Proposing a Sampling Method to Build Effective Bankruptcy Prediction Models for North American Companies

BY RACHEL CARDARELLI

ADVISOR • Dr. Son Nguyen

EDITORIAL REVIEWER • Dr. Rick Gorvett

Submitted in partial fulfillment of the requirements for graduation
with honors in the Bryant University Honors Program

APRIL 2020

Proposing a Sampling Method to Build Effective Bankruptcy Prediction Models for North American Companies

Bryant University Honors Program
Honors Thesis
Student's Name: Rachel Cardarelli
Faculty Advisor: Son Nguyen
Editorial Reviewer: Rick Gorvett
April 2020

Table of Contents

Abstract

Bankruptcy prediction is a widely researched topic. However, few studies focus on dealing with the imbalance problem. This paper proposes a new technique that applies a bagging undersampling procedure to balance the data and compares it to random undersampling and five oversampling techniques. The performance of the algorithm is evaluated by a random forest's balanced accuracy, sensitivity, and specificity. The results show that models trained after applying the oversampling techniques are prone to overfitting, and the model trained after applying the proposed method had the highest balanced accuracy without overfitting.

Introduction

The study of bankruptcy prediction aims to provide risk assessment in order to reduce the likelihood of financial distress for individual companies and the macroeconomy as a whole. It provides creditors and investors with insight when making financial decisions, and the timely recognition of the potential for bankruptcy is important for mitigating its potential costs to many parties.

For decades, this subject has been flooded with new research, benefiting from the continuously evolving field of data science, and because of this has seen numerous advancements including the implementation of highly accurate modeling, feature selection, and ensemble techniques. However, one facet of bankruptcy prediction that has not been researched as thoroughly is the imbalance problem. If there are 100 observations in a dataset, with 99 of them being positive and 1 being negative, a model could predict that the entire dataset is positive with 99% accuracy. However, this disregards the minority, negative observation altogether. And as the ratio of

majority to minority samples increases, this problem worsens. So, this paper proposes a new technique to address the imbalance problem.

The proposed method undersamples the training data at various levels of imbalance (defined as the ratio of the number of majority points to the number of minority points). It performs this on 200 bagged samples, and each resulting training set is used to train a random forest. Then, a majority voting procedure determines the final prediction.

The Background section contains a literature review and concludes with the motivation for this study. The Methodology section contains descriptions of and the rationale for this study's experimental settings as well as the details of the proposed Bagging Undersampling method. The Results section provides an evaluation of each model's performance and the support for the proposed sampling method. Finally, the Conclusions section discusses the results and suggests ideas for future study.

Background

Literature Review

In this review of bankruptcy prediction literature, there is a discussion of commonly used predictive models, feature selection techniques, model tuning, ensemble learning, and other distinctive topics studied in the literature. The review concludes with an overview of methods used to handle the imbalance problem.

Predictive Models

Bankruptcy prediction models can be divided into two main classes. The first consists of statistical methods which began with Beaver in 1966 followed by Altman in 1968 who applied univariate discriminant analysis and multivariate linear regression respectively. It continued with

stochastic models such as logit regression (Ohlson, 1980) and probit regression (Zmijewski, 1984). The second class consists of artificial intelligence (AI) methods. This class has been used in a large amount of studies and in application to bankruptcy prediction since the 1990's. AI methods include decision tree[1], genetic algorithm (GA)[2], support vector machine (SVM), and several kinds of neural networks such as BPNN (back propagation trained neural network), PNN (probabilistic neural networks), and ANN (artificial neural networks) (Jeong, Min, 2009).

According to a 2014 review written by Sun, Li, Huang, and He, AI models have dominated more recent studies because of their superior accuracy and mapping abilities (Sun et. al, 2014) (Kumar, Ravi, 2007). In a review of studies from 1968-2005, the authors found that SVM and NN (especially BPNN) are objectively more powerful than other methods, especially statistical

Model	Formula	Variable	Description
Altman (1968) Multiple-discriminant analysis	$Z = \beta' X$ where Z is the MDA score and X represents the variables listed	$X1$	= Net working capital/total assets
		$X2$	= Retained earnings/total assets
	Cutoff point:	$X3$	= Earnings before interest and taxes/total assets.
	$Z \geqslant 2.675$, classified as non-bankrupt	$X4$	= Market value of equity/book value of total liabilities.
	$Z < 2.675$, classified as bankrupt	$X5$	Sales/total assets.
Ohlson (1980) Logit model	$P = (1 + \exp\{-\beta' X\})^{-1}$ where P is the probability of bankruptcy and X represents the variables listed. The logit function maps the value of $\beta' X$ to a probability bounded between 0 and 1.	Ohlsonsize	= Log(total assets/GNP price-level index). The index assumes a base value of 100 for 1968.
		TLTA	= Total liabilities divided by total assets.
		WCTA	= Working capital divided by total assets.
		CLCA	= Current liabilities divided by current assets.
		OENEG	= 1 If total liabilities exceed total assets, 0 otherwise.
		NITA	= Net income divided by total assets.
		FUTL	= Funds provided by operations (income from operation after depreciation) divided by total liabilities.
		INTWO	= 1 If net income was negative for the last 2 years, 0 otherwise.
		CHIN	= $(NI_t - NI_{t-1})/(\|NI_t\| + \|NI_{t-1}\|)$, where NI_t is net income for the most recent period. The denominator acts as a level indicator. The variable is thus intended to measure the relative change in net income.
Zmijewski (1984) Probit model	$P = \Phi(\beta' X)$ where P is the probability of bankruptcy and X represents the variables listed, and $\Phi(\bullet)$ represents the cumulative normal distribution	NITI	= Net income divided by total liabilities.
		TLTA	= Total liabilities divided by total assets.
		CACL	= Current assets divided by current liabilities.

Figure 1 – Descriptions of well-known statistical models in bankruptcy prediction (Wu, Gaunt, Gra 2010)

[1] Visually, tree-like model. Each branch represents an event and its likelihood.
[2] Inspired by the process of natural selection, a model that includes operations such as mutation, crossover, and selection. Generally used for optimization and search problems.

methods. They also suggest that decision trees, while less powerful, are underused but recommended due to their "if-then" rule-based interpretation (Kumar, Ravi, 2007). Another review of corporate failure and financial failure from 2014 agrees with these assertions and adds that decision trees (DT) are easy to interpret and powerful, especially in combination with an ensemble method, but they only work best for short term use and are easy to overfit (Sun et. al, 2014). Evolutionary algorithms (EA) such as genetic algorithms (GA) are rule-based and more easily interpreted, but usually do not perform as well as NN and SVM. (Sun et. al, 2014). The review also argues that statistical single classifier methods such as Altman's and Beaver's, which require normality, as well as the logit regression model which requires independent variables (i.e. no multicollinearity, which is hard to achieve with accounting data) are not preferred due to the assumptions that they require (Sun et. al, 2014).

Finally, on the note of linear models and linear mapping in general, one study by Barboza, et. al compares statistical methods versus AI and ensemble methods confirmed what the review stated (i.e. that statistical methods cannot perform as well as machine learning models) and drew other conclusions, including that linear models perform worse as the number of variables increases and SVM with linear kernels perform just as badly as linear models, thus confirming that models with more complex mapping abilities are preferred (Barboza et. al, 2017).

Feature Selection

Feature selection is another facet of bankruptcy prediction that has been widely studied. Features can be selected empirically with the methods for doing so falling under two categories: wrapper and filter feature selection. Wrapper feature selection methods assess the subsets of features according to their usefulness to a given predictor, by following a searching process for a good

feature subset. They are best suited for smaller datasets because the searching process becomes more challenging as the size increases and becomes prone to overfitting. The other type of feature selection is filter feature selection which attempts to select the most "relevant" features from a larger feature set. Unlike wrapper methods, this selection process is independent from the classifier used to build the prediction model and can be implemented when the feature set is large. However, these methods do not always help yield the most accurate prediction because "the filter approach ignores the interaction with the classification algorithm used to build the predictor," and "do[es] not model the feature dependency" (Lin, Liang, Yeh, Huang, 2014).

There are widely used methods such as LASSO, which one study by Tian et. al uses to select variables for two previous studies' hazard models to address the multicollinearity problem when using accounting variables. The authors also say that LASSO is good for stabilization in the face of "perturbations in the data" and its results give insight into the relative importance of the variables that it selects (Tian et. al, 2015). However, other studies often chose the variables that the models found to be most important and by majority voting. The following are examples of this.

Using an AdaBoost ensemble of ANN, one study by Fedorova, Gilenko, and Dovzhenko selected variables that were the two most significant variables received from multivariate discriminant analysis (MDA), classification and regression tree, and logit regression (LR) (Fedorova, Gilenko, Dovzhenko, 2013). A similar study employed "several statistical methods including independent sample t-test, discriminant analysis, logistic regression (stepwise), decision tree, and factor analysis" to select their features (Jeong, Min, 2009). Arjana and Masten studied CART-based selection of bankruptcy predictors for the logit model. They first generated the principal

component of each type of variable. Taking the principal components with highest loadings and with statistical significance, they used CART decision trees with dummy variables to select predictors (Arjana, Masten, 2012). And yet another study by Wang, Ma, and Yang selected features using a boosted decision tree ensemble (Wang, Ma, Yang, 2014).

Another approach is to combine expert opinion and empirical methods. One study by Lin, Liang, Yeh, and Huang implements a feature selection method that combines expert knowledge and wrapper feature selection. It categorizes financial features into seven classes according to their "financial semantics" based on experts' domain knowledge from the literature. It then applies the wrapper method to search for "good" feature subsets that contain the top candidates from each feature class. The search space of the wrapper method effectively shrinks because features have been pre-classified before the wrapper method was applied (Lin, Liang, Yeh, Huang, 2014).

Model Tuning

Empirical methods for model tuning are also a common theme in the study of bankruptcy prediction. The previously mentioned 2014 review also details studies that employed hybrid classifiers, which use other methods or algorithms to optimize the classifier being studied, to optimize parameters for the other classification algorithm. One study that performed well created a GA-SVM hybrid model that used GA to optimize both feature selection and SVM parameters (Sun et. al, 2014). Another study by Jeonga, Minb, and Kim uses a hybrid of grid search to find local optimum and GA to find global optimum to select the number of hidden nodes and the weight decay parameter of a NN (Jeonga, Minb, Kim, 2012).

Ensemble Methods

Ensemble is a machine learning concept where multiple learners are trained to solve the same problem (Wang, Yang, 2014). Three of the most common ensemble methods are boosting,

bagging and random subspace. Boosting constructs a composite classifier by training base classifiers while increasing weight on their misclassified observations. The observations that are incorrectly classified are chosen more often than examples that were correctly classified to produce new classifiers that are better able to predict observations that previously led to poor performance. Boosting combines predictions with weighted majority voting by giving more weights to more accurate predictions (Kim, Kang, 2010). Bagging, which stands for Bootstrap Aggregating is an intuitive and simple ensemble method. Diverse bags are obtained by bootstrapping different training data sets (i.e. randomly drawn with replacement). Subsequently, a classifier is built for each training data set and their predictions are combined by a majority vote (Abellán, Mantas, 2014). The random subspace ensemble consists of several classifiers that use randomly chosen sets of the original dataset and combines the classifiers into a final decision rule by majority vote. However, each single classifier uses only a subset of the available features in the data set for training and testing, and these features are chosen uniformly at random from the full set of features (Abellán, Mantas, 2014)

In terms of ensemble techniques, the 2014 review on corporate failure and financial distress discusses parallel and serial ensemble methods. Beginning with parallel methods, which use majority voting on the results of multiple classifiers, there are ensembles with different algorithms, ensembles with one algorithm under different samples or features, and ensembles with one algorithm under different parameters. Serial methods arrange several base classifiers in sequence and selects the result of one base classifier as the final output according to certain principles They include well-studied techniques like bagging, boosting and AdaBoost (although, the study stresses that AdaBoost performs best with weak learners) (Sun et. al, 2014).

Employing bagging and boosting decision tree ensembles, Kima and Kang compare ensemble methods to a NN tuned by back propagation learning algorithm. The ensemble methods removed generalization error and reduced overfitting. Overall, the bagging ensemble performed the best followed by the boosting ensemble (Kima, Kang, 2010).

One study by Abellán and Mantas uses credal decision trees (CDT) that are, "based on imprecise probabilities (more specifically, on the Imprecise Dirichlet Model (IDM)," in random subspace and bagging ensembles to improve their accuracy. They found that, because the split criterion used to build a credal decision tree has a different treatment of the imprecision than the one used for the classic split criteria, it improved the ensemble's ability to learn (Abellán, Mantas, 2014).

Other Topics in Bankruptcy Studies

Bankruptcy prediction literature knows no bounds and the facets that researchers have studied are varied. The 2014 review suggests that future studies look at the problem of bankruptcy on a spectrum from mild financial distress or a temporary cash flow difficulty to business failure or bankruptcy, and that future studies broaden their definition of bankruptcy from a dichotomous outcome to a metric that encompasses the timeframe and intensity of financial distress before bankruptcy (Sun et. al, 2014). The following studies explore these topics and other miscellaneous topics.

One study by Režňáková and Karas attempted to construct dynamic indicators of bankruptcy, assuming that the development of financial ratios is not very dynamic over the years (Režňáková, Karas, 2014). To avoid assuming that their data is normal and does not have meaningful outliers, they used non-parametric boosted decision trees ("non-parametric" meaning it does not need assumptions about parameters) to select predictors for an LDA. It used boosted

decision trees because they are nonparametric (i.e. they do not need to be normally distributed, allow for outliers, and can capture non-linear relationships between inputs). They found that, despite not seeing improved accuracy, the models were able to identify, "bankruptcy symptoms," before actual bankruptcy (Režňáková, Karas, 2014).

Another study by Tinoco and Wilson endeavored to detect financial distress as opposed to bankruptcy. The timeframe for a company to be considered bankrupt can be a lengthy process where the "legal" date of failure can differ from the "economic"/ "real" date of failure by up to two years. This model looks for financial distress because it can still cost a firm a lot of money and it includes the time that a company cannot meet its financial needs even before bankruptcy, potentially as a warning that bankruptcy is eminent. However, financial distress does not always end in bankruptcy as, "there are several stages that a firm can go through before it is defined as dead: financial distress, insolvency, filing of bankruptcy, and administrative receivership (in order to avoid filing for bankruptcy), for instance." Using accounting, macroeconomic and market variables, the study compares five models: one using only accounting variables, one using accounting and macroeconomic variables, one using only market variables, one using market and macroeconomic and one using all variables. It found that macroeconomic variables improved the model marginally yet positively and that market variables considerably increased model accuracy (Tinoco, Wilson, 2013).

Along the same lines, one study conducted by Zhou, Tam, and Fujita uses multi-class classification to predict the listing status of Chinese listed companies using ensembles of binary classifiers, random undersampling, and VIF feature selection process. Their feature selection process aided in correctly classifying delisted status rather than correctly classifying healthy

status, and their selected variables performed well with several different ensemble methods (retaining low standard deviation across the board) (Zhou, Tam, Fujita, 2016).

The Imbalance Problem

The imbalance problem is of great significance to bankruptcy prediction studies because, while there are generally more instances of healthy companies rather than bankrupt companies in a given dataset, "any degree of imbalance somewhat damages a method's prediction capacity (Veganzones, Séverin, 2018)." Below are examples of how past studies have dealt with this issue.

In a study that compares the performance of 5 models, Wu, Gaunt, and Gra balance the data by "matched-pair" methods. A pair of points (one from the minority class and the other from the majority class) are matched by the size of the firm (Wu, Gaunt, Gra 2010).

One study by Zhou compares several sampling techniques when considering the imbalance problem. For oversampling techniques, it looked at random oversampling with replication (ROWR) and Synthetic Minority Over-sampling Technique (SMOTE). SMOTE oversamples the minority class by taking each minority class sample and introducing synthetic examples along the line segments joining any of the k minority class nearest neighbors. For undersampling techniques, it looked at random undersampling (RU), Undersampling Based on Clustering from Nearest Neighbor (UBOCFNN) and Undersampling Based on Clustering from Gaussian Mixture Distribution (UBOCFGMD). UBOCFNN partitions the points in a sample space into k clusters and selects the point which is the nearest to the central point of each cluster to represent the whole cluster. (UBOCFGMD) is clustering based on Gaussian Mixture distribution using the Expectation Maximization (EM) algorithm to estimate the parameters. In a Gaussian mixture

model with k components for data in the majority class, the EM algorithm will estimate its parameters and then that data will be partitioned into k clusters in terms of the k components of Gaussian mixture distribution (Zhou, 2013). In this study, k in UBOCFGMD and UBOCFNN was equal to the size of the minority population. UBOCFNN selected the point nearest to the central point of each cluster and resulted in a training set that was double the size of the minority population. For UBOCFGMD, in the case that no points are partitioned into one cluster, for each cluster with points, one point is randomly selected to represent that cluster and for the clusters without points, one point was randomly selected from the nearest cluster with points. In general, the study found that undersampling techniques (RU and UBOCFGMD with NN and SVM) are generally better for datasets with larger minority populations and oversampling techniques (logistic regression with SVM) are generally better for datasets with smaller minority populations (Zhou, 2013).

When considering the imbalance problem, one study by Kim, Kang, and Kim examined the difference between arithmetic and geometric means in boosting algorithms to minimize the decision boundary of the majority class invading the decision boundary of the minority class. Their method starts by sampling five portions of the data with different imbalance ratios. It then performs classification experiments using AdaBoost, cost-sensitive boosting and GMBoost of SVM classifiers and compares the performances after using SMOTE to balance the imbalanced sets. Cost-sensitive boosting uses a cost to punish for positive and negative misclassifications, and normally these costs are the same, but in this algorithm, they are different and calculated using geometric accuracy rather than arithmetic accuracy. The results showed that in the AdaBoost ensemble, higher imbalance ratios led to higher arithmetic accuracy for the majority

population but lower arithmetic accuracy for the minority population. The inverse effect was seen for geometric accuracy. In the CostBoost ensemble, the geometric accuracy was higher than that of AdaBoost, but it had a lower arithmetic accuracy. Its average accuracies for the two most imbalanced sets were lower than a random guess. However, the GMBoost ensemble had better, more stable arithmetic accuracies than the previous methods as well as a better geometric accuracy. The geometric accuracy was significantly better when comparing the results of the most imbalanced samples (Kim, Kang, Kim, 2014)**.**

Cluster-based undersampling was used by Lin, Tsai, Hu, and Jhang to deal with data imbalance in a general sense (i.e. not just with bankruptcy data). The study's algorithm set the number of clusters in the majority class equal to the number of data points in the minority class. Then, one version of the algorithm (referred to as "centers") used the cluster centers as the representative of the majority class, but another version (referred to as "centers_nn") used the nearest neighbors of the cluster centers to represent the cluster. Using clustering as a replacement for another undersampling strategy has two main benefits. The first is that the elements of the clusters are relatively homogeneous. The second is that it preserves some of the information that may have been lost if another undersampling technique was used to remove some of the data points. When applying these two algorithms to multilayer perceptron classifier and decision tree classifier AdaBoost ensembles, the centers_nn algorithm used in combination with the decision tree ensemble produced the highest rate of classification accuracy. Furthermore, the results showed statistically significant improvements in both small-scale and large-scale datasets (Lin, Tsai, Hu, Jhang, 2017).

With the goal of reducing noise and overcoming imbalance between and within classes, Douzas, Bacao, and Last implemented a hybrid k-means and SMOTE clustering algorithm to deal with the imbalance issue. This study's algorithm uses k-means to generate clusters and then, "a filter step chooses clusters to be oversampled and determines how many samples are to be generated in each cluster (Douzas, Bacao, Last, 2018)." It then applies SMOTE to oversample the minority class.

Motivation

The literature surrounding bankruptcy prediction encompasses a variety of topics. Some studies focus on variable selection, others the performance of different models, etc. However, although the imbalance issue is prevalent in any real-world dataset, where the count of non-bankrupt companies is almost always greater than the count of bankrupt companies, the literature regarding how to deal with the imbalance issue is noticeably sparse. Many studies barely mention the issue or rely on empirical methods to hand-pick the points in the training sets (e.g. the "matched-pair" method used by Wu, Gaunt, and Gra). So, this study aims to further the research on how to handle the imbalance issue in a way that can be applied globally to any study, regardless of its purpose.

Methodology

The Data

The data comes from COMPUSTAT which is a database of financial, statistical and market information on active and inactive global companies. The data provides quarterly SEC filing information for companies in the USA and Canada from 2009-2019. Half of the dataset contains information for the companies within the same year and the other half contains information for one year in advance. The variables include the target variable, which is whether a company is

bankrupt or not, and the most popular variables used in bankruptcy prediction models: Return on Assets, Net Income, Retained Earnings, Total Assets, Working Capital, Earnings Before Interest and Taxes, Net Sales, Cash, Current Assets, Total Stockholder's Equity, Total Debt, and Total Current Liabilities. The target variable is severely imbalanced as each dataset has 78,321 non-bankrupt entries, but the dataset for one year in advance has 426 bankrupt points while the dataset within the same year has 353 bankrupt points.

Pre-Existing Sampling Techniques Used in This Study

There are two main categories of sampling methods: undersampling and oversampling. Undersampling reduces the size of the majority class to match the size of the minority class. The most basic form of undersampling, random undersampling, picks a subset of the majority class points randomly in order to have the same amount of points as the minority class. The other type of sampling, oversampling, increases the size of the minority class to meet the size of the majority class. Random oversampling duplicates points from the minority population randomly until there are just as many minority points as there are majority points. While this study compares the results of the proposed method to both random undersampling and random oversampling, it compares SMOTE-based oversampling techniques as well.

One of the most popular oversampling methods is Synthetic Minority Over-sampling Technique, or SMOTE, which creates new samples of the minority class based on the other minority samples around it. The algorithm starts by selecting a minority point and drawing lines between it and the closest minority points around it. It then imagines new points along those connecting lines.

SMOTE is widely used, even outside the field of bankruptcy prediction, for the following reasons. First, as with oversampling in general, none of the information in any of the majority

points is discarded, like in undersampling. Next, SMOTE has been shown to be less prone to overfitting than random oversampling. With random oversampling, points that already exist in the dataset are duplicated, and in the case of bankruptcy data, a model would only understand the type of companies that are represented by these points. But, because SMOTE generates points that are different than the points in the original minority class with features that other similar companies may have, a model would have a better chance of recognizing not only the companies in the original minority population of the dataset but also bankrupt companies that are similar.

The downsides to SMOTE are as follows. First, it tends to generate noise. If the points that SMOTE generates are not useful nor representative of the minority population, they may lead inaccurate predictions. Second, while SMOTE is not as prone as random oversampling to overfitting, it is still possible for it to lead to overfitting, especially with extremely imbalanced data.

This study also compares the Bagging Undersampling results with several other algorithms that are variations of SMOTE. Each of the following algorithms follows the aforementioned SMOTE procedures apart from the changes that follow. Adaptive Neighbor SMOTE (ANS) changes the number of lines drawn from each minority point depending on number of minority points around it. So, if a point is an outlier and does not have any minority points close to it, ANS will only draw one line between it and its closest point. The goal of this is to reduce noise while still including the information captured in outliers (Siriseriwan, Sinapiromsaran, 2017). Borderline SMOTE (BLSMOTE) generates more points from minority points that are near majority points. The algorithm takes each “pocket” of minority points and selects the points on the border of these pockets to generate more points from them. Because these borderline points are closer to

non-bankrupt points, they are often more easily misclassified as majority class by a model. The goal is to produce more of these points so that the model can learn that they are minority points (Franco, 2009). Finally, Density Based SMOTE (DBSMOTE), starts by clustering the minority points and then generates samples on the line between each point and the closest cluster's centroid. The goal is to segment the bankrupt points into clusters to capture points with similar features within these clusters. Then by only drawing one line between each point and the cluster center, noise is reduced because the points generated along this line should be within the cluster that, ideally, represents a type of company that the algorithm will see again (Bunkhumpornpat, Sinapiromsaran, Lursinsap, 2012).

Metrics for Model Performance

Each of the metrics uses information from the following confusion matrix.

		Actual	
		Bankrupt	Non-Bankrupt
Prediction	Non-Bankrupt	True Positives (TP)	False Positives (FP)
	Bankrupt	False Negatives (FN)	True Negatives (TN)

Sensitivity, otherwise known as the true positive rate, shows how well the model predicted on the minority population by taking the number of points that were predicted to be positive and are in fact positive and dividing it by the number of points that are actually positive regardless of what the model predicted.

$$Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Specificity, or the true negative rate, shows how well the model predicted on the majority population by taking the number of points that were predicted to be negative and are in fact negative and dividing it by the number of points that are actually negative regardless of what the model predicted.

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives}$$

Balanced accuracy is a metric that evaluates both sensitivity and specificity by averaging them. It is used to account for how well the model predicts for both the majority and minority populations.

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2}$$

Model Selection

There are three types of models that are used in bankruptcy prediction studies: statistical models, AI models, and ensemble methods. However, because statistical methods and linear models do not perform as well as machine learning models, this study will focus on machine learning models (Barboza et. al. 2017) (Sun et. al. 2014) (Kumar, Ravi, 2007). While numerous models have been tested in bankruptcy studies, one that the review in 2014 praised highly for its power and interpretability is the decision tree, and the study suggested using them in ensemble methods (Sun set. al. 2014). This is one of the reasons why this study chooses to implement a random forest, which is an ensemble of decision trees after bagging. Bagging, or bootstrap aggregating entails sampling from the data with replacement (bootstrapping) and then training and testing decision trees on these samples, taking a majority vote to make a prediction for each testing point

(aggregating). Random forest uses a combination of the results of many decision trees as its final prediction with the hope of increasing predictive ability. The more bags, or samples, that are taken, and the more models that are trained, the less likely the random forest is to miss out on information from the data. The only thing this model forfeits is the interpretability that is characteristic of decision trees. However, its predictive ability is much higher than that of a decision tree.

To solidify the decision to use a random forest in this study, a random forest is compared to several other machine learning algorithms including a fast implementation of AdaBoost, logistic regression, naïve Bayes, decision tree, and support vector machine. Figure 2 shows the balanced accuracy on the testing set of each model after applying random undersampling to the training set, and fastAdaBoost and random forest performs the best.

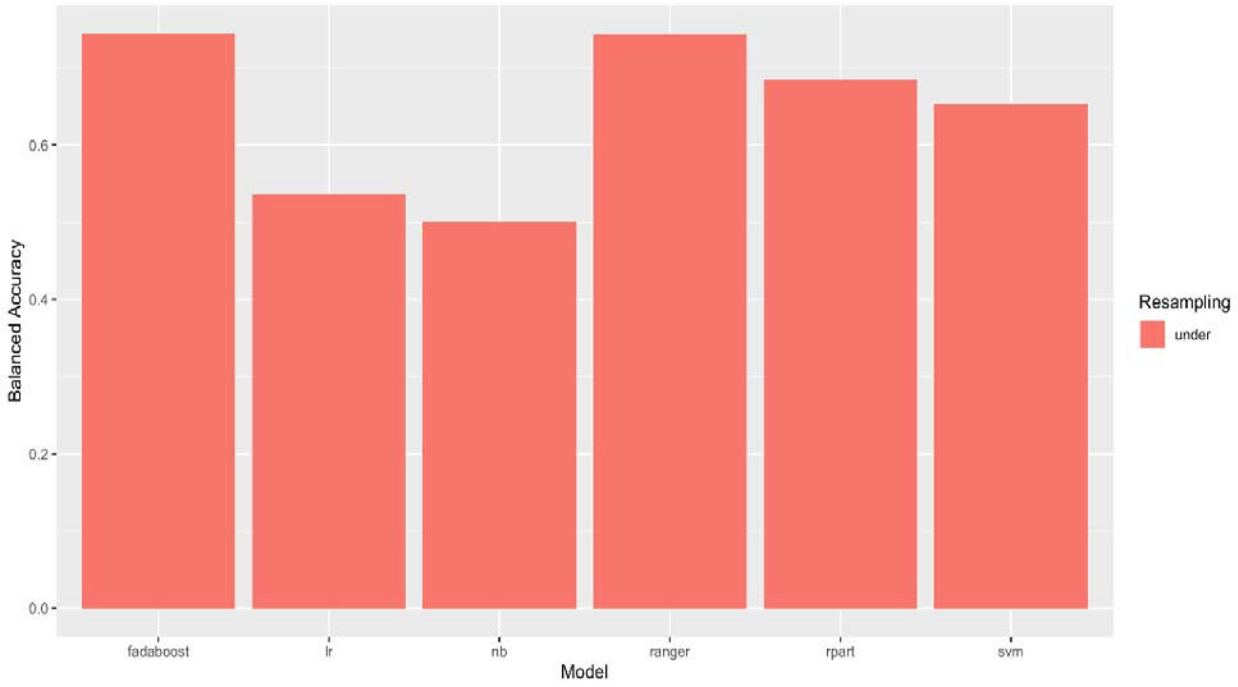

Figure 2 Balanced accuracy of several models after random undersampling.

However, as seen in Figure 3, the run time for fastAdaBoost is considerably higher than that of any other model, including random forest, so this solidified the decision to use a random forest.

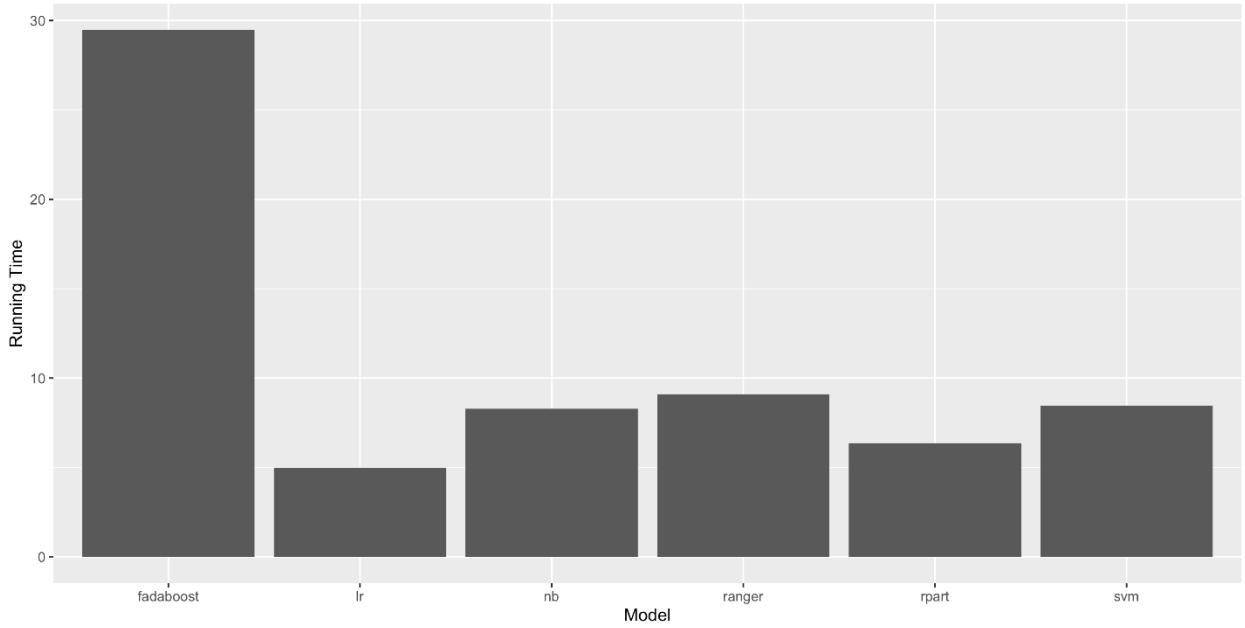

Figure 3 The run time of each model.

Proposed Sampling Method

After splitting the data with a 70/30 training/testing split, the study compares the five oversampling techniques (random oversampling, SMOTE, and the three SMOTE variations) and random undersampling. It also tests these with different imbalance ratios, meaning the sampling algorithms are used to generate different ratios of majority to minority points, and compares the results of the random forest across different levels of imbalance. The range of the imbalanced ratios used was 1 to 183, with 183 being the original imbalance ratio of the full dataset. The proposed Bagging Undersampling technique undersamples the majority population to yield a certain imbalance ratio and then repeats this on 200 bagged samples from the training set. The resulting 200 training set samples are used to train and test 200 random forests, and the resulting predictions are yielded by a majority voting procedure.

Algorithm

For imbalance ratio ($\frac{\sum majority\ observations}{\sum minority\ observations}$) = 1,2,…mm where mm equals the original imbalance ratio of the dataset:

1. Apply random undersampling to yield an imbalance ratio of k 200 times yielding 200 trainings sets.
2. Train 200 random forests on the 200 training subsets, and then receive a prediction on the testing set from each of the random forests.
3. Perform majority vote, where the most common prediction among the 200 models is used as the final prediction for each observation in the testing set.
4. Next k.

Results

Oversampling Techniques

The oversampling methods were prone to overfitting. Figure 4 shows the balanced accuracy (i.e. the y-axis) for each imbalance ratio (i.e. x-axis) after applying each oversampling technique on the portion of the data that is trying to predict bankruptcy within the same year. The balanced accuracy on the training set is in blue and on the testing set is in orange. The trend of the training balanced accuracy to slope upward while the testing balanced accuracy slopes downward is indicative of overfitting. ANS is the best of these methods, but it is still overfit.

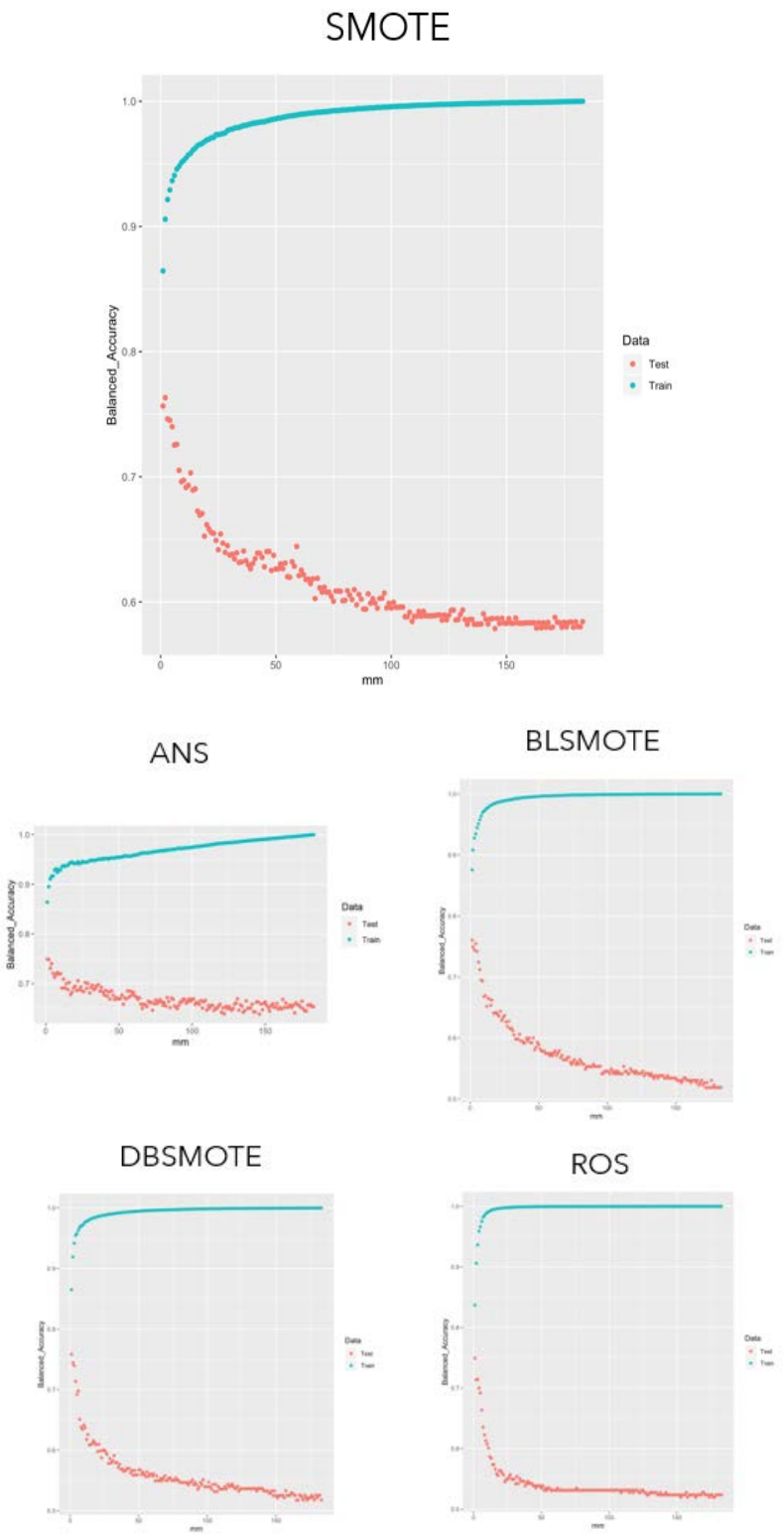

Figure 4 Balanced accuracy for models predicting bankruptcy within the same year after applying oversampling techniques. Note that the axes for each of the diagrams is the same.

Bagging Undersampling

Figure 5 shows the balanced accuracy after applying Bagging Undersampling for predictions within the same year on the left and for one year in advance on the right. For an imbalance ratio of 1.2, the model yields the best balanced accuracy and is not yet overfit because the curve of the testing points (orange) starts as an upwards slope, increasing simultaneously with the training balanced accuracy (blue). It peaks at an imbalance ratio of 1.2 and afterwards starts to slope downwards while the training curve continues to increase. In other words, it starts to overfit.

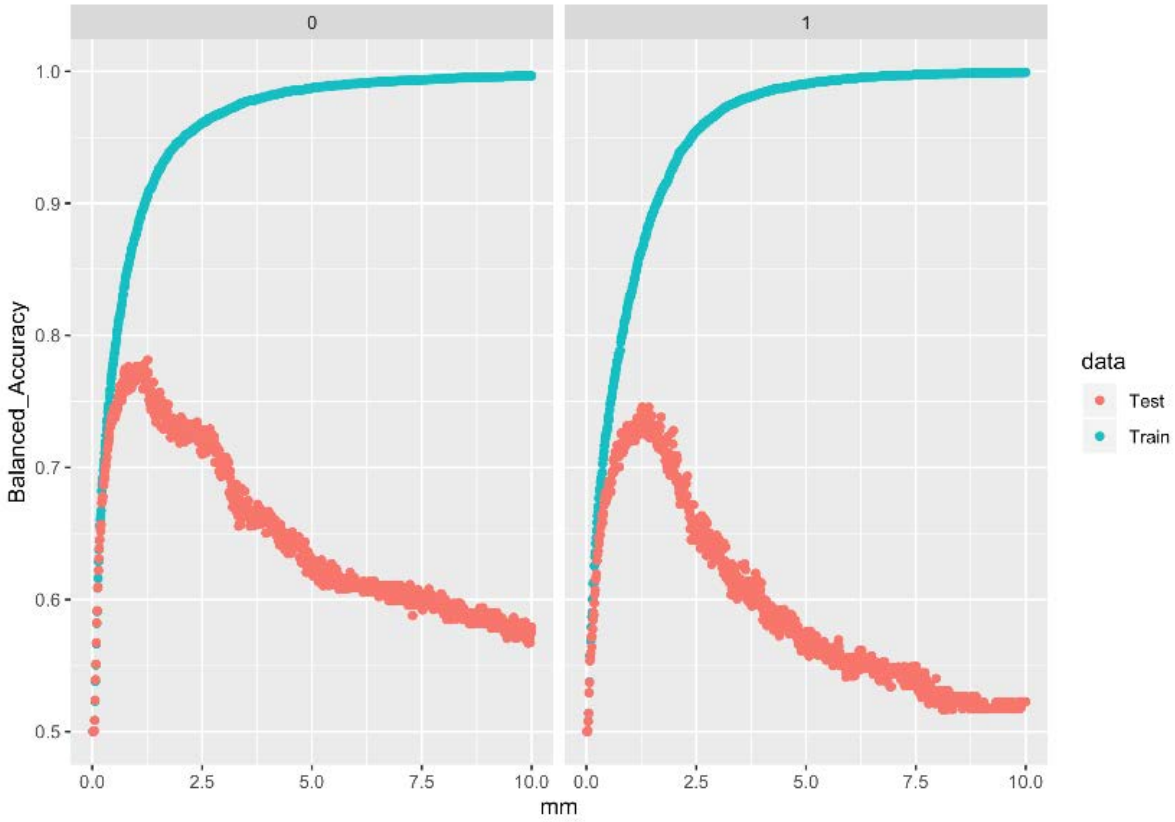

Figure 5 Training and testing balanced accuracy for models predicting within the same year and for one year in advance after applying Bagging Undersampling to the training data.

As seen in Figure 6, the sensitivity for Bagging Undersampling with an imbalance ratio of 1.2 is roughly .7, and the specificity is about .75. This means that models that used Bagging Undersampling are successful in predicting both the bankrupt and non-bankrupt populations.

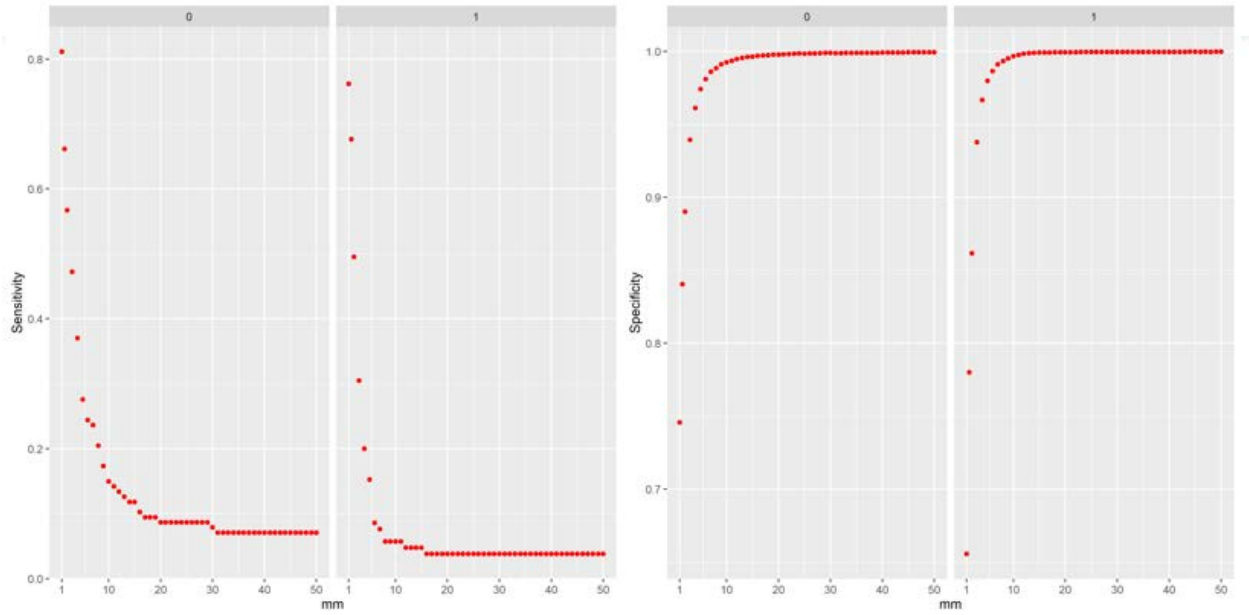

Figure 6 Testing sensitivity (left) and specificity (right) for models predicting within the same year and for one year in advance after applying Bagging Undersampling to the training data.

Figure 7 compares the performance of the aforementioned oversampling techniques with random undersampling and Bagging Undersampling. It shows the balanced accuracy on the testing set for each of these methods. For both the models that predicted bankruptcy within the same year (orange bars) and the models that predicted one year in advance (blue bars), Bagging Undersampling yields the best balanced accuracies.

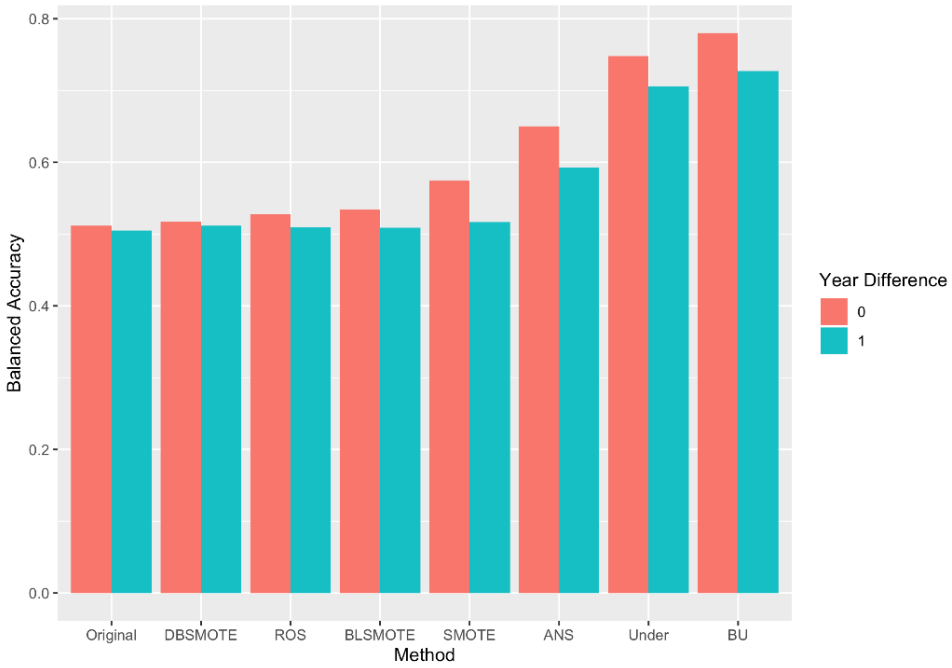

Figure 7 Comparison of balanced accuracies on the testing sets after applying each of the sampling techniques.

Conclusions

The first conclusion that this study came to is that the oversampling techniques led to overfitting, with ANS yielding the least overfit models. But more importantly, the study found that the proposed method was able to avoid overfitting and yield a notable balanced accuracy. For future study, these results suggest the use and exploration of undersampling techniques as there are many other options that this study did not investigate. Additionally, other ensemble techniques could be experimented with as well.

References

Abellán, Joaquín. Mantas, Carlos J. (2014). Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications, 41*(8), 3825-3830. doi: 10.1016/j.eswa.2013.12.003

Barboza, Flavio. Kimura, Herbert. Altman, Edward. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications, 83*, 405-417. doi: 10.1016/j.eswa.2017.04.006

Bauer, Julian. Agarwal, Vineet. (2014). Are hazard models superior to traditional bankruptcy prediction approaches? A comprehensive test. *Journal of Banking and Finance, 40*, 432-442. doi: 10.1016/j.jbankfin.2013.12.013

Brezigar-Masten, Arjana. Masten, Igor. (2012). CART-based selection of bankruptcy predictors for the logit model. *Expert Systems with Applications, 39*(11), 10153-10159. doi: 10.1016/j.eswa.2012.02.125

Bunkhumpornpat, C., Sinapiromsaran, K. & Lursinsap, C. (2012). DBSMOTE: Density-Based Synthetic Minority Over-sampling TEchnique. *Appl Intell* 36**,** 664–684. doi:10.1007/s10489-011-0287-y

Douzas, Georgios. Bacao, Fernando. Last, Felix. (2018) Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Information Sciences, 465*, 1-20. doi: 10.1016/j.ins.2018.06.056

Du Jardin, Philippe. (2010). Predicting bankruptcy using neural networks and other classification methods: The influence of variable selection techniques on model accuracy. *Neurocomputing, 73*(10-12), 2047-2060. doi: 10.1016/j.neucom.2009.11.034

Du Jardin, Philippe. (2016). A two-stage classification technique for bankruptcy prediction. *European Journal of Operational Research, 254*(1), 236-252. doi: 10.1016/j.ejor.2016.03.008

Fedorova, Elena. Gilenko, Evgenii. Dovzhenko, Sergey. (2013). Bankruptcy prediction for Russian companies: Application of combined classifiers. *Expert Systems with Applications, 40*(18), 7285-7293. doi: 10.1016/j.eswa.2013.07.032

Franco, H. (2009, April 4). LinkedIn. Retrieved April 11, 2020, from https://www.slideshare.net/hecfran/borderline-smote

Hernandez Tinoco, Mario. Wilson, Nick. (2013). Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables. *International Review of Financial Analysis, 30,* 394-419. doi: 10.1016/j.irfa.2013.02.013

Jabeur, Sami Ben. (2017). Bankruptcy prediction using Partial Least Squares Logistic Regression. *Journal of Retailing and Consumer Services, 36,* 197-202. doi: 10.1016/j.jretconser.2017.02.005

Jeong, Chulwoo. Min, Jae H. Kim, Suk Myung. (2012). A tuning method for the architecture of neural network models incorporating GAM and GA as applied to bankruptcy prediction. *Expert Systems with Applications, 39*(3),3650-3658. doi: 10.1016/j.eswa.2011.09.056

Kim, Myoung-Jong. Kang, Dae-Ki. (2010). Ensemble with neural networks for bankruptcy prediction. *Expert Systems with Applications, 37*(4), 3373-3379. doi:10.1016/j.eswa.2009.10.012

Kim, Myoung-Jong. Kang, Dae-Ki. Kim, Hong Bae. (2015). Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy

prediction. *Expert Systems with Applications, 42,* 1074-1082. doi: 10.1016/j.eswa.2014.08.025

Kumar, P. Ravi. Ravi V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review. *European Journal of Operational Research, 180*(1), 1-28. doi: 10.1016/j.ejor.2006.08.043

Le, Tuong. Lee, Mi Young. Park, Jun Ryeol. Baik, Sung Wook. (2018). Oversampling Techniques for Bankruptcy Prediction: Novel Features from a Transaction Dataset. *Symmetry, 10*(4), 79. doi:10.3390/sym10040079

Lin, Fengyi. Liang, Deron. Yeh, Ching-Chiang. Huang, Jui-Chieh. (2014). Novel feature selection methods to financial distress prediction. *Expert Systems with Applications,* 41(5), 2472-2483. doi: 10.1016/j.eswa.2013.09.047

Lin, Wei-Chao. Tsai, Chih-Fong. Hu, Ya-Han. Jhang, Jing-Shang. (2017). Clustering-based undersampling in class-imbalanced data. *Information Sciences, 409-410,* 17-26. doi: 10.1016/j.ins.2017.05.008

Mattsson, B. Steinert, O. (2017). Corporate Bankruptcy Prediction Using Machine Learning Techniques (Bachelor's thesis). Retrieved from https://gupea.ub.gu.se/bitstream/2077/54283/1/gupea_2077_54283_1.pdf

Min, Jae H. Jeong, Chulwoo. (2009). A binary classification method for bankruptcy prediction. *Expert Systems with Applications, 36*(3), 5256-5263. doi: 10.1016/j.eswa.2008.06.073

Režňáková, Mária. Karas, Michal. (2014). Bankruptcy Prediction Models: Can the prediction power of the models be improved by using dynamic indicators? *Procedia Economics and Finance, 12,* 565-574. doi:10.1016/S2212-5671(14)00380-3

Siriseriwan, W., & Sinapiromsaran, K. (2017). Adaptive neighbor synthetic minority oversampling technique under 1NN outcast handling. *Songklanakarin Journal of Science and Technology, 39*(5), 565. doi: 10.14456/sjst-psu.2017.70

Sun, Jie. Li, Hui. Huang, Qing-Hua. He, Kai-Yu. (2014). Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowledge-Based Systems, 57,* 41-56. doi: 10.1016/j.knosys.2013.12.006

Tian, Shaonan. Yu, Yan. Guo, Hui. (2015). Variable selection and corporate bankruptcy forecasts. *Journal of Banking and Finance, 52,* 89-100. doi: 10.1016/j.jbankfin.2014.12.003

Veganzones, David. Séverin, Eric. (2018). An investigation of bankruptcy prediction in imbalanced datasets. *Decision Support Systems, 112,* 111-124. doi: 10.1016/j.dss.2018.06.011

Wang, Gang. Ma, Jian. Yang, Shanlin. (2014). An improved boosting based on feature selection for corporate bankruptcy prediction. *Expert Systems with Applications*, *41*(5), 2353-2361. doi: 10.1016/j.eswa.2013.09.033

Wickham, Hadley. Grolemund, Garrett. (2016). *R For Data Science*. Retrieved from: https://r4ds.had.co.nz/index.html

Wu, Y. Gaunt, C. Gray, S. (2010). A comparison of alternative bankruptcy prediction models. *Journal of Contemporary Accounting & Economics*, *6*(1), 34-45. doi:10.1016/j.jcae.2010.04.002

Zhou, Ligang. (2013). Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods. *Knowledge-Based Systems, 41*, 16-25. doi: 10.1016/j.knosys.2012.12.007

Zhou, Ligang. Tam, Kwo Ping. Fujita, Hamido. (2016). Predicting the listing status of Chinese listed companies with multi-class classification models. *Information Sciences, 328*, 222-236. doi: 10.1016/j.ins.2015.08.036