Effective Bankruptcy Prediction Models for North American Companies

E

Rachel Cardarelli
Bryant University, USA

Son Nguyen
Bryant University, USA

Rick Gorvett
Bryant University, USA

John Quinn
Bryant University, USA

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. 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.

DOI: 10.4018/978-1-7998-9220-5.ch108

BACKGROUND

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 number of studies and in application to bankruptcy prediction since the 1990's. AI methods include decision tree, genetic algorithm (GA), 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) (Min and Jeong, 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 et al., 2007). In a review of studies from 1968-2005, the authors found that SVM and neural network (NN) (especially BPNN) are objectively more powerful than other methods, especially statistical methods. They also suggest that decision trees, while less powerful, are underused but recommended due to their "if-then" rule-based interpretation (Kumar et al., 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