Company Bankruptcy Predictive Modeling

Noah Shannon
Towson University
Towson, USA
nshann3@students.towson.edu

Anthony Gillis
Towson University
Towson, USA
agilli12@students.towson.edu

Devere Anthony Weaver
Towson University
Towson, USA
dweave8@students.towson.edu

***Abstract*—Companies have long tried to better understand their financial status and predict how market changes will affect their business. This project seeks to assist with that by creating a machine learning model that will predict if a company will become bankrupt. We will utilize a dataset from Kaggle that contains bankruptcy data from Taiwan's Economic Journal between 1999-2009. After the predictive model has been trained we will analyze the accuracy of its results.**

***Index Terms*—Artificial Intelligence (AI), Predictive, Machine Learning Model**

I. Introduction

Companies have long tried to better understand market trends and predict its short term, as well as, any long term affects on their business's health. However, this information is not enough on its own to predict how a company's internal finances are performing. Many companies have financial experts that track how the company is performing as well as the company's shortcomings. This project aims to predict if a company is going to be bankrupt to aid financial experts in making decisions that will be better guide the finances of businesses. This will be done using a dataset from Taiwan's Economic Journal between 1999 and 2009. This dataset will be used by our machine learning model to make financial predictions. We also aim to find correlations and meanings of different features in the dataset to find which features are the most important in predicting financial outcomes.

II. Description of Data

The multivariate dataset being used in this project is originally from a research report made in 2014 called "Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study" which gathered company bankruptcy data from the Taiwan Economic Journal form 1999-2009. The requirements for a company to be classified as bankrupt is defined by the business regulations found in the Taiwanese Stock Exchange. The dataset spans 96 attributes with 6819 instances. The dataset on Kaggle is taken from the UC Irvine Machine Learning Repository [1] which uses the same data from the aforementioned research survey. Kaggle lists the definition for each attribute as well as they're values. Accessing this dataset can be done by downloading it as a single ZIP file or by making a request to the Kaggle API. The dataset has only one output feature in the form of the Bankruptcy class label. The other attributes are all input features. The dataset contains no missing values and the attributes are either continuous or integers. All of the attributes, except for bankruptcy, are features. The bankruptcy attribute is the target attribute. This dataset is usually used for classification purposes.

III. Preliminary Literature Review

In a paper by Liang et al [2] the authors examine how combining financial ratios (FRs) and corporate governance indicators (CGIs) can improve bankruptcy prediction models. The research uses data from Taiwanese companies and explores seven categories of FRs and five categories of CGIs to determine the best predictors.

In the paper, the authors used different feature selection methods to reduce the dimensionality of the dataset. The study highlighted that without key features like solvency or board structure, prediction models performed worse. The methodology in this paper can be used as a baseline to help us with feature selection using the same dataset and we can even attempt to improve on their work by implementing different feature selection techniques.

To further aid in our understanding of the feature selection process, we consulted a paper by Guyon et al [3]. The paper provides a comprehensive overview of the methods and challenges associated with selecting relevant variables in machine learning models, especially when dealing with large datasets.

While we're not specifically dealing with a large dataset, some of the techniques are still applicable to our smaller dataset, for example how to use different validation techniques such as cross-validation to avoid overfitting. It also includes the goals of feature selection and how to select them based on the goals of our analysis.

IV. Proposed Methodology

To build a predictive model for predicting bankruptcy status of a company, we'll start by partitioning the dataset into testing and training datasets and then performing an exploratory analysis on the dataset and its variables. The exploratory data analysis (EDA) will help us characterize the statistical properties of each variable and any potential relationships that exist between them.

Each variable will be characterized using univariate methods graphically (e.g. density plots) and numerically (e.g. five-number summary). Then, to determine if any relationships exist, bi-variate methods will be used.

After examining the dataset variables, we'll then test for and select the variables that are likely to have the most predictive

power. To determine which set of variables will likely produce the best results, a number of statistical methods will be potentially utilized including statistical correlation, dimensionality reduction (e.g. PCA, SVD, LDA, etc.), regularization, stepwise selection, etc. However, the exact statistical techniques to use on the given dataset won't be known until the exploratory data analysis phase is completed.

After selecting the "best" subset of variables to use to build the model, it is likely that we'll need to handle class imbalances in the dataset. Financial data, to include bankruptcy data, will often have severe class imbalances due to the nature of financial systems. The amount of class imbalance can be observed in the EDA phase of analysis. To address this potential imbalance, we can choose one of several different sampling techniques to see which gives a desired result.

The three most common sampling techniques that are used in classification imbalances include random over-sampling, random under-sampling, and synthetic over-sampling. Each method has benefits and drawbacks but the decision of which to be used will be based on the results of the EDA phase.

After establishing the best features for the predictive model and handling the class imbalance, we can do a training and testing split on the original dataset. We can move onto building the mode once we've held out data for evaluation of our models. Due to the financial importance of building a good predictive model, we'll build multiple different classification models to include, but not necessarily limited to, K-nearest neighbors (KNN), logistic regression and decision trees.

All model performances will be compared to determine which will maximize model recall since we want a model that classifies the greatest number of true positives (i.e. companies that actually go bankrupt).

V. Anticipated Outcome

As mentioned in the Proposed Methodology section, there will be a comparison of all the models that have been used for this dataset based upon their performance. Although no specific models have been chosen yet, the expectation is that one model should have the highest recall and will, in-turn, be the best model for predicting bankruptcy.

References

[1] "Taiwanese Bankruptcy Prediction," UCI Machine Learning Repository, 2020. [Online]. Available: https://doi.org/10.24432/C5004D.

[2] D. Liang, C.-C. Lu, C.-F. Tsai, and G.-A. Shih, "Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study," European Journal of Operational Research, vol. 252, no. 2, pp. 561–572, Jul. 2016, doi: 10.1016/j.ejor.2016.01.012.

[3] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of Machine Learning Research, vol. 3, pp. 1157–1182, Mar. 2003, doi: 10.5555/944919.944968.