ratings change. The lag of financial statement variables is expected to be forty-five days for quarterly data.[8] The lagging of financial statement variables is consistent with prior literature. Since our study constructs the trailing twelve month financial information from quarterly data; forty-five days is considered a sufficient lag (see Altman and Rijken, 2006). The justification for constructing trailing twelve month financial information is as follows. It is reasonable to assume that ratings changes may not occur exactly three month after the fiscal year end. Therefore, rating agencies must have sufficient financial information to implement a firm's rating changes. For example, if a company's rating is changed from "AA" to "A" three weeks prior to the actual fiscal year end for this company, rating agencies would have constructed trailing 12 month information based on the previous four quarters. Hence, our income statement variables for the trailing 12 months are constructed from previous four quarter and balance sheet variables are obtained from the fourth quarter.[9,10] We believe this approach represents an improvement on the sampling methodology adopted by previous literature.[11]

[7]The Global Company Key (GVKEY) is a unique identifier that represents each company throughout Xpressfeed. All company data records are identified by a GVKEY. S&P's Compustat® Xpressfeed Understanding the Data (2007).

[8]The Securities and Exchange Commission (SEC) requires all registrants (accelerated filers have to file within forty days) to file the quarterly financial statements within forty-five days after the fiscal quarter end.

[9]S&P Corporate Ratings Criteria (2006) displays a set of 7 key ratios that are used in the rating process (p.43).

[10] If ratings change date > fiscal quarter end date (fqenddt) of Quarter 1, Quarter 2 and Quarter 3 by 45 days and <=135 days then financial variables are linked with corresponding ratings. If initial ratings or ratings change date > fqenddt of Quarter 4 by 90 days and <=180 then financial variables are linked with corresponding ratings.

[11]See Altman and Rijken (2006) and Amato and Furfine (2004).

Variable Selection

For the purposes of evaluating the performance of alternative binary classification models, we examine a range of conventional performance variables that have been widely tested in previous literature (see e.g., Altman, 1968; Altman, Haldeman, and Narayanan, 1977; Kaplan and Urwitz, 1979; Ohlson, 1980; Zmijewski, 1984; Ederington Yawitz, 1987; Iskandar-Datta and Emery, 1994; Blume, Lim and MacKinlay, 1998; Shumway, 2001; Altman and Rijken, 2004; Amato and Furfine, 2004; Jones and Hensher, 2004; Hensher, Jones and Greene, 2007; Jorion et al., 2009); including Standard and Poor's corporate ratings criteria (2006). These variables include: liquidity and solvency ratios (current ratio, the acid test ratio, interest cover, working capital to total assets); earnings and profitability measures (ROE, ROA, ROI, EBIT to total assets); cash flow performance (cash flow to total assets), firm age (number of years since a rating was issued); firm size (market capitalization and log of total assets), leverage (total debt to assets and total debt to equity) and activity (sales to total assets). Appendix 1 provides a definition of each indicator and the expected sign in the direct of rating changes (ie positive means that a higher value of the predictor variable is expected to be associated with an upward revision in ratings and vice versa). Consistent with previous literature, we expect ratings changes to be correlated with improving/deteriorating financial performance of the firm. For instance, companies with stronger liquidity and solvency indicators (such as working capital, cash flow, interest cover ratios) should be positively associated with upward revisions in ratings and vice versa. Firms with higher earnings and profitability, as measured by indicators such as ROA, ROE and return on capital employed should also be positively associated with upward revisions in ratings and vice versa.

In addition to financial variables, we also include market variables, firm size and age variables. The smaller the ratio, the higher the risk of insolvency and therefore the greater the likelihood of a ratings downgrade (Altman, 2002; Arora et al., 2005). Firm size is proxied by market value of equity. An alternate measure of firm size proxied by total assets is also included in the analysis (Blume et al., 1998; Altman and Rijken, 2004). Age is measured as the duration since a firm was first assigned a rating. Consistent with Altman and Rijken (2004), age variable is assigned a value of ten for those firms with values greater than ten and for firms that previously had ratings

at the beginning of the sample period in 1981. This age covariate is distinctly different from the "aging effect" studies (for e.g., Carty and Fons, 1994; Kavvathas (2001); Lando and Skodeberg (2002) which examined the duration dependence for different rating categories. Higher value for this variable is expected to lead to improved ratings (Altman and Rijken, 2004). Other variables are defined in Appendix 1.

Empirical Context

In order to provide a robust basis for model comparison, the predictive performance of all classifiers is tested on eight permutations of the dataset, including: (1) out-of-sample predictive accuracy, using both randomized cross-sectional and longitudinal holdout samples; (2) cross-sectional and longitudinal predictive performance both with and without variance stabilizing variable transformations; (3) cross-sectional and longitudinal predictive performance both with and without missing value imputation (using singular value decompositions).

Validation samples. For the cross sectional validation sample, fifty percent of the total sample is randomly allocated to the training data (estimation sample) and fifty percent is allocated randomly to the validation sample. For the cross sectional validation sample, no consideration is given to the temporal order of the data. For instance, it is possible to have a year 2006 observation in the estimation sample, and a 1981 observation in the validation sample. In order to compare the performance of classifiers in a more realistic forecasting setting, we also use a *longitudinal* validation sample. For the purposes of this study, the longitudinal estimation sample included all observations from the 1981 until 2000 inclusive. The longitudinal validation sample included all observations from 2001 to 2006 inclusive.

Predictive performance. We use ROC curves to test out-of-sample predictive accuracy on both the cross sectional and longitudinal validation samples. ROC curves are an established technique for comparing predictive performance across alternative classifiers (see e.g., Swets et al., 2000). For a binary classifier, the ROC curve plots the true positive rate (sensitivity) relative to the false positive rate (1 − specificity), as its discrimination threshold or cut-off score is varied (for the binary classifiers, this is the predicted probability of class membership). A random guess

describes a horizontal curve through the unit interval and has an area under the curve (AUC) of .5. As a minimum, classifiers are expected to perform better than random guessing, whereas an AUC score of 1 represents perfect classification accuracy. AUCs greater than .9 demark a very strong classifier, exhibiting an excellent balance between sensitivity and specificity across different probability thresholds; whereas AUCs between .8 and .9 are indicative of a good or useful classifier. For AUCs under .8, the performance of the classifier is regarded as fair, and does not display a particularly good balance between sensitivity and specificity over different probability thresholds.

Variable transformation. A major challenge in all modelling exercises, particularly when using parametric models, is dealing with variance stability issues in the data. Non-normalness in the data can arise from many possible data issues, such as the affects of outliers, skewness and kurtosis. These issues can significantly affect model estimation and predictive performance. We are interested in knowing which classifiers are more sensitive to variance stability issues, and whether appropriate variable transformations (to induce more normalness or 'better behaved' data) can actually improve out-of-sample predictive performance. It is common practice to transform data by taking the natural logarithm of a predictor variable (and occasionally the response variable) of interest. However, this approach is arbitrary and not always effective. Box and Cox (1964) suggested a more robust procedure which entails examining a family of possible transformations, and selecting the transformation that most normalises the data. The Box Cox transformation is formally defined as:

$$y'_\lambda = \frac{y^\lambda - 1}{\lambda}$$

Where y'_λ is the transformed data, y is the original data and λ is the lambda value indicating the power that the original data should be raised. The λ estimate is determined through maximum likelihood estimation. The Box-Cox transformation searches for values of λ between -5 and +5 until the best value for data normalization is found. The λ estimate is usually rounded to a whole number for ease of calculation. For instance, where λ = 0, the natural logarithm of the variable is taken. If λ = -2, the appropriate transformation reduces to $\frac{1}{y^2}$. While the Box Cox procedure is designed to find values of λ which minimises variances in the data, all variable transformations

were visually inspected to ensure the transformation did in fact improve the normalness of the data.[12]

Missing values imputation. Missing values are another common problem in most datasets used in accounting and finance research, including credit risk research. Many empirical studies deal with missing values through a simple process of case wise deletion, which can significantly reduce sample size and complicate model estimation if the missing data distorts the observed data in some way. This study compares the performance of classifiers using the conventional approach to missing values (case wise deletion) versus a more sophisticated and widely used missing value imputation technique known singular value decomposition (SVD) (see Strang, 1980). SVD assumes missing data is *missing at random* (MAR) or *missing completely at random* (MCAR).[13] SVD is an orthogonal linear transformation that optimally captures the underlying variance in the data.[14]

Model over-fitting. The tell tale signs of over fitting is where the classifier achieves excellent classification accuracy on the training/estimation sample, but poor accuracy on the validation sample. This is a fundamental property of all statistical models regardless of the dataset or the properties of a particular classifier (James et al., 2013). As the flexibility of the model increases, we expect to observe a monotone increase in the mean squared error of the estimation sample, and a monotone decrease in the accuracy of the validation sample (Hastie et al., 2009). Over fitting can also be an issue when the number of predictors is high relative to the sample size. This study uses a number of techniques to limit the effects of potential model over fitting. Every classifier utilises some form of subset selection, penalization or hyper parameter selection which prevents model over fitting (see Appendix 2 for more details). Some classifiers, such as generalised boosting and random forests have been shown to be resistant to model over fitting (see Appendix 2). More importantly, classification accuracy is tested on both cross sectional and

[12]Box Cox only works for positive numbers, but this is easily rectified by adding a constant to ensure all data was positive before transformation.
[13]Both MAR and MCAR assumes that the missing data mechanism does not cause our training data to give a distorted picture of the true population.
[14]SVD essentially performs a coordinate rotation that aligns the transformed axes with the directions of maximum variance in the data.

longitudinal validation samples. Strong out-of-sample predictive performance is probably the strongest indication that a classifier has not been over fitted on the estimation sample.

Model selection. In order to provide a valid and robust test of classifier performance, each classifier is tested on the full dataset and using all the variables available in Appendix 1. As stated previously, each classifier is trained on the full datasets using its own inbuilt optimal variable selection procedure. This also limits any (unintended) biases arising from the researcher interfering with model selection process (such as 'data snooping'). Importantly, this study views variable selection for all classifiers as a type of hyper-parameter selection (see Appendix 2). For example, with penalised likelihood methods we do not explicitly select input variables but we do select hyper-parameters that then shrink the coefficients for a subset of input variables to zero (this is a type of variable selection). For stepwise regression, we include all variables in the model but then select a subset of input variables by reducing the full model (with all input variables) in a stepwise fashion based on AIC or BIC. For the best subset regression methods, the classifier automatically selects a subset of inputs which is also a type of hyper-parameter selection. For neural networks, we include all input variables but then reduce the influence of many of them through the selection of weights for the hidden layer. For neural networks, this process is aided by including a weight decay penalty and a restriction on the number of hidden layers (both of these were chosen using a computationally intensive cross validation process). For generalised boosting, AdaBoost and random forests, this process is aided by selecting the optimal number of trees and the tree depth. Hence, we select hyper-parameters (using cross-validation) and these hyper-parameters are used by the algorithm to select the relative contribution of the input variables. With regards to GAMs and MARS models, they are both linear additive models. GAMs use basis function expansion to create a "richer" feature space (input variable space), this allows nonlinear effects of variables to be automatically incorporated in the model. MARS uses "hinge" functions to achieve the same result (see Appendix 2 for a full review of classification models).

Empirical Models

Appendix 2 describes the functional form of all classifiers examined in this study. One way to conceptualise the properties and characteristics of the different classifiers is in terms of the trade-off between flexibility and interpretability as is shown in Figure 1 below.

Insert Figure 1 about Here

Classifiers which are highest on the interpretability scale (*y*-axis of Figure 1) tend to be the most rigid or inflexible (see *x*-axis of Figure 1). For instance, highly linear models that are designed to accommodate a smaller range of explanatory variables (or reduce a large number of predictor variables to a small optimal set of predictors) are the most rigid on the scale, but are the most interpretable in terms of understanding the role and influence of explanatory variables on the response outcome (for instance, through parameter estimates and marginal effects). Two approaches for selecting subsets of predictor variables is *best subset* and *stepwise* procedures. With best subset selection, the classifier fits a separate least squares regression for each combination of *p* predictors (or maximum likelihood for binary classifiers). The classifier will fit all *p* models that contain exactly one predictor, then all models that contain exactly two predictors and so on. The final selection can be based on several criteria, but for the purposes of this study we use AIC and BIC which formally penalises model fits for over-parameterisation. Backward stepwise models work in a similar way, but are far less computationally intensive. Backward stepwise begins with a model containing all parameters, and then iteratively removes the least useful variables, one at a time until the optimal model is found based on AIC and BIC.

Penalised models or shrinkage methods (such as ridge regression and lasso) are an alternative to best subset classifiers. Rather than using least squares (or maximum likelihood for binary classifiers) to find a subset of variables, ridge regression uses all variables in the dataset but constrains or regularises the coefficient estimates or "shrinks" the coefficient estimates of unimportant variables to zero. Shrinking the parameter estimates can significantly reduce their variance while having little effect on the bias of the classifier. A weakness of ridge regression is that all variables are included in the model making the model difficult to interpret. The lasso has

a similar construction to ridge regression but the penalty or shrinkage term forces the parameter estimate of very unimportant variables to equal zero (hence the lasso has a variable selection feature and can produce parsimonious models). For this study, we use the *elastic net* technique (Zou and Hastie, 2005) which combines the strengths of both ridge regression and the lasso. The *elastic net* method allows very unimportant variable parameters to be shrunk to zero (a kind of subset selection), while variables with small importance will be shrunk to some small (non zero) value.

Insert Figure 1 about here

There next cluster of classifiers in Figure 1 are standard form logit and probit models, and linear disriminant models (LDA). The logit model is more flexible than both probit and LDA in terms of underlying statistical assumptions. The logistic classifer assumes IID on the error structure, but the explanatory variables are assumed to be distribution free. By contract, the probit model and LDA classifiers both assume multivariate normality for predictor variables and IID on the error structure. While these assumptions are certainly restrictive, they often do not adversely impact on the performance of binary classifiers, particularly if data is 'well behaved' (see Greene, 2008).[15]

Towards the middle of Figure 1, we have classifiers that can be thought of as "half way houses" between highly rigid model structures and fully general nonlinear classifiers. Three examples are multivariate adaptive regression splines (MARS); generalised additive models (GAMs); and mixed models (such as mixed logit/probit). A conventional way to extend linear regression functions to capture nonlinear relationships is to replace the linear model with a high degree polynomial function. MARS is more general (powerful) than this and works by dividing the range of X into R distinct regions (or knots). Within each region, a lower degree polynomial function can be fitted to the data and constrained so that they join to the region boundaries through knots. This can lead to better fits, more stable parameter estimates and frequently better out-of-sample prediction.

[15]However, differences in these models can become more apparent in multinomial settings (ie more than two outcomes), where a range of computational and estimation issues can lead to significant differences in model performance (Jones and Hensher, 2004).

As indicated in Appendix 2, mixed logit/probit models are also more general than standard logit/probit model. For instance, the mixed logit model completely relaxes the IID condition, allowing the error structure to be correlated across outcomes and with the underlying parameters of the model. The major difference between a standard logit and a mixed logit is that the standard model only contains fixed parameter estimates. With a mixed logit, the role and influence of explanatory variables can be described with up to four parameter estimates: fixed parameters; random parameters; heterogeneity in mean parameters; and heterogeneity in variance parameters. Previous research has shown that mixed logit models provide better fits and out-of-sample predictive success relative to standard models in multinomial contexts (Jones and Hensher, 2004).

GAMs were developed as a blend of generalized linear models and additive models. GAMs are more general than MARS, because this classifier can be estimated with any number of smoothed functions for each predictor variable; hence GAM models can automatically model many nonlinear relationships not captured in standard linear models. This provides further potential to improve model fits and predictive accuracy. A weakness of GAMs is that they can lead to over fitting if too many smoothing parameters are set by the analyst. GAMs are also computationally intensive and can be challenging to interpret. As pointed out in Appendix 2, MARS, mixed logit and GAM models are only *partially* nonlinear: ultimately these classifiers are all constrained by the additivity condition. While nonlinearity is introduced through the functions that can be fitted to predictors, the additivity condition imposes a strict linearity on the overall relationship between the predictors and the response variable; and this ultimately limits the extent to which nonlinear relationships and their interactions can be modelled.

Neural networks, support vector systems, generalised boosting models (and its main variant, AdaBoost) and random forests, can be characterised as *fully nonlinear* models – they tend to have maximum flexibility but usually this is achieved at the expense of interpretability. Neural networks are sometimes described as nonlinear discriminant models, representing essentially a two stage regression or classification model. For a typical single hidden layer model, there are any number of input variables (X), one hidden layer (Z) and two output classes the case of a

binary classifier (Y_k). Derived features Z_m are created from linear combinations of the inputs X, and then the target Y_k is modelled as function of the linear combinations of Z_m. Similar to other nonlinear approaches, the strength of neural networks is that they do not rely on any assumptions about the relationship between variable inputs and the response variable. Neural networks are uniquely designed to handle latent and highly complex nonlinear relationships in the data. The major limitation, particularly with backpropagational methods, is that they are the penultimate 'black box'. Apart from defining the general architecture of a network, the researcher has little other role to play once all the input variables are selected. Similar to other statistical learning techniques, neural networks provide no equations or coefficients defining a relationship (beyond its own internal mathematics). Relative to other fully nonlinear classifiers (particularly boosting models), neural networks have less capacity to handle large numbers of potentially irrelevant inputs and to handle data of mixed type (categorical and continuous). Computational intensity and scaleability (to very large numbers of observations and predictors) have also be seen as limitations of neural networks (see Appendix 2).

Support Vector Systems (SVS) differ from conventional classification techniques such as logit/probit in that they are non-probabilistic and strictly binary linear classifiers. SVSs are based on the concept of a separating *hyperplane.* A hyperplane divides *p*-dimensional space into two halves; where a good separation in the training set is achieved where the hyperplane that has the largest distance to the nearest training data point of any class. *Support Vector Machines* (SVM) enlarge the feature space of an SVS to deal with nonlinear decision boundaries – this is achieved by using various types of kernel functions. A widely kernel is the *radial kernel* which is used for the SVM classifier estimated in this study. Similar to neural networks, the major limitation of SVMs is lack of interpretability and lack of calibration of membership class. More recently, deep conceptual relationships between SVMs and the logit model have been demonstrated (James et al., 2013), which might lead us to suspect that the two classifiers might yield similar predictive performance. Similar to neural networks, a weakness of SVMs is that they have less capacity to handle large numbers of potentially irrelevant inputs and to handle data of mixed type (categorical and continuous). Computational intensity and scalability (to very large numbers of observations and predictors) are also known limitations with SVMs.

Generalised boosting (at its variant AdaBoost) is one of the most important developments in statistical learning in recent years and one of the most intensively researched (see Hastie et al., 2009). As stated by Schapire and Freund (2012): *"Boosting refers to a general and provably effective method of producing a very accurate prediction rule by combining rough and moderately inaccurate rules"* (p.4). See also Hastie et al (2009, pp.337-339, and Friedman et al., 2000). Many studies have found that generalised boosting is resistant to over fitting and yields impressive out-of-sample predictive accuracy (Schapire and Freund, 2012). The method is considered a major improvement on traditional tree based methods such as CART and recursive partitioning. The idea behind boosting is to combine the outputs of many weak classifiers to produce a powerful overall 'voting' committee. The weighted voting is based on the quality of the weak classifiers, and every additional weak classifier improves the prediction outcome. The weak learning algorithm is forced to focus on examples where the previous rules of thumbs provided *inaccurate* predictions. The intuition here is straight forward. The first classifier is trained on the data where all observations receive equal weights. Some observations will be misclassified by the first weak classifier. A second classifier is developed to focus on the trainings errors of the first classifier. The second classifier is trained on the same dataset, but misclassified samples receive a higher weighting while correctly classified observations receive less weight. The re-weighting occurs such that first classifier gives 50% error (random) on the new distribution. Iteratively, each new classifier focuses on ever more difficult samples. The algorithm keeps adding weak classifiers until some desired low error rate is achieved. As noted in Appendix 2, generalised boosting has a number of appealing features relative to neural networks and SVMs. For instance, generalized boosting has impressive computational scaleability (it can handle many thousands of predictors); has high capacity to deal with irrelevant inputs; tends to be better at handling data of mixed (continuous and categorical) type. Generalised boosting also scores a little better on interpretability. For instance, generalized boosting provides relative influence metrics and marginal effects which show which explanatory variables contributed most to overall model performance.

Random forests works in a similar way to generalised boosting (ie it works on a similar voting or committee approach) but works on the concept of de-correlated trees (see Breiman, 2001). As with the bagging technique, random forests build a number of trees based on bootstrapped

training samples. The intuition behind random forests is evident. In a bagged tree process, a particularly strong predictor in the dataset (along with some moderately strong predictors) will be used by most if not all the trees in the top split. Consequently, all the bagged trees will look quite similar to each other, hence the predictions from bagged trees will be highly correlated. A significant reduction in error can be achieved by averaging uncorrelated quantities as opposed to averaging many highly correlated quantities. Random forests overcome this problem by forcing each split to utilise only a small subset of predictors. The generalised boosting classifier differs from random forests in that it performs an exhaustive search for which trees to split on; while random forests choose a small subset. Boosting grows trees in sequence, with the next tree dependent on the last. Random forests grow trees in parallel independently of each others. Random forests also provides a methodology for measuring variable influence in the model (see Breiman, 2001); and shares many of the advantages as generalised boosting; including resilience to over fitting and computational scalability.

4. Empirical Results

This section outlines the descriptive statistics and empirical performance of alternative classifiers.

Descriptive Analysis: Ratings Changes

Table 1 describes the sampled firms by initial ratings and rating changes for the period 1981 to 2006. Amongst the investment grade companies (ie companies with a rating of at least BBB- or higher), only 0.5% (n = 20) of initial ratings or rating changes have been AAA over the sample period. The percentage of AA and A rated firms have decreased from 27% and 38% in 1980 to 0.9% and 13.9% respectively in 1993, largely due to the aftermath of early 1990s recession (July 1990 to March 1991). The early 2000s recession (March 2001 to November 2001) signalled the end of the decade long economic boom of the 1990s. As a result of the 2001 recession and the "dot-com bubble", AA and A rated firms bottomed out at 0% and 2.3% respectively in 2003 and and 1.2% (AA) and 2.5% (A) respectively at the end of the sample period in 2006. In contrast, a

sharp uptrend in ratings is observable with *speculative grade* firms (firms with a rating lower than BBB-). The percentage of BB and B rated firms increased to 39.8% and 24.1% respectively in 1993 (from 6.7% and 2.2% in 1980 respectively). This trend continued and peaked in 2006 with BBs and Bs rated firms at around 35.4% each. The percentage of default firms peaked at 15% in 2001 (i.e. the year of the 'dot com' recession). The time lags between rating changes and recession years can be construed as "through-the-cycle" rating methodology at work. In other words, rating agencies seem to react slowly during the economic downturns in adjusting the ratings except for the 'default' category. Another interpretation of above findings could be that rating agencies avoid rating reversals to minimize rating bounce (Loffler, 2002; Cantor and Mann, 2003b). In summary, the findings appear consistent with credit rating agencies' mandate to achieve a balance between stability and timeliness of ratings.

Insert Table 1 here

Ratings Downgrades and Upgrades

Table 2 documents the number of downgrades (negative changes) and upgrades (positive changes) for the sample. It is evident that downgrades (36.5% of the sample or 1393 observations in total) have outnumbered upgrades (17.4% or 664 observations in total) by two-to-one over the sample period 1981 and 2006. For 22 out of 27 sampled years, downgrades have been greater than upgrades reaching a peak of eight-to-one ratio in 2001. Years following the 2001 recession also evidenced significant numbers of downgrades relative to upgrades with five-to-one and three-to-one ratios in 2002 and 2003 respectively.

Insert Table 2 here

The total sample comprises 39.1% or 1429 speculative grade companies and 60.9% or 2224 investment grade companies. Table 3 displays the distribution of investment grade vs speculative grade firms over the sample period. We now turn to the empirical performance of alternative classifiers.

Insert Table 3 about here

Model Performance

Comparisons of predictive performance across classifiers are based on ROC curves, using the area under the curve (AUC) as the basis of statistical comparison. Figures 2 and 3 below display the box plots graphs of the AUCs for each classifier across all datasets. Figure 2 displays the AUC performance on the longitudinal validation sample and Figure 3 displays AUC performance on the cross sectional validation sample. For Figures 2 and 3 (and all Tables below) 'Original Data' represents the untransformed data with no missing value imputation (missing values are deleted case wise); and 'Original Imputed Data' is the original data but with missing values imputed using the SVD method. 'Transformed Data' represents the Box Cox transformed data but where no missing values are imputed (missing values are deleted case wise); while 'Transformed Imputed Data' is the Box Cox transformed with missing values imputed using SVD.

Insert Figures 2 & 3 about here

Table 4 below summarises (1) the average overall AUC performance across all classifiers and datasets; and (2) the average overall AUC performance of all classifiers across the longitudinal and cross sectional validation samples. Tables 5 and 6 provide a detailed breakdown of AUC performance across the individual datasets for both the longitudinal and cross sectional validation samples. Table 7 provides a breakdown of mean differences and significance levels in AUCs across classifiers, and over all datasets for both the longitudinal and cross sectional validation samples.

Insert Table 4 about here

The overall results displayed in Table 4 indicate that the 'new age' statistical learning classifiers such as generalised boosting, its major variant AdaBoost, and random forests have outperformed

all other models, both on the longitudinal and cross-sectional validation samples, and across all permutations of the datasets, both transformed and untransformed, and with and without missing value imputation. The standout model is generalised boosting, which has consistently outperformed all other models. Averaging the AUCs across the eight different versions of the dataset, generalised boosting is ranked first with an overall average AUC of .9469. This represents very strong classification accuracy and an excellent balance of sensitivity and specificity over the different probability cut-off regions of the ROC curve. Over all datasets, the AdaBoost classifier ranked second with an overall average AUC of .9433; while random forests is ranked third with an overall average AUC of .9297. Table 4 indicates that the next three highest performing classifiers are Logistic_GAM, Probit_GAM (general additive models); and mixed logit. The Probit_GAM model has an overall AUC of .9181; while Logistic_GAM has an overall AUC score of .9159; and the overall AUC for mixed logit is .9125. Other popular statistical learning techniques, such as neural networks (overall AUC of .9046) and support vector machines (overall AUC of .8973) performed quite strongly, but nevertheless significantly under-performed the generalised boosting and AdaBoost classifiers.

A somewhat surprising result is that highly inflexible but interpretable models such as probit stepwise (overall AUC of .8986) and logistic stepwise (overall AUC of .8961) performed strongly on the overall results. Based on the overall AUCs, LDA was the best performing of the simple or basic model structures with an overall AUC of .9002. The worst performing classifer in Table 4 is quadratic discriminant analysis (QDA), however even this classifier still scored a respectable AUC of .8797. It is interesting to observe that the logit and probit classifiers (both for restricted and flexible models forms such as GAMs and MARS) performed equally well on the validation samples, which is consistent with previous literature (Greene, 2008).

The results in Table 4 average out AUC performance over the cross sectional and longitudinal validation samples. However, a more realistic forecasting context for the classifiers is the longitudinal validation sample. In the longitudinal tests, the classifiers are predicting ahead of time which is a more robust test of the model's temporal validity (Jones and Hensher, 2004). Prediction models are developed to forecast unobserved events in the future. As might be expected, Table 4 indicates that most of the classifiers performed a little worse on the

longitudinal validation sample. Generalised boosting, AdaBoost, and random forests still outperformed all other models on the longitudinal sample, and their respective performance ranks have not changed. The next three best performing models on the longitudinal validation sample are Probit_GAM, mixed logit, and Logistic_GAM, in that order. Table 4 indicates that the worst performing models on the longitudinal validation sample are neural networks, SVMs, and the logit/probit MARS classifiers. Table 4 indicates that most of the classifiers have evidenced a deterioration in AUC performance of around 2% (compared to the cross section validation sample), but some classifiers have performed noticeably worse – particularly neural networks (the AUC is around 4.9% worse on the longitudinal sample, while the AUC for SVMs is around 4% worse compared to the cross sectional sample). Many of the simpler classifiers performed very well on the longitudinal validation sample with only a small deterioration in AUC performance compared to the results on the cross sectional validation sample. For example, LDA only performed 1.2% worse on the longitudinal sample, logistic stepwise performed 1.85% worse, while the probit stepwise classifier only performed 1.1% worse. The predictive performance of more complex classifiers, such as Probit_MARS and Logistic_MARS deteriorated by larger margins on the longitudinal sample (around 3.8%). The only anomalous result in Table 4 is the significant improvement of the Probit_Subset classifier which went from an AUC of .7745 (a poor or fair classifier) to an AUC of .8980 (an excellent classifier) on the longitudinal validation sample.

Tables 5 and 6 provides detailed ROC analysis breakdown across different permutations of the dataset. Table 7 summarises the mean AUC differences across models with significance levels (based on two tailed Z statistic).

Insert Tables 5, 6 and 7 about here

Table 5 displays the AUC breakdowns over the longitudinal validation sample, while Table 6 shows the AUC breakdowns for the cross sectional validation sample. Tables 5 and 6 also provide lower and upper bound confidence intervals for each AUC at the 95% level.

The ‘original data’ in Tables 5, 6 and 7 represents what a typical dataset might look like in accounting and finance. Missing values are typically not imputed using techniques such as SVD, but are deleted case wise. While there might be some variable transformation carried out by the researcher to normalise the data, use of such transformations tend to be restricted to a limited number of predictors and transformation types (natural logarithm being the most common). In many cases, outliers are removed/winsorised to assist normalisation of the data rather than directly transforming the variable of interest.

The results for ‘original data’ reported in Table 5 show that the three top performing models on the longitudinal validation sample are again generalised boosting (AUC = .9262), AdaBoost (AUC = .9243) and random forests (AUC = .9065), respectively. Table 7 Panel A shows the significance levels of mean AUC differences across classifiers for ‘original data’ (note that longitudinal results are shaded and above the diagonal while cross sectional results are below the diagonal). It can be seen from Table 7 Panel A that generalized boosting has statistically outperformed all other classifiers except AdaBoost (the mean AUC difference between generalized boosting and AdaBoost is only .002 and is not statistically significant). Table 7 indicates that random forests has statistically underperformed generalised boosting and Adaboost but has statistically outperformed most other classifiers.

The next two strongest classifiers reported in Table 5 are mixed logit (AUC=.8978) and Logistic_GAM (AUC=.8825). Table 7 indicates that on the ‘original data’ mixed logit has statistically outperformed a number of classifiers including all the standard form logit and probit models, both GAM and MARS, including LDA, and conventional data mining techniques such as SVM and neural networks. It is noteworthy that some of the more sophisticated statistical learning techniques such as neural networks and SVMs performed the worst on ‘original data’ (AUCs of .8573 and .8560 respectively). Table 5 indicates that these classifiers also have wider confidence intervals than the better performing models such as generalised boosting. In fact, the entire range of the AUC confidence interval for neural networks (.8271 to .8905) is outside the lower bound of the confidence interval for generalised boosting and AdaBoost, and only just within the lower bound confidence interval for random forests, suggesting it is an inferior classifier relative to these models. Table 5 indicates that the performance of SVM was slightly

worse than neural networks. By comparison, simple classifiers such as LDA, logistic stepwise, logistic subset, probit stepwise and probit subset, all performed better than neural networks and SVMs on the original data (Table 5 shows that the AUCs are .8596, .8722, .8733, .8818, .8814 respectively for these classifiers). The worst performing models in Table 5 are the Logistic_MARS and Probit_MARS models. These results are confirmed in Table 7 Panel A. Simple classifiers (such as LDA) performed quite strongly in statistical terms. For instance, while LDA was comprehensively outperformed by the 'new age' techniques, it performed on par with other sophisticated techniques such as SVM and neural networks, as well as the GAM and MARS classifiers. The same observation is true for other simple classifiers, such as logistic subset and probit subset. Table 7 Panel A indicates that neural networks and SVM tended to underperform several other classifiers, including several simple classifiers (for instance neural networks statistically underperformed probit stepwise, probit subset, logistic subset and the mixed logit classifiers; and did not significantly outperform any other classifier). SVM only performed slightly better than neural networks and also failed to statistically outperform any other classifier on the original data.

Once again, the results point to the relative strength of the new age classifiers, and the surprisingly robust predictive performance of the more basic conventional classifiers. While the AUCs on the original data are quite strong overall, they are noticeably weaker than the AUCs reported on the Box Cox transformed and missing value imputed dataset ('Transformed Imputed Data' in Tables 5, 6 and 7).

It can be seen from Table 5, that for most of the classifiers the out-of-sample predictive performance on the longitudinal sample has improved as a result of data transformation; and the AUC confidence intervals are generally narrower. Comparing the original data with transformed imputed data, it can be seen that the rankings of the three highest performing classifiers (generalised boosting, AdaBoost and random forests) remains unchanged, but their AUC performance has only marginally improved as a result of variable transformation and missing value imputation. Hence, the performance of these classifiers appears quite insensitive to the shape and structure of the data. Table 7 Panel D (transformed and imputed data) indicates that generalised boosting again statistically outperforms all other models except AdaBoost. AdaBoost

slightly underperforms generalised boosting relative to other classifiers but significantly outperforms all classifiers (except the GAM models).

Table 5 indicates that most of the simple classifiers have significantly improved their predictive performance on the 'transformed imputed data', particularly LDA (the AUC of LDA has improved by 6% or more), logistic penalised (AUC has improved by 4.3%), and logistic stepwise, where the AUC has improved by around 4.3%. Several other classifiers have improved their AUCs by around 3% or better. It is also noteworthy that many of the more sophisticated models, including neural networks, SVMs, MARS and GAMS models, evidenced improved AUCs by between 4% and 6% as a result of transformation. While the performance of all classifiers improved as a result of transformation, and the AUCs were generally strong across all classifiers, Table 7 Panel D shows that some of the more sophisticated classifiers such as neural networks, SVMs did not statistically outperform simple classifiers such as LDA, logistic stepwise and probit stepwise models. In fact, SVM underperformed some of the simple classifiers and neural networks only statistically outperformed one classifier (QDA) on the transformed imputed data.

A closer analysis of Table 5 indicates that the Box Cox power transformations have had a much stronger impact on improving predictive accuracy than missing value imputation. In fact, for 10 of the 17 models reported in Table 5, the forecasting accuracy slightly deteriorated on the 'original imputed data'. Only the top three performing models actually improved as a result of the combined effects of transformation *and* missing value imputation. The Logistic_MARS and Probit_MARS classifiers also significantly improved through the combined effects of Box Cox transformation and missing value imputation. However, for all other classifiers, predictive performance was enhanced more by the single effect of Box Cox transformation than by the combined effect of both Box Cox transformation and missing value imputation.

Cross sectional validation sample. Tables 6 and 7 (lower diagonal) displays classification performance on the cross sectional validation sample. As might be expected, nearly all classifiers performed better on the cross sectional sample compared to the longitudinal sample; and the AUC confidence intervals tended to be narrower (see Table 6). Similar to the

longitudinal sample, we find that the three top performing models on the cross sectional sample are generalised boosting, AdaBoost and random forests. On the original data, Table 7 Panel A indicates that generalised boosting statistically outperforms all other classifiers, followed by AdaBoost and random forests. Similar performance is displayed in Panels B, C and D of Table 7. The relative performance of neural networks and SVM has improved on the cross sectional original data, but overall these classifers are statistically inferior predictors relative to the 'new age' classifiers. On the original data, neural networks statistically outperformed only 4 of the 17 classifiers, while SVM only outperformed 2 classifiers; while significantly underperforming several others. However, neural networks showed some noticeable improvement on the cross sectional transformed imputed dataset.

Consistent with Table 5 results, the cross sectional results indicate that the top three classifiers are little changed by either the affects of Box Cox transformation, missing value imputation or the combined affect of both procedures. Table 6 also indicates that missing value imputation slightly lowers predictive accuracy for many classifiers on the cross sectional sample (relative to the original data). At best missing value imputation seems to have a negligible impact on overall predictive performance. Similar to Table 5 results, Box Cox transformation has had a positive impact on the predictive performance of most classifiers, particularly neural networks, SVM and many of the more basic classifiers such as LDA, probit subset (AUC improved dramatically by 31%), probit stepwise and QDA (also dramatically improved by 9.9% following transformation). Other than the mixed logit model (where the AUC has improved by around 2.6% after transformation) many of the other more sophisticated classifiers were not strongly affected by variable transformation on the cross sectional validation sample. One result from the cross sectional analysis that does differ from the longitudinal results is that the classifiers improved slightly more from the combined effects of variable transformation and missing value imputation. Generally, all the classifiers improved their performance after transformation and missing value imputation relative to Box Cox transformation alone. However, the improvement is modest. The best AUC improvement was only around 1.66% for the Probit_GAM model.

Conclusions and Directions for Future Research

In this paper we compare the predictive performance of 17 binary classifiers on US credit ratings data between 1981 and 2006. The classifiers are empirically tested on eight permutations of the dataset involving cross sectional and longitudinal validation samples; and comparing predictive performance across transformed and untransformed datasets, both with and without missing value imputation. A widely used technique to compare predictive performance across different classifiers is the receiver operating characteristic (ROC) curve which plots the true positive fraction (model sensitivity or accuracy) against the false positive fraction (model specificity) over different probability cut-offs or thresholds. We statistically compared areas under the curve (AUC) for each classifier's ROC curve. The ROC analysis indicates that 'new age' classifiers, generalised boosting, AdaBoost and random forests, statistically outperformed all other classifiers on the cross sectional and longitudinal validation samples, and on all permutations of the dataset. The results suggest that these classifiers may hold significant promise for future research and practice in this field. This conclusion is further galvanised by the many appealing statistical properties of these classifiers. While generalised boosting and related classifiers are fully nonlinear models (and therefore are among the most flexible model structures available), these classifiers afford some level of interpretability in terms of understanding the role and influence of predictor variables on the response outcome. For instance, these classifiers provide relative influence statistics, including marginal effects, which identify which explanatory variables and their magnitude have contributed to overall model performance. The results of this study also suggest that such classifiers may require minimal data intervention as their predictive performance appears largely immune to the shape and structure of data. Other benefits of these classifiers have been well documented in the literature and include scalability (they can handle many thousands of predictor variables), greater capacity to handle irrelevant inputs and greater ability to handle variables of mixed type (Hastie et al., 2009; Schapire and Freund, 2012).

A second finding is that the performance of all classifiers was improved, in some cases quite substantially, through Box Cox power transformations of predictor variables. For instance, the longitudinal AUCs of the LDA and neural network classifiers improved by over 7% as a result of the transformations. However, missing value imputation using the singular value decomposition (SVD) approach contributed little to the overall predictive performance of the classifiers. A third finding of this study is that quite simple classifiers (such as logit/probit stepwise/subset models

and LDA) performed quite strongly on the validation samples, and frequently did as well if not slightly better than more complex classifiers such as neural networks and SVMs. A well accepted principle in statistical modelling, following the Occam's razor principle, is that if two classifiers predict comparably well, the simpler more interpretable model should be preferred to a more complex and less interpretable classifier.

With respect to the more complex fully nonlinear classifiers examined in this study, generalised boosting, AdaBoost and random forests appear to hold the most promise for future research in terms of the trade-off between flexibility and interpretability. These classifiers not only seem to predict exceptionally well on holdout samples, but they have a number of appealing characteristics and properties that could make them highly amenable to credit risk research (for instance, generalised boosting is relatively straight forward to implement and the influence of predictor variables on model performance can be diagnosed). However, simple model structures performed surprisingly well in this study, and represent a viable alternative to more sophisticated approaches if statistical inference and interpretability is a major objective of the modelling exercise.

There are several possible directions for future research. First, we have compared the performance of binary classifiers with a major focus on out-of-sample predictive performance using ROC curves. It is possible to extend this study to compare model performance in multinomial settings. Previous research has shown that more flexible model structures (such as mixed logit) can outperform more restrictive model structures in multinomial settings (Jones and Hensher, 2004). It would be useful to see if the conclusions reached in this study hold in more complex predictive settings involving multiple outcome domains. Second, the study can be usefully extended to many other empirical contexts involving binary or multinomial outcomes, such as corporate bankruptcies and bond default prediction. In fact, the findings of this study can be potentially tested on a range of accounting and finance related research questions and problems involving a discrete dependent variable. Another direction for research is to combine the forecasts of sophisticated classifiers such as generalised boosting with expert opinions (ie where the researcher chooses predictor variables based on prior knowledge or theory) to assess whether predictive performance can be enhanced. There are also numerous possible ways to

examine the predictive performance of binary classifiers across different types of data structures and transformations. Our preliminary results indicate that more sophisticated variance stabilizing techniques, such as Box Cox power transformations, can have a significant impact on out-of-sample predictive performance of *all* classifiers, irrespective of their functional form or complexity. While we have used established techniques in this study, there are several other methods available for transforming data and imputing missing values that could be further investigated in future research.

References

Acharya, V. V., and M. Richardson, 2009, Restoring Financial Stability, John Wiley & Sons, Inc.

Akin, J. S., and D. K. Guilkey, and R. Sickles, 1979, A random coefficient Probit model with an application to a study of migration, Journal of Econometrics 11, 233-246

Allen, Linda, and Anthony Saunders, January 2003, A survey of cyclical effects in credit risk measurement models, Bank for International Settlements, Working paper, No. 126

Altman, E. I., September 2002, Corporate distress prediction models in a turbulent economic and Basel II environment, Risk Books, London

Altman, E. I., 2002, Bankruptcy, credit risk, and high yield junk bonds, Blackwell Publishers, MA, USA

Altman, E. I., 1998, The importance and subtlety of credit rating migration, Journal of Banking & Finance 22, 1231-1247

Altman, E. I., Brooks Brady, Andrea Resti, and Andrea Sironi, 2005, The link between default and recovery rates: Theory, empirical evidence, and implications, Journal of Business Vol 78, No. 6

Altman, E. I. and Duen Li Kao, June 1992, The implications of corporate bond ratings drift, Financial Analysts Journal

Altman, E. I. and Vellore M. Kishore, November/December 1996, Almost everything you wanted to know about recoveries on defaulted bonds, Financial Analysts Journal

Altman, E. I., and H.A. Rijken, 2004, The effects of rating through the cycle on rating stability, rating timeliness and default prediction performance, Working paper, New York University

Altman, E. I., and H.A. Rijken, 2005, Are outlooks and rating reviews capable to bridge the gap between the agencies through-the-cycle and short-term point-in-time perspectives?, Working paper, New York University

Altman, E. I., and H.A. Rijken, 2004, How rating agencies achieve rating stability, Journal of Banking & Finance 28, 2679-2714

Altman, E. I., and H.A. Rijken, March 2005, The impact of the rating agencies' through-the-cycle methodology on rating dynamics, Economic Notes by Banca dei Paschi di Siena SpA, Vol. 34, No. 2-2005, pp. 127-154

Altman, E. I., and H.A. Rijken, 2006, A Point-in-Time Perspective on Through-the Cycle Ratings, Financial Analysts Journal, CFA Institute, Vol. 62, No. 1, 54-70

Altman, E. I., and Anthony Saunders, 1998, Credit risk measurement: Developments over the last 20 years, Journal of Banking & Finance 21, 1721-1742

Amato, J. D., and C.H. Furfine, 2004, Are credit ratings procyclical? Journal of Banking & Finance 28, 2641-2677

Anderson, R. C., Sattar A. Mansi, and David M. Reeb, 2004, Board characteristics, accounting report integrity, and the cost of debt, Journal of Accounting and Economics 37 315-342

Apoteker, T., and Sylvia Barthelemy, Predicting financial crisis in emerging markets using a composite non-parametric data mining model, Working paper

Arora, N., Jeffrey R. Bohn, and Fanlin Zhu, 2005, Reduced form vs. structural models of credit risk: A case study of three models, Journal of Investment Management, Vol. 3, No. 4

Ashbaugh, H., Daniel W. Collins, and Ryan LaFond, December 2004, Corporate governance and the cost of equity capital, Working paper

Ashbaugh, H., Daniel W. Collins, and Ryan LaFond, June 2004, The effects of corporate governance on firms' credit rating, Working paper

Ashcraft, A., P. Goldsmith-Pinkham, and J. Vickery, May 2010, MBS ratings and the mortgage credit boom, Federal Reserve Bank of New York, Staff Report No. 449

Ashenfelter, O., P. B. Levine, and D. J. Zimmerman, 2003, Statistics and econometrics, methods and applications, John Wiley & Sons, Inc.

Atiya, Amir F., July 2001, Bankruptcy prediction for credit risk using neural networks: A survey and new results, IEEE Transactions on Neural Networks, Vol. 12, No. 4

Bahar, Reza, and K. Nagpal, March/June 2001, Dynamics of rating transition, Algo Research Quarterly, Vol. 4 Nos. ½

Baker, H. K., and Sattar A. Mansi, November/December 2002, Assessing credit rating agencies by bond issuers and institutional investors, Journal of Business Finance and Accounting, 29(9) & (10), 0306-686X

Balcaen, S., and Hubert Ooghe, 2006, 35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems, The British Accounting Review 38 63-93

Bangia, A., F.X. Diebold, A. Kronimus, C. Schagen, and T. Schuermann, 2002, Ratings migration and the business cycle, with application to credit portfolio stress testing, Journal of Banking & Finance 26, 445-474

Barney, D. K., and Janardhanan A. Alse, 2001, Predict LDC debt rescheduling: performance evaluation of OLS, Logit, and neural networks models, Journal of Forecasting, 20, 603-615

Barniv, R., and James B. McDonald, 1999, Review of categorical models for classification issues in accounting and finance, Review of Quantitative Finance and Accounting, 13 39-62

Basel Committee on Banking Supervision, June 2004, International convergence of capital measurement and capital standards

Becker, W. E., and William H. Greene, 2004, Using the Nobel Laureates in economics to teach quantitative methods, Becker/Watts books 6-23

Becker, B., and T. Milbourn, 2010, How did increased competition affect credit ratings?, Harvard Business School, Working paper 09-051

Benos, A., and George Papanastasopoulos, June 2005, Extending the Merton model: A hybrid approach to assessing credit quality, Working paper

Bernanke, B. S., Mark Gertler, and Simon Gilchrist, 1999, The financial accelerator in a quantitative business cycle framework, Handbook for Macroeconomics, Vol. 1

Bhojraj S., and Partha Sengupta, 2003, Effects of corporate governance on bond rates and yields: The role of institutional investors and outside directors, Journal of Business, Vol 76, No. 3

Bissoondoyal-Bheenick, E., 2004, Determinants and impact of credit ratings: Australian evidence, Working paper, RMIT University, Melbourne, Australia

Blume, M. E., F. Lim, and A.C. Mackinlay, 1998, The Declining Credit Quality of U.S. Corporate Debt: Myth or Reality? The Journal of Finance Vol.53 No. 4, 1389-1413

Blochlinger, A., and Markus Leippold, 2006, Economic benefit of powerful credit scoring, Journal of Banking & Finance, 30 851-873

Bolton, P., X. Freixas, and J. Shapiro, April 2010, The credit ratings game, Working paper, Columbia Business School.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society*, Series B, *26*, 211-252.

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

Camanho, N., P. Deb, and Z. Liu, March 2010, Credit rating and competition, Working paper

Cantor, R., 2001, Moody's investors service response to the consultative paper issued by the Basel Committee on Bank Supervision "A new capital adequacy framework", Journal of Banking & Finance, 25 171-185

Cantor, R., 2004, An introduction to recent research on credit ratings, Journal of Banking & Finance, 28 2565-2573

Caouette, J. B., E. I. Altman, P. Narayanan, and R. W. J. Nimmo, 2008, Managing credit risk, The great challenge for global financial markets, 2nd Edition, Wiley Frontiers in the Finance

Caouette, J. B., E. I. Altman, and P. Narayanan, 1998, Managing credit risk, The next great financial challenge, Wiley Frontiers in the Finance

Capuano, C., Jorge Chan-Lau, Giancarlo Garsha, Carlos Medeiros, Andre Santos, and Marcos Souto, 2009, Recent advances in credit risk modelling, IMF Working paper, WP/09/162

Carey, M., and Mark Hrycay, 2001, Parameterizing credit risk models with rating data, Journal of Banking & Finance, 25 197-270

Carrasco, J. A., and J. D. Ortuzar, 2002, Review and assessment of the nested Logit model, Transport Review, Vol. 22, No. 2, 197-218

Carty, L. V., July 1997, Moody's rating migration and credit quality correlation, 1920-1996, Moody's Investors Service, Report No. 25097

Catarineu-Rabell, E., Patricia Jackson, and Dimitrios P. Tsomocos, 2005, Procyclicality and the new Basel Accord – banks' choice of loan rating system, Economic Theory, 26, 537-557

Chen, Z., A.A. Lookman, N. Schurhoff, and D. J. Seppi, Dec 2010, Why ratings matter: Evidence from the Lehman Brothers' index rating redefinition, Working paper, University of Neuchatel

Cheung, S., 1996, Provincial credit ratings in Canada: An ordered Probit analysis, Working paper 96-6

Christiansen, J. H. E., Ernst Hansen, and David Lando, 2004, Confidence sets for continuous-time rating transition probabilities, Journal of Banking & Finance, 28 2575-2606

Coleman, M. S., and Stephen L'Heureux, 2003, Simulating conditional ratings transition matrices for credit risk analysis, Chatham Research Alliance, Arlington, MA, USA

Creighton, A., April 2004, Credit ratings and market dynamics, Reserve Bank of Australia Bulletin

Cremers, K. J. M., V. B. Nair, and C. Wei, October 2004, The congruence of shareholder and bondholder governance, Working paper

Crosbie, P., and Jeff Bohn, December 2003, Modelling default risk, Moody's K.M.V

Crouhy, M., D. Galai, and R. Mark, 2000, A comparative analysis of current credit risk models, Journal of Banking & Finance, 24 59-117

Crouhy, M., D. Galai, and R. Mark, 2000, A comparative analysis of current credit risk models, Journal of Banking & Finance, 25 47-95

Coval, J., J. Jurek, and E. Stafford, 2009, The economics of structured finance, Journal of Economic perspectives Vol. 23, No. 1

Cunha, F., J. J. Heckman, and S. Navarro, The identification and economic content of ordered choice models with stochastic thresholds, International Economic Review, Vol. 48, No. 4

Czarnitzki, D., and Kornelius Kraft, 2007, Are credit ratings valuable information?, Centre for European Economic Research, Discussion paper No. 04-07

Damodaran, A., 2002, Investment evaluation, Tools and techniques for determining the value of *any* asset, 2nd edition, Wiley Finance Series

Daniels, K. N., and Malene Shin Jensen, Dec 2005, The effect of credit ratings on credit default swap spreads and credit spreads, Working paper

Das, S. R, R. Fan and G. Geng, Dec 2002, Bayesian migration in credit ratings based on probabilities of default, The Journal of Fixed Income

Delianedis, G., and R. Geske, Feb 2003, Credit risk and risk neural default probabilities: Information about rating migrations and defaults, The Anderson School, UCLA

Dimson, E., 1979, Risk measurement when shares are subject to infrequent trading, Journal of Financial Economics, 7 197-226

Doumpos, M. and F. Pasiouras, 2005, Developing and Testing Models for Replicating Credit Ratings: A Multicriteria Approach, Computational Economics, 25, 327-341

Du, Y., 2003, Predicting Credit Rating and Credit Rating Changes: A New Approach, Queen's School of Business, March 2003 Draft

Du, Y. and W. Suo, 2004, Assessing Credit Quality from Equity Markets: Can Structural Approach Forecast Credit Ratings?. RBC Financial Group and Queen's University

Duffie, D., and K. Wang, 2004, Multi-period corporate failure prediction with stochastic covariates, Stanford University

Duffie, D., and K. J. Singleton, 2004, Credit risk, Pricing, Measurements, and Management, Princeton University Press

Elizalde, A., Dec 2005, Do we need to worry about credit risk correlation?, CEMFI and UPNA, Madrid, Spain

Elton, E., M. J. Gruber, D. Agrawal, C. Mann, 2004, Factors affecting the valuation of corporate bonds, Journal of Banking & Finance, 28 2747-2767

Eluru, N., C. R. Bhat, and D. A. Hensher, 2008, A mixed generalized ordered response for examining pedestrian and bicyclist injury severity level in traffic crashes, Accident Analysis and Prevention, 40 1033-1054

Farnsworth, H., and T. Li, Sept 2007, The dynamics of credit spreads and ratings migrations, Journal of Financial and Quantitative Analysis, Vol. 42, No. 3 pp 595-620

Figlewski, S., H. Frydman, and W. Liang, 2008, Modelling the effect of macroeconomic factors on corporate default and credit rating transitions, New York University

Frino, A., and S. Jones, Sept/Oct 2005, The impact of mandated cash flow disclosure on bid-ask spreads, Journal of Business Finance and Accounting, 32(7) & (8), 0306-686X

Frino, A., S. Jones, and W. J. Boon, Market behaviour around bankruptcy announcements: evidence from the Australian stock exchange, University of Sydney

Fulghieri, P., G. Strobl, H. Xia, Nov 2010, The economics of unsolicited credit ratings, Preliminary draft

Gentry, J. A., D. T. Whitford, and P. Newbold, Aug 1988, Predicting industrial bond ratings with a PROBIT model and funds flow components, The Financial Review, Vol. 23, No. 3

Gonzalez, F., F. Haas, R. Johannes, M. Persson, L. Toledo, R. Violi, M. Wieland, and C. Zins, June 2004, Market dynamics associated with credit ratings, European Central Bank, No. 16

Gordy, M. B., 2000, A comparative anatomy of credit risk models, Journal of Banking and Finance, 24, 119-149

Gordy, M. B., and E. Heitfield, June 2004, Of Moody's and Merton: a structural model of bond rating transitions, Working paper

Gordy, M. B., and B. Howells, Dec 2003, Procyclicality in Basel II: Can we treat the disease without killing the patient?, Journal of Financial Intermediation

Greene, W. H., Feb 2011, Econometric analysis, 7th edition, Prentice Hall

Greene, W. H., and D. A. Hensher, 2003, A latent class model for discrete choice analysis: contrasts with mixed Logit, Transportation Research, Part B 37, 681-698

Greene, W. H., D. A. Hensher, and J. Rose, 2006, Accounting for heterogeneity in the variance of unobserved effects in mixed Logit models, Transportation Research, Part B 40, 75-92

Greene, W. H., and D. A. Hensher, 2010, Does scale heterogeneity across individuals matter? An empirical assessment of alternative Logit models, Transportation, 37:413-428

Greene, W. H., and D. A. Hensher, Sept 2010, Ordered choices and heterogeneity in attribute processing, Journal of Transport Economics and Policy, Vol. 44, Part 3, pp. 331-364

Greene, W. H., and D. A. Hensher, 2010, Modelling ordered choices, A primer, Cambridge University Press, Cambridge, United Kingdom

Greene, W. H., M. N. Harris, B. Hollingsworth, and P. Maitra, April 2008, A bivariate latent class correlated generalized ordered Probit model with an application to modelling observed obesity levels, Working paper, Monash University

Griffin, J. M., and D. Y. Tang, Oct 2010, Did subjectivity play a role in CDO credit ratings?, McCombs Research Paper Series No. FIN-04-10

Gujarati, D. N., 2003, Basic econometrics, 4th edition, McGraw-Hill Companies

Hamilton, D. T., and R. Cantor, Sept 2004, Rating transition and default rates conditioned on outlooks, Moody's Investors Services, New York City

Hanson, S, and T. Schuermann, 2006, Confidence intervals for probabilities of default, Journal of Banking & Finance, 30 2281-2301

Hanson, S. G., M. H. Pesaran, and T. Schuermann, 2008, Firm heterogeneity and credit risk diversification, Journal of Empirical Finance, 15 583-612

Harris, R. S., J. F. Stewart, D. K. Guilkey, and W. T. Carleton, Characteristics of acquired firms: Fixed and random coefficients Probit analyses, University of North Caroline, USA

He, J., J. Qian, and P. E. Strahan, Dec 2010, Credit ratings and the evolution of the mortgage-backed securities market, Working paper, University of Georgia

Hensher, D. A., 2001, The valuation of commuter travel time savings for car drives: evaluating alternative model specifications, Transportation 28: 101-118, 2001

Hensher, D. A., 2006, The signs of the times: Imposing a globally signed condition on willingness to pay distributions, University of Sydney, Australia

Hensher, D. A., and W. H. Greene, 2002, Specification and estimation of the nested Logit model: alternative normalisations, Transportation Research Part B 36 1-17

Hensher, D. A., and S. Jones, 2007, Forecasting corporate bankruptcy: optimizing the performance of the mixed Logit model, Abacus Vol. 43, No. 3, pp. 241-264

Hensher, D. A., S. Jones, and W. Greene, 2007, An Error Component Logit Analysis of Corporate Bankruptcy and Insolvency Risk in Australia, The Economic Record, vol.83:260, pp. 86-103

Hensher, D. A., S. Jones, and W. H. Greene, March 2007, An error component Logit analysis of corporate bankruptcy and insolvency, The Economic Record Vol. 83, No. 260

Hensher, D. A., and S. Jones, 2008, Mixed logit and error component models of corporate insolvency and bankruptcy risk in Advances in Credit Risk Modelling and Corporate Bankruptcy Prediction, ed. S Jones and D.A Hensher, Cambridge University Press, Cambridge, United Kingdom, pp. 44-79

Hensher, D. A., J. M. Rose, and W. H. Greene, 2005, Applied choice analysis, A Primer, Cambridge University Press, Cambridge, United Kingdom

Hensher, D. A., and J. M. Rose, 2005, Respondent behavior in discrete choice modelling with a focus on the valuation of travel time savings, Journal of Transportation and Statistics, V8, N2

Hill, R. C., W. E. Griffiths, and G. G. Judge, 2001, Undergraduate econometrics, 2nd edition, John Wiley & Sons, Inc.

Hilsche, J., and M. Wilson, Aug 2010, Credit ratings and credit risk, Working paper, Brandeis, MA, USA

Horrigan, J. O., The determination of long term credit standing with financial ratios, University of Notre Dame, Working paper

Hu, Y., and J. Ansell, 2005, Developing financial distress prediction models, a study of US, Europe and Japan retail performance, Working paper

Huang, Z., H. Chen, C.J. Hsu, W.H. Chen and S. Wu, 2004, Credit rating analysis with support vector machines and neural networks: a market comparative study, Decision Support Systems, 37, 543-558

Hull, J., M. Predescu, and A.White, 2004, The relationship between credit default swap spreads, bond yields, and credit rating announcements, Journal of Banking & Finance, 28 2789-2811

Hunter, W., and S. D. Smith, 2002, Risk management in the global economy: A review essay, Journal of Banking & Finance, 26 205-221

Hurd, T., and A. Kuznetsov, Affine Markov chain models of multiform credit migration, Working paper

Hsiao, C., 2004, Analysis of panel data, 2nd edition, Econometric Society Monographs

Jackson, P., and W. Perraudin, 2000, Regulatory implications of credit risk modelling, Journal of Banking & Finance, 24 1-14

Jafry, Y., and T. Schuermann, 2004, Measurements, estimation and comparison of credit migration matrices, Journal of Banking & Finance, 28 2603-2639

Jarrow, R. A., and S. M. Turnbull, 2000, The intersection of market and credit risk, Journal of Banking & Finance, 24 271-299

Jones, S., and Hensher D. A., 2004, Predicting firm financial distress: A mixed logit model, The Accounting Review, vol.79:4, pp. 1011-1038

Jones, S., and Hensher D. A., 2007, Evaluating the behavioural performance of alternative logit models: An Application to Corporate Takeovers Research, Journal of Business Finance and Accounting, vol.34:7, pp. 1193-1220

Jones, S., and Hensher D. A., 2007, Modelling corporate failure: A multinomial nested logit analysis for unordered outcomes, British Accounting Review, vol.39:1, pp. 89-107

Jones, S., and Hensher D.A. 2008, An evaluation of open- and closed-form distress prediction models: The nested logit and latent class models in Advances in Credit Risk Modelling and Corporate Bankruptcy Prediction, ed. S Jones and D.A Hensher, Cambridge University Press, Cambridge, United Kingdom, pp. 80-113

Jones, S., and M. Peat, 2008, Credit derivatives: Current practices and controversies in Advances in Credit Risk Modelling and Corporate Bankruptcy Prediction, ed. S Jones and D.A Hensher, Cambridge University Press, Cambridge, United Kingdom, pp. 207-241

Jorion, P. and G. Zhang, 2006, Information effects of bond rating changes: The role of the rating prior to the announcement, University of California at Irvine

Jorion, P., C. Shi, and S. Zhang, 2009, Tightening credit standards: the role of accounting quality, University of California at Irvine

Juttner, D. J., and J. Mc Carthy, 2009, Modelling a rating crisis, Macquarie University, Sydney

Kaavvathas, D., Estimating credit rating transition probabilities for corporate bonds, Working paper

Kahle, K. M. and R. A. Walkling, Sept 1996, The impact of industry classification on financial research, Journal of Financial and Quantitative Analysis, Vol. 31 No. 3

Kaplan, R. S., and G. Urwitz, 1979, Statistical models of bond ratings: A methodological inquiry, Journal of Business, Vol. 52, No. 2

Kliger, D., and O. Sarig, Dec 2000, The information value of bond ratings, The Journal of Finance, Vol. LV No. 6

Klock, M. S., S. A. Mansi, and W. F. Maxwell, Jul 2004, Does corporate governance matter to bondholders?, Journal of Financial and Quantitative Analysis

Koopman, S.J., and A. Lucas, 2005, Business and default cycles for credit risk, Journal of Applied Econometrics 20, 311-323

Koskela, E., and R. Stenbacka, 2004, Agency cost of debt and credit market imperfections: A bargaining approach, Bulletin of Economic Research 56:4, 0307-3378

Kraft, P., 2010, Do rating agencies cater? Evidence from rating-based contracts, Working paper, New York University

Krahnen, J. P., and M. Weber, 2001, Generally accepted rating principles: A primer, Journal of Banking and Finance 25 3-23
Krishnan, P., 1993, Random parameters and self-selection models, Empirical Economics, 18:197-213

Krishnan, C. N. V., P. H. Ritchken, and J. B. Thomson, Nov 2003, On credit spread slopes and predicting bank risk, Federal Reserve Bank of Cleveland, Working paper, 03-14

Kuhner, C., Jan 2001, Financial rating agencies: Are they credible? – Insights into the reporting incentives of rating agencies in times of enhanced systemic risk, Schmalenbach Business Review, Vol. 53

Lando, D., 2004, Credit risk modelling, Theory and applications, Princeton Series in Finance

Lando, D. and T. M. Skødeberg, 2002, Analysing rating transitions and rating drift with continuous observations, Journal of Banking and Finance, 26, 423-444

Lau, A. H., 1987, A five-state financial distress prediction model, Journal of Accounting Research, Vol. 25, No. 1

Leclere, M. J., 1999, The interpretation of coefficients in N-Chotomous qualitative response models, Contemporary Accounting Research, 16, 4

Lewis, M., 2010, The big short, inside the doomsday machine, W. W. Norton & Company Ltd.

Liao, T. F., 1994, Interpreting probability models Logit, Probit, and other generalized linear models, A SAGE University paper

Löffer, G., 2004, An anatomy of rating through the cycle, Journal of Banking and Finance, 28, 695-720

Löffer, G., 2004, Rating verses market-based measures of default risk in portfolio governance, Journal of Banking and Finance, 28, 2715-2746

Löffler, G., 2005, Avoiding the rating bounce: why rating agencies are slow to react to new information, Journal of Economic Behavior & Organization 56, 365-381

Long, J. A. and M. Prigmore, 2003, A Comparison of the Effectiveness of Neural and Wavelet Networks for Insurer Credit Rating Based on Publicly Available Financial Data, South Bank University

Long, J.A. and A. Raudys, 2000, Modelling Company Credit Ratings Using a Number of Classification Techniques, South Bank University

Long, J. S., 1997, Regression models for categorical and limited dependent variables, SAGE Publications, Inc.

Lopez, J. A., and M. R. Saidenberg, 2000, Evaluating credit risk models, Journal of Banking and Finance, 24, 151-165

Louviere, J. J., D. A. Hensher, and J. D. Swait, 2000, Stated choice methods, analysis and application, Cambridge University Press

Louviere, J. J., and D. Street, 2002, Dissecting the random component of utility, Marketing Letters 13:3, 177-193

Lowe, P., Sept 2002, Credit risk measurement and procyclicality, Bank for International Settlements, Working paper, No. 116

Lucas, A., and P. Klaassen, 2005, Discrete verses continuous state switching model for portfolio credit risk, Journal of Banking and Finance

Maddala, G. S., 1999, Limited-dependent and qualitative variables in econometrics, Econometric Society Monographers, No. 3

Magidson, J., and J. K. Vermunt, Latent class models, Statistical Innovations, Inc.

Maher, J. J., Oct 1987, Pension obligations and the bond credit market: An empirical analysis of accounting numbers, The Accounting Review, Vol. LXII, No. 4

Manso, G., Feb 2011, Feedback effects of credit ratings, Working paper, MIT Sloan School of Business

Metz, A, and R. Cantor, Nov 2006, Moody's credit rating prediction model, Moody's Investors Services

Mingo, J. J., 2000, Policy implication of the Federal Reserve study of credit risk models at major US banking institutions, Journal of Banking and Finance, 24, 15-33

Moody's, Mar 2005, Moody's KMV internal rating platform and the Basel II RB approaches

Nickell, P., W. Perraudin, and S. Varotto, 2000, Stability of rating transitions, Journal of Banking & Finance 24, 203-227

Norden, L., and M. Weber, 2004, Informational efficiency of credit default swap and stock markets: The impact of credit rating announcements, Journal of Banking and Finance, 28, 2813-2843

Opp, C. C., and M. M. Opp, and M. Harris, 2010, Ratings agencies in the face of regulation, rating inflation and regulatory arbitrage, Working paper, The Wharton School

Parnes, D., 2007, A density-dependent model for credit ratings migration dynamics, Working paper, University of South Florida, USA

Partnoy, F., 2001, The paradox of credit ratings, Law and Economics Research Paper No. 20, University of San Diego, USA

Peat, M., 2008, Non-parametric methods for credit risk analysis: Neural networks and recursive partitioning techniques in Advances in Credit Risk Modelling and Corporate Bankruptcy Prediction, ed. S Jones and D.A Hensher, Cambridge University Press, Cambridge, United Kingdom, pp. 137-153

Perraudin, W., and A. P. Taylor, 2004, On the consistency of ratings and bond market yields, Journal of Banking and Finance, 28, 2769-2788

Pesaran, M. H., April 2005, Macroeconomics dynamics and credit risk: A global perspective, Working paper, University of Cambridge, UK

Pfarr, C., A. Schmid, and U. Schneider, June 2010, Estimating ordered categorical variables using panel data: a generalised ordered Probit model with an autofit procedure, Munich Personal RePEc Archive

Wiemann, M., Rating triggers in loan contracts – how much influence have credit rating agencies on borrowers – in monetary terms, Frankfurt School of Finance and Management, Preliminary draft

Pluto, K., and D. Tasche, July 2005, Thinking positively, Deutsche Bundesbank, Germany

Poon, W. P. H., 2003, Are unsolicited credit ratings biased downward?, Journal of Banking and Finance, 27, 593-614

Powers, D. A., and Y. Xie, 2000, Statistical methods for categorical data analysis, Academic Press

Prigmore, M., and J. A. Long, 2003, A comparison of the effectiveness of neural and wavelet networks for insurer credit rating based on publicly available financial data, South Bank University, UK

Sabourin, P., Analysing and forecasting credit ratings: Some Canadian evidence, Working paper 99-02

Sengupta, P, Oct 1998, Corporate disclosure quality and the cost of debt, The Accounting Review, Vol. 73, No. 4, pp. 459-474

Shaw, K. W., Managerial entrenchment, incentive compensation, and credit ratings, Working paper, University of Missouri

Shumway, T., 2001, Forecasting bankruptcy more accurately: A simple hazard model, Journal of Business, Vol. 74, No. 1

Sillano, M., and J. D. Ortuzar, 2004, Willingness-to-pay estimation with mixed Logit models: Some new evidence, Environment and Planning A Vol. 37, pages 525-550

Simon, C. P., and L. Blume, 1994, mathematics for economists, W. W. Norton & Companies, Inc.

Standard & Poor's, 2003, Compustat user's guide, The McGraw-Hill Companies, Inc.

Standard & Poor's, 2003, Compustat North America user's guide, The McGraw-Hill Companies, Inc.

Standard & Poor's, 2006, Corporate ratings criteria, The McGraw-Hill Companies, Inc.

Standard & Poor's, July 2009, The devil is in the details: Understanding the variation in corporate default rates and rating transitions, The McGraw-Hill Companies, Inc.

Stefanescu, C., R. Tunaru, and S. Turnbull, Oct 2008, The credit rating process and estimation of transition probabilities: A Bayesian approach, Working paper, London Business School, UK

Strahan, P. E., Oct 2010, Do regulations based on credit ratings affect a firm's cost of capital? Oxford University Press.

Strang, G. 1980. Linear Algebra and Its Applications, 2nd ed., Academic Press, New York,

Sy, A. N. R., 2004, rating the rating agencies: Anticipating currency crisis or debt crises?, Journal of Banking & Finance 28, 2845-2867

Sylla, R., 2001, A historical primer on the business of credit ratings, Working paper, Stern School of Business, New York University

Swets, J.A, Dawes, R.M., and J. Monahan, 2000. 'Better Decisions Through Science', *Scientific American*, 283: 82-87.

Terza, J., 1985, Ordered Probit: A generalization, Communications in Statistics – A. Theory and Methods, 14, 1-11

Treacy, W. F., and M. Carey, 2000, Credit risk rating systems at large US banks, Journal of banking & Finance 24 167-201

Truck, S., and S. T. Rachev, 2005, Credit portfolio risk and PD confidence sets through the business cycle, Journal of Credit Risk, Vol. 3, No. 2

Wendin, J., and A. J. McNeil, Jan 2006, Dependent credit migrations, Working paper

White, H. 2000. 'A Reality Check for Data Snooping', Econometrica, 68 (5): 1097-1126.

Wilson, T. C., Oct 1998, Portfolio credit risk, FRBNY Economic Policy Review

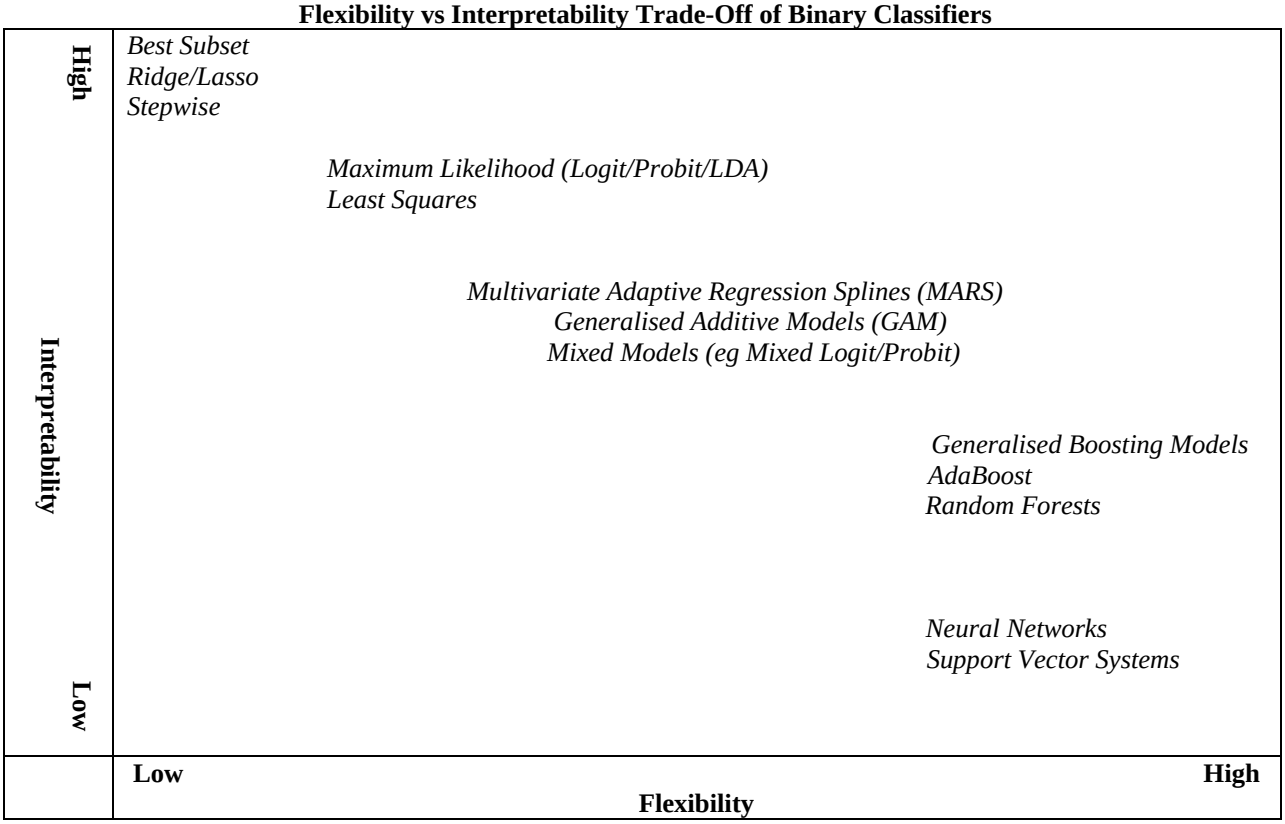
William, R., 2006, Generalized ordered Logit/partial proportional odds models for ordinal dependent variables, The Stata Journal, 6 No. 1 pp. 58-82

Zheng, H., 2006, Interaction of credit and liquidity risks: Modelling and valuation, , Journal of Banking & Finance 30, 391-407

Ziebart, D. A., and S. A. Reiter, 1992, Bond ratings, bond yields and financial information, Contemporary Accounting Research, Vol. 9, No. 1, pp. 252-282

Zou, H., and T. Hastie (2005). 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society, Series* B 67, *Part* 2: 301–320.

Figure 1
Flexibility vs Interpretability Trade-Off of Binary Classifiers



Adapted from Hastie et al., 2009.

Figure 2: AUC Performance of All Binary Classifiers on the Longitudinal Validation Sample

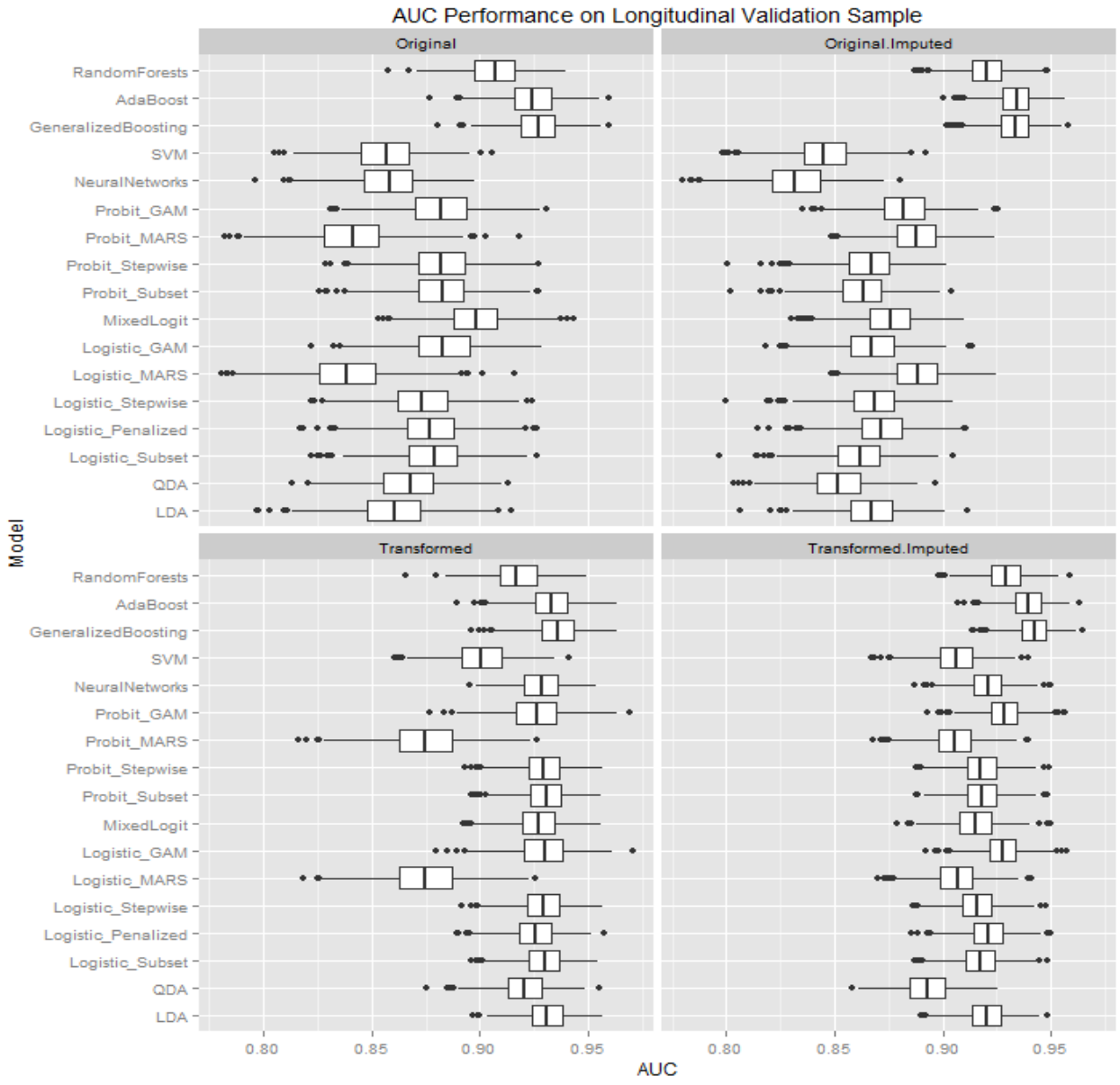


Figure 3: AUC Performance of All Binary Classifiers on Cross Sectional Validation Sample

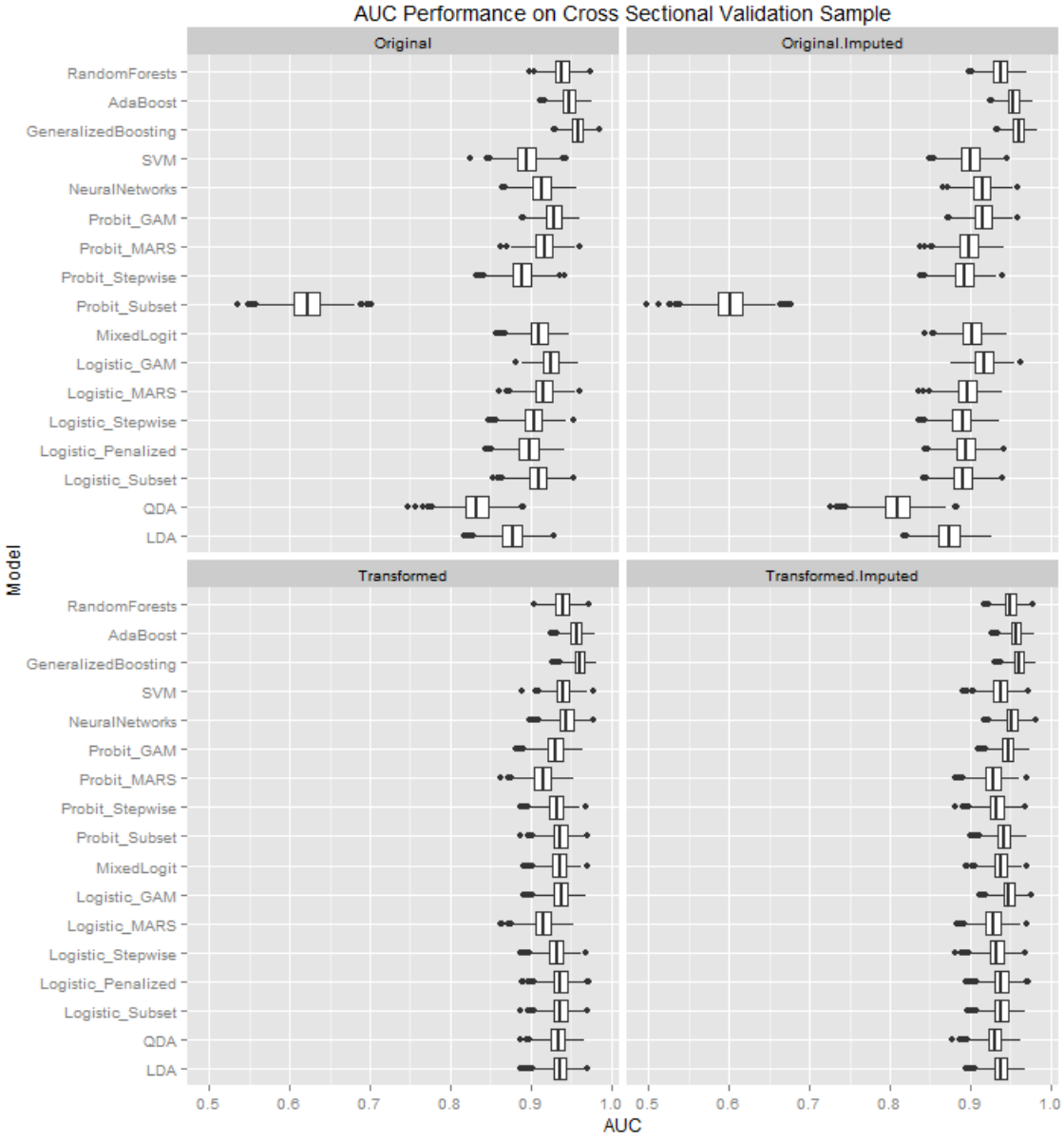


Table 1 Distribution of sample firms by S&P initial ratings/rating changes and year

Year	AAA	AA	A	BBB	BB	B	CCC	CC/C	D	**Total**
1980	6	24	34	16	6	2	1	0	0	89
1981	1	7	16	5	4	5	0	0	0	38
1982	0	1	14	16	5	1	0	0	0	37
1983	0	12	10	8	9	4	1	0	0	44
1984	0	7	21	14	9	8	0	0	0	59
1985	0	5	27	31	13	14	0	0	0	90
1986	2	10	24	27	18	14	9	0	3	107
1987	2	8	16	20	25	14	2	0	0	87
1988	2	0	16	20	14	19	1	0	0	72
1989	0	7	17	24	20	4	2	0	0	74
1990	0	8	9	17	12	7	2	0	0	55
1991	0	4	18	18	22	8	2	4	0	76
1992	2	1	14	28	27	15	2	2	0	91
1993	0	1	15	22	43	26	1	0	0	108
1994	1	4	10	23	26	25	3	0	0	92
1995	0	3	26	38	31	28	5	0	1	132
1996	0	30	26	41	55	45	8	2	1	181
1997	0	4	31	59	63	75	8	1	0	241
1998	2	5	30	57	80	85	23	2	0	284
1999	2	5	33	56	68	62	27	10	26	289
2000	0	6	26	58	61	76	26	9	22	284
2001	0	4	19	48	51	62	51	21	45	301
2002	0	1	12	36	63	60	41	13	20	246
2003	0	0	5	23	69	69	26	8	17	217
2004	0	1	5	27	40	53	15	3	10	154
2005	0	1	12	29	58	65	23	4	12	204
2006	0	2	4	24	57	57	13	1	3	161
Total	20	134	490	785	949	903	292	80	160	3813

Table 2 Distribution of sample firms by S&P initial ratings/rating changes and year

Year	Negative Change	No Change	Positive Change	**Total**
1980	0	89	0	89
1981	2	30	6	38
1982	6	29	2	37
1983	10	24	10	44
1984	12	39	8	59
1985	24	54	12	90
1986	39	50	18	107
1987	26	44	17	87
1988	21	28	23	72
1989	20	21	33	74
1990	23	18	14	55
1991	28	34	14	76
1992	25	44	22	91
1993	19	60	29	108
1994	24	51	17	92
1995	27	78	27	132
1996	39	107	35	181
1997	47	141	53	241
1998	68	172	44	284
1999	124	126	39	289
2000	123	127	34	284
2001	190	86	25	301
2002	140	77	29	246
2003	114	62	41	217
2004	69	51	34	154
2005	101	56	47	204
2006	72	58	31	161
Total	1393	1756	664	3813

Table 3 Distribution of sample firms by S&P initial ratings/rating changes and year

Year	Investment Grade	Speculative Grade	**Total**
1980	80	9	89
1981	29	9	38
1982	31	6	37
1983	30	14	44
1984	42	17	59
1985	63	27	90
1986	63	41	104
1987	46	41	87
1988	38	34	72
1989	48	26	74
1990	34	21	55
1991	40	36	76
1992	45	46	91
1993	38	70	108
1994	38	54	92
1995	67	64	131
1996	70	110	180
1997	94	147	241
1998	94	190	284
1999	96	167	263
2000	90	172	262
2001	71	185	256
2002	49	177	226
2003	28	172	200
2004	33	111	144
2005	42	150	192
2006	30	128	158
Total	1429	2224	3653

Table 4: ROC Curve Analysis for All Binary Classifiers for Longitudinal and Cross Sectional Validation Sample

Overall Performance			Overall Performance (Longitudinal Sample)		Overall Performance (Cross Section Sample)	
Model	**AUC**	**Rank**	**AUC**	**Rank**	**AUC**	**Rank**
RandomForests	.9297	3	.9179	3	.9416	3
AdaBoost	.9433	2	.9328	2	.9539	2
GeneralisedBoosting	.9469	1	.9343	1	.9595	1
SVM	.8973	13	.8768	16	.9177	9
NeuralNetworks	.9046	11	.8800	14	.9292	6
Probit_GAM	.9181	4	.9061	4	.9302	5
Probit_MARS	.8959	14	.8768	16	.9149	11
Probit_Stepwise	.9046	10	.8986	7	.9106	14
Probit_Subset	.8362	17	.8980	9	.7745	17
Mixed Logit	.9125	6	.9034	5	.9215	7
Logistic_GAM	.9159	5	.9015	6	.9304	4
Logistic_MARS	.8957	15	.8766	17	.9149	12
Logistic_Stepwise	.9053	9	.8961	11	.9146	13
Logistic_Penalised	.9078	7	.8984	8	.9172	10
Logistic_Subset	.9076	8	.8961	10	.9191	8
QDA	.8797	16	.8827	13	.8768	16
LDA	.9002	12	.8943	12	.9062	15

Table 4 displays rankings of overall mean AUC performance for all binary classifiers, including the breakdown of the rankings of overall mean AUC performance on the longitudinal and cross sectional validation samples. AUC is the 'area under the curve' for the receiver operating characteristic (ROC) curve. Table 4 indicates that the top three performing models are generalised boosting, AdaBoost and random forests respectively.

Table 5: ROC Curve Analysis for Longitudinal Validation Sample

AUC Summaries for Longitudinal Validation Sample Across Datasets																
Classifiers:	**Original Data AUCs**	**AUC Rank**	**UCI**	**LCI**	**Original Imputed Data AUCs**	**AUC Rank**	**UCI**	**LCI**	**Transformed Data AUCs**	**AUC Rank**	**UCI**	**LCI**	**Transformed Imputed Data AUCs**	**AUC Rank**	**UCI**	**LCI**
RandomForests	.9065	3	.8806	.9335	.9185	3	.8993	.9392	.9164	14	.8934	.9429	.9301	3	.9135	.9501
AdaBoost	.9243	2	.9025	.9487	.9337	1	.9173	.9516	.9333	2	.9137	.9557	.9399	2	.9246	.9580
GeneralisedBoosting	.9262	1	.9048	.9498	.9332	2	.9172	.9515	.9356	1	.9167	.9577	.9422	1	.9276	.9604
SVM	.8560	15	.8250	.8876	.8453	16	.8157	.8747	.9000	15	.8740	.9261	.9059	15	.8858	.9287
NeuralNetworks	.8573	14	.8271	.8905	.8212	17	.7905	.8551	.9276	10	.9067	.9482	.9138	13	.8950	.9348
Probit_GAM	.8774	8	.8446	.9138	.8866	6	.8603	.9175	.9324	3	.9099	.9581	.9280	4	.9101	.9466
Probit_MARS	.8404	16	.8030	.8802	.8872	5	.8618	.9138	.8746	16	.8416	.9105	.9050	16	.8835	.9273
Probit_Stepwise	.8818	6	.8527	.9135	.8661	12	.8406	.8941	.9290	8	.9096	.9501	.9176	9	.8991	.9370
Probit_Subset	.8814	7	.8525	.9136	.8624	13	.8357	.8899	.9300	5	.9112	.9503	.9180	8	.8988	.9382
Mixed Logit	.8978	4	.8712	.9288	.8765	7	.8508	.9038	.9254	11	.9042	.9475	.9139	12	.8929	.9326
Logistic_GAM	.8825	5	.8497	.9180	.8670	10	.8381	.8977	.9290	8	.9057	.9541	.9273	5	.9096	.9459
Logistic_MARS	.8381	17	.7991	.8776	.8876	4	.8621	.9142	.8745	17	.8421	.9092	.9061	14	.8848	.9282
Logistic_Stepwise	.8722	11	.8406	.9061	.8676	9	.8417	.8958	.9285	9	.9093	.9493	.9159	11	.8969	.9355
Logistic_Penalised	.8770	10	.8464	.9098	.8711	8	.8448	.8976	.9249	12	.9055	.9463	.9204	6	.9023	.9403
Logistic_Subset	.8773	9	.8468	.9108	.8608	14	.8338	.8893	.9292	6	.9106	.9501	.9170	10	.8983	.9360
QDA	.8677	12	.8367	.9024	.8508	15	.8229	.8793	.9195	13	.8986	.9409	.8927	17	.8673	.9161
LDA	.8596	13	.8276	.8954	.8668	11	.8419	.8939	.9305	4	.9118	.9511	.9201	7	.9024	.9394

Table 5 displays the AUC performance and rankings for all binary classifiers on the longitudinal validation sample. AUC is the 'area under the curve' for the receiver operating characteristic (ROC) curve. AUC performance is displayed for 'Original Data', 'Original Imputed Data', 'Transformed Data' and 'Transformed Imputed Data'. 'Original Data' represents the untransformed data with no missing value imputation (missing values are deleted case wise). 'Original Imputed Data' is the original data but with missing values imputed using the single value decomposition or SVD method. 'Transformed Data' represents the Box Cox transformed data with no missing value imputation (missing values are deleted case wise). 'Transformed Imputed Data' represents the Box Cox transformed data with missing values imputed using SVD. Also provide are the upper and lower confidence intervals (UCI and LCI) at the 95% level.

Table 6: ROC Curve Analysis for the Cross Sectional Validation Sample

AUC Summaries for Cross Sectional Holdout Samples Across Datasets																
Models	**Original Data AUCs**	**AUC Rank**	**UCI**	**LCI**	**Original Imputed Data AUCs**	**AUC Rank**	**UCI**	**LCI**	**Transformed Data AUCs**	**AUC Rank**	**UCI**	**LCI**	**Transformed Imputed Data AUCs**	**AUC Rank**	**UCI**	**LCI**
RandomForests	.9381	3	.9146	.9633	.9386	3	.9155	.9629	.9393	5	.9170	.9630	.9504	3	.9308	.9715
AdaBoost	.9476	2	.9270	.9699	.9532	2	.9352	.9737	.9566	2	.9396	.9754	.9581	2	.9417	.9768
GeneralisedBoosting	.9590	1	.9408	.9779	.9588	1	.9416	.9777	.9596	1	.9436	.9779	.9606	1	.9442	.9791
SVM	.8943	13	.8618	.9291	.8992	10	.8678	.9316	.9403	4	.9186	.9639	.9371	12	.9140	.9616
NeuralNetworks	.9132	8	.8831	.9446	.9098	6	.8818	.9407	.9442	3	.9226	.9692	.9495	5	.9292	.9720
Probit_GAM	.9230	5	.8975	.9500	.9152	4	.8868	.9448	.9329	13	.9083	.9637	.9495	5	.9285	.9713
Probit_MARS	.9162	6	.8890	.9459	.9010	8	.8692	.9327	.9145	17	.8877	.9442	.9280	17	.9021	.9560
Probit_Stepwise	.8887	14	.8588	.9267	.8892	14	.8580	.9240	.9316	15	.9100	.9576	.9328	13	.9113	.9586
Probit_Subset	.6207	17	.5708	.6632	.5981	17	.5513	.6466	.9367	8	.9145	.9617	.9423	7	.9228	.9671
Mixed Logit	.9100	9	.8812	.9416	.9020	7	.8718	.9352	.9365	9	.9156	.9620	.9375	10	.9170	.9634
Logistic_GAM	.9267	4	.9019	.9535	.9141	5	.8855	.9452	.9338	11	.9100	.9633	.9470	6	.9254	.9688
Logistic_MARS	.9158	7	.8880	.9463	.9008	9	.8693	.9327	.9149	16	.8874	.9447	.9280	17	.9027	.9560
Logistic_Stepwise	.9038	11	.8748	.9375	.8903	13	.8593	.9260	.9319	14	.9099	.9580	.9324	14	.9101	.9582
Logistic_Penalised	.8977	12	.8663	.9321	.8949	11	.8639	.9311	.9374	6	.9159	.9621	.9387	9	.9181	.9638
Logistic_Subset	.9093	10	.8805	.9420	.8912	12	.8594	.9258	.9370	7	.9149	.9618	.9387	9	.9185	.9641
QDA	.8338	16	.7949	.8780	.8095	16	.7678	.8589	.9333	12	.9110	.9600	.9305	15	.9093	.9576
LDA	.8772	15	.8460	.9155	.8740	15	.8420	.9117	.9363	10	.9148	.9627	.9374	11	.9172	.9627

Table 6 displays the AUC performance and rankings for all binary classifiers on the cross sectional validation sample. AUC is the 'area under the curve' for the receiver operating characteristic (ROC) curve. AUC performance is displayed for 'Original Data', 'Original Imputed Data', 'Transformed Data' and 'Transformed Imputed Data'. 'Original Data' represents the untransformed data with no missing value imputation (missing values are deleted case wise). 'Original Imputed Data' is the original data but with missing values imputed using the single value decomposition or SVD method. 'Transformed Data' represents the Box Cox transformed data with no missing value imputation (missing values are deleted case wise). 'Transformed Imputed Data' represents the Box Cox transformed data with missing values imputed using SVD. Also provide are the upper and lower confidence intervals (UCI and LCI) at the 95% level.

Table 7: Mean AUC Differences and Significance Levels across Binary Classifiers (Longitudinal Validation Sample is above the Diagonal and Cross Sectional Validation Sample is below the Diaganol)

Panel A: AUC Differences Original Data	RandomForests	AdaBoost	GeneralizedBoosting	SVM	NeuralNetworks	Probit_GAM	Probit_MARS	Probit_Stepwise	Probit_Subset	MixedLogit	Logistic_GAM	Logistic_MARS	Logistic_Stepwise	Logistic_Penalised	Logistic_Subset	QDA	LDA
RandomForests	NA	-0.017 ****	-0.020 ****	0.051 ****	0.050 ****	0.025	0.066 ****	0.025 **	0.025 **	0.009	0.024	0.068 ****	0.034 ****	0.030 **	0.029 **	0.040 **	0.047 ****
AdaBoost	-0.009 **	NA	-0.002	0.068 ****	0.067 ****	0.042 ***	0.083 ****	0.042 ****	0.042 ****	0.026 ****	0.041 **	0.085 ****	0.051 ****	0.047 ****	0.046 ****	0.057 ****	0.064 ****
GeneralizedBoosting	-0.02 ****	-0.011 ****	NA	0.070 ****	0.069 ****	0.045 ***	0.086 ****	0.044 ****	0.044 ****	0.028 ****	0.044 **	0.088 ****	0.053 ****	0.049 ****	0.048 ****	0.059 ****	0.066 ****
SVM	0.044 ****	0.053 ****	0.064 ****	NA	-0.001	-0.025	0.016	-0.026 *	-0.026 *	-0.042 ****	-0.026	0.018	-0.017	-0.021	-0.022	-0.011	-0.004
NeuralNetworks	0.025 **	0.034 ****	0.045 ****	-0.019 *	NA	-0.024	0.016	-0.025 **	-0.025 **	-0.041 ***	-0.026	0.019	-0.016	-0.020	-0.021 *	-0.010	-0.003
Probit_GAM	0.01	0.02 **	0.03 ****	-0.033 **	-0.015	NA	0.041 **	-0.001	0.000	-0.016	-0.001	0.043 **	0.008	0.004	0.004	0.015	0.021
Probit_MARS	0.022 **	0.032 ****	0.042 ****	-0.021	-0.003	0.012	NA	-0.041 **	-0.041 ***	-0.057 ****	-0.042 **	0.002 **	-0.033 **	-0.036 **	-0.037 **	-0.026	-0.019
Probit_Stepwise	0.05 ****	0.059 ****	0.07 ****	0.006	0.025 *	0.04 ***	0.028 **	NA	0.000	-0.016	-0.001	0.044 **	0.009	0.005	0.004	0.015	0.022 **
Probit_Subset	0.317 ****	0.326 ****	0.337 ****	0.273 ****	0.292 ****	0.307 ****	0.295 ****	0.267 ****	NA	-0.016 *	-0.001	0.043 ***	0.009	0.005	0.004 *	0.015	0.022 **
MixedLogit	0.028 **	0.038 ****	0.048 ****	-0.015	0.003	0.018	0.006	-0.021 ****	-0.289 ****	NA	0.015	0.059 ****	0.025 ****	0.021 *	0.020 **	0.031 **	0.038 ***
Logistic_GAM	0.014	0.023 **	0.034 ****	-0.03 **	-0.011	0.004	-0.008	-0.036 ***	-0.303 ****	-0.014	NA	0.044 **	0.009	0.006	0.005	0.016	0.023
Logistic_MARS	0.023 **	0.032 ****	0.043 ****	-0.021	-0.002	0.012	0	-0.027 *	-0.294 ****	-0.006	0.009	NA	-0.035 **	-0.039 **	-0.039 ****	-0.029	-0.022
Logistic_Stepwise	0.035 **	0.044 ****	0.055 ****	-0.009	0.01	0.025 *	0.013	-0.015 **	-0.282 ****	0.007	0.021	0.012	NA	-0.004	-0.005	0.006	0.013
Logistic_Penalised	0.041 ***	0.05 ****	0.061 ****	-0.003	0.016	0.031 **	0.019	-0.009	-0.276 ****	0.013 *	0.027 **	0.018	0.006	NA	-0.001	0.010	0.017 *
Logistic_Subset	0.03	0.039	0.05	-0.014	0.005	0.019	0.007	-0.02	-0.287	0.001	0.016	0.007	-0.005	-0.011	NA	0.011	0.018*

	**	****	****					***	****					**			
QDA	0.105 ****	0.115 ****	0.126 ****	0.062 ****	0.08 ****	0.095 ****	0.083 ****	0.056 ****	-0.212 ****	0.077 ****	0.092 ****	0.083 ****	0.071 ****	0.064 ****	0.076 ****	NA	0.007
LDA	0.061 ****	0.071 ****	0.082 ****	0.018	0.03 **	0.051 ****	0.039 ***	0.012 *	-0.256 ****	0.033 ***	0.048 ***	0.039 **	0.027 ****	0.02 **	0.032 ****	-0.044 ****	NA

Table 7 -continued

Panel B: AUC Differences Original Imputed Data	RandomForests	AdaBoost	GeneralizedBoosting	SVM	NeuralNetworks	Probit_GAM	Probit_MARS	Probit_Stepwise	Probit_Subset	MixedLogit	Logistic_GAM	Logistic_MARS	Logistic_Stepwise	Logistic_Penalised	Logistic_Subset	QDA	LDA
RandomForests	NA	-0.013 ***	-0.013 ***	0.075 ****	0.089 ****	0.039 ***	0.033 ***	0.054 ****	0.058 ****	0.045 ****	0.053 ****	0.032 ***	0.052 ****	0.049 ****	0.059 ****	0.069 ****	0.053 ****
AdaBoost	-0.016 ****	NA	0.001	0.088 ****	0.102 ****	0.052 ****	0.046 ****	0.067 ****	0.071 ****	0.058 ****	0.066 ****	0.045 ****	0.066 ****	0.062 ****	0.072 ****	0.082 ****	0.067 ****
GeneralizedBoosting	-0.022 ****	-0.006 **	NA	0.087 ****	0.102 ****	0.051 ****	0.045 ****	0.066 ****	0.070 ****	0.057 ****	0.066 ****	0.045 ****	0.065 ****	0.061 ****	0.072 ****	0.082 ****	0.066 ****
SVM	0.038 ***	0.054 ****	0.06 ****	NA	0.014	-0.036 **	-0.042 ***	-0.021 *	-0.017	-0.030 **	-0.022 *	-0.042 ***	-0.022 *	-0.026 **	-0.016	-0.006	-0.021 *
NeuralNetworks	0.024 **	0.04 ****	0.046 ****	-0.014	NA	-0.050 ***	-0.056 ***	-0.035 **	-0.031 **	-0.044 ***	-0.036 **	-0.057 ***	-0.037 ***	-0.040 ***	-0.030 **	-0.020	-0.035 ***
Probit_GAM	0.022 *	0.038 ****	0.044 ****	-0.016	-0.002	NA	-0.006	0.015	0.019	0.006	0.015 *	-0.006	0.014	0.010	0.021	0.031 **	0.015
Probit_MARS	0.04 ***	0.055 ****	0.061 ****	0.001	0.016	0.017	NA	0.021	0.025 *	0.012	0.021	0.000	0.020	0.016	0.027 *	0.036 **	0.021
Probit_Stepwise	0.045 ***	0.061 ****	0.067 ****	0.007	0.022 *	0.023	0.006	NA	0.004	-0.009	-0.001	-0.022	-0.002	-0.005	0.005	0.015 *	0.000
Probit_Subset	0.337 ****	0.353 ****	0.359 ****	0.299 ****	0.313 ****	0.315 ****	0.297 ****	0.292 ****	NA	-0.013	-0.004	-0.025 *	-0.005	-0.009	0.002	0.011	-0.004
MixedLogit	0.036 ***	0.052 ****	0.057 ****	-0.003	0.012	0.013	-0.004	-0.01	-0.301 ****	NA	0.008	-0.012	0.008	0.004	0.014 *	0.024 **	0.009
Logistic_GAM	0.021 *	0.037 ****	0.042 ****	-0.017	-0.003	-0.001	-0.019	-0.025 *	-0.316 ****	-0.015	NA	-0.021	-0.001	-0.004	0.006	0.016	0.000
Logistic_MARS	0.042	0.057	0.063	0.003	0.018	0.019	0.002	-0.004	-0.295	0.006	0.021	NA	0.020	0.016	0.027	0.037	0.021

	***	****	****				**		****						*	**	
Logistic_Stepwise	0.048 ****	0.064 ****	0.07 ****	0.01	0.024 *	0.026 *	0.008	0.003	-0.289 ****	0.012 *	0.027 *	0.006	NA	-0.004	0.007	0.017 **	0.001
Logistic_Penalised	0.043 ****	0.059 ****	0.065 ****	0.005	0.02	0.021	0.004	-0.002	-0.294 ****	0.008	0.023	0.002	-0.005	NA	0.010	0.020 **	0.005
Logistic_Subset	0.047 ****	0.062 ****	0.068 ****	0.008	0.023 *	0.024 *	0.007	0.001	-0.29 ****	0.011	0.026 *	0.005	-0.001	0.003	NA	0.010	-0.006
QDA	0.129 ****	0.145 ****	0.15 ****	0.091 ****	0.105 ***	0.107 ****	0.089 ****	0.083 ****	-0.208 ****	0.093 ****	0.108 ****	0.087 ****	0.081 ****	0.085 ****	0.082 ****	NA	-0.016
LDA	0.064 ****	0.08 ****	0.086 ****	0.026	0.04 ***	0.042 ***	0.025	0.019 ****	-0.273 ****	0.029	0.043 ****	0.023	0.016	0.021 ***	0.018 ***	-0.065 ****	NA

Table 7 -continued

Panel C: AUC Differences Transformed Data	RandomForests	AdaBoost	GeneralizedBoosting	SVM	NeuralNetworks	Probit_GAM	Probit_MARS	Probit_Stepwise	Probit_Subset	MixedLogit	Logistic_GAM	Logistic_MARS	Logistic_Stepwise	Logistic_Penalised	Logistic_Subset	QDA	LDA
RandomForests	NA	-0.016 ****	-0.018 ****	0.016 *	-0.011	-0.009	0.042 ***	-0.013	-0.013	-0.010	-0.012	0.042 ***	-0.012	-0.008	-0.013	-0.003	-0.014
AdaBoost	-0.016 ****	NA	-0.003	0.032 ***	0.005	0.007	0.058 ***	0.003	0.002	0.006	0.003	0.058 ***	0.004	0.007	0.003	0.013	0.002
GeneralizedBoosting	-0.02 ****	-0.004	NA	0.035 ****	0.007	0.010	0.061 ****	0.006	0.005	0.009	0.006	0.061 ****	0.006	0.010	0.006	0.015 *	0.005
SVM	-0.001	0.016 **	0.02 ***	NA	-0.027 ***	-0.025 *	0.026	-0.029 ***	-0.030 ***	-0.026 ***	-0.029 **	0.026	-0.028 ***	-0.025 **	-0.029 ***	-0.019 *	-0.030 ***
NeuralNetworks	-0.004	0.012	0.016 **	-0.004	NA	0.002	0.053 ***	-0.002	-0.002	0.001	-0.001	0.053 ***	-0.001	0.003	-0.002	0.008	-0.003
Probit_GAM	0.01	0.026 **	0.03 **	0.011	0.014	NA	0.051 ***	-0.004	-0.005	-0.001	-0.004	0.051 ***	-0.003	0.000	-0.004	0.006	-0.005
Probit_MARS	0.025 **	0.042 ****	0.046 ****	0.026 **	0.03 **	0.015	NA	-0.055 ***	-0.056 ****	-0.052 ***	-0.055 ***	0.000	-0.054 ***	-0.051 ***	-0.055 ****	-0.045 ***	-0.056 ****
Probit_Stepwise	0.008	0.025 ***	0.029 ****	0.009	0.013 *	-0.002	-0.017	NA	-0.001	0.003	0.000	0.055 ****	0.001	0.004	0.000	0.009	-0.001
Probit_Subset	0.003	0.02	0.024	0.004	0.008	-0.007	-0.022	-0.005	NA	0.004	0.001	0.056	0.001	0.005	0.001	0.010	0.000

		**	***				*					***				*	
MixedLogit	0.004	0.02 ***	0.024 ***	0.005	0.008	-0.006	-0.021 *	-0.004	0.001	NA	-0.003	0.052 ***	-0.002	0.001	-0.003	0.007	-0.004
Logistic_GAM	0.002	0.019	0.023 *	0.003	0.007	-0.008	-0.023	-0.006	-0.001	-0.002	NA	0.055 ***	0.000	0.004	0.000	0.009	-0.001
Logistic_MARS	0.025 **	0.041 ****	0.045 ****	0.026 **	0.029 **	0.015 *	0	0.016	0.022 *	0.021 *	0.022	NA	-0.054 ***	-0.051 ***	-0.055 ****	-0.045 ***	-0.056 ****
Logistic_Stepwise	0.008	0.024 ***	0.028 ****	0.009	0.012	-0.002	-0.017	0	0.005	0.004	0.006	-0.017	NA	0.004	-0.001	0.009	-0.002
Logistic_Penalised	0.003	0.019 ***	0.023 ***	0.003	0.007	-0.007	-0.023 *	-0.006	-0.001	-0.001	0	-0.022 *	-0.005	NA	-0.004	0.005	-0.005
Logistic_Subset	0.003	0.019 **	0.023 ***	0.004	0.007	-0.007	-0.022 **	-0.006	0	-0.001	0.001	-0.022 *	-0.005	0	NA	0.010	-0.001
QDA	0.007	0.023 ***	0.027 ***	0.007	0.011	-0.003	-0.018	-0.002	0.004	0.003	0.004	-0.018	-0.001	0.004	0.004	NA	-0.011
LDA	0.004	0.02 **	0.024 ***	0.005	0.008	-0.006	-0.021 *	-0.005	0.001	0	0.001	-0.021 *	-0.004	0.001	0.001	-0.003	NA

Table 7 -continued

Panel D: AUC Differences Transformed and Imputed Data	RandomForests	AdaBoost	GeneralizedBoosting	SVM	NeuralNetworks	Probit_GAM	Probit_MARS	Probit_Stepwise	Probit_Subset	MixedLogit	Logistic_GAM	Logistic_MARS	Logistic_Stepwise	Logistic_Penalised	Logistic_Subset	QDA	LDA
RandomForests	NA	-0.010 ***	-0.013 ***	0.023 ***	0.008	0.001	0.024 **	0.011	0.011	0.014 **	0.002	0.023 **	0.013 *	0.009	0.012	0.036 ***	0.009
AdaBoost	-0.008 *	NA	-0.003	0.034 ****	0.019 ***	0.011	0.034 ****	0.022 ***	0.021 ***	0.024 ****	0.012 *	0.033 ***	0.023 ****	0.019 ***	0.022 ***	0.047 ****	0.019 ***
GeneralizedBoosting	-0.011 **	-0.003	NA	0.036 ****	0.021 ****	0.014 **	0.037 ****	0.024 ****	0.024 ****	0.026 ****	0.015 **	0.036 ****	0.026 ****	0.021 ****	0.025 ****	0.049 ****	0.022 ***
SVM	0.013	0.02 **	0.024 ***	NA	-0.015 **	-0.022 **	0.000	-0.012	-0.012	-0.010	-0.022 **	-0.001	-0.010	-0.015 **	-0.011	0.013	-0.015 *
NeuralNetworks	-	0.006	0.009	-0.014	NA	-0.007	0.016	0.003	0.003	0.005	-0.007	0.015	0.005	0.000	0.004	0.028	0.001

	0.002			*												***	
Probit_GAM	0.003	0.011	0.014 *	-0.009	0.005	NA	0.023 **	0.010	0.010	0.013	0.001	0.022 **	0.012	0.008	0.011	0.035 ****	0.008
Probit_MARS	0.022 **	0.029 **	0.033 ****	0.009	0.023 **	0.018	NA	-0.012	-0.013	-0.010	-0.022 **	-0.001 **	-0.011	-0.015	-0.012	0.013	-0.015
Probit_Stepwise	0.017 **	0.025 ****	0.028 ****	0.005	0.019 ***	0.014	-0.004	NA	0.000	0.002	-0.010	0.011	0.002 ***	-0.003	0.000	0.025 ***	-0.003
Probit_Subset	0.009	0.016 **	0.02 ***	-0.004	0.01 *	0.005	-0.013	-0.009 **	NA	0.003	-0.009	0.012	0.002 ***	-0.002 ***	0.001	0.025 ***	-0.002
MixedLogit	0.013	0.021 ***	0.024 ***	0	0.015 **	0.01	-0.009	-0.004	0.004	NA	-0.012	0.009	-0.001	-0.005	-0.002	0.023 **	-0.005
Logistic_GAM	0.002	0.01	0.013 *	-0.01	0.004	-0.001	-0.02 *	-0.015	-0.007	-0.011	NA	0.021 **	0.011	0.007	0.010	0.035 ***	0.007
Logistic_MARS	0.022 **	0.029 ***	0.033 ****	0.009	0.023 **	0.018	0	0.004	0.013	0.009	0.019 *	NA	-0.010	-0.014	-0.011	0.014	-0.014
Logistic_Stepwise	0.018 **	0.025 ****	0.029 ****	0.005	0.019 ***	0.014	-0.004	0	0.009 **	0.005	0.016	-0.004	NA	-0.004	-0.001	0.023 **	-0.004
Logistic_Penalised	0.012	0.019 ***	0.022 ***	-0.001	0.013 **	0.008	-0.01	-0.006 *	0.003	-0.001	0.009	-0.01	-0.006 *	NA	0.003	0.028 ***	0.000
Logistic_Subset	0.011	0.019 ****	0.022 ***	-0.001	0.01	0.008	-0.01	-0.006	0.003	-0.002	0.009	-0.01	-0.006 *	0	NA	0.025 ***	-0.003 *
QDA	0.02 **	0.028 ***	0.031 ****	0.007	0.022 **	0.017	-0.002	0.002	0.011 *	0.007	0.018	-0.002	0.002	0.008	0.008	NA	-0.028 ***
LDA	0.013	0.021 ***	0.02 ***	0	0.005	0.01	-0.009	-0.005	0.004 *	0	0.011	-0.009	-0.005	0.001	0.001	-0.007	NA

Table 7 displays mean differences in AUC performance and significance levels across all paired groups of binary classifiers. AUC is the 'area under the curve' for the receiver operating characteristic (ROC) curve. Panel A of Table 7 displays mean AUC differences and significance levels for the 'Original Data'. Panels B, C and D displays the same results for 'Original Imputed Data', 'Transformed Data' and 'Transformed Imputed Data' respectively. 'Original Data' represents the untransformed data with no missing value imputation (missing values are deleted case wise). 'Original Imputed Data' is the original data but with missing values imputed using the single value decomposition or SVD method. 'Transformed Data' represents the Box Cox transformed data with no missing value imputation (missing values are deleted case wise). 'Transformed Imputed Data' represents the Box Cox transformed data with missing values imputed using SVD. Mean differences in AUCs for the longitudinal validation sample are shown above the diagonal (and are shaded) while the cross sectional validation results are shown below the diagonal. Significance levels are shown in asterisks where *p<.1 ** p<.05 *** p<.01 **** p<.001. Significance levels are based on a two tailed Z test.

Appendix 1 Variable Definitions

Expected(Predicted) Sign	Variable Acronym	Definition of Variables
+ve	Crta	Cash Resources / Total Assets
+ve	Reta	Retained Earnings / Total Assets
+ve	Ebitta	EBIT / Total Assets
+ve	Mebvtl	Market Equity / Book Value of Total Liabilities
-ve	Tdta	Total Debt/ Total Assets
+ve	Sta	Sales / Total assets
+ve	Size1	Market Equity
+ve	Age	Age i.e. Years since the firm was first rated by an agency. The Age variable is set to 10 for Age values greater than 10 and for firms already rated at the beginning of the dataset in 1980.
	UtilityD	Takes a value of one if the firm belongs to the Utilities group and zero otherwise
	IndustD	Takes a value of one if the firm belongs to the Industrials group and zero otherwise

	TransD*	Takes a value of one if the firm belongs to the Transportations group and zero otherwise
+ve	IntCov	(Operating Income after Depreciation(EBIT) + Interest Expense)/Interest Expense
+ve	OpMar	Operating Income before Depreciation/Net Sales
-ve	Ldta	Long-term Debt/ Total Assets
+ve	Size2	Natural log of Total Assets
+ve	FFOTd	Funds From Operations (FFO) / Total Debt

+ve	OCFtd	Operating Cash Flows / Total Debt
+ve	Roc	EBIT/ Invested Capital
-ve	Td_Tdte	Total Debt / Total Debt + Equity
-ve	Tdebitda	Total Debt / EBITDA
+ve	OCFta	Net Operating Cash Flows / Total Assets
+ve	FFOta	Funds From Operations (FFO) / Total Assets
-ve	TdOCF	Total Debt / (Gross) Net Operating Cash Flows
-ve	TdFFO	Total Debt/ Funds From Operations (FFO)
-ve	Tdte	Total Debt / Total Equity
-ve	Tlte	Total Liabilities / Total Equity
+ve	CFOCov1	Net Operating Cash Flows/Interest Expense
+ve	CFOCov2	Funds From Operations (FFO) / Interesst Expense
+ve	Crcl	Cash Resources / Current Liabilities
+ve	CurR	Current Assets / Current Liabilities
-ve	Tlta	Total Liabilities / Total Assets
+ve	Roe	Net Income(Loss)/ Common Equity Total
+ve	Roa	Net Income(Loss)/ Total Assets

*Transportations dummy (TransD) is suppressed to avoid dummy trap

Appendix 2: Description of Binary Classifiers		
Model Type	**Specification**	**Explanation/Pros/Cons**
Logit	$\text{Prob}[Y_i = 1\|x_i] = \frac{e^{\beta' x_i}}{1 + e^{\beta' x_i}}$ where $\beta' x_i$ is a vector of parameter estimates and explanatory variables.	The logit model is conceptualised as log-odds which converts a binary outcome domain (0,1) to the real line ($-\infty, \infty$). For the logit model this index or link function is based on the logistic distribution. The error structure is assumed to be IID while explanatory variables have distribution free assumptions. Parameters are estimated using maximum likelihood.
Probit	$\text{Prob}[Y_i = 1\|x_i] = \Phi(\beta' x_i)$, where Φ is the inverse of the cumulative normal distribution and $\beta' x_i$ is a vector of parameter estimates and explanatory variables.	The link function for a probit model is the inverse of the cumulative normal distribution Φ. The explanatory variables and error structure of a probit model are assumed to be IID, which makes the model more restrictive and computationally more intensive. Parameters are estimated using the maximum likelihood. The standard probit model has a similar conceptualisation to the logit model. While the probit classifier has more restrictive assumptions, both classifiers normally produce consistent parameter estimates and have comparable predictive accuracy (Greene, 2008).
Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA)	The Bayesian linear discriminant classifier is defined as follows: $\delta_k(x) = x'\sum^{-1}\mu_k - \frac{1}{2}\mu_k'\sum^{-1}\mu_k + \log(\pi_k)$ where parameters to be estimated are: μ_k which is a class specific mean vector; $\sum$ which is a covariance matrix that is common to all *K* classes; and π_k which denotes the prior probability that a randomly chosen observation comes from the *k*th class. An observation *X* = *x* is assigned to a class where this equation is largest.	The LDA classifier assumes that the observations in the *k*th class are drawn from a multivariate normal distribution and all classes share a common covariance matrix (ie the variance is the same for all *K* classes). For QDA, predictor variables appear in the discriminant function as a *quadratic* function. Like LDA, the quadratic discriminant classifier (QDA) makes the same assumption. However, unlike LDA, QDA assumes that each class has its own covariance matrix. As a rule of thumb, LDA can lead to improve modelling performance if the sample sizes are small (and hence reducing variance is important). However, if the dataset is relatively large, QDA is preferred because this method usually fits the data better and can handle a greater range of data issues (such as nonlinearity in the data). While LDA is based on quite restrictive statistical assumptions, Greene (2008) observes that these concerns have been exaggerated in the literature and the performance differences between logit/probit classifiers and LDA are usually not exceptional. However, Greene (2008) states that the core assumption on which LDA is conceptually based is naive: that class membership of an observation will be in one class or the other – as if class membership is 'preordained'. Whereas logit/probit models assume that an observation can be in either class up to a *probability* of an event occurring, conditional on the underlying parameters.
Logit/Probit – Best Subset Selection	Let M_0 denote the null model having no predictors. (1) For *k*=1,2,...,*p* predictors, fit all $\binom{p}{k}$ models that contain exactly *k*	Two popular approaches to selecting subsets of predictors is (1) best subset selection and (2) stepwise procedures. With best subset selection, the classifier fits a separate least squares regression for each combination of *p* predictors. The classifier fits all *p* models that contain exactly one predictor, then all models that

	predictors. Pick among these $\binom{p}{k}$ models based on smallest RSS and denote it M_k. (2) Select a single best model from among $M_0,\dots,M_p$ based on AIC and BIC criterion.	contain exactly two predictors and so on. The algorithm then examines the resulting models and identifies the best model (ie the best model among the single predictor models, the best among the two predictor models etc) based on the smallest residual sum of the squares (RSS) or similarly highest R^2. For logistic models, the deviance is used instead of RSS. Deviance is defined as -2 multiplied by the maximised log likelihood function (the smaller the deviance the better the fit). The next step is to select the best final subset using a criterion (for this study we use AIC and BIC).
Logit/Probit – Backward Stepwise Model	Let M_p denote the full model containing all p predictors. (1) For k=p, p-1,...,1 predictors, consider all k models that contain all but one of the predictors in M_k for a total of k-1 predictors. (1) Pick the best among these k models and call it M_{k-1} based on smallest RSS. (2) Select a single best model from among $M_0,\dots,M_p$ based on AIC and BIC criterion.	The major limitation of best subset procedure is computation complexity which rapidly escalates for large numbers of predictors. Generally, there are 2^p models that involve subsets of p predictors (so if p=20, there are 1000,000 models to estimate). The higher search space can lead to over fitting and high variance in parameter estimates. Stepwise explores a much more restricted set of models. Backward stepwise (used for this study) begins with a model containing all parameters, then sequentially removes less useful predictors, one at a time. Stepwise models have a number of significant limitations, including potential overstatement of model-fits, biased parameters, inconsistency in model selection; and deletion of variables which potentially carry signal. However, some of these limitations are mitigated against by using out-of-sample prediction tests to evaluate overall model performance.
Logit/Probit – Penalised Models	The *elastic net* penalty (Zou and Hastie, 2005) is set out as follows: $$\sum_{j=1}^{p}\left(\alpha\left\|\beta_j\right\| + (1-\alpha)\beta_j^2\right)$$ where α is the penalty parameter, β_j are the estimate coefficients. If α = 0, we have a ridge regression penalty; if α = 1, we have a lasso penalty.	Penalised models or shrinkage methods are an alternative to subset procedures. Two popular techniques are ridge regression and the lasso. A relative new technique (elastic net) combines the strengths of both techniques. Rather than using OLS to find a subset of variables, ridge regression uses all variables in the dataset but constrain or regularises the coefficient estimates or they "shrink" the coefficient estimates towards zero for non important variables. Shrinking the parameter estimates can significantly reduce their variance with only a small increase in bias. A weakness of ridge regression is that all variables are included in the model making the model difficult to interpret. The lasso has a similar construction but the penalty forces some parameters to equal zero (so the lasso has a variable selection feature and produce parsimonious models). By setting α=0.5, this allows very unimportant variable parameters to be shrunk to zero (a kind of subset selection), while variables with small importance will be shrunk to some small (non zero value).
Logit/Probit Models with Multiple Adapting Regressive Splines (MARS).	$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \beta_{R+3} b_{R+3}(x_i) + \varepsilon_i$, which represents a cubic spline with K knots; parameters $\beta_0, \beta_1, \beta_2$ and β_{R+3} are estimated over different regions of X (ie knots); and	The standard way to extend regression functions for nonlinear relationships is to replace the linear model with a polynomial function. MARS is a more general technique. It works by dividing the range of X into R distinct regions (or splines/knots). Within each region, a lower degree polynominal function can be fitted to the data with the constraint functions join to the region boundaries through

	$b_1, b_1, \ldots, b_{R+3}$ are *basis* functions.	knots. This can provide more stable parameter estimates and frequently better predictive performance than fitting a high degree polynomial over the full range of *X*. In estimating logit and probit models with a MARS feature, we followed convention of placing knots uniformly, and using a cross validation to determine the number of knots. A limitation is that regression splines can have high variance on the outer range of the predictors (when *X* takes on very small or large values). This can be rectified by imposing boundary constraints. A further limitation is the additivity condition, hence the model is only partially non-linear.
Mixed Logit	$L_i(\eta) = \exp(\beta' x_i + \eta_i) / \sum_j \exp(\beta' x_j + \eta_j)$ Where η is a stochastic part of the error term correlated over alternative outcomes. Models of this form are called *mixed logit* because the outcome probability $L_i(\eta)$ is a mixture of logits with f as the mixing distribution. The mixing distribution is typically assumed to be continuous; meaning that η can have an infinite set of values, that are used to obtain mixed logit probability through weighted averaging of the logit formula, evaluated at different values of η, with the weights given by the density $f(\eta|\Omega)$ (Train 2003).	A highly restrictive assumption of the standard logit (and probit model) is the IID condition. It is assumed the error structure is independently and identically distributed across outcomes. The mixed logit model completely relaxes the IID condition and allows for correlated predictor variables. The key idea behind the mixed logit model is to partition the stochastic component (the error term) into two additive (i.e., uncorrelated) parts. One part is correlated over alternative outcomes and heteroscedastic, and another part is IID over alternative outcomes and firms as follows: $Y_i = \beta' x_i + (\eta_i + \varepsilon_i)$. The main improvement is that mixed logit models include a number of additional parameters which capture observed and unobserved heterogeneity both within and between firms. A limitation of mixed logit models is the distribution of random parameters is unknown and has to be assumed by the researcher. Furthermore, mixed logit models are relatively complex to interpret and are time intensive to compute. These models are only partially non-linear because of the additivity condition.
Logit/Probit – General Additive Model (GAM)	$\log\left(\frac{p(X)}{1-p(X)}\right) = f_0 + f_1(X_1) + f_2(X_2) + \cdots f_p(X_p)$, for a logistic model, and similar for probit moel.	General additive models (GAM) is a non-parametric technique for extending the linear framework by allowing non-linear smooth functions of each of the explanatory variables while maintaining the additivity condition. GAMs are estimated using a backfitting algorithm. Hence, the linear relationship between predictors ($\beta_1 X_1$) above can be replaced by a smooth nonlinear function $f_1(X_1)$. GAM is called an additive model because we calculated a separate f_j for each X_j and then sum together their collective contributions. Hence, GAM models automatically capture nonlinear relationships not reflected in standard linear models. This flexibility to allow non-parametric fits with relaxed assumptions on the actual relationship between response and predictor, provides the potential for better fits to data than purely parametric models, but arguably with some loss of interpretablity. As with MARS and mixed logit, the major limitation of GAMs is the additivity condition, which can result in many important interactions being missed.
Neural Networks	For a typical single hidden layer binary	Neural networks are sometimes described as non-linear discriminant models –

	neural network classifier, there inputs (X), one hidden layer (Z) and two output classes (Y). Derived features Z_m are created from linear combinations of the inputs, and then the target Y_k is modelled as function of the linear combinations of the Z_m, $Z_m = \alpha(\alpha_{om} + \alpha_m^T X), m = 1, \dots, M.$ $T_k = \beta_{ok} + \beta_k^T Z, k = 1, \dots, K.$ $f_k(X) = g_k(T), k = 1, \dots, K.$ Where Z = $(Z_1, Z_2, Z_3, \dots, Z_M)$, and T = $(T_1, T_2, T_3, \dots, T_K)$. The activation function $\alpha(v)$ is typically the sigmoid $\alpha(v) = \frac{1}{1+e^{-v}}$. The output function $g_k(T)$ allows a final transformation of the vector of outputs *T*. For *K*-class classification the identify function $g_k(T)$ is estimated using the *softmax* function: $g_k(T) = \frac{e^{T_k}}{\sum_{\ell=1}^{K} e^{T_\ell}}.$	essentially neural networks are a two stage regression or classification model. For typical hidden layer, there are three inputs (X), one hidden layer (Z) and two output classes (Y). Derived features Z_m are created from linear combinations of the inputs, and then the target Y_k is modelled as function of the linear combinations of the Z_m. This is the same transformation as the multinomial logit model and provides positive estimates that sum to one. The units in the middle of the network computing the derived features of Z_m are called hidden or latent units as they are not directly observable. Note that if α is the identify function, then the entire model collapses to a linear model in the inputs. Hence, a neural network can be thought of as a nonlinear generalisation of a linear model, both for regression and classification. NNs are good at dealing with dynamic nonlinear relationships. The major limitation is that backpropagational neural networks are essentially 'black boxes'. Apart from defining the general architecture of a network, the researcher has little role to play other than observe the classification performance (ie., NNs provide no parameters or algebraic expressions defining a relationship (as in regression) beyond the classifiers own internal mathematics). Further, this classifier generally has less capacity to handle large numbers of irrelevant inputs, data of mixed type, outliers and missing values. The computational scaleability (in terms of sample size and number of predictors) is also a potential limitation.
Support Vector Machines	Support Vector Systems are a solution to the optimization problem: maximise M $\beta_0, \beta_1, \dots, \beta_p, \varepsilon_1 \dots \varepsilon_n$ subject to $\sum_{j=1}^{p} \beta_j^2 = 1$, $y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$ $\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C,$ Where *C* is a nonnegative tuning parameter. *C* bounds the sum of the $\epsilon_i's$ so	Support Vector Systems (SVS) differ from conventional classification techniques such as LDA and logit/probit through the use of a separating *hyperplane.* A hyperplane divides *p*-dimensional space into two halves; where a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (ie we would expect that the larger the margin the lower the out-of-sample classification error). Classification is based on the sign of the test observation. If $\epsilon_i > 0$, then the *i*th observation is on wrong side of the margin. If $\epsilon_i > 1$, *i*th observation is on wrong side of the hyperplane. The tuning parameter *C* bounds the sum of the $\epsilon_i's$ and sets the tolerance level for misclassification. Larger values of *C* indicate larger tolerance. Importantly, *C* controls the bias-variance trade off. Higher C indicates a higher margin (ie many observations violate the margin) and so there many support vectors. In this situation, many observations are involved in determining the hyperplane, leading to low variance but high bias (the

	sets the tolerance level for misclassification. *M* is the width of the margin which we want to make as large as possible. ϵ_i are slack variables which allow observations to be on the wrong side of the margin or hyperplane (ie "soft classifier" approach).	reverse is true when the *C* is set at lower values). *Support Vector Machines* (SVM) enlarge the feature space to deal with nonlinear decision boundaries – support vector machines do this in an automatic way using various types of kernels (mainly for ease of computation). A widely kernel is the *radial kernel* which is used for this study. A major advantage of the SVM classifier is that the classification is quite robust to observations far away from the hyperplane – support vectors are based on a subset of the observations. Other techniques (such as LDA, logit and probit) are more sensitive to outliers. SVM suffers many of the limitations of NNs, particularly in terms of computational scaleability, lack of interpretability and ability to handle irrelevant inputs and data of mixed type.
Generalized Boosting/AdaBoost	The GBM classifier (and its variant, AdaBoost) is initiated through the following steps (see Schapire and Freund, 2012): 1. Train weak learner using distribution D_t. 2. Get weak hypothesis or classifier $h_t : X \rightarrow \{-1, +1\}$ 3. Select weak classifier h_t to minimise weighted error. 4. Choose $\alpha_t = \frac{1}{2} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$. Where α_t is the parameter importance assigned to the weak classifier h_t. 5. Update, for $i = 1,..., m$: $D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & if\ h_t(x_i) = y_i \\ e^{\alpha_t} & if h_t(x_i) \neq y_i \end{cases}$ Output the final hypothesis or strong classifier: $$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right).$$ where $H(x)$ is the linear combination of weak classifiers computed by Generalised	The idea behind boosting is to combine the outputs of many weak classifiers to produce a powerful overall 'voting' committee. The weighted voting is based on the quality of the weak classifiers, and every additional weak classifier improves the prediction outcome. The first classifier is trained on the data where all observations receive equal weights. Some observations will be misclassified by the first weak classifier. A second classifier is developed to focus on the trainings errors of the first classifier. The second classifier is trained on the same dataset, but misclassified samples receive a higher weighting while correctly classified observations receive less weight. The re-weighting occurs such that first classifier gives 50% error (random) on the new distribution. Iteratively, each new classifier focuses on ever more difficult samples. The algorithm keeps adding weak classifiers until some desired low error rate is achieved. More formally, generalised boosting methodology, and its main variant AdaBoost, is set out in Schapire and Freund (2012). A number of attractive features have been associated with this classifier. For instance, this classifier has been shown to be resistant to over fitting and has impressive computational scaleability in terms of the classifiers capacity to handle many thousands of predictors. This classifier is also robust to outliers and monotone transformations of variables; has a high capacity to deal with irrelevant inputs; and is better at handling data of mixed (continuous and categorical) type. Another attractive feature is that the generalised boosting classifier has some level of interpretability as the algorithm provides a ranking of variable influences and their magnitude on prediction outcomes.

	Boosting or AdaBoost. AdaBoost differs from GBM only with respect to the loss function (AdaBoost uses the exponential loss function; wheras GBM uses the Bernoulli loss function).	
Random Forests	1. For b = 1 to *B* training sets: (a) Draw a bootstrap sample $\boldsymbol{Z}^*$ of size N from the training data. (b) Grow a random forest tree T_b to the bootstrapped data by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached. (i) Select *m* variables at random from the *p* variables. (ii) Pick the best variable/split point among *m*. (iii) Split the node into two daughter nodes. 2. Output the ensemble of trees $\{T_b\}_1^B$. For a discrete outcome variable, let $\hat{C}_b(x)$ be the class prediction of the *b*th random forest tree. Then $\hat{C}_{rf}^B(x)$ = *majority vote* $\{\hat{C}_b(x)\}_1^B$.	Random forests are an improvement of the CART system (binary recursive partitioning) and bagged tree algorithms which tend to suffer from high variance (ie if a training sample is randomly split into two halves, the fitted model can vary significantly across the samples); and weaker classification accuracy. Random forests maintain advantages of CART and bagged tree methodology by de-correlating the trees and using the 'ensemble' or maximum votes approach of generalised boosting. Does not require true pruning for generalisation. As in bagging, random forests build a number of decision trees based on bootstrapped training samples. But when building these decision trees, each time a split in the tree is considered, a random sample of *m* predictors is chosen as split candidates from the full set of *p* predictors. The split is only allowed to use one of these *m* predictors. A fresh sample of *m* predictors is taken at each split and typically we choose $m \approx \sqrt{p}$ which suggests that at each split, we consider the square root of the total number of predictors (ie if 16 predictors, no more than 4 will be selected). By contract, if a random forest is built where the predictor size subset *m* = the number of predictors *p*, which simply amounts bagging. The intuition behind random forests is clear. In a bagged tree process, a particularly strong predictor in the dataset (along with some moderately strong predictors) will be used by most if not all the trees in the top split. Consequently, all the bagged trees will look quite similar to each other. Hence, the predictions from bagged trees will be highly correlated. But averaging many highly correlated quantities does not lead to such a significant reduction in error as averaging uncorrelated quantities. Random forests overcome this problem by forcing each split to consider only a subset of predictors. Therefore, on average, $\frac{p-m}{p}$ of the splits will not even consider the strong predictor and so other predictors will have more of a chance. By de-correlating the trees, the averaging process will be less variable and more reliable. If there is strong correlation among the predictors, *m* should be small. Furthermore, random forests typically do not over fit if we increase *B* (the number of bootstrapped training sets) and in practice a sufficiently large *B* should be used for the test error rate to settle down. Both random forests and generalised boosting share the 'ensemble' approach. Where the two methods differ is that boosting performs exhaustive search for which trees to split on, whereas random forest chooses a small subset. Generalised boosting grows trees in sequence (with the next tree dependent on the

		last); however random forests grow trees in parallel independently of each other. Hence, random forests can be computationally more attractive for very large datasets.