Midway Report - Bankruptcy Prediction

Analysis and prediction for company bankruptcy based on company asset attributes

TaeJoon Kim
Computer Science Department
North Carolina State University
tkim12@ncsu.edu

Nick Thompson
Computer Science Department
North Carolina State University
nathomp3@ncsu.edu

John Widdifield
Computer Science Department
North Carolina State University
jfwiddif@ncsu.edu

1 INTRODUCTION AND BACKGROUND

1.1 Problem Statement

These days, a lot of new companies come into industry and sometimes world-class top companies collapse against the challenges of new competition. So, there cannot be the forever top company in the world and bankruptcy is always a possibility. Inspired by the rise of Tesla and the collapse of old traditional company Nokia after challenges by Apple, Samsung, and Huawei from all around the world, we wanted to make prediction model for company bankruptcy so that each individual could make wise decisions when it comes to choosing a company to work for, investing money in the stock market, and also the company could make actions before the collapse by improving the few factors that could contribute to the bankruptcy the most and focus on them to prevent bankruptcy.

The goal of this project is to predict whether a company is likely to go bankrupt based on their company assets and economic activity.
There has already been many researches to the prediction of company bankruptcy. And the random forest tree has been found to be the best model to use for this case [4]. However, we are going to use other ensemble models to compare the accuracy of the prediction with the result of the random forest model and we will also discover the top features that will influence the chance of bankruptcy the most.

1.2 Related Work

Feature selection in bankruptcy prediction [1]
In this paper, t-test, Principal Component Analysis, Factor Analysis, Correlation Matrix, and stepwise regression were used for feature selection and they found stepwise regression to perform the best based on average accuracy and factor analysis to perform better than the others in terms of recall rate. In this project, Linear Regression, Lasso Regression, Correlation Matrix, and Factor Analysis are going to be used and the one with best recall will be chosen as feature selection for our random forest classifier to predict bankruptcy since we don't know that Factor Analysis will work the best for our dataset.

Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study [2]
This paper goes into detail about attribute selection. The authors first split attributes into two sub categories. The first category of attributes were financial ratios and the second category were corporate governance indicators. Basically, they compared the difference of model accuracy (across different markets) when selecting a various number features which belong to the different categories. The results they found, which are of interest to us, are that in Chinese markets (because our data comes from the taiwan stock exchange) the inclusion of attributes belonging to corporate governance indicators don't provide value to the model.

Machine learning models and bankruptcy prediction [4]
This paper details the application of different machine learning models to predict bankruptcy. The models which they tested were: Bagging, Boosting, Random Forest, SVM, ANN, Logistic Regression, and MDA. What they found was that bagging, boosting, and random forest performed the best with

random forest being the top performer. This research helps guide us in decision making with respect to which classifiers to focus on for lowest prediction error as well as which classifiers have not been thoroughly tested.

2 METHODOLOGY

2.1 Approach

1. Preprocessing

Linear Regression, Lasso Regression Correlation Matrix, and Factor Analysis will be used for feature selection. They will be used to eliminate features that are unnecessary or have overlapping influence on the classification. Z-score normalization will also be used prior to the feature selection to balance the influence of each company's asset features that have different units of measure.

2. Model Training

According to the research paper, among various ensemble methods, the random forest method works the best to predict the bankruptcy. For this reason, the Random forest model will be used for classifiers. Also, it will be compared with other conventional classifiers that’s not belong to ensemble methods. For this purpose, KNN classifiers will be used to evaluate the performance of the model. The accuracy of the two models will be compared to choose the best model for prediction. Also KNN classifiers will be used to find out the top most significant factors to the bankruptcy prediction.

3. Model Tuning/Evaluation

LOOCV(Leave One Out Cross Validation), Hold Out, and K-Folds Validation are going to be used for model evaluation. In the evaluation of the model, recall is the more important factor than precision or accuracy since the cost is way greater than the others when we falsely classify that company will not go bankrupt when it should be classified to go bankrupt.

4. Novelty

While the prediction of the bankruptcy model is a famous topic and there are many researches about it. The novelty of this project lies in the fact that it is comparing one of the ensemble methods that work the best among others with conventional classifiers, KNN that predicts bankruptcy by measuring similarity between factors potentially contributing to the same classification. Also we will find out the top 10 influential factors that could lead to bankruptcy by using various feature selection techniques before training the model and compare the accuracy and recall of trained random forest and KNN models on the new test dataset.

2.2 Rationale

Multiple feature selection methods including Linear Regression, Lasso Regression, Correlation Matrix, Factor Analysis will be compared and used since the bankruptcy could be influenced by an unexpected combination of multiple factors so we should not rely on the single feature selection technique for dimensionality reduction. In addition to, embedded feature selection in the random forest model will be used in the process of model training.

Random forest methods were used as a model to predict bankruptcy since ensemble method is an aggregation of simple classifiers that are relatively simple to implement yet it is the best method to reduce the generalization error which matters the most when it comes to accuracy of the prediction. Among various ensemble methods, random forest was chosen since it is proven by the research paper[4] to show the best performance in terms of accuracy. However, we are not just simply using random forest methods for the prediction since there’s still a chance other non ensemble methods might perform better than the random forest model. Therefore, KNN classifiers will be compared with the random forest model to choose the best model for prediction.

For model evaluation, the precision rate will be compared and LOOCV, Hold out, K-Folds validation will be used to evaluate the performance of the model.

3 EXPERIMENT

3.1 Dataset

The Company Bankruptcy Prediction dataset [3] that we will be using was obtained from Kaggle. The data itself was obtained from the Taiwan Economic Journal during 1999 to 2009. Bankruptcy was defined by the Taiwan Stock Exchange. The dataset includes 95 attributes, one class label (bankruptcy), and 6819 instances. Many of the attributes are ratios and percentages displayed as a decimal, however there are a few attributes that are not, such as operating profit per person. All attributes would be categorized as financial ratios to the exclusion of corporate governance indicators. This, however, should not impact the training of our models much as there should be little disparity when excluding corporate governance indicators [2]. The dataset does not include any missing values.

3.2 Hypothesis

The main question that we pose is: Can we predict (with high level of confidence) whether a company is likely to go bankrupt based on their company assets and economic activity.

A second hypothesis that we would like to test is whether or not there are improvements to be made by testing different models comparing their results and determining which classifier works best on the dataset we have.

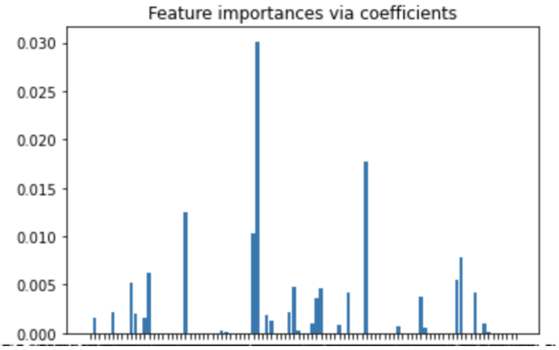
A third and final question we would like to answer is what attributes do companies need to pay the most attention to. Or rather, what specific financial ratios are prevalent in predicting bankruptcy so that companies who receive analysis can focus their resources toward.

3.3 Experimental Design

We start by testing the third hypothesis. The result from this test will determine what form of preprocessing we will attempt to apply to our first hypothesis. Here plan on using a few different techniques to determine which attributes to retain (and that are predictive of bankruptcy). We applied a lasso regression to the data and used the attributes which it finds are significant. We plan on comparing these attributes to the other attributes found by different means of attribute selection and then unioning them together to get the best attributes to use for our models.

4 RESULTS

4.1 Partial Results



According to the lasso regression model. Among 98 variables, top 10 variables that had the most influence on the bankruptcy are Debt ratio %, Operating Funds to Liability, Revenue Per Share (Yuan ¥). Total debt/Total net worth, CFO to Assets, Cash flow rate, Cash Flow to Liability. Non-industry income and expenditure/revenue, Total Asset Turnover, Revenue per person. At least these 10 variables are going to be used for training prediction models. The z-score normalization process was done when we chose a training data set for the lasso regression model. Repository:https://github.com/mark7588/Company_Bankruptcy_Prediction.git

5 PROPOSED WORK

5.1 Design of Future Experiments

1. In addition to the linear regression that we have done we plan to do SVC and linear regression to the dataset in order to get valid attributes. From these two additional feature selection techniques we will union all three results to determine the final attributes that we will use to train our model.
2. After this, we will test our second hypothesis using the now dimensionality reduced dataset. To do this we will compare results using recall from the following models:

- Random-Forest
- KNN
- Linear Regression

The recall is going to be used to find the best model since the cost of falsely predicting the company not to go bankrupt is higher than any other factors to consider. Also, the precision should not matter much since if the company is predicted to go bankrupt, it is better for the company to take action if there's a small chance of going bankrupt and it is also better for the individual not to invest money on the company that has a chance of going bankrupt even when the company might not go bankrupt.

We will estimate the generalization error by comparing results from the LOOCV, Holdout, and K-Fold results from each model.

3. Then we will use the best model (as given to us by our answer to the second hypothesis) with the most significant attributes (as given to us by our answer to the third hypothesis) to create a powerful model to predict whether a company will go bankrupt or not. This model should predict bankruptcy consistently, satisfying our first hypothesis and if it does not we will reflect on decisions which could have been made to improve it.

5.2 Plan of Activities

All of the activities will be equally done by the team members together at the meeting. However, a certain member has a responsibility for leading each designated process before the meeting and throughout the meeting,

1. Preprocessing (TaeJoon, Nick)

Linear Regression, Correlation matrix, Factor Analysis will be done to the dataset. For Linear Regression, alpha value will be decided based on the multiple experiments with the different values. Correlation Matrix will be built to find out the factors that have an overlapping impact on the prediction and those features will be removed before the model training, For factor analysis, PCA(Principal Component Analysis) will be used for factor extraction and then Varimax rotation will be used for factor rotation.

2. Model Training (John, TaeJoon)

Random forest and KNN models will be trained with the features selected by the preprocessing. For the random forest, John will take lead in training the model. Parameters such as max_features, n_estimators and random_state are going to be decided based on the performance of the model. For KNN classifiers, TaeJoon will lead the model training. K value will be decided on the performance of the model based on the recall.

3. Model Evaluation / Selection (John, Nick)

John and Nick are going to compare the performance of each model based on the recall by constructing a confusion matrix based on the result of the testing model with the random test dataset from the data. For the selection of the test dataset K-Fold, hold out, LOOCV(Leave One Out Cross Validation) will be used and each related variables such as K value for K-Fold, proprition of test and training data set for hold out will be decided based on the recall of the prediction.

4. Analysis of result and the applicability of model (TaeJoon, John, Nick)

Everybody in the team member will be working together on 4/26 meeting to analyze the performance of the selected model by using a randomly selected data set and apply the model to the dataset consisting of actual companies which collapse in the past or arise from the failure.

6 COLLABORATION

Collaboration will take place throughout the week regularly updating team members with our individual progress on deliverables which we collective set as a group during meetings. These updates will occur via a discord server. Our virtual group meetings will occur via zoom on the following schedule:

● April 12th	6:00p - 7:00p
● April 14th	6:00p - 7:00p
● April 19th	6:00p - 7:00p
● April 21th	6:00p - 7:00p
● April 26th	6:00p - 9:00p
● April 28th	6:00p - 7:00p

7 REFERENCES

[1] Chih-Fong Tsai. 2008. Feature selection in bankruptcy prediction. (August 2008). Retrieved March 23, 2021 from https://www.sciencedirect.com/science/article/abs/pii/S0950705108001536

[2] Deron Liang, Chia-Chi Lu, Chih-Fong Tsai, and Guan-An Shih. 2016. Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. (January 2016). Retrieved March 23, 2021 from https://www.sciencedirect.com/science/article/abs/pii/S0377221716000412

[3] Fedesoriano. 2021. Company Bankruptcy Prediction. (February 2021). Retrieved March 23, 2021 from https://www.kaggle.com/fedesoriano/company-bankruptcy-prediction

[4] Flavio Barboza, Herbert Kimura, and Edward Altman. 2017. Machine learning models and bankruptcy prediction. (April 2017). Retrieved March 23, 2021 from https://www.sciencedirect.com/science/article/pii/S0957417417302415#eco m0001