8-2018

Predicting Changes in Earnings: A Walk Through a Random Forest

Joshua Hunt
University of Arkansas, Fayetteville

Predicting Changes in Earnings: A Walk Through a Random Forest

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Business Administration with a concentration in Accounting

by

Joshua O'Donnell Sebastian Hunt
Louisiana Tech University
Bachelor of Science in Mathematics, 2007
Louisiana Tech University
Master of Arts in Teaching, 2011
University of Arkansas
Master of Accountancy, 2013
University of Arkansas
Master of Science in Statistics and Analytics, 2017

August 2018
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council.

_____
Vern Richardson, Ph.D.
Dissertation Director

_____
James Myers, Ph.D.
Committee Member

_____
David Douglass, Ph.D.
Committee Member

_____
Cory Cassell, Ph.D.
Committee Member

Abstract

This paper investigates whether the accuracy of models used in accounting research to predict categorical dependent variables (classification) can be improved by using a data analytics approach. This topic is important because accounting research makes extensive use of classification in many different research streams that are likely to benefit from improved accuracy. Specifically, this paper investigates whether the out-of-sample accuracy of models used to predict future changes in earnings can be improved by considering whether the assumptions of the models are likely to be violated and whether alternative techniques have strengths that are likely to make them a better choice for the classification task. I begin my investigation using logistic regression to predict positive changes in earnings using a large set of independent variables. Next, I implement two separate modifications to the standard logistic regression model, stepwise logistic regression and elastic net, and examine whether these modifications improve the accuracy of the classification task. Lastly, I relax the logistic regression parametric assumption and examine whether random forest, a nonparametric machine learning technique, improves the accuracy of the classification task. I find little difference in the accuracy of the logistic regression-based models; however, I find that random forest has consistently higher out-of-sample accuracy than the other models. I also find that a hedge portfolio formed on predicted probabilities using random forest earns larger abnormal returns than hedge portfolios formed using the logistic regression-based models. In subsequent analysis, I consider whether the documented improvements exist in an alternative classification setting: financial misstatements. I find that random forest's out-of-sample area under the receiver operating characteristic (AUC) is significantly higher than the logistic-based models. Taken together, my findings suggest that the accuracy of classification models used in accounting

research can be improved by considering the strengths and weaknesses of different classification models and considering whether machine learning models are appropriate.

Acknowledgements

I would like to thank my mother, Catherine Hunt, who not only taught me how to read, but also instilled in me the importance of education and cultivated my love of learning from an early age.

Table of Contents

1. Introduction

The goal of this paper is to show that accounting researchers can improve the accuracy of classification (using models to predict categorical dependent variables) by considering whether the assumptions of a particular classification technique are likely to be violated and whether an alternative classification technique has strengths that are likely to make it a better choice for the classification task. Accounting research makes extensive use of classification in a variety of research streams. One of the most common classification techniques used in accounting research is logistic regression. However, logistic regression is not the only classification technique available and each technique has its own set of assumptions and its own strengths and weaknesses. Using a data analytics approach, I investigate whether the out-of-sample accuracy of predicting changes in earnings can be improved by considering limitations found in a logistic regression model and addressing those limitations with alternative classification techniques.

I begin my investigation by predicting positive versus negative changes in earnings for several reasons. First, prior accounting research uses statistical approaches to predict changes in earnings that focus on methods rather than theory, providing an intuitive starting point for my investigation (Ou and Penman 1989a, 1989b; Holthausen and Larcker 1992). While data analytics has advanced since the time of these papers, the statistical nature of their approach fits in well with a data analytics approach. Data analytics tends to take a more statistical, results-driven approach to prediction tasks relative to traditional accounting research. Second, changes in earnings are a more balanced dataset in regard to the dependent variable relative to many of the other binary dependent variables that accounting literature uses (e.g., the incidence of fraud, misstatements, going concerns, bankruptcy, etc.). Positive earnings changes range from 40% to 60% percent prevalence in a given year for my dataset. Logistic regression can achieve high

accuracy in unbalanced datasets but this accuracy may have little meaning because of the nature of the data. For example, in a dataset of 100 observations that only have 5 occurrences of a positive outcome, one can have high accuracy (95 percent for this example) without correctly classifying any of the positive outcomes. Third, focusing on predicting changes in earnings allows me to use a large dataset which, in turn, allows me to use a large set of independent variables. Lastly, changes in earnings are also likely to be of interest to investors and regulators because of their relationship to abnormal returns (Ou and Penman 1989b; Abarbenell and Bushee 1998).

Logistic regression is the first algorithm I investigate because of its prevalent use in accounting literature. Logistic regression uses a maximum likelihood estimator, an iterative process, to find the parameter estimates. Logistic regression has several assumptions.[1] First, logistic regression requires a binary dependent variable. Second, logistic regression requires that the model be correctly specified, meaning that no important variables are excluded from the model and no extraneous variables are included in the model. Third, logistic regression is a parametric classification algorithm, meaning that the log odds of the dependent variable must be linear in the parameters.

I use a large number of independent variables chosen because of their use in prior literature.[2] This makes it more likely that extraneous variables are included in the model, violating the second logistic regression assumption. To address this potential problem, I implement stepwise logistic regression, following prior literature (Ou and Penman 1989b; Holthausen and Larcker

[1] I only discuss a limited number of the assumptions for logistic regression here. More detail is provided on all of the assumptions in the logistic regression section.
[2] Ou and Penman (1989b) begin with 68 independent variables and Holthausen and Larcker (1992) use 60 independent variables. My independent variables are based on these independent variables as well as 11 from Abarbenell and Bushee (1998).

1992; Dechow, Ge, Larson, and Sloan 2011). The model begins with all the input variables and each variable is dropped one at a time. The Akaike information criterion (AIC) is used to test whether dropping a variable results in an insignificant change in model fit, and if so, it is permanently deleted. This is repeated until the model only contains variables that change the model fit significantly when dropped.[3]

While stepwise logistic regression makes it less likely that extraneous variables are included in the model, it has several weaknesses. First, the stepwise procedure performs poorly in the presence of collinear variables (Judd and McClelland 1989). This can be a concern with a large set of independent variables. Second, the resulting coefficients are inflated, which may affect out-of-sample predictions (Tibshirani 1996). Third, the measures of overall fit, z-statistics, and confidence intervals are biased (Pope and Webster 1972; Wilkinson 1979; Whittingham, Stephens, Bradbury, and Freckleton 2001).[4]

I implement elastic net to address the first two weaknesses of stepwise logistic regression (multicollinearity and inflated coefficients). Elastic net is a logistic regression with added constraints. Elastic net combines Least Absolute Shrinkage and Selection Operator (lasso) and ridge regression constraints. Lasso is an L1 penalty function that selects important variables by shrinking coefficients toward zero (Tibshirani 1996).[5] Ridge regression also shrinks coefficients, but uses an L2 penalty function and does not zero out coefficients (Hoerl and Kennard 1970).[6]

[3] This is an example of backward elimination. Stepwise logistic regression can also use forward elimination or a combination of backward and forward elimination. I use backward elimination because it is similar to what has been used in prior literature (Ou and Penman 1989b; Holthausen and Larcker 1992; Dechow et al. 2011).
[4] Coefficients tend to be inflated because the stepwise procedure overfits the model to the data. The procedure attempts to insure only those variables that improve fit are included based on the current dataset and this causes the coefficients to be larger than their true parameter estimates. Similarly, the model fit statistics are inflated. The z-statistics and confidence intervals tend to be incorrectly specified due to degrees of freedom errors and because these statistical tests are classical statistics that do not take into account prior runs of the model.
[5] A L1 penalty function penalizes the model for complexity based on the absolute value of the coefficients.
[6] A L2 penalty function penalizes the model for complexity based on the sum of the squared coefficients.

Lasso performs poorly with collinear variables while ridge regression does not. Elastic net combines the L1 and L2 penalties, essentially performing ridge regression to overcome lasso's weaknesses and then lasso to eliminate irrelevant variables.

Logistic regression, stepwise logistic regression, and elastic net are all parametric models subject to the assumption that the independent variables are linearly related to the log odds of the dependent variable (the third logistic regression assumption). Given that increasing (decreasing) a particular financial ratio may not equate to a linear increase (decrease) in the log odds of a positive change in earnings, it is not clear that the relationship is linear. To address this potential weakness, I implement random forest, a nonparametric model. The basic idea of random forest was first introduced in 1995 by Ho (1995) and the algorithm now known as random forest was implemented in 2001 by Brieman (2001). Since then it has been used in biomedical research, chemical research, genetic research, and many other fields (Díaz-Uriarte and De Andres 2006; Svetnik, Liaw, Tong, Culberson, Sheridan, and Feuston 2003; Palmer, O'Boyle, Glen, and Mitchell 2007; Bureau, Dupuis, Falls, Lunetta, Hayward, Keith, and Van Eerdewegh 2005).

Random forest is a decision tree-based algorithm that averages multiple decision trees. Decision trees are formed on random samples of the training dataset and random independent variables are used in forming the individual decision trees.[7] Many decision trees are formed with different predictor variables and these trees remain unpruned.[8] Each tree is formed on a different bootstrapped sample of the training data.

These procedures help ensure that the decision trees are not highly correlated and reduce variability. Highly correlated decision trees in the forest would make the estimation less reliable

[7] A training data set refers to the in-sample data set used to form estimates to test on the out-of-sample data set. In my setting, I use rolling 5 year windows as training set and test out-of-sample accuracy on the 6th year.
[8] Pruning a decision tree refers to removing branches that have little effect on overall accuracy. This helps reduce overfitting.

due to the same information being available. Random forest also provides internal measures of variable importance formed from the training set. These measures are constructed by using the out-of-bag error rate from each tree that has been formed in the forest.[9]

Random forest has several advantages relative to the logistic models. First, this method tends to be an accurate classifier due to its ensemble nature.[10] Second, it performs well with a large set of independent variables, even in the presence of collinear variables, and computes variable importance measures. Third, it is a nonparametric method (i.e., it does not have distributional assumptions). The biggest disadvantage is that random forest tends to over-fit data with noisy classification (i.e. the set of independent variables does a poor job classifying the outcome variable). However, of the four models, random forest is the least restrictive and may improve out-of-sample prediction accuracy.

To predict changes in earnings, I use the change in diluted earnings per share from time t to t+1. I classify those companies that experience a future increase in earnings per share as a positive change and those that do not as a negative change.[11] I use independent variables based primarily on those variables found in Ou and Penman (1989b) and Abarbanell and Bushee (1998). I eliminate variables that are not present for at least 50% of the sample, leaving 71 input variables.[12,13] I use these inputs to predict whether earnings changes will be positive.

[9] Out-of-bag error is the mean prediction error on the training sample from the bootstrapped subsamples.
[10] Ensemble means that a model uses multiple learning algorithms. In this case, random forest uses multiple decision trees.
[11] I do not adjust for the trend in earnings as Ou and Penman (1989b) and Houlthausen and Larcker (1992) do in order to preserve the largest possible set of data. All else equal, more data leads to more robust model selection and evaluation.
[12] If all variables are required to be present for all of the sample, the sample becomes very small. I examine several cutoffs 40, 50, 60, and 70%. The 50% and lower cutoffs leave the sample and the number of variables large. Several variables are dropped because they are not available in the later years of the sample due to the inclusion of the statement of cash flows. I also examined only taking variables with at least 50% availability for later years 1995, 1999, 2000, 2005, and 2015 to examine the extent of look ahead bias. The variables left in the sample are fairly static, whether examining the entire sample or later years.
[13] I use independent variables and input variables interchangeably throughout the paper.

Following Holthausen and Larcker (1992), I rank the probabilities of changes in earnings in order to have more balanced cutoffs (i.e., I split the samples based on ranked probability cutoffs of 50/50, 60/40, 70/30, 80/20, 90/10, and 95/05). Using this methodology not only balances the top and bottom groups but keeps the number of observations consistent for each model and cutoff. Using raw probability cutoffs yields different sample sizes and unbalanced top and bottom groups.[14] I evaluate the out-of-sample accuracy of the classification models and the abnormal returns generated by trading strategies formed using the predictions from each of the models.

I find that random forest yields better out-of-sample accuracy than the three methods based on logistic regression. Interestingly, the three methods based on logistic regression perform similarly, with elastic net lagging behind logistic regression and stepwise logistic regression. The results suggest that the data may be highly complex because the penalty functions force elastic net to find a simpler model. If logistic regression cannot capture the relation between the independent variables and the outcome, then using an algorithm that forces a simpler relation will almost certainly perform worse.

Random forest has higher out-of-sample accuracy for all samples. Specifically, I find that random forest improves out-of-sample classification accuracy over the next closest model by 2.3 for the 50/50 split, 3.5 percent for the 60/40 split, 4.4 percent for the 70/30 split, 4.2 percent for the 80/20 split, 2.2 percent for the 90/10 split, and 2.1 percent for the 95/05 split.

In subsequent tests, I examine the effect that different models have on abnormal returns using the 95/05 split sample. I find that returns are 3 percent larger for random forest than for the next highest return model. This suggests that improving out-of-sample accuracy of the classification

[14] All inferences remain qualitatively similar for raw probabilities.

of changes in earnings allows investors to earn larger abnormal returns. Because the models use ratios from financial statements, this also provides evidence that financial statements continue to provide information that is not fully reflected in security prices.

I also investigate whether out-of-sample accuracy of classification models can be improved by using a novel cross-validation method. Machine learning algorithms are trained using cross-validation. I use cross-validation in this paper to find the weights for the lasso and ridge regression penalties and to find the number of input variables to use with random forest. Cross-validation allows a researcher to estimate out-of-sample accuracy rates but does not typically take time into account. The main results presented in this paper use traditional K-fold cross-validation (see the methodology section for details). I adapt rolling window, a cross-validation technique used in time-series data, and incorporate it in a pooled cross-sectional data setting. To my knowledge, this is the first paper to implement a cross-validation method that incorporates a time component in pooled cross-sectional data. I find that for a majority of the years in my sample, the out-of-sample accuracy using this cross-validation technique is more similar to the estimated out-of-sample accuracy relative to the typical K-fold cross-validation, though out-of-sample accuracy based on ranking probabilities does not improve.

In further analysis, I consider whether the documented improvements exist in an alternative classification setting: financial misstatements. I use the same algorithms as described above: logistic regression, step-wise logistic regression, elastic net, and random forest. I define misstatements as big misstatements if they are disclosed in an 8-K or 8-KA. These reissuance restatements address a material error that requires the reissuance of past financial statements. I drop all other misstatements. I classify those companies that experience big misstatement as a 1 and those that do not as a 0. I use independent variables based primarily on those variables found

in Perols, Bowen, and Zimmerman (2017). I eliminate variables that are not present for at least 25% of the sample, leaving 77 input variables. I use random forest to impute the remaining missing values.

I next implement an unsupervised variable reduction technique called variable clustering.[15] Variable clustering will find groups or clusters of variables that are highly correlated among themselves and less correlated with variables in other clusters. I then reduce the number of variables by taking those that have the highest correlation with its own cluster and the lowest correlation with other clusters, this reduces the number of inputs to 32.[16] I use these inputs to predict whether big misstatements will occur in a given year.

Because big misstatements are rare, approximately 5% in my sample, I implement three sampling techniques to help with prediction in the presence of an unbalanced dataset. I implement down-sampling, up-sampling, and SMOTE. Down-sampling balances the data set by taking a random sample of the majority class that is equal size to the less prevalent class. Up-sampling randomly samples the less prevalent class with replacement to match the size of the majority class. SMOTE down samples the majority class and synthesizes new observations for the less prevalent class. I follow Perols et al (2017) and use AUC to assess out-of-sample performance of the misstatement prediction models.

I find that random forest yields a better out-of-sample AUC (0.7462) than the three methods based on logistic regression. Interestingly, the three methods based on logistic regression perform similarly to each other, with AUC not being statistically different for the original sample at approximately 0.70. The results show that the sampling techniques do not help the logistic

[15] Unsupervised refers to an algorithm that does not consider a dependent variable.
[16] Results are qualitatively similar without using variable clustering, but computation time is greatly increased. Variable clustering was also implemented with changes in earnings with similar results.

models, in fact most of them degrade the fit. Random forest up-sampling performs as good as the original sample random forest. Random forest significantly out-performs the logistic-based models in predicting big misstatements.

I make two main contributions to the literature. First, I provide evidence that the assumptions of the logistic regression may be too restrictive in certain accounting settings and that using a nonparametric machine learning algorithm may improve out-of-sample accuracy.[17] Second, I introduce a novel cross-validation method to the machine learning area that should be of particular interest to accounting researchers due to its panel data nature. I also present a new method to accounting research for assessing the fit of binary predictions called a separation plot (Greenhill 2011). This method allows me to visualize how often high probabilities match actual occurrences and how often low probabilities match nonoccurrences.

While I focus on predicting changes in earnings and financial misstatements, improving the accuracy of classification is likely to benefit other binary outcomes examined in the accounting literature as well. These outcomes include bankruptcy and financial distress (Ohlson 1980; Beaver, McNichols, and Rhie 2005; Campbell, Hilscher, and Szilagyi 2008; Beaver, Correia, and McNichols 2012), goodwill impairments (Francis, Hannah, and Vincent 1996; Hayn and Hughes 2006; Gu and Lev 2011; Li, Shroff, Venkataraman, and Zhang 2011; Li and Sloan 2017), write-offs (Francis et al. 1996), restructuring charges (Francis et al. 1996; Bens and Johnston 2009), initial public offerings (Friedlan 1994; Pagano, Panetta, and Zingales 1998; Teoh, Welch, and Wong 1998; Brau, Francis, and Kohers 2003; Boehmer and Ljungqvist 2003; Brau and Fawcett 2006), seasoned equity offerings (McLaughlin, Safieddine, and Vasudevan 1996; Guo and Mech 2000; Jindra 2000; DeAngelo, DeAngelo, and Stulz 2009; Alti and Sulaeman 2012; Deng,

[17] Accuracy also refers to AUC for subsequent misstatement analysis.

Hrnjic, Ong 2012), and Accounting and Auditing Enforcement Releases (Dechow, Sloan, and Sweeney 1996; Beasley 1996; Beneish 1999; Erickson, Hanlon, and Maydew 2006; Dechow et al. 2011; Feng, Ge, Luo, and Shevlin 2011; Price, Sharp, and Wood 2011; Hribar, Kravet, and Wilson 2013).

2. Algorithms

2.1 Logistic Regression

Logistic regression is the most common classification algorithm in the accounting literature. Logistic regression coefficients are estimated using maximum likelihood estimation, which uses an iterative process to find coefficients that produce a number that corresponds as closely as possible to the observed outcome. Equation 1 is the formula for the maximum likelihood estimation. This method finds β such that the log likelihood is maximized.

$$\log P(y|\beta, x) = \sum_{i=1}^{m} y_i \log\left(\frac{1}{1+\exp(-x_i\beta)}\right) + (1-y_i)\log\left(\frac{\exp(-x_i\beta)}{1+\exp(-x_i\beta)}\right) \qquad (1)$$

Logistic regression does not have the same set of assumptions as ordinary least squares (OLS). First, logistic regression does not assume that error term is normally distributed. Second, it does not assume linearity between the dependent variable and the independent variables. Third, it does not assume homoscedasticity.

Logistic regression is subject to several other assumptions, however. First, the dependent variable must be a categorical variable that represents categories that are mutually exclusive and exhaustive. Second, the model should be properly specified. Related to this assumption, logistic regression performs poorly in the presence of multicollinearity and in the presence of outliers. Third, while linearity between the dependent variable and independent variables is not assumed, linearity between the log odds of the dependent variable and the independent variables is assumed. Fourth, similar to OLS, the error terms are assumed to be uncorrelated. Fifth, it is

assumed that an adequate number of observations for each category of the dependent variable are available.[18]

In my first setting, the dependent variable takes a value of one when the change in earnings from year t-1 to year t is positive, and zero otherwise, where earnings are measured as diluted earnings per share. This coding represents two mutually exclusive and exhaustive groups, satisfying the first assumption.[19,20]

Most techniques assume that the model is correctly specified, but misspecification may be a more serious problem for logistic regression (Mood 2010). Excluding relevant variables results in an omitted variable bias similar to OLS, with the added complication that this bias affects all of the independent variables even if the variable that is omitted is unrelated to the variable of interest (Wooldridge 2002; Mood 2010; Gail, Wieand, and Piantadosi 1984). Including irrelevant variables also creates a problem, depending on the correlation between the irrelevant variables and the other independent variables (Menard 2008). Specifically, the inclusion of irrelevant variables can inflate the standard errors of the irrelevant variables and those of the other independent variables that are correlated with them.

Further, misspecification relates not only to the inclusion/exclusion of variables, but also to the measurement error and multicollinearity of the variables that are included in the final model. The mismeasurement of variables induces bias in coefficient estimates. The measurement error can also come from misclassifications in the dependent variable which can lead to significant amounts of bias in coefficient estimates (Hausman 2001). Outliers are also a concern.

[18] These assumptions are broad and the ordering is not relevant. For more detailed discussions of the assumptions and how to test them, see Hosmer et al (2013) and Menard (2008).
[19] Although I am dichotomizing a continuous variable, changes in earnings, I am not interested in the magnitude of the change. That is, I don't predict large changes vs small changes. I predict a positive change in earnings and, in subsequent analyses, I examine whether that prediction is related to abnormal returns.
[20] Dichotomizing a continuous dependent variable at the median, mean, or any other cutoff results in a loss of information, which affects the power of the test and increases the false positive rate (Austin and Brunner 2004).

Similar to OLS, outliers affect the coefficient estimates and model fit, and can be assessed with traditional methods such as leverage and dfbetas (Menard 2008). Multicollinearity causes inflated standard error estimates and can be assessed using the correlation matrix and variance inflation factors (Menard 2008).

As mentioned above, the third assumption is that the parameters are linear in the logit or log odds of the dependent variable (though linearity between the dependent variable and independent variables is not assumed). Menard (2008) finds that the failure of this assumption is similar to an omitted variable and will bias coefficients. Similar to OLS, a researcher can include transformations of independent variables in order to assess whether nonlinearities exist or examine a plot of the logit against the independent variables.[21]

The fourth assumption is similar to OLS. The error terms are assumed to be uncorrelated. Correlated error terms result when data are related over time and/or space. It may also be related to mismeasurement if the data include non-random measurement error. If this assumption fails, then standard errors tend to be inflated. This assumption is not easily tested and must be considered when designing the tests. If the data have a time/space component, then error terms are not likely to be independent.

The fifth assumption is that there are an adequate number of observations for each category of the dependent variable. The most extreme form of this potential problem results in zero cells and complete separation. A zero cell occurs whenever the dependent variable is invariant for one or more levels of an independent variable. This will result in a probability of 1 or 0 for an entire group, causing high standard errors and uncertainty related to the coefficient

[21] Menard (2008) offers further discussion on the topic of detecting nonlinearity in the logit.

estimate associated with that independent variable (Menard 2008).[22] Complete separation refers to perfectly predicting the dependent variable with a given set of input variables. This can create problems even in less extreme forms, when a given set of input variables predict the dependent variable with extremely high accuracy, but not perfectly (quasi-separation). Both complete and quasi-separation can result in coefficients and standard errors being extremely large.

In this paper, I focus on the assumptions that are likely to affect the accuracy of classification. In particular, the second assumption (model specification) and the third assumption (linearity between the input variables and the logit) are likely to affect out-of-sample accuracy. While the first assumption can also affect accuracy, the binary dependent variable assumption is generally easily satisfied. Violations of the remaining assumptions can cause problems, such as inflated standard errors and misspecified test statistics but these are unlikely to affect out-of-sample accuracy, the focus of this paper.

Concerns about model specification relate primarily to the inclusion/exclusion of variables, multicollinearity, and outliers. These concerns are likely justified in my setting because of the large number of variables included in the analysis. This makes it likely that irrelevant variables are included in the model. Multicollinearity is a concern because the majority of the variables are based on common financial ratios that are likely to be related. Outliers are also a common concern when using financial data. The third assumption may not be satisfied because it isn't clear that forcing every financial ratio to be linearly related to the log odds of a

[22] The zero cell assumption only affects dichotomous and nominal variables because continuous and ordered categorical variables have an assumed distributional relationship with the dependent variable and the gaps can be estimated.

positive change in earnings is a realistic assumption (i.e., the parametric assumption may be too strong).[23] If it is not satisfied, then the effect is similar to an omitted variable bias.

2.2 Stepwise Logistic Regression

In order to address the model specification assumption, I begin with stepwise logistic regression. In my setting I start with a large set of variables, which may suffer from the inclusion of irrelevant variables.[24] Backward stepwise logistic regression begins with all of the variables included and iteratively removes the least helpful predictor (James, Witten, Hastie, and Tibshirani 2013). The Akaike information criterion (AIC) is used to test whether dropping the variable gives an insignificant change in model fit, and if so, the variable is permanently deleted.[25] This is repeated until the model only contains variables that change the model fit significantly when dropped. Hosmer, Lemeshow, and Sturdivant (2013) state that stepwise logistic regression provides an effective data analysis tool because it can provide an effective way to screen a large number of inputs in a new setting.

However, stepwise logistic has several weaknesses. First, the stepwise procedure performs poorly in the presence of multicollinear variables (Judd and McClelland 1989). The deletion of the collinear variables becomes random and it is possible to include noise variables (Hosmer 2013). This can be a concern when using a large set of independent variables. Second, the resulting coefficients are inflated, which may affect out-of-sample predictions (Tibshirani 1996). The coefficients tend to be inflated because the model is overfit to the sample data. This causes the coefficients to be high for that sample and the coefficients are biased high relative to

[23] In my setting the parametric assumption is difficult to test because the relation between input variables and target variable may change over time and I examine 45 years.
[24] Perols et al. (2017) investigate best subset selection as a method for finding relevant variables. Best subset selection may have statistical issues and overfit the data if the set of variables is large (James et al. 2013).
[25] Asymptotically, minimizing the AIC is equivalent to minimizing the error generated from cross-validation estimation (Stone 1977). Other metrics can be used to select variables such as Bayesian information criterion (BIC), pseudo R-squared, Mallows c statistic, etc.

the true parameter. Third, the measures of overall fit, z-statistics, and confidence intervals are biased. The test statistics are biased because of multiple testing and because these classical statistics tests were designed for single tests. Fourth, stepwise logistic regression does not guarantee the best model from the subset of total variables because not every combination is tested, and it proceeds with one deletion at a time. Interestingly, the residuals tend to be close to other methods that do iterate through all possible combinations (James et al. 2013).

2.3 Elastic Net

Next, I implement elastic net, a shrinkage method that is based on logistic regression, in order to address the weaknesses of stepwise logistic regression that may affect out-of-sample accuracy (multicollinearity, inflated coefficients, and selecting noise variables). Elastic net still allows the researcher to investigate associations but it should increase out-of-sample accuracy as well. Elastic net is a combination of ridge regression and lasso.[26]

2.3.1 Ridge Regression

Ridge regression and Lasso are methods that constrain coefficient estimates. Ridge regression is very similar to standard logistic regression, except that the coefficients are estimated by maximum likelihood with an added constraint, namely the square of the coefficients (James et al. 2013). Equation 2 shows how the estimation of logistic regression is related to ridge regression. Here we minimize the negative log likelihood with the added L2 constraint.

$$\log P(y|\beta, x) =$$

$$-\left(\sum_{i=1}^{m} y_i \log\left(\frac{1}{1+\exp(-x_i\beta)}\right) + (1-y_i)\log\left(\frac{\exp(-x_i\beta)}{1+\exp(-x_i\beta)}\right)\right) + \frac{\varphi}{2}\sum_{i=1}^{k}\beta_i^2 \qquad (2)$$

[26] Random forest does not allow for specific association rules to be examined.

The penalized maximum likelihood estimation includes a tuning parameter or shrinking penalty, φ, where higher values increase the penalty and lower values decrease the penalty, all while still finding the maximum likelihood. When the tuning parameter is zero then the model is a standard logistic regression, but as the tuning parameter approaches infinity the coefficients approach zero (James et al. 2013). Because ridge regression shrinks coefficients and coefficient size is dependent on their scale, the inputs must be standardized. I use a standard z-score standardization, where the independent variables are demeaned and scaled by standard deviation each year.

Standard logistic regression will have low bias but high variance in the presence of many inputs (if the distributional assumption holds). Therefore, a small change in sample may result in a large change in coefficients. Ridge regression has the benefit of reducing the variance of the models produced. That is, if the sample changes, then the model coefficients will change very little. However, ridge regression increases the bias (within an acceptable range) because it shrinks coefficients that have a small effect on the dependent variable close to zero. Ridge regression is also robust to multicollinearity due to the shrinkage penalty. Multicollinearity causes the coefficients to change wildly with small sample changes. The shrinkage function causes coefficients to be more stable while biasing them towards zero. I use cross-validation to identify the best shrinkage parameter (discussed in more detail in section 2.4).

2.3.2 Lasso

The main disadvantage of using ridge regression is that it does not select a subset of variables like stepwise logistic regression.[27] To address this, elastic net incorporates lasso in

[27] Ridge regression also performs poorly in the presence of outliers.

addition to ridge regression. Lasso is very similar to ridge regression with the exception that the penalty added to the maximum likelihood is the absolute value of the coefficients.

$$\log P(y|\beta, x) =$$

$$-\left(\sum_{i=1}^{m} y_i \log\left(\frac{1}{1+\exp(-x_i\beta)}\right) + (1-y_i)\log\left(\frac{\exp(-x_i\beta)}{1+\exp(-x_i\beta)}\right)\right) + \frac{\delta}{2}\sum_{i=1}^{k}|\beta_i| \qquad (3)$$

This is an L1 constraint, where ridge regression uses an L2 constraint. This constraint allows for coefficients to equal zero, effectively selecting the more important variables.[28] Lasso contains a tuning parameter, δ, that controls the amount of shrinkage, similar to the ridge regression. Again, I use cross-validation to identify the best tuning parameter (discussed in more detail in section 2.4).

Although lasso addresses ridge regression's main disadvantage by reducing the number of variables, it has weaknesses of its own. If the number of variables is greater than the size of the sample (i.e., a large number of variables but a small sample size n), the number of variables that lasso will select is limited by the size of the sample. This is usually not an issue in accounting research given the typically large data sets used. Lasso also performs poorly in the presence of multicollinearity. If there is a group of multicollinear variables, lasso tends to select one from the group and ignore the rest.

2.3.3 Elastic Net

Elastic net is designed to address many of the weaknesses of ridge regression and lasso. Elastic net uses both the L1 and L2 shrinkage constraints (Zou and Hastie 2005). This allows for the strengths of each of the two methods (ridge regression and Lasso) to overcome the

[28] For a detailed discussion of why the L1 penalty results in zeroed out coefficients and the L2 does not, see James et al 2013. The geometric explanation is that the absolute value is not a smooth function and when the optimum coefficient is found it can be at the peak of the function allowing for zero coefficients.

weaknesses of the other. The ridge regression penalty addresses multicollinearity and the lasso penalty eliminates nonessential variables.

$$\log P(y|\beta, x) =$$

$$-\left(\sum_{i=1}^{m} y_i \log\left(\frac{1}{1+\exp(-x_i\beta)}\right) + (1-y_i)\log\left(\frac{\exp(-x_i\beta)}{1+\exp(-x_i\beta)}\right)\right) + \frac{\varphi}{2}\sum_{i=1}^{k}\beta_i^2 + \frac{\delta}{2}\sum_{i=1}^{k}|\beta_i| \qquad (4)$$

Elastic net is subject to the basic assumptions of the logistic regression. The main weakness is the parametric assumption present in the logistic regression. It also requires that the variables be standardized. The algorithm shrinks coefficients and if the coefficients do not have the same scale then it will perform poorly.

2.4 Cross-Validation

I use cross-validation to identify the two tuning parameters for elastic net (φ and δ). Cross-validation is a resampling technique. Resampling techniques such as cross-validation and bootstrapping are useful when forming an estimate of the implementation error rate and when adjusting tuning parameters.

In order to describe cross-validation, first consider a traditional validation approach that uses a simple random data split of 60-40, where 60% is the training sample or training set and 40% is the out-of-sample or hold out set. The machine learning methods are fit to the training set and their respective fits are assessed on the hold out set. This traditional validation method suffers from two main drawbacks. First, the out of sample error rate can be highly variable because of the random 60-40 split. If the same methods are performed on a different random 60-40 split, the out of sample error rate can be quite different. Second, the original complete data set is subset to form two smaller data sets. Because statistical methods tend to perform worse on smaller datasets, holding all other factors constant, the estimated error rate tends to overestimate the implementation error rate (James et al. 2013).

Cross-validation addresses the two weaknesses of a traditional validation method. K-fold cross-validation divides the training sample into k non-overlapping random samples.[29] It then uses each of the k samples as the hold out sample set and uses the other k-1 samples to fit the model. The hold out sample error rate is averaged over the k hold out sets as tuning parameters are investigated. The final model that is selected is validated using the original complete sample. The advantage to k-fold cross validation is that all of the observations are used for the training and hold out sets, and each observation is used exactly once for the hold out set. The biggest disadvantage is that each statistical method must be run from scratch k times, which increases the computational burden.

I use five-fold cross-validation for my main tests. Each training data set includes a five-year period. Five random samples are drawn from each training data set and four of the five random samples are used to identify the optimal weights of the elastic net penalty functions. The weights of the penalty functions are randomly generated and tested on the fifth random sample and the accuracy for each random weight is measured. This process is completed four more times using a different random sample each time but using the same initial weights. The test sample accuracy is averaged over each of the five folds and the random weights that produce the highest accuracy are chosen.[30] The model is then run on the entire training sample with the chosen weights. This model is used to form the probability of a positive change in earnings.

2.5 Random Forest

While elastic net addresses several weaknesses of logistic regression, it still assumes that independent variables are linear in the logit, which may be an inaccurate assumption. I address this potential weakness by implementing random forest, a nonparametric model. In order to

[29] K-fold cross-validation and cross-validation refer to the same technique.
[30] Other metrics can be used to select the best tuning parameter such as area under the ROC or specificity.

describe random forest, I begin by explaining the components of the model: decision trees and tree bagging.

Decision trees are a set of binary splits. Each split creates an internal node or step that represents a value of one of the input variables. For example, the root node may be the size of a company with the condition that if total assets are greater than 10 million, then split. From this node, it may split again if cash flows are greater than 4 million, and so on. This is a greedy process and is recursive, meaning that it continues to split the data.[31] The first split is based on purity or how well the split separates the data into distinct classes. Every variable and every possible split is considered until the split with the highest purity is found. This happens at each node and continues until a stopping criteria is reached (James et al. 2013).[32] New observations are classified by passing down the tree to a terminal node or leaf.

Decision trees have several strengths. First, because they are nonparametric, there are no distributional assumptions. Second, if the trees are small, then they are easily interpreted. Third, decision trees are robust to outliers and collinear variables. Fourth, they can handle missing data. The main disadvantage is high variability, meaning that a small change in the sample can cause a large change in the final tree (James et al. 2013). This disadvantage leads to the decision tree being a poor classifier. Decision trees tend to overfit the training data and perform poorly out-of-sample.

Tree bagging helps decision trees overcome this weakness. Bagging is a bootstrap aggregation method and is a general purpose tool in machine learning used to reduce model variance. If the prediction method has a lot of variance, then bagging can improve accuracy

[31] A greedy algorithm solves for a local optimum with the hope of finding a global optimum. In the case of decision trees, it finds a variable that forms the best split but does not consider future splits.
[32] I do not have a stopping criteria. The trees are allowed to grow as large as possible.

(Breiman 1996). This fits particularly well with decision trees, but can also be applied to other methods. Tree bagging forms decision trees on bootstrapped samples (with replacement) taken from the complete training data set. This allows for different trees to form on each sample. The trees are then averaged (i.e., the classification is accomplished by majority vote).[33] Tree bagging improves classification by reducing the variance, but at the cost of losing the simple tree structure. The bootstrapped samples help ensure that the trees are different, forming a better average. However, tree bagging becomes less effective when the trees are very similar (James et al. 2013).

Random forest addresses this weakness by forming less similar trees. Random forest takes tree bagging one step further by randomly choosing a subset of input variables at each decision tree split. This is done for each tree grown on a bootstrapped sample. For example, if the chosen number of input variables is four, then four variables are chosen at random at each split of the decision tree. The number of variables to be chosen is a tuning parameter. Similar to the other models, I use cross-validation to choose the best tuning parameter for random forest. Specifically, I try a random set of possible numbers limited only by the total number of variables available and choose the number that produces the best cross-validation accuracy.

Random forest tends to be an accurate classifier due to its ensemble nature. Ensemble methods combine the results from different models and can perform better than each of the individual models. Tree bagging is also an ensemble method with the weakness that the combination of multiple trees is moot if the trees are correlated. Because random forest uses random variables at each split, the resulting trees are not highly correlated by construction. Random forest inherits the strengths from decision trees in that it performs well in the presence

[33] To classify a new observation the observation is run down every tree in the forest. Each tree has a vote on whether the outcome is positive or not. The forest chooses the outcome that has the most votes.

of outliers and highly correlated variables. It also performs well with a very large set of predictor variables and computes variable importance measures. The importance of a variable is estimated using the mean decrease in node impurity (i.e., the important variables aide the most in classification). Random forest and other tree methods also do not require any variable transformations, unlike many other machine learning algorithms, including elastic net. Random forest can be applied to data sets with missing data, can be used to find outliers, and can be used to find natural clusters in the data.[34]

Random forest also has weaknesses. Random forest will over-fit data with noisy classification (i.e., the set of input variables does a poor job classifying the outcome variable). Its greatest strength can also be a weakness. Random forest is nonparametric. This allows for complex relationships between the input variables and outcome. Splits are performed on single input variables rather than combinations of input variables and trees can miss relationships, particularly those that logistic regression may capture (Shmueli, Patel, and Bruce 2010).

Logistic regression will outperform nonparametric models, including random forest, if the logistic regression assumptions hold. However, if the parametric assumption fails, then random forest will outperform logistic regression-based models. In sum, random forest is robust to common logistic regression weaknesses and less restrictive in its distributional assumptions and likely to outperform logistic regression-based models in certain settings.

3. Data and Methods

I use independent variables based primarily on variables found in Ou and Penman (1989b) and Abarbanell and Bushee (1998). Ou and Penman (1989b) include levels, changes,

[34] For a detailed discussion of what all random forest offers, see https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

and percent changes of financial ratios, but I only include levels and changes.[35] The sample period is between 1965 and 2014 inclusive.[36] In order to preserve the sample, all of the Ou and Penman (1989b) and Abarbenell and Bushee (1998) variables that were not present for at least 50% of the sample are not included. Out of 96 independent variables this left a total of 71 independent variables.[37] All of the variables are constructed from Compustat. Each model is run at the largest available sample that meet the above conditions, which leaves a sample of 101,905 company year observations. The sample consists of December year end firms that have the probabilities available as well as CRSP data, leaving 41,094 company year observations (Ou and Penman 1989a, 1989b; Holthausen and Larcker 1992; Abarbenell and Bushee 1998).[38]

I use five-year rolling windows as my training sample to predict changes in earnings for the out-of-sample sixth year. For example, my first training sample is 1965-1969 inclusive and I use this sample to predict 1970. The out-of-sample accuracy obtained in 1970 is the metric of interest. Each year the window is rolled forward.[39] I use all 71 input variables for each model. The dependent variable takes a value of one when the change in earnings from year t-1 to year t is positive, and zero otherwise, where earnings are measured as diluted earnings per share. Following Holthausen and Larcker (1992), I rank the probabilities of changes in earnings in order to have more balanced cutoffs, i.e. I rank probabilities for each model and split the sample based on 50/50, 60/40, 70/30, 80/20, 90/10, and 95/05, effectively making percentiles. The 50/50 split halves the dataset and the 95/05 split takes the top 5 percent and bottom five percent of the

[35] I only include levels and changes because elastic net requires that the independent variables be standardized and standardizing a percent is nonsensical, but I want each model to contain the same variables. This leads me to include only levels and changes of Ou and Penman (1989) variables.
[36] The data is too sparse to begin my sample earlier than 1965.
[37] See appendix for variable definitions.
[38] I also require companies to have a stock price at the end of year greater than or equal to five dollars.
[39] Ou and Penman 1989a, 1989b, and Holthausen and Larcker 1992 use five year blocks. For my time period that means using 1965-1969 inclusive to predict 1970-1974 inclusive and rolling the block forward. In untabulated results every model performs significantly worse with five year blocks relative to what is presented in the paper.

probability. Using this methodology not only balances the top and bottom groups, but keeps the number of observations consistent for each model and cutoff. Using raw probability cutoffs yields different sample sizes and unbalanced top and bottom groups.[40]

4. Results

4.1 Main Analyses

The focus of this section is maximizing out-of-sample accuracy. Table 1 presents the total number of observations per data split and the accuracy for each split and model. Logistic regression, stepwise logistic regression, and elastic net present nearly identical results for the first 3 splits. Stepwise logistic regression begins to meaningfully outperform logistic regression at the 90/10 split and the 95/05 split (by 1.2 percent and 1.7 percent, respectively). Interestingly, elastic net performs similarly to logistic regression for the first four splits but underperforms logistic regression for the 90/10 split and the 95/05 split (by 2.3 percent and 3.6 percent, respectively). Remember that elastic net is a logistic regression with two added constraints. Since logistic regression and stepwise logistic regression perform better than elastic net, this may indicate that the data is complex and that using logistic regression is not sufficiently capturing the pattern of the data. If this is the case, then the models are underfitting the data and adding constraints makes the problem worse.

The logistic regression based models are very similar in terms of accuracy for the first three splits. Addressing potential failed assumptions does not improve out-of-sample accuracy within the parametric models. Loosening the distributional assumption with random forest, however, shows an improvement over all of the parametric models (the improvement over logistic regression is as large as 4.4 percent). Random forest performs better (in terms of out-of-

[40] All inferences remain qualitatively similar for raw probabilities of future changes in earnings.

sample accuracy) for the whole sample in all splits. Because logistic regression will perform better than random forest when the distributional assumption holds, this suggests that the parametric assumption implicit in the logistic regressions may be too strong in this setting. Random forest is able to better capture the more complex relations between the input variables and the output variable.

In table 2, I examine out-of-sample accuracy for the 95/05 split in different five-year time periods. I look at five-year time periods beginning with 1970-1974 and ending with 2010-2014, inclusive. Random forest has the highest accuracy for 6 of the 9 time periods. Stepwise logistic regression has the highest accuracy in 1970-1974, 1980-1984, and 1995-1999. Interestingly, the logistic models perform very similarly in all time periods except 2005-2009. This suggests that the differences between the logistic regression-based models in table 1 may be largely due to the 2005-2009 time-period. Stepwise outperforms random forest in 3 time periods (1970-1974, 1980-1984, and 1995-1999), which may indicate that the complexity of the relation between input variables and the outcome variable changes over time. Random forest is consistently the most accurate from 2000 through 2014, the most recent 15 years. This time period includes the dotcom bubble and the financial crisis, which may be why a model that can handle more complex relationships outperforms the other models. The highest accuracy overall accuracy is 79.1 percent during the 2005-2009 time-period.

Table 3 investigates which input variables are most important. Table 3 presents the ten input variables chosen most often for stepwise, elastic net, and random forest, and presents the number of years that each respective variable is chosen (45 is the largest possible number of years). Because random forest outperforms the logistic models, it arguably chose best. Random forest chose current year earnings and effective tax rate for every year and capital expenditures

44 times. Elastic net chose capital expenditures, change in sales scaled by total assets, and net income scaled by total assets every year. The most frequent variable selected by stepwise logistic regression is net income scaled by total assets. Elastic net has three input variables in common with random forest: capital expenditures, change in inventory scaled by total assets, and current year earnings. Interestingly, stepwise logistic regression did not have any variables in common with random forest.

4.2 Additional Analyses

4.2.1 Abnormal Returns

Though out-of-sample accuracy is the focus of this paper, following prior literature that classifies earnings changes, I also investigate the abnormal returns that can be earned using these methods for the 1970-2014 time period (Ou and Penman 1989b; Holthausen and Larcker 1992; Abarbenell and Bushee 1997). The data corresponds with the accuracy results. Trading begins four months after fiscal year-end (i.e., when current-year results would be widely available for all firms). I present size adjusted abnormal returns held for 12 months. I examine abnormal returns from the 95/05 split for each model because abnormal returns are most likely to be found in the extremes of the distribution.

Table 4 presents the abnormal returns results. Panel A presents results using logistic regression, Panel B presents results using stepwise, Panel C presents results using elastic net, and Panel D presents results using random forest. Each panel includes the hedge portfolio return as well as the abnormal returns generated by subsets of the sample: observations predicted positive (PP), those predicted negative (PN), true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The number of observations that fall in each of these categories is presented in the fourth column within each respective panel.

Table 4 also presents fit metrics in the lower half of each panel. Accuracy is the main metric of interest in this paper but other fit metrics may provide insight into what affects accuracy. Kappa is a measure of how well the classifier performed as compared to how well it would have performed simply by chance. Kappa is not sensitive to class unbalance and can be compared across models. A kappa of 0 corresponds with 50 percent accuracy and a kappa of 100 corresponds with 100 percent accuracy.[41] Sensitivity is also called the true positive rate and recall. Sensitivity measures the proportion of 1's that are correctly classified. Specificity is also called the true negative rate and it measures the proportion of 0's that are correctly classified. Prevalence is a measure of how often 1's occur in the sample. Detection rate is the ratio of true positives to the total number of observations. The detection prevalence is the ratio of predicted positives to the total number of observations.

Logistic regression and stepwise logistic regression perform similarly in terms of the hedge return (14.4 percent and 14.2 percent, respectively) though stepwise outperforms logistic regression for all performance metrics. Elastic net performs the worst with an abnormal return of 5.1 percent. In light of the results presented in table 2, this may be because of poor performance in the 2005-2009 period. The relatively low abnormal return generated using elastic net is likely due to false negatives, which are much larger in number than the other methods. Random forest performs the best both in terms of the hedge return and the performance metrics. The hedge return is 17.4, 3.2 percent higher than logistic regression. It outperforms all other models both for accuracy and kappa. Specificity is particularly large for random forest relative to the other models at 76.2 percent. It classifies the true negatives at a much higher rate than the other models, with the next highest being stepwise logistic regression at 73.2 percent. The difference

[41] For a detailed discussion of kappa, see Landis and Koch (1977).

in the hedge return appears to be primarily driven by the lower return for predicted negatives (-9.9 percent for random forest versus -6.7 percent for logistic regression).

4.2.2 Incorporating time into Cross-validation

Next, I investigate whether incorporating time into cross-validation in a pooled cross-sectional data setting improves expected accuracy estimation. Because of the time series nature of the data, the k-fold process can be adapted to include a time component. I accomplish this by setting the five training sets to include only the first four years of each five year rolling window and the five hold out sets to include only the fifth year (rolling window cross-validation). This allows me to simulate true implementation conditions during the training phase.

It is an empirical question as to whether rolling window cross-validation will improve the accuracy expectation relative to traditional cross-validation. Traditional cross-validation does not take the order in which the observations occur into account. It takes random samples from the training set to form its k-folds. By forcing the test fold to be the fifth year in the k-fold process, I incorporate a time component in the assessment of the accuracy of the models. I take the accuracy generated during the rolling window cross-validation and compare it to the out-of-sample accuracy. If the relation between the input variables and the outcome are more or less time invariant, then cross-validation should produce a good estimate of expected accuracy. However, if the relation changes over time, then incorporating time into cross-validation could improve the estimate of expected accuracy. I use random forest to discuss the expected accuracy results and present the difference in expected accuracy produced by both methods.

Table 5 presents in-sample accuracy and out-of-sample accuracy for traditional cross-validation (CV) and for rolling window cross-validation (RWCV) for a 50/50 split random forest model. Table 5 also presents the difference between in-sample and out-of-sample accuracy for

each of the two validation methods. The last column of table 5 presents the comparison between CV and RWCV. The column compares the absolute value of the in-sample versus out-of-sample difference for CV to the absolute value of the in-sample versus out-of-sample difference for RWCV. The method that produces the smaller absolute difference is the superior model in that year ("Smaller" indicates that CV outperformed RWCV while "Larger" indicates that CV underperformed RWCV). The results show that RWCV outperforms CV in 33 out of 45 years. RWCV likely performs worse when the fifth year of the window is very different from the following year. For example, RWCV performs worse during the dot.com bubble and following the financial crisis in 2009.

Interestingly, the improved accuracy expectation does not translate into higher out-of-sample accuracy for the 95/05 split. Table 6 presents the accuracy for five year groups for the 95/05 split for random forest. Cross-validation and rolling window cross-validation produce similar out-of-sample accuracy figures. RWCV is higher for only two groups, 1990-1994 and 1995-1999, for the 95/05 split. Rolling window cross-validation outperforms traditional cross-validation in terms of accuracy expectation for the 50/50 split, but not in terms of out-of-sample accuracy at the 95/05 split.

4.2.3 Separation Plots

Next, I present separation plots to assist in analyzing the earnings change data. Separation plots allow users to see the predicted probabilities and the number of instances the actual 1's and 0's occur. Figure 1 shows the separation plot for random forest formed using traditional cross-validation. The gray color represents the 1's and the white color represents the 0's. Moving from left to right along the x-axis should correspond with more occurrences of 1's. The y-axis presents the raw probability of a positive earnings change in year t+1. The black line represents the raw

probability associated with each observation ordered from lowest probability to highest probability. An ideal separation plot would be white towards the left of the graph and get increasingly gray towards the right. Figure 1 indicates that most of the raw probabilities are between 40 and 80 percent.

Figure 2 presents the separation plot for random forest formed using RWCV. RWCV appears to tighten the distribution of raw probabilities relative to traditional cross-validation. RWCV also shows more white color towards the left of the graph, suggesting a better fit. This is consistent with the table 5 results.

Figure 3 is a separation plot prepared using ranked probabilities for random forest CV and follows the results presented this paper. The black line represents the rank of raw probability for each observation and is straight by construction. The overall darker right side of figure 3 (relative to figure 1) indicates that ranked probabilities perform better than raw probabilities.

Figure 4 is a separation plot prepared using ranked probabilities for random forest RWCV. Consistent with the results from table 6, comparing figure 3 and figure 4 indicates that CV performs better than RWCV, particularly in the extremes. The overall darker right side of figure 3 (relative to figure 4) indicates that CV performs better (in terms of accuracy) than RWCV.

Greenhill, Ward, and Sacks (2011) describe three main advantages to using separation plots. First, they allow for the actual 1's and 0's to be observed. Second, they allow for the range of the predicted probabilities to be visualized. Third, they allow for the relation between predicted probabilities and actual data to be visualized (i.e., probabilities of 1's relative to actual 1's). These plots are applicable can be used in any binary classification setting and can be compared across models.

4.3 Additional Misstatements Analysis

4.3.1 Data and Methods

Next, I investigate whether the documented improvements exist in an alternative classification setting: financial misstatements. I use independent variables based on variables found in Perols et al (2017). Perols et al. (2017) predict the occurrence of fraud or Accounting and Auditing Enforcement Releases (AAERs) and draw their inputs from other related literature that predicts AAERs (Cecchini, Aytug, Koehler, and Pathak 2010; Dechow et al. 2011; and Perols (2011). The sample period is between 2004 and 2014 inclusive. In order to preserve the sample, all of Perols et al. (2017) variables that are not present for at least 25% of the sample are not included.[42] After eliminating variables based on missing observations, I am left with 85 independent variables.[43] I follow Perols et al. (2017) and impute missing values with mean and mode for continuous and dichotomous variables respectively. I run the model at the largest available sample that meet the above conditions, which leaves a sample of 60,873 company year observations.

I use five-year rolling windows as my training sample to predict restatements for the out-of-sample next year. For example, my first training sample is 2000-2004 and I use this sample to predict 2005.[44] The out-of-sample AUC obtained in 2005 is the metric of interest. Each year the window is rolled forward. I follow Perols et al. (2017) in using AUC as my metric of interest. Because the dataset is unbalanced, with restatement occurring approximately 2.26% of the time, using accuracy would be inappropriate. Using my dataset, I could achieve over 97.74% accuracy by guessing no restatement will occur every time, but I would miss every time a restatement did

[42] Perols et al. (2017) states that imputation is not advised if more than 25% of the observations are missing per independent variable.
[43] See appendix for variable definitions.
[44] I use 5-fold cross-validation and use AUC to tune algorithms within the cross-validation.

occur. I define a big restatement as those that are filed in 8-K's. These reissuance restatements address a material error that requires the reissuance of past financial statements. The dependent variable takes a value of one when there is a current big restatement and zero otherwise.

Because the sample may be unbalanced, I examine three sampling techniques designed to help machine learning algorithms in the presence of rare events: down-sampling, up-sampling, and SMOTE. Theoretically, unbalanced data should not affect the logistic models as long as there are enough observations of the less prevalent class. The maximum likelihood estimation suffers from small-sample bias. This bias is strongly dependent on the number of observations in the less prevalent class. In my setting there are 1,622 cases of big restatements, which should be enough for estimation. However, 2.26% prevalence may be a rare event in the case of random forest.

Each of the sampling methods aim to make the training dataset more balanced in order for the machine learning algorithms to perform better. Unbalanced datasets tend to cause machine learning algorithms to perform well at predicting the majority class, but suffer at predicting the minority class. Down-sampling and up-sampling are essentially opposites of each other. Down-sampling balances the data set by taking a random sample of the majority class that is equal size to the less prevalent class. Up-sampling randomly samples the less prevalent class with replacement to match the size of the majority class. These approaches are less sophisticated than other approaches that are used for balancing datasets, while SMOTE sampling is a more sophisticated sampling method. SMOTE down samples the majority class and synthesizes new observations for the less prevalent class (Chawla, Bowyer, Hall, and Kegelmeyer 2002). SMOTE uses a nearest neighbor approach to synthesize the observations. SMOTE finds observations that are close to one another (nearest neighbors) in the feature space and takes the difference between

these neighbors and multiplies that difference by a random number between 0 and 1 to generate the synthetic observations. The number of neighbors, the amount of down-sampling, and the amount of new observations can be chosen by the researcher.[45]

4.3.2 Results

The focus of this section is maximizing out-of-sample AUC. Table 7 presents each model with each sampling method with its corresponding AUC. Logistic regression, stepwise logistic regression, and elastic net present nearly identical results. The logistic regression based models perform more poorly when the training sample is balanced using sampling methods. This result corresponds with conventional wisdom on logistic regression. The AUC for logistic original sample is 0.6998, with down-sampling at 0.5852, up-sampling at 0.6115, and SMOTE at .05992. The AUC for stepwise logistic original sample is 0.7026, with down-sampling at 0.5852, up-sampling at 0.6111, and SMOTE at 0.5992. The AUC for elastic net original sample is 0.7030, with down-sampling at 0.5914, up-sampling at 0.6162, and SMOTE at 0.5992. Random forest original sample performed significantly better with an AUC of 0.7462. Random forest up-sampling had a fit that was not statistically different from the original sample fit, 0.7458, while the other two were significantly worse with down-sampling at 0.6621 and SMOTE sampling at 0.6903. This may be an indication that although the dataset appeared to be unbalanced it was not an issue for random forest.

Figures 5, 6, and 7 present the separation plots for logistic regression, random forest original sample and random forest up-sampling. This provides a more complete picture of the out-of-sample predictions. Figure 4 shows that although the logistic regression has an AUC of 0.6998, there are quite a few observations on the left half of the graph and the probability

[45] I use 5 nearest neighbors and 200% down-sampling and synthesizing.

distribution is very tight. Looking at figures 6 and 7 show that although the AUC's are not statistically different, the probability distribution for random forest up-sampling is almost double that of random forest original sample fit. Both of the random forest separation plots show that there are more dark lines on right side of the plot than logistic regression, suggesting that the probabilities map better to realized occurrences of big restatements.

In table 8, I examine out-of-sample AUC in different five-year time periods. I look at five-year time periods beginning with 2000-2004 and ending with 2010-2014, inclusive. The period 2000-2004 has 33,841 company-year observations with 2,829 occurrences of big restatements, 2005-2009 has 29,895 company-year observations with 1,245 occurrences of big restatements, and 2010-2014 has 27,669 company-year observations with 377 occurrences of big restatements. Big restatements are declining a throughout the full sample period. Random forest has the highest AUC for all of the time periods and methods. Random forest original sample fit continues to be the best performing model and random forest up-sampling is the second best model. The logistic regression based models all perform very similarly in all time periods. The highest AUCs for all of the models original fits occur during the 2010-2014 time-period.

Table 9 investigates which input variables are most important for predicting big restatements. Table 9 presents the ten input variables chosen most often for stepwise, elastic net, and random forest, and presents the number of years that each respective variable is chosen (15 is the largest possible number of years). Because random forest outperforms the logistic models, it arguably chooses best. Random forest chose return on assets, ppe (property, plant, and equipment) scaled by assets, and long-term debt scaled by common equity for every year. Elastic net chose total accruals scaled by assets, sales growth, and demand for financing 12 of the 15 years. The most frequent variable selected by stepwise logistic regression is receivables scaled

by sales at 11 years. Elastic net has two input variables in common with random forest: gross and change in return on assets. Stepwise logistic regression has 3 variables in common with random forest: gross, financing, and total accruals scaled by assets.

5. Conclusion

The goal of this paper is to show that accounting researchers can improve the accuracy of classification (using models to predict categorical dependent variables) by considering whether the assumptions of a particular classification technique are likely to be violated and whether an alternative classification technique has strengths that are likely to make it a better choice for the classification task. I show that considering a model's weaknesses and addressing those weaknesses can yield increased accuracy. I find that greater out-of-sample accuracy can be obtained from using a nonparametric model that is less restrictive than logistic regression-based models. Random forest outperforms logistic regression, stepwise logistic regression, and elastic net in predicting changes in earnings. Random forest also earns three percent larger hedge returns than the next closest model. My evidence suggests that logistic regression-based models underfit the data in my setting.

I also examine model performance for different time periods. Although elastic net seems to lag behind the other logistic-based models, examining the performance in different time periods suggests that elastic net experiences most of its poor performance in the 2005-2009 time period. Otherwise, the logistic models perform similarly. I also find that although random forest performs better over the entire sample period, stepwise logistic regression outperforms random forest in three of the nine time periods examined. This suggests that the relation between the input variables and output variable changes over time. Random forest consistently outperforms the other models in the most recent 15 years.

I also find that current year earnings, effective tax rate, and capital expenditures are the most important input variables for random forest when predicting changes in earnings. Elastic net chose three input variables from its top ten list of important variables in common with random forest's top ten list. Stepwise logistic regression did not have any input variables in common with random forest's top ten list.

In additional analysis, I investigate a novel cross-validation method that incorporates a time component. This rolling window method is similar to time-series cross-validation, but it is implemented in a pooled cross sectional data sample. To my knowledge, this is the first paper to examine a different cross-validation method in an accounting setting that incorporates a time component. I find that rolling window cross-validation outperforms traditional cross-validation for a majority of the sample years. However, this does not translate to higher out-of-sample accuracy for the 95/05 split of the ranked probabilities.

I also find that greater out-of-sample AUC can be obtained from using a nonparametric model that is less restrictive than logistic regression-based models. Random forest outperforms logistic regression, stepwise logistic regression, and elastic net in predicting financial restatements. Random forest continues to outperform the logistic regression based models for each of the different time periods examine, with the latest time period (2010-2014) showing the best out-of-sample AUC at 0.7452. Random forest finds that variables return on assets, ppe scaled by assets, and long-term debt scaled by common equity are the most important variables for predicting financial restatements.

While I only investigate one nonparametric method, others would likely also be useful in this setting. I use random forest because it is easily understood relative to other machine learning methods and it does not require any data preparation. Better performance might be achieved

from another nonparametric method, but I leave that to future research. I only examine one method of incorporating time into cross-validation and other time-series methods exist that might improve expected accuracy estimation. I believe that this is an important topic for accounting researchers and should be examined more closely. While I focus solely on predicting changes in earnings and financial restatements, the improved accuracy of these models could benefit other binary outcomes examined in the accounting literature as well.

References

Abarbanell, J. S., and B. J. Bushee. 1997. Fundamental analysis, future earnings, and stock prices. *Journal of Accounting Research* 35(1), 1-24.

Abarbanell, J. S., and B. J. Bushee. 1998. Abnormal returns to a fundamental analysis strategy. *Accounting Review* 73(1), 19-45.

Alti, A., and J. Sulaeman. 2012. When do high stock returns trigger equity issues? *Journal of Financial Economics* 103(1), 61-87.

Austin, P. C., and L. J. Brunner. 2004. Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. *Statistics in Medicine* 23(7), 1159-1178.

Beasley, M. 1996. An empirical analysis of the relation between the board of director composition and financial statement fraud. *The Accounting Review* 71(4), 443-465.

Beaver, W. H., M. Correia, and M. McNichols. 2012. Do differences in financial reporting attributes impair the predictive ability of financial ratios for bankruptcy? *Review of Accounting Studies* 17(4), 969-1010.

Beaver, W. H., M. McNichols, and J. W. Rhie. 2005. Have financial statements become less informative? Evidence from the ability of financial ratios to predict bankruptcy. *Review of Accounting Studies* 10(1), 93-122.

Beneish, M. 1999. The detection of earnings manipulation. *Financial Analysts Journal* 55(5), 24-36.

Bens, D. and R. Johnston. 2009. Accounting discretion: Use or abuse? An analysis of restructuring charges surrounding regulator action. *Contemporary Accounting Research* 26(3), 673-699.

Boehmer, E. and A. Ljungqvist. 2004. On the decision to go public: Evidence from privately-held firms. Working paper, Texas A&M University.

Brau, J. and S. Fawcett. 2006. Initial public offerings: An analysis of theory and practice. *The Journal of Finance* 90(1), 399-436.

Brau, J., B. Francis, and N. Kohers. 2003. The choice of IPO versus takeover: Empirical evidence. *Journal of Business* 76(4), 583-612.

Breiman, L. (1996). Bagging predictors. *Machine learning* 24(2), 123-140.

Breiman, L. (2001). Random forests. *Machine learning* 45(1), 5-32.

Bureau, A., J. Dupuis, K. Falls, K. L. Lunetta, B. Hayward, T. P. Keith, and P. Van Eerdewegh. 2005. Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology* 28(2), 171-182.

Campbell, J. Y., J. Hilscher, and J. Szilagyi. 2008. In search of distress risk. *The Journal of Finance* 63(6), 2899-2939.

Cecchini, M., H. Aytug, G. J. Koehler, and P. Pathak. 2010. Detecting management fraud in public companies. *Management Science* 56(7), 1145-1160.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic minority oversampling technique. *Journal of artificial intelligence research* 16, 321-357.

DeAngelo, H., L. DeAngelo, R. M. Stulz. 2010. Seasoned equity offerings, market timing, and the corporate lifecycle. *Journal of Financial Economics* 95(3), 275-295.

Dechow, P., R. Sloan, and A. Sweeney. 1996. Causes and consequences of earnings misstatement: An analysis of firms subject to enforcement actions by the SEC. *Contemporary Accounting Research* 13(1), 1-36.

Dechow, P., W. Ge, C. Larson, and R. Sloan. 2011. Predicting material accounting misstatements. *Contemporary Accounting Research* 28(1), 17-82.

Deng, X., E. Hrnjic, and S. Ong. 2012. Investor sentiment and seasoned equity offerings. Working paper, National University of Singapore.

Díaz-Uriarte, R., & De Andres, S. A. 2006. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 7(1), 3.

Erickson, M., M. Hanlon, and E. Maydew. 2006. Is there a link between executive equity incentives and accounting fraud? *Journal of Accounting Research* 44(1), 113-143.

Feng, M., W. Ge, S. Luo, and T. Shevlin. 2011. Why do CFOs become involved in material accounting manipulations? *Journal of Accounting and Economics* 51, 21-36.

Francis, J., D. Hanna, and L.Vincent. 1996. Causes and effects of discretionary asset write-offs. *Journal of Accounting Research* 34(Supplement), 117-134.

Friedlan, J. 1994. Accounting choices of issues of initial public offerings. *Contemporary Accounting Research* 11(1), 1-31.

Gail, M.H., S. Wieand, and S. Piantadosi. 1984. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika,* 71(3), 432-444.

Greenhill, B., M. D. Ward, and A. Sacks. 2011. The separation plot: A new visual method for evaluating the fit of binary models. *American Journal of Political Science* 55(4), 991-1002.

Gu, F. and B. Lev. 2011. Overpriced shares, ill-advised acquisitions, and goodwill impairment. *The Accounting Review* 86(6), 1995-2022.

Guo, L., and T. Mech. 2000. Conditional event studies, anticipation, and asymmetric information: the case of seasoned equity issues and pre-issue information releases. *Journal of Empirical Finance* 7, 113-141.

Hausman, J. 2001. Mismeasured variables in econometric analysis: problems from the right and problems from the left. *The Journal of Economic Perspectives*, 15(4), 57-67.

Hayn, C. and P. Hughes. 2006. Leading indicators of goodwill impairment. *Journal of Accounting, Auditing and Finance* 21, 223-265.

Ho, T. K. 1995. Random decision forests. *Document Analysis and Recognition.* (Proceedings of the Third International Conference on Document Analysis and Recognition) 1, 278-282.

Hoerl, A. E., and R. W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12(1), 55-67.

Holthausen, R. W., and D. F. Larcker. 1992. The prediction of stock returns using financial statement information. *Journal of Accounting and Economics* 15(2-3), 373-411.

Hosmer, D. W., S. Lemeshow, and R.X. Sturdivant. 2013. Applied logistic regression. Hoboken, NJ: Wiley.

Hribar, P., T. Kravet, and R. Wilson. 2014. A new measure of accounting quality. *Review of Accounting Studies* 19(1), 506-538.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An introduction to statistical learning* (6). New York: springer.

Jindra, J. 2000. Seasoned equity offerings, overvaluation, and timing. Working paper, Ohio State University.

Judd, C. M., G.H. McClelland. 1989. Data analysis: A model-comparison approach Harcourt Brace Jovanovich. *San Diego*.

Landis, J., G. Koch. (1977). The Measurement of observer agreement for categorical Data. *Biometrics*, 33(1), 159-174.

Li, K. and R. Sloan. 2017. Has goodwill accounting gone bad? *Review of Accounting Studies* 22(2), 964-1003.

Li, Z., P. Shroff, R. Venkataraman, and X. Zhang. 2011. Causes and consequences of goodwill impairment losses. *Review of Accounting Studies* 16, 745-778.

McLaughlin, R., A. Safieddine and G. Vasudevan. 1996. The operating performance of seasoned equity issuers: Free cash flow and post-issue performance. *Financial Management* 25(4), 41-53.

Menard, S. 2008. Applied logistic regression analysis. Thousand Oaks, Calif.: Sage.

Mood, C. 2010. Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review* 26(1), 67-82.

Ohlson, J. A. 1980. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research* 18(1), 109-131.

Ou, J. A., S. H. Penman. 1989. Accounting measurement, price-earnings ratio, and the information content of security prices. *Journal of Accounting Research* 27, 111-144.

Ou, J. A., S. H. Penman. 1989. Financial statement analysis and the prediction of stock returns. *Journal of Accounting and Economics* 11(4), 295-329.

Pagano, M., F. Panetta, and L. Zingales. 1998. Why do companies go public? An empirical analysis. *The Journal of Finance* 53(1), 27-64.

Palmer, D. S., N. M. O'Boyle, R. C. Glen, and J. B. Mitchell. 2007. Random forest models to predict aqueous solubility. *Journal of Chemical Information and Modeling* 47(1), 150-158.

Perols, J. 2011. Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practice & Theory* 30(2), 19-50.

Perols, J., B. Bowen, and C. Zimmerman. 2017. Finding needles in a haystack: Using data analytics to improve fraud prediction. *The Accounting Review* 92(2), 221-245.

Pope, P. T., J. T. Webster. 1972. The use of an F-statistic in stepwise regression procedures. *Technometrics* 14(2), 327-340.

Price, R., N. Sharp, and D. Wood. 2011. Detecting and predicting accounting irregularities: A comparison of commercial and academic risk measures. *Accounting Horizons* 25(4), 755-780.

Shmueli, G., N. Patel, P. Bruce. 2010. *Data Mining for Business Intelligence*. Hoboken.

Stone, M. 1977. An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society*. Series B (Methodological), 39(1), 44-47.

Svetnik, V., A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston. 2003. Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences* 43(6), 1947-1958.

Teoh, S., I. Welch, and T. Wong. 1998. Earnings management and the long-run performance of initial public offerings. *Journal of Finance* 53(6), 1935-1974.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*. Series B (Methodological), 267-288.

Whittingham, M. J., P.A. Stephens, R.B. Bradbury, and R. P. Freckleton. 2006. Why do we still use stepwise modeling in ecology and behaviour? *Journal of Animal Ecology* 75(5), 1182-1189.

Wilkinson, L. 1979. Tests of significance in stepwise regression. *Psychological Bulletin* 86(1), 168.

Wooldridge, J. M. 2002. *Econometric analysis of cross section and panel data*. MIT press.

Zou, H. and T. Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*. Series B (Statistical Methodology) 67(2), 301-320.

Appendices

Appendix A
Change in earnings variable definitions

ADJEPSFX	= Diluted earnings per share.
GM_AB	= ((COGS-ECOGS_AB)-(SALE-ESALE_AB)); where COGS represents cost of goods sold, ECOGS_AB represents the average of the past two years COGS, SALE represents total sales, ESALE_AB represents the average of the past two years SALE.
AR_AB	= ((SALE-ESALE_AB)-(RECT-ERECT_AB)); where SALE represents total sales, ESALE_AB represents the average of the past two years SALE, RECT represents receivables, ERECT_AB represents the average of the past two years RECT.
CAPX_AB	= ((CAPX-ECAPX_AB) / ECAPX_AB)-((AVGINDCAPX-EAVGINDCAPX_AB) / EAVGINDCAPX_AB); where CAPX represents capital expenditures, ECAPX_AB represents the average of the past two years CAPX, AVGINDCAPX represents the average 2 digit SIC industry year CAPX, EAVGINDCAPX_AB represents the average of the past two years AVGINDCAPX_AB.
SGA_AB	= ((SALE-ESALE_AB)-(XSGA-ESGA_AB)); where SALE represents total sales, ESALE_AB represents the average of the past two years SALE, XSGA represents Selling, General and Administrative Expense, EXSGA represents the average of the past two years XSGA.
ETR_AB	= ((ETR-E_ETR)*((EPSPX - LAG1_EPSPX) / LAG1_PRCC_F)); where ETR is (TXT / (PI+AM)) and TXT is income taxes, PI is pretax income, and AM is amortization of intangibles, EPSPX is earnings per share, LAG1_EPSPX is prior year's earnings per share, LAG1_PRCC_F prior years stock price at the end of the year.
LF_AB	= (((SALE/EMP)-(LAG1_SALE / LAG1_EMP)) / (LAG1_SALE / LAG1_EMP)); where SALE is total sales, EMP number of employees, LAG1_SALE prior year's total sales, LAG1_EMP prior year's number of employees.

Appendix A (cont.)

CURRENT_OP	= ACT / LCT; where ACT is total current assets, LCT is total current liabilities.
CHG_CURRENT_OP	= CURRENT_OP less prior year's CURRENT_OP.
QUICK_OP	= (ACT-INVT) / LCT; where ACT is total current assets, INVT is total current inventory, LCT is total current liabilities.
CHG_QUICK_OP	= QUICK_OP less prior year's QUICK_OP.
DAYSAR_OP	= (SALE / ((RECT-LAG1_RECT) / 2)); where SALE is total sales, RECT is receivables, LAG1_RECT is prior year's receivables.
CHG_DAYSAR_OP	= DAYSAR_OP less prior year's DAYSAR_OP.
INVTO_OP	= (SALE / ((INVT-LAG1_INVT) / 2)); where SALE is total sales, INVT is total inventory, LAG1_INVT is prior year's INVT.
CHG_INVTO_OP	= INVTO_OP less prior year's INVTO_OP.
INVTAT_OP	= INVT / AT; where INVT is total inventory, AT is total assets.
CHG_INVTAT_OP	= INVTAT_OP less prior year's INVTAT_OP.
INVT_OP	= INVT; where INVT is total inventories.
CHG_INVT_OP	= INVT_OP less prior year's INVT_OP.
SALE_OP	= SALE; where SALE is total sales.
CHG_SALE_OP	= SALE_OP less prior year's SALE_OP.
DP_OP	= DP; where DP is depreciation and amortization.
CHG_DP_OP	= DP_OP less prior year's DP_OP.
DVPSX_OP	= DVPSX_F; where DVPSX_F is dividends per share.
CHG_DVPSX_OP	= DVPSX less prior years DVPSX.

Appendix A (cont.)

DPPPEGT_OP	= (((DP / PPEGT)-(LAG1_DP / LAG1_PPEGT))); where DP is depreciation and amortization, LAG1_DP prior year's DP, PPEGT is total property, plant, and equipment, LAG1_PPEGT.
CHG_DPPPEGT_OP	= DPPPEGT_OP less prior years DPPPEGT_OP.
ROE_OP	= NI / SEQ; where NI is net income, SEQ is total stockholders' equity.
CHG_ROE_OP	= ROE_OP less prior years ROE_OP.
CAPXAT_OP	= CAPX / AT; where CAPX is capital expenditures, AT is total assets.
CHG_CAPXAT_OP	= CAPXAT_OP less prior year's CAPXAT_OP.
LAG1_CAPXAT_OP	= prior year's CAPXAT_OP.
LAG1_CHG_CAPXAT_OP	= LAG1_CAPXAT_OP less the 2 years prior CAPXAT_OP.
LTCEQ_OP	= LT / CEQ; where LT is total liabilities, CEQ is common equity.
CHG_LTCEQ_OP	= LTCEQ_OP less prior year's LTCEQ_OP.
DLTTCEQ_OP	= DLTT / CEQ; where DLTT is long term debt, CEQ is common equity.
CHG_DLTTCEQ_OP	= DLTTCEQ_OP less prior year's DLTTCEQ_OP.
SEQPPENT_OP	= SEQ / PPENT; where SEQ is total stockholders' equity, PPENT is total property, plant, and equipment.
CHG_SEQPPENT_OP	= SEQPPENT_OP less prior year's SEQPPENT_OP.
COVER_OP	= (XINT+PI) / XINT; where XINT is interest and related expense, PI is pretax income.
CHG_COVER_OP	= COVER_OP less prior year's COVER_OP.

Appendix A (cont.)

SALEAT_OP	= SALE / AT; where SALE is total sales, AT is total assets.
CHG_SALEAT_OP	= SALEAT_OP less prior year's SALEAT_OP.
PIAT_OP	= PI / AT; where PI is pretax income, AT is total assets.
CHG_PIAT_OP	= PIAT_OP less prior year's PIAT_OP.
NISALE_OP	= NI / SALE; where NI is net income, SALE is total sales.
CHG_NISALE_OP	= NISALE_OP less prior year's NISALE_OP.
SALECHE_OP	= (SALE / CHE; where SALE is total sales, CHE is cash and short-term investments.
CHG_SALECHE_OP	= SALECHE_OP less prior year's SALECHE_OP.
SALERECT_OP	= SALE / RECT; where SALE is total sales, RECT is total receivables.
CHG_SALERECT_OP	= SALERECT_OP less prior year's SALERECT_OP.
SALEINVT_OP	= SALE / INVT; where SALE is total sales, INVT is total inventories.
CHG_SALEINVT_OP	= SALEINVT_OP less prior year's SALEINVT_OP.
SALEWCAP_OP	= SALE / WCAP; where SALE is total sales, WCAP is working capital.
CHG_SALEWCAP_OP	= SALEWCAP_OP less prior year's SALEWCAP_OP.
SALEPPENT_OP	= SALE / PPENT; where SALE is total sales, PPENT is total property, plant, and equipment.
CHG_SALEPPENT_OP	= SALEPPENT_OP less prior year's SALEPPENT_OP.
COGS_OP	= COGS; where COGS is cost of goods sold.
CHG_COGS_OP	= COGS_OP less prior year's COGS_OP.
AT_OP	= AT; where AT is total assets.

Appendix A (cont.)

CHG_AT_OP	= AT_OP less prior year's AT_OP.
CSHDBT_OP	= (IB+DP) / (DLTT+DLC); where IB is income before extraordinary items, DP is depreciation and amortization, DLTT is long term debt, DLC debt in current liabilities.
CHG_CSHDBT_OP	= CSHDBT_OP less prior year's CSHDBT_OP.
WCAPAT_OP	= WCAP / AT; where WCAP is working capital, AT is total assets.
CHG_WCAPAT_OP	= WCAPAT_OP less prior year's WCAPAT_OP.
OIADPAT_OP	= OIADP / AT; where OIADP is operating income after depreciation, AT is total assets.
CHG_OIADPAT_OP	= OIADAT_OP less prior year's OIADAT_OP.
DLTT_OP	= DLTT; where DLTT is long term debt.
CHG_DLTT_OP	= DLTT_OP less prior year's DLTT_OP.
WCAP_OP	= WCAP; where WCAP is working capital.
CHG_WCAP_OP	= WCAP_OP less prior year's WCAP_OP.
NIIBDP_OP	= (NI / (IB+DP)); where NI is net income, IB is income before extraordinary items, DP is depreciation and amortization.
CHG_NIIBDP_OP	= NIIBDP_OP less prior year's NIIBDP_OP.

Appendix B
Misstatements Variable Definitions

BIGMISS	= 1 if the misstatement filing is an 8-K or 8-K/A and zero otherwise.
EMPPROD	= (SALE/EMP – LAG1_SALE/LAG1_EMP) / (LAG1_SALE/LAG1_EMP); where SALE is total sales, LAG1_SALE is the prior year total sales, LAG1_EMP is the prior year number of employees, EMP is the number of employees.
GEOSALEGROW	= ((SALE / LAG3_SALE)*(1/4))-1; where SALE is total sales, LAG3_SALE is 3 years prior total sales.
ABNPCLTINT	= PRCNTCHGLIAB-INDPRCNTCHGLIAB; where PRCNTCHGLIAB =(LT-LAG1_LT) / LAG1_LT and LT is total liabilities, LAG1_LT is prior year total liabilities. INDPRCNTCHGLIAB is the two-digit yearly mean of PRCNTCHGLIAB.
PRCNTCHGEXPENSES	= (XOPR-LAG1_ XOPR) / LAG1_ XOPR; where XOPR is operating expenses, LAG1_XOPR is prior year operating expenses.
PRCNTCHGSALEAT	= ((SALE / AT) -(LAG1_SALE / LAG1_AT)) / (LAG1_SALE / LAG1_AT); where SALE is total sales, AT is total assets, LAG1_SALE is prior year sales, LAG1_AT is prior year total assets.
CHGLIAB	= LT-LAG1_LT; where LT is total liabilities LAG1_LT is prior year total liabilities.
DEMANDFIN	= 1 if (((OANCF-(LAG3_ CAPX +LAG2_ CAPX +LAG1_ CAPX) / 3) / ACT) < (-.05)) and zero otherwise; where OANCF is operating cash flows, CAPX is capital expenditures, ACT is current assets. LAG3_CAPX, LAG2_CAPX, and LAG1_CAPX refer to 3, 2, and 1 year prior CAPX respectively.
GROSS	= (SALE-COGS) / SALE; where SALE is total sales and COGS is cost of goods sold.
CHGSALE	= SALE-LAG1_SALE where SALE is total sales and LAG1_SALE is prior year total sales.

Appendix B (cont.)

CHGINVSALE	= ((INVT) / (SALE)) - ((LAG1_INVT) / (LAG1_SALE)); where INVT is total inventories, SALE is total sales, LAG1_INVT is prior year total inventories, LAG1_SALE is prior year total sales.
RECTSALE	= RECT / SALE; where RECT is receivables and SALE is total sales.
ASSETS	= AT; where AT is total assets.
RECTDUM	= 1 if (RECT / LAG1_RECT) > 1.1, and zero otherwise; where RECT is total receivables, LAG1_RECT is prior year receivables.
SALEMP	= SALE/EMP; where SALE is total sales and EMP is number of employees.
LEVERAGE	= DLTT/AT; where DLTT is long term debt and AT is total assets.
PRCNTCHGASS	= (AT-LAG1_AT) / LAG1_AT; where AT is total assets.
FIN	= (IVST + IVAO) - (DLTT + DLC + PSTK); where IVST is short term investments, IVAO is investment and advances, DLTT is long term debt, DLC is debt in current liabilities, and PSTK is preferred stock.
CHGROA	= (NI/AT) - (LAG1_NI / LAG1_AT); where NI is net income, AT is total assets, LAG1_NI is prior year net income, LAG1_AT is prior year total assets.
ROAAT	= (LAG1_NI / LAG1_AT) / AT; AT is total assets, LAG1_NI is prior year net income, LAG1_AT is prior year total assets.
ISSUE	= 1 if SSTK>0 or DLTIS>0, and zero otherwise; where SSTK is sale of stock and DLTIS is long term debt issuance.

Appendix B (cont.)

PRCNTCHGCASHMAR	= ((1-(COGS+(INVT-LAG1_INVT)) / (SALE-(RECT-LAG1_RECT)))-(1-(LAG1_COGS+(LAG1_INVT - lag2_INVT)) / (LAG1_SALE-(LAG1_RECT-lag2_RECT)))) / (1-(LAG1_COGS+(LAG1_INVT-lag2_INVT)) / (LAG1_SALE-(LAG1_RECT-lag2_RECT))); where COGS is cost of goods sold, INVT is inventories, SALE is total sales, and RECT is total receivables. LAG2 and LAG1 refer to prior 2 and 1 year.
CHGSALEAT	= (SALE / AT) - (LAG1_SALE / LAG1_AT); where SALE is total sales, AT is total assets, LAG1_SALE is prior year total sales, LAG1_AT is prior year total assets.
RETONEQ	= (NI / CEQ); where NI is net income and CEQ is common equity.
GROSSDUM	= 1 if (((SALE-COGS) / SALE) / ((LAG1_SALE - LAG1_COGS)/LAG1_SALE)) > 1.1, and zero otherwise; where SALE is total sales, COGS is cost of goods sold, LAG1_SALE is prior year total sales, LAG1_COGS is prior year cost of goods sold.
INVTSALE	= INVT / SALE; where INVT is total inventories, SALE is prior year total sales.
LTCEQ	= LT / CEQ; where LT is total liabilities and CEQ is common equity.
PRCNTCHGATLT	= ((AT/LT)-(LAG1_AT/LAG1_LT)) / (LAG1_AT/LAG1_LT); where AT is total assets, LT is total liabilities, LAG1_AT is prior year total assets, LAG1_LT is prior year total liabilities.
PPENTAT	= PPENT/AT; where PPENT is property, plant, and equipment and AT is total assets.

Appendix B (cont.)

PRCNTCHGRETONSALE	= ((NI / SALE)-(LAG1_NI / LAG1_SALE)) / (LAG1_NI / LAG1_SALE); where NI is net income, SALE is total sales, LAG1_NI is prior year net income, LAG1_SALE is prior year total sales.
TOTACCAT	= (IB-OANCF) / AT; where IB is income before extraordinary items, OANCF is operating cash flows and AT is total assets.
PRCNTCHGLIAB	= (LT-LAG1_LT) / LAG1_LT; where is LT is total liabilities, LAG1_LT is prior year total liabilities.
NETSALE	= SALE; where SALE is total sales.

Tables

Table 1 Model Accuracy
1970-2014

Split	*N*	*Logistic Accuracy*	*Stepwise Accuracy*	*Elastic Net Accuracy*	*Random Forest Accuracy*
50/50	41094	0.57	0.57	0.577	**0.6**
60/40	32881	0.588	0.586	0.589	**0.624**
70/30	24663	0.601	0.601	0.598	**0.645**
80/20	16445	0.621	0.623	0.612	**0.665**
90/10	8223	0.659	0.671	0.636	**0.693**
95/05	4113	0.697	0.714	0.661	**0.735**

Table 1 shows the percentile splits, corresponding sample size, and accuracy. The percentile splits are taken from ranking raw probabilities formed from each respective model. Accuracy represents how correctly each model classifies a positive change in earnings. The bold numbers represent the largest accuracies.

Table 2 Five year groups
95/05 split

Years	*Logistic Regression*	*Stepwise Logistic*	*Elastic Net*	*Random Forest*
1970-1974	0.631	**0.644**	0.601	0.605
1975-1979	0.644	0.63	0.647	**0.692**
1980-1984	0.703	**0.715**	0.69	0.71
1985-1989	0.748	0.743	0.735	**0.756**
1990-1994	0.709	0.724	0.702	**0.729**
1995-1999	0.753	**0.761**	0.732	0.728
2000-2004	0.717	0.728	0.702	**0.748**
2005-2009	0.685	0.754	0.549	**0.791**
2010-2014	0.667	0.672	0.655	**0.739**

Table 2 presents 5 year groups. The corresponding out-of-sample accuracy is given. The bolded numbers represent the largest accuracies.

Table 3 Top ten most important variables

Random Forest Variables	*Freq*	*Elastic Net Variables*	*Freq*	*Stepwise Variables*	*Freq*
ADJEPSFX	45	Z_CAPX_AB	45	PIAT_OP	44
ETR_AB	45	Z_CHG_SALEAT_OP	45	CHG_CURRENT_OP	42
CAPX_AB	44	Z_PIAT_OP	45	OIADPAT_OP	42
CHG_DAYSAR_OP	43	Z_CHG_INVTAT_OP	44	INVTAT_OP	38
CHG_SALECHE_OP	43	Z_INVTAT_OP	43	CHG_QUICK_OP	37
LAG1_CHG_CAPXAT_OP	43	Z_CHG_INVT_OP	42	CHG_SALE_OP	37
CHG_SALERECT_OP	40	Z_DVPSX_OP	42	SALE_OP	35
LF_AB	39	Z_OIADPAT_OP	42	CHG_PIAT_OP	34
CHG_INVTAT_OP	38	Z_ ADJEPSFX	41	CHG_SALEAT_OP	34
DAYSAR_OP	38	Z_CHG_CURRENT_OP	41	CHG_INVT_OP	33

Table 3 shows the top ten most chosen independent variables over the sample period 1970-2014 inclusive. The numbers represent the corresponding number of times chosen, with 45 being the largest possible number. Variables are defined in the appendix.

Table 4 Abnormal returns

Panel A Logistic Regression				Panel B Stepwise Logistic Regression				Panel C Elastic Net				Panel D Random Forest			
Abnormal Returns															
Confusion Matrix	BHAR	PValue	N	Confusion Matrix	BHAR	PValue	N	Confusion Matrix	BHAR	PValue	N	Confusion Matrix	BHAR	PValue	N
Hedge	0.144	0.000	4113	Hedge	0.142	0.000	4113	Hedge	0.051	0.005	4113	Hedge	0.174	0.000	4113
PP	0.077	0.000	2079	PP	0.066	0.000	2079	PP	0.054	0.000	2079	PP	0.075	0.000	2079
PN	-0.067	0.000	2034	PN	-0.077	0.000	2034	PN	0.002	0.848	2034	PN	-0.099	0.000	2034
TP	0.12	0.000	1520	TP	0.12	0.000	1582	TP	0.083	0.000	1455	TP	0.123	0.000	1650
TN	-0.153	0.000	1347	TN	-0.159	0.000	1356	TN	-0.068	0.000	1265	TN	-0.189	0.000	1371
FP	-0.038	0.085	559	FP	-0.106	0.000	497	FP	-0.016	0.454	624	FP	-0.112	0.000	429
FN	0.103	0.000	687	FN	0.087	0.000	678	FN	0.118	0.000	769	FN	0.086	0.000	663

Panel A Logistic Regression		Panel B Stepwise Logistic Regression		Panel C Elastic Net		Panel D Random Forest	
Fit Metrics							
Metric	**Value**	**Metric**	**Value**	**Metric**	**Value**	**Metric**	**Value**
Accuracy	0.697	Accuracy	0.714	Accuracy	0.661	Accuracy	0.735
Kappa	0.394	Kappa	0.428	Kappa	0.322	Kappa	0.468
Sensitivity	0.689	Sensitivity	0.700	Sensitivity	0.654	Sensitivity	0.713
Specificity	0.707	Specificity	0.732	Specificity	0.670	Specificity	0.762
Prevalence	0.537	Prevalence	0.549	Prevalence	0.541	Prevalence	0.562
Detection Rate	0.370	Detection Rate	0.385	Detection Rate	0.354	Detection Rate	0.401
Detection Prevalence	0.505	Detection Prevalence	0.505	Detection Prevalence	0.505	Detection Prevalence	0.505

Table 4 presents abnormal returns and supplemental fit data. The confusion matrix column represents data available in a confusion matrix for the 95/05 data split. PP represents those predicted to be a positive change, PN represents those predicted to be a negative change, TP represent the true positives, TN represents the true negatives, FP represents false positives, and FN represents false negatives. BHAR represents the 12 month abnormal size adjusted returns, PValue represents the significance for the abnormal returns, and N is the number of observations. Fit Metrics are accuracy, Kappa which represents how well the classifier performs relative to random chance, Sensitivity is the true positive rate, Specificity is the true negative rate, prevalence is the number of positive occurrences, detection rate is the number of true positives relative to the total, and detection prevalence is the number of predicted positive relative to the total.

Table 5 Cross-validation vs. rolling window cross-validation by year

Year	*In-sample CV*	*Out-of-Sample CV*	*Difference CV*	*In-sample RWCV*	*Out-of-Sample RWCV*	*Difference RWCV*	*CV compared to RWCV*
1970	0.59	0.575	0.015	0.552	0.573	-0.021	Smaller
1971	0.595	0.535	0.06	0.593	0.557	0.037	Larger
1972	0.615	0.509	0.106	0.609	0.525	0.084	Larger
1973	0.644	0.586	0.058	0.715	0.581	0.134	Smaller
1974	0.647	0.543	0.104	0.624	0.573	0.05	Larger
1975	0.658	0.584	0.074	0.584	0.581	0.003	Larger
1976	0.685	0.572	0.113	0.691	0.572	0.119	Smaller
1977	0.682	0.591	0.09	0.646	0.6	0.045	Larger
1978	0.678	0.591	0.087	0.674	0.617	0.057	Larger
1979	0.665	0.57	0.095	0.652	0.594	0.058	Larger
1980	0.66	0.569	0.091	0.549	0.59	-0.041	Larger
1981	0.655	0.549	0.106	0.578	0.576	0.003	Larger
1982	0.633	0.612	0.021	0.536	0.605	-0.069	Smaller
1983	0.631	0.569	0.062	0.613	0.568	0.045	Larger
1984	0.628	0.604	0.024	0.567	0.603	-0.037	Smaller
1985	0.632	0.617	0.014	0.617	0.626	-0.009	Larger
1986	0.637	0.564	0.073	0.637	0.594	0.043	Larger
1987	0.64	0.585	0.055	0.605	0.572	0.033	Larger
1988	0.629	0.555	0.074	0.571	0.587	-0.016	Larger
1989	0.613	0.589	0.023	0.594	0.61	-0.016	Larger
1990	0.619	0.575	0.044	0.607	0.586	0.02	Larger
1991	0.607	0.554	0.053	0.573	0.579	-0.006	Larger
1992	0.6	0.596	0.005	0.589	0.586	0.003	Larger
1993	0.602	0.568	0.034	0.598	0.578	0.02	Larger

Table 5 (cont.)

Year	*In-sample CV*	*Out-of-Sample CV*	*Difference CV*	*In-sample RWCV*	*Out-of-Sample RWCV*	*Difference RWCV*	*CV compared to RWCV*
1994	0.605	0.575	0.03	0.568	0.575	-0.008	Larger
1995	0.61	0.59	0.02	0.6	0.599	0.001	Larger
1996	0.615	0.551	0.064	0.605	0.564	0.041	Larger
1997	0.61	0.57	0.04	0.58	0.582	-0.002	Larger
1998	0.604	0.578	0.026	0.564	0.615	-0.051	Smaller
1999	0.6	0.551	0.05	0.61	0.591	0.018	Larger
2000	0.602	0.579	0.023	0.587	0.616	-0.028	Smaller
2001	0.604	0.61	-0.006	0.604	0.622	-0.017	Smaller
2002	0.609	0.588	0.021	0.617	0.589	0.028	Smaller
2003	0.619	0.568	0.05	0.607	0.58	0.027	Larger
2004	0.623	0.553	0.07	0.588	0.571	0.017	Larger
2005	0.615	0.582	0.033	0.6	0.577	0.023	Larger
2006	0.618	0.582	0.037	0.634	0.587	0.047	Smaller
2007	0.631	0.563	0.067	0.633	0.571	0.062	Larger
2008	0.61	0.805	-0.195	0.493	0.639	-0.146	Larger
2009	0.617	0.597	0.021	0.624	0.583	0.04	Smaller
2010	0.624	0.591	0.033	0.609	0.582	0.027	Larger
2011	0.628	0.569	0.06	0.604	0.592	0.013	Larger
2012	0.63	0.566	0.064	0.601	0.611	-0.009	Larger
2013	0.643	0.636	0.007	0.613	0.59	0.023	Smaller
2014	0.643	0.591	0.053	0.611	0.609	0.001	Larger

Table 5 shows presents accuracy for the 50/50 split of the data. The cross-validation (CV) in-sample accuracy, out-of-sample, and difference is compared with the rolling window cross-validation (RWCV) method. The method that produces the smallest absolute difference performs the best in this setting. The last column highlights whether the difference from CV is larger than RWCV.

Table 6 Out-of-sample accuracy

Years	*Random Forest CV*	*Random Forest RWCV*
1970-1974	**0.605**	0.592
1975-1979	**0.692**	0.687
1980-1984	**0.71**	0.698
1985-1989	**0.756**	0.735
1990-1994	0.729	**0.762**
1995-1999	0.728	**0.732**
2000-2004	**0.748**	0.731
2005-2009	**0.791**	0.779
2010-2014	**0.739**	0.716

Table 6 presents the accuracy for 5 year groups for the 95/05 split for Random forest CV and Random forest RWCV. The bold numbers represent the largest value.

Table 7 Model AUC
2000-2014

Sampling	*Logistic AUC*	*Stepwise AUC*	*Elastic Net AUC*	*Random Forest AUC*
Original	0.6720	0.6722	0.6768	**0.7175**
Down	0.5917	0.5915	0.5702	**0.6622**
Up	0.5653	0.5977	0.5818	**0.7480**
SMOTE	0.5859	0.5837	0.5715	**0.6736**

Table 7 shows the sampling methods and AUC. The sampling methods are Down-sampling, Up-sampling, and SMOTE. AUC represents the out-of-sample area under the roc curve. The bold numbers represent the largest AUC.

Table 8 Five year groups

Sampling	*Logistic Regression*	*Stepwise Logistic*	*Elastic Net*	*Random Forest*
2005-2009				
Original	0.5775	0.5763	0.5791	**0.6390**
Down	0.5669	0.5741	0.5539	**0.6261**
Up	0.5486	0.5681	0.5663	**0.6808**
SMOTE	0.5577	0.5542	0.5537	**0.6257**
2010-2014				
Original	0.6667	0.6667	0.6667	**0.7409**
Down	0.5974	0.5814	0.5667	**0.7093**
Up	0.6057	0.6099	0.5762	**0.7644**
SMOTE	0.6072	0.6023	0.5824	**0.7149**

Table 8 presents five year groups by sampling method. The corresponding out-of-sample AUC is given. The bold numbers represent the largest AUC.

Table 9 Top ten most important variables

Random Forest Variables	*Freq*	*Elastic Net Variables*	*Freq*	*Stepwise Variables*	*Freq*
WC	10	SOFTASS	10	PPENTAT	10
SALEEMP	10	SALEAT	10	SOFTASS	9
SALEAT	10	RECTSALE	10	NCO	9
PPENTAT	10	PPENTAT	10	LVLFIN	9
MKTVOLATILITY	10	POSACC	10	FIN	9
HOLDRET	10	NETSALE	10	CHGLIABB	9
GROSS	10	NCO	10	CHGASS	9
FIN	9	ISSUE	10	RECTSALE	8
SOFTASS	9	LVLFIN	10	POSACC	8
RECTAT	9	GEOSALEGROW	10	PRCNTCHGSALE	7

Table 9 shows the top ten most chosen independent variables for misstatements over the sample period 2005-2014 inclusive. The numbers represent the corresponding number of times chosen, with 10 being the largest possible number. Variables are defined in the appendix.

Figures

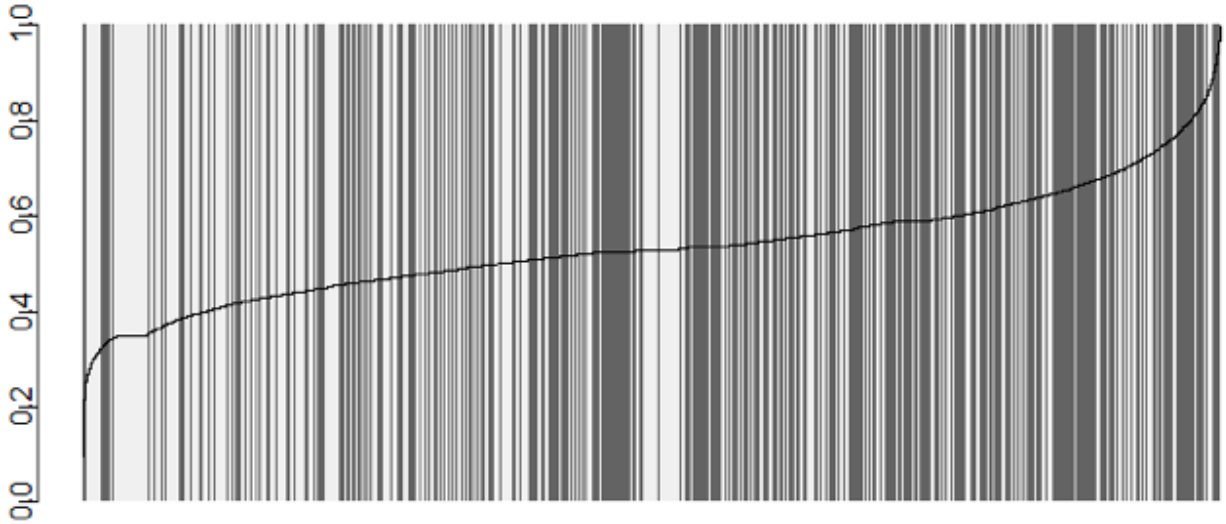

Figure 1. Separation plot of raw probabilities of traditional cross-validation random forest.

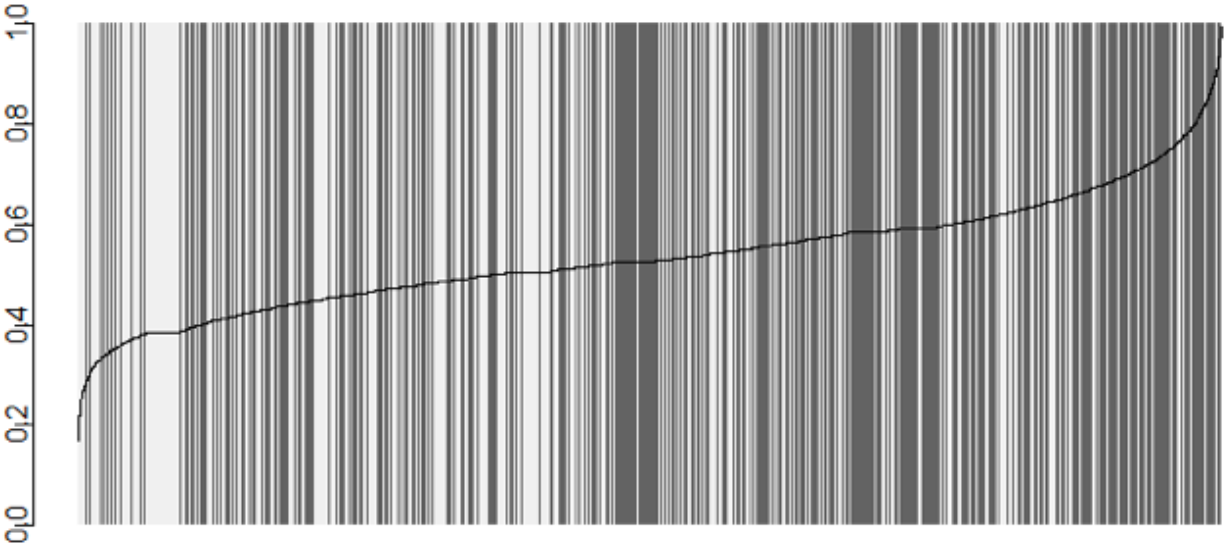

Figure 2. Separation plot of raw probabilities of rolling window cross-validation random forest.

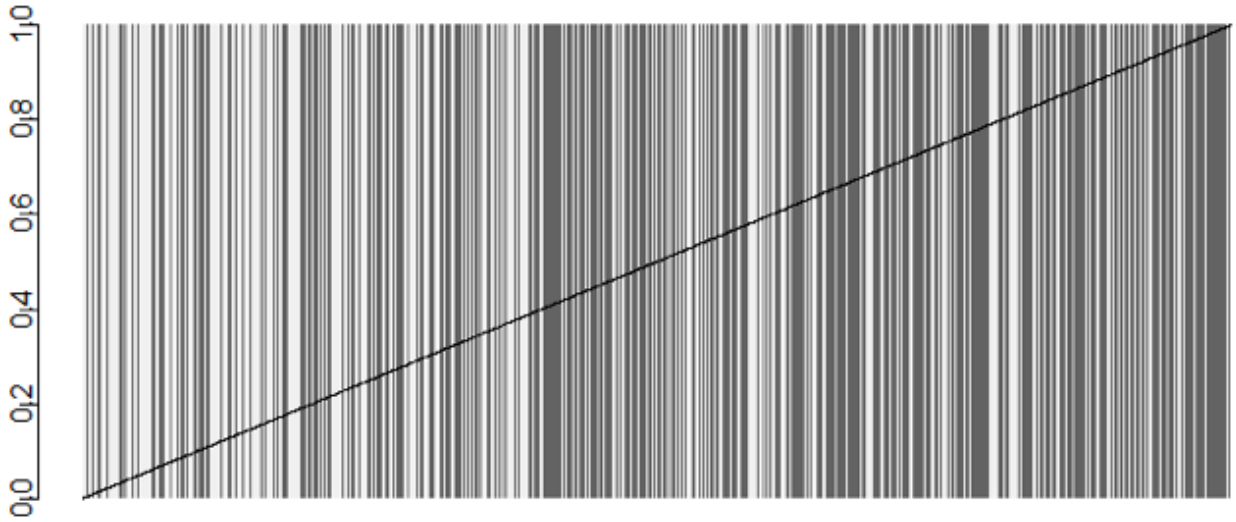

Figure 3. Separation plot of ranked probabilities of traditional cross-validation random forest.

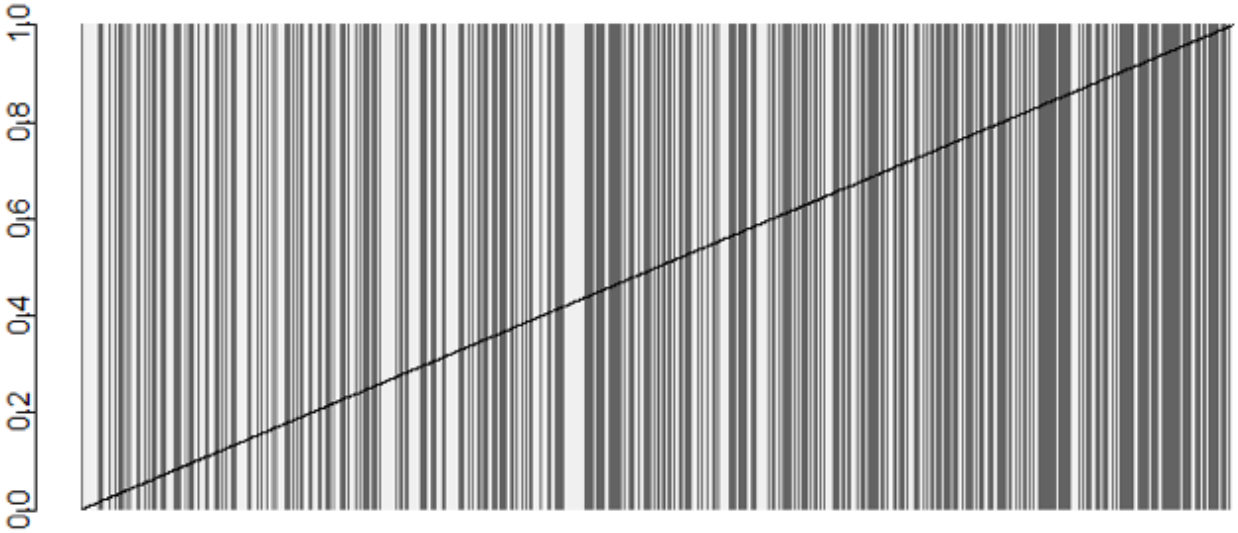

Figure 4. Separation plot of ranked probabilities of rolling window cross-validation random forest.

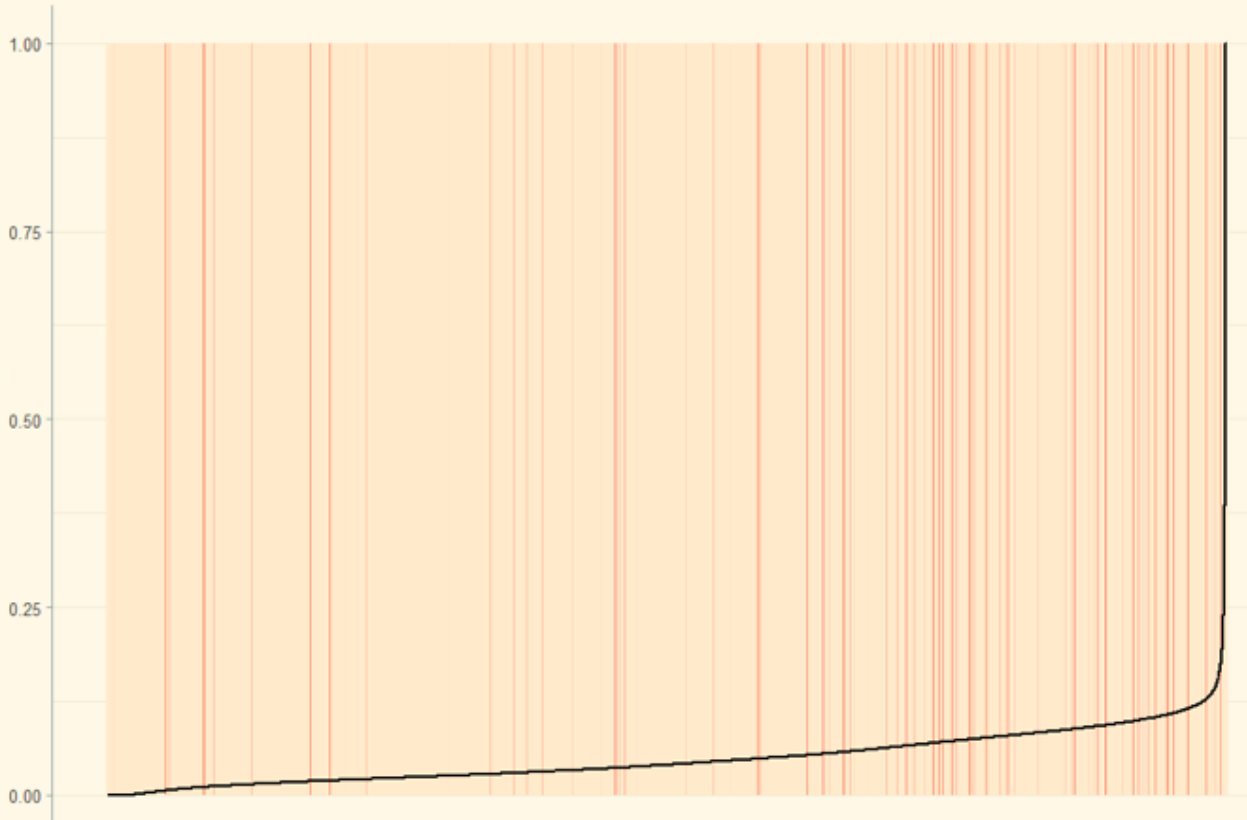

Figure 5. Separation plot of ranked probabilities of logistic regression for the original sample (AUC = 0.6998).

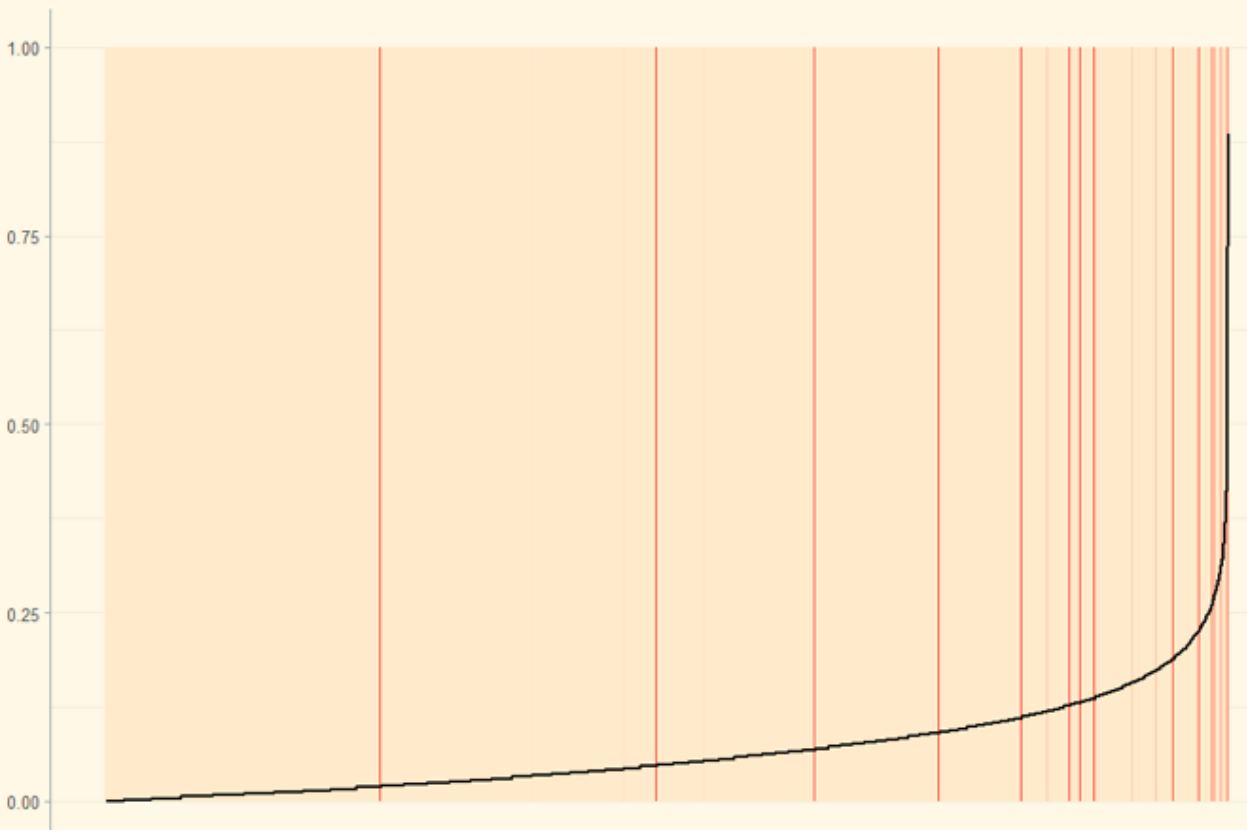

Figure 6. Separation plot random of ranked probabilities of random forest for the original sample (AUC = 0.7462).

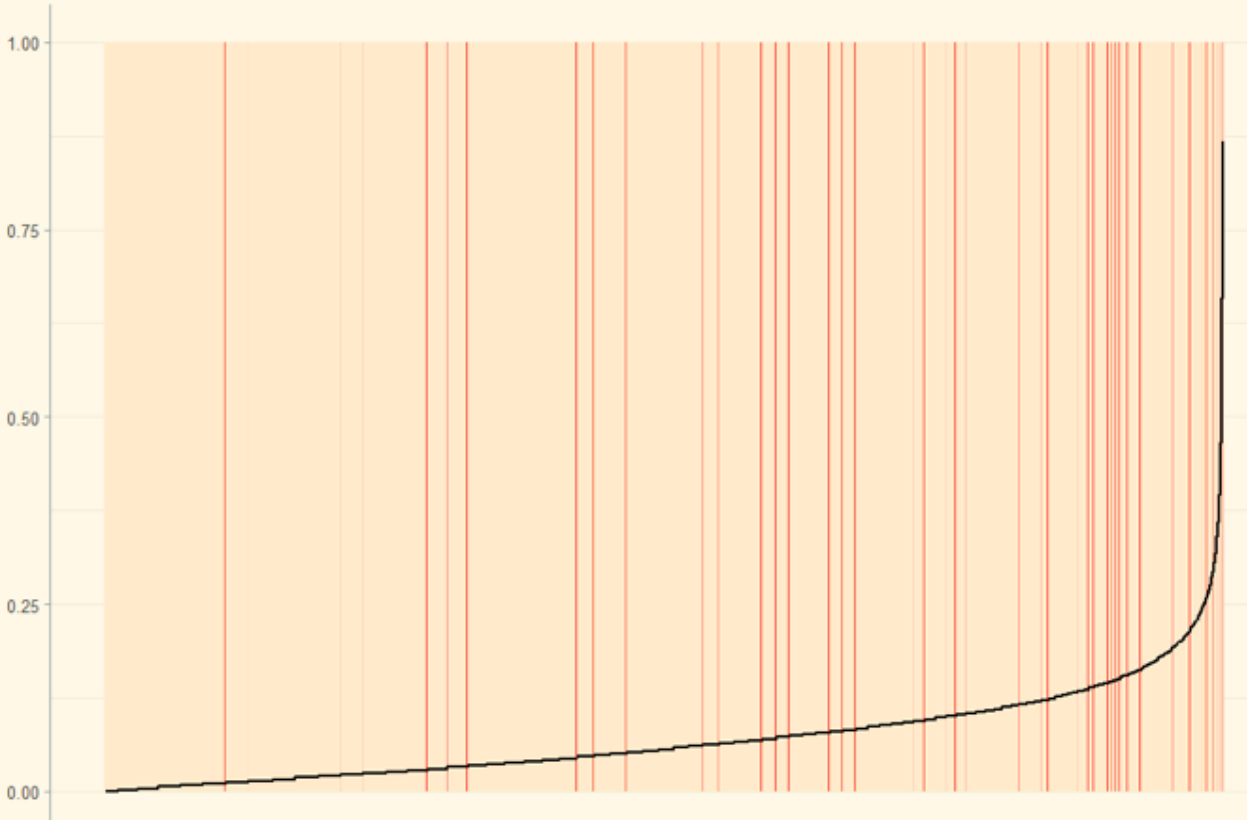

Figure 7. Separation plot of ranked probabilities of random forest up-sampling fit (AUC = 0.7458).[46]

[46] Wharton Research Data Services (WRDS) was used in preparing this manuscript. This service and the data available thereon constitute valuable intellectual property and trade secrets of WRDS and/or its third-party suppliers.