ISyE 6740 – Summer 2023

Project Proposal

Team Member Names: Enhong Liu, Yong Yan Zhu, Wen Yu Ho

Project Title: Predicting Company Bankruptcy

Problem Statement

The occurrence of company bankruptcies poses a critical problem with far-reaching consequences, including job losses, economic instability, and financial setbacks for investors. Recent statistics indicate that over 6 million Americans between May 2020 and June 2021 were unable to work because employer closed or lost business, as reported by the U.S. Bureau of Labor Statistics. Additionally, research published by the University of Pennsylvania reveals that the bankruptcy of a company leads to a loss of approximately 20-45% for shareholders (Reindl). Given these detrimental effects, it is imperative to develop effective methods for predicting company bankruptcy, as it would aid in risk assessment, loss mitigation, and stabilization of employment and economics. Furthermore, the urgency to address this issue has been heightened in the wake of the COVID-19 pandemic. The global economy has experienced unprecedented disruption, with numerous industries facing challenges they have never encountered before. Reduced consumer spending, supply chain disruptions, and temporary closures have had a profound negative effect on the financial stability of businesses. As a result, a significant number of companies have faced financial difficulties, unable to sustain their operations and leading to a surge in bankruptcies or financial restructuring. Given the widespread impact of the pandemic on the global economy, predicting company bankruptcy has become even more critical. It is essential to develop robust and accurate models that can assess the financial health of companies in these unprecedented times.

Much effort has been put into developing prediction models. However, predicting company bankruptcy is still a complex and challenging task, due to the numerous factors and variables involved. Traditional financial analysis methods, such as financial ratio analysis, cash flow analysis, and qualitative assessment (e.g., multiple discriminant analysis), have been the primary approaches employed in the field. However, these methods have certain limitations and may not adequately capture the dynamic nature of financial data and the intricate interrelationships between variables. To address the limitations of traditional modeling methods, we propose utilizing advanced machine learning techniques and predictive modeling algorithms for predicting company bankruptcy. Recent advancements in machine learning algorithms and increased computing power have made it possible to model complex financial data more accurately and efficiently. Additionally, the availability of large-scale financial datasets and alternative data sources provides a rich resource for training and validating predictive models.

Various machine learning models are being explored in literature for bankruptcy prediction, each with its own strengths and limitations in terms of accuracy and applicability (Narvekar).

- Logistic regression assumes a linear relationship between features and bankruptcy probabilities, but it may struggle to capture complex nonlinear relationships present in the data.
- Random forests, on the other hand, can handle nonlinearity and capture intricate interactions among variables, but they are susceptible to overfitting and can be computationally expensive for large datasets.

- Gradient boosting can model complex relationships and achieve high predictive accuracy, but it requires careful regularization to avoid overfitting.
- Support Vector Machines offer flexibility in handling linear and nonlinear relationships through different kernels, although the choice of kernel function and parameter tuning affects their performance.
- Lastly, neural networks excel at learning complex patterns and extracting hierarchical features, but they can be computationally intensive, require substantial resources, and lack interpretability due to their black-box nature.

Selecting the most appropriate model for bankruptcy prediction depends on the specific dataset characteristics and the trade-offs between accuracy, interpretability, and computational constraints. Thus, the overarching goal of this project is to develop and implement other advanced machine learning techniques and predictive modeling algorithms to improve further the accuracy and efficiency of predicting company bankruptcy.

Successful implementation of the proposed predictive model will have significant impacts across various sectors. It will enable businesses to proactively identify financial distress and take corrective actions to avoid bankruptcy, thus ensuring their survival and preserving jobs. Investors will be empowered to make more informed decisions, mitigating the risk of financial loss. Financial institutions can enhance their risk assessment processes and make sound lending decisions, which ultimately contribute to overall financial stability.

Data Source:

Our dataset originates from Kaggle Company Bankruptcy Prediction, which the author further sources the data from UCI Machine Learning Repository. It was used in the study "Financial Ratios and Corporate Governance Indicators in Bankruptcy Prediction: A Comprehensive Study "published in the European Journal of Operational Research in 2016. The data contains financial information and bankruptcy status of ~7000 Taiwanese companies that were on the Taiwan Stock Exchange from 1999-2009. The companies belonged to various industries including manufacturing, electronics, service, and more. For each company, there are 95 input variables including total asset turnover, current liability to assets, equity to liability, and more. Each of the variables are normalized into the range from 0 to 1. Additionally, there is one dependent variable indicating the bankruptcy status.

Methodology:

Exploratory Data Analysis

We will start with exploratory data analysis to get some big picture of our datasets. Common tools such as correlation map, label counts, and null value investigations are to be expected. This step is generally an iterative process with our data pre-processing and model buildings.

Data Pre-processing

As only 2-3% of the companies in the dataset are bankrupted, there is a class imbalance. If this problem is not addressed before building models, the models will naturally achieve 96% prediction accuracy. One method to handle imbalanced data is Synthetic Minority Oversampling Technique (SMOTE). SMOTE oversamples the minority class by generating synthetic samples (Chawla). Hence, before we build any of the models, we will balance the class distribution.

Another issue we need to consider is the correlation of the data. Companies' financial information (our input variables) are often correlated with each other. For example, different financial ratios often share a financial variable or often all depend on some specific market factors such as tax or interest rate. A business with high revenue often has high expenses such as operation cost. Additionally, if a business has good cash flow, it often has high working capital as well. Businesses in the same sector face similar challenges including market conditions and regulatory environments. Businesses across sectors are impacted by macro-economic factors such as interest rates and inflation. The correlation of these variables will often violate some model assumptions or bring a lot of noise to model interpretation.

To address the correlation issue, we will perform variable selection or shrinkage techniques (i.e., elastic net) as a part of data preprocessing or model building, depending on the specific models we are intending to use. We will also deploy cross validation as a part of data pre-processing. Additional steps include up/down sampling, feature standardization, or even extracting top eigenvectors from PCA (Principal Component Analysis). PCA is a dimensional reduction method that can identify the most influential features (e.g., feature extraction). PCA can be used to address multicollinearity in the dataset. If the data is nonlinear, we can use ISOMAP instead of PCA.

Model Building/Feature Selection

Elastic Net Variable Selection and Logistic Regression: We will be attempting to perform logistic regression on scaled full variable dataset and top principal components from our PCA results. For the full dataset, we will perform elastic net, which combines Lasso and Ridge regression. Elastic net can perform feature selection and regularization. It also can be used to handle multicollinearity in datasets. We will use elastic net to select features and generate a smaller dataset, which will allow us to build a logistic regression model that is easier to interpret. We will of course check the logistic regression assumptions are met, and if not, we might have to transform the data. Besides combining elastic net with logistic regression, we will also pair PCA with logistic regression.

Random Forest: We will build random forest models with the unscaled full dataset and the top principal components from PCA. Unlike logistic regression, the random forest model does not assume there is a linear relationship between the predictor variables and the log-odds of the response variable. Random forest can model nonlinear relationships if that is the case with our data.

K-nearest neighbors (KNN): We will use the KNN model to separate companies into two groups, bankrupt and viable companies. New data/companies will be classified depending on the status of its neighbors. We will identify the optimal number of neighbors (k) using cross-validation.

Evaluation and Results

Using feature selection techniques, we can identify the most relevant variables that impact the bankruptcy prediction, and under the help of random forest model, we can identify the most influential features that contribute significantly to the predicted outcomes. Understanding these important features provides valuable insights for investors, analysts, and policymakers in assessing the financial health of companies.

One of the key evaluation metrics for such predictions is the confusion matrix, which provides insights into the performance of the bankruptcy prediction model. It outlines four main outcomes: true positives (correctly predicted bankrupt companies), true negatives (correctly predicted viable companies),

false positives (viable companies incorrectly classified as bankrupt), and false negatives (bankrupt companies incorrectly classified as viable). Analyzing these outcomes helps assess the accuracy, precision, recall, and other prediction performance of models.

Furthermore, the receiver operating characteristic (ROC) curve and the corresponding area under the curve (AUC) are other crucial indicators of the bankruptcy prediction model's performance. The ROC curve represents the trade-off between true positive rate (sensitivity) and false positive rate (1 - specificity) at various classification thresholds. A higher AUC value indicates a better discriminatory power of the model, meaning it can distinguish between bankrupt and viable companies more effectively. We hope to see AUC values close to 1, which indicates a high prediction accuracy.

We expect our initial models to perform well on training data because of overfitting but perform worse on test data since machine learning models usually perform worse on financial data or human-related variables. Furthermore, we anticipate the high correlation of the input variables will bring down our overall model performance even after our variable selection techniques. However, our model will still provide important insight to decision makers.

Citations

1. Bureau of Labor Statistics, U.S. Department of Labor, *The Economics Daily*, 6.2 million unable to work because employer closed or lost business due to the pandemic, June 2021 at https://www.bls.gov/opub/ted/2021/6-2-million-unable-to-work-because-employer-closed-or-lost-business-due-to-the-pandemic-june-2021.htm (visited *June 19, 2023*).

2. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

3. Liang, D., Lu, C. C., Tsai, C. F., & Shih, G. A. (2016). Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European journal of operational research*, *252*(2), 561-572.

4. Narvekar, A., & Guha, D. (2021). Bankruptcy prediction using machine learning and an application to the case of the COVID-19 recession. *Data Science in Finance and Economics*, *1*(2), 180-195.

5. Reindl, J., Stoughton, N., & Zechner, J. (2017). Market implied costs of bankruptcy. *Available at SSRN 2324097*.