Final 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 in the unstable globalized market 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 much research into 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 classification 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 of 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 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 is much research 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.

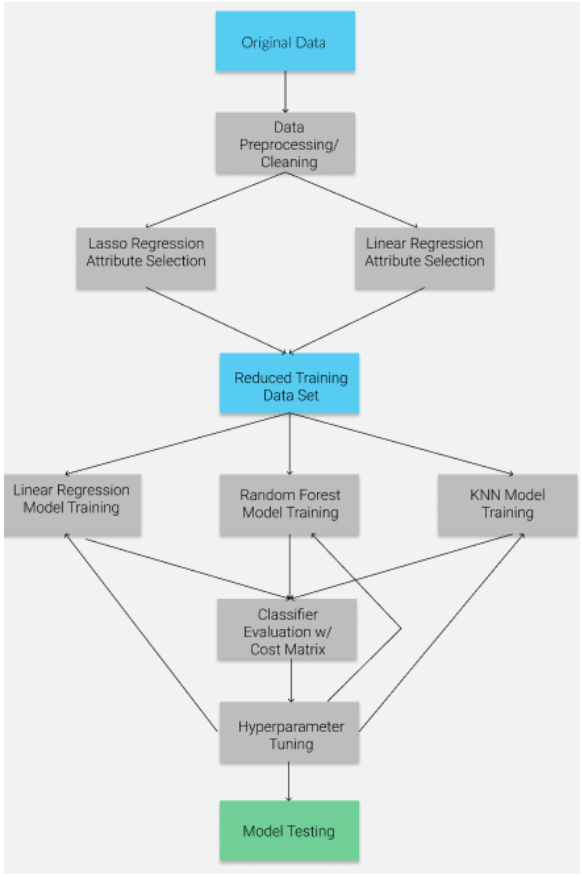

Figure 1 - Pipeline Design

4 RESULTS

4.1.1 Feature Selection Results

From the preprocessing stage, we received the following results for important attributes from the coefficients returned by the lasso classifier in order from most important to least:

Liability-Assets Flag
Long-term fund suitability ratio (A)
Total debt/Total net worth
Current Liability to Current Assets
Cash Flow to Sales
Cash Flow to Equity
Inventory Turnover Rate (times)
Current Liability to Assets
Fixed Assets Turnover Frequency
Interest-bearing debt interest rate
Quick Assets/Current Liability
Total Asset Growth Rate
Working capital Turnover Rate
CFO to Assets
Total Asset Turnover

And, we received the following attributes from the linear regression classifier:

Debt ratio %
Net worth/Assets

So, we chose the most important attribute that each attribute selection method found. And, on a rotating basis selected the top one from each. This resulted in a new dataset containing only the following attributes (excluding the class label):

Debt ratio %
Liability-Assets Flag
Long-term fund suitability ratio (A)
Net worth/Assets
Total debt/Total net worth

According to our feature selection algorithms, these attributes are the most prevalent in predicting bankruptcy answering our 3rd hypothesis.

4.1.2 Model Comparison Results

So, we trained our three models (Linear Regression, KNN, and Random-Forest) using the new dataset. And, tested them which gave us the following evaluation metrics on each model. The results are displayed on table 1 below:

	Accuracy	Precision	Recall	F1
Linear Reg	96.7%	38.9%	23.7%	29.5%
KNN	95.6%	21.8%	20.3%	21.1%
Random Forest	97.2%	75%	5.1%	9.5%

Table 1

These results were derived after optimizing our models (hyperparameter tuning) based on cost. This cost was created by the cost matrix in Table 2.

Cost Matrix		Predicted Class	
		-	+
Actual Class	-	-1	1
	+	25	-50

Table 2

Passing all of our results through the cost matrix resulted in the following costs for our models:

	Cost
Linear Regression	-1518
KNN	-1326
Random-Forest	-735

Table 3

The lower the cost, the better the model did when results are weighted by importance by the cost matrix. This means that Linear Regression did the best, then KNN, and then Random-Forest did the worst. This answers our second hypothesis with Linear Regression as it performed the best on our dataset.

4.2 Discussion

Our goal of the prediction model is to predict bankruptcy as accurately as possible and then minimize the misclassification of bankruptcy when the actual result is positive. Therefore, the penalty for false-negative should be way higher than false positive since in reality, a lot more things are at stake when the model predicts the company will not go bankrupt when it is going towards bankruptcy. On the other hand, when the model predicts that it would go bankrupt when it turns out not to go bankrupt. There's not much to lose in the perspective of an investor or prospective employer. However, the model should still aim for minimizing any

misclassification so we had to penalize a little bit for false positives too. For a true positive, we will have to put on the highest weight since predicting bankruptcy correctly is the main goal of our model. For a true negative, we put on little less weight than a true positive since catching the company goes to bankruptcy is more important than catching that it is not. However, the true negative has a relatively high weight compared to false negative and false positive since catching a company that does not go bankrupt will be highly beneficial to the individuals investing money in the company.

Oddly enough, our results do not follow the research. The research says that random forest model should be superior to KNN and it has been undocumented whether or not it would be better than linear regression. However because linear regression is typically not used in classification, it would be reasonable to assume that it would perform poorly on such a complex classification. Our odd results originate from the fact that we have a very sparse dataset. There are only very few positive attributes in our data. This makes it much more difficult to predict with accuracy the positive attribute. While KNN should perform better on data which has rare positive class attributes it still did not do well simply because there are so many variables and the positive attributes are very sparse. We were able to get accuracy above 97% however this also was because of our rare class attribute. Originally, many of our models simply classified everything as not going bankrupt and nearly had perfect accuracy. However, to combat this we needed to edit the parameters to minimize our cost function so as to get more meaningful predictions. This in turn makes our models seem worse but this is not the case, rather, they now produce much more meaningful results.

We somewhat supported our first hypothesis. We can, with some accuracy, predict whether a company will go bankrupt or not. And this should be somewhat generalizable. However I believe that we could have done much better given a better dataset that is not so sparse. Because most businesses will go bankrupt however it is odd that there is not more data. This could be a result of the fact that bankrupt companies have little incentive to publicize their data and that they are removed from the stock exchange where there can be no more data from them.

Our second hypothesis was supported. In the research we could find, there were no pipelines using linear regression for this specific application. And, given a dataset with sparse positive class attributes that you would like to use to predict bankruptcy, linear regression seems to be the best option out of the three models we tried.

Finally, our third hypothesis was supported in that we retrieved attributes that lead to 97% accuracy (even though this isn't the most important metric). Those attributes which are important to optimize to improve your chances to not go bankrupt are:

Debt ratio %
Liability-Assets Flag
Long-term fund suitability ratio (A)
Net worth/Assets
Total debt/Total net worth

<Reasoning behind random forest model>
Based on the confusion matrix, the random forest model has a high accuracy rate but relatively low precision and recall rate. Comparing the result to the other model, the others were selected to be used as the final model since it shows low recall rate, which is an important goal of our prediction model. Also embedded feature selection function was not practical to our project as we planned in the project proposal since we already reduced the dimensionality of features by using preprocessing methods such as linear Regression and Lasso Regression. The potential reason for low recall rate could be less dataset for the bankrupt company and limited number of

features that determined the classification. Also the poor performance of the random forest model could be explained by the fact that there were only 5 features to train the model. Multiple decision trees constructed for the random forest model have based on those 5 features. However, generally the performance of the random forest model is maximized when the complexity of features is high enough to get a good amount of complex decision trees to select from.

5 CONCLUSION

Throughout this report we have demonstrated the use of three models to predict bankruptcy.

- We prioritized true positives and penalized false negatives by use of a cost matrix. After applying the cost matrix, we determined linear regression to be the best model.
- We addressed our three hypotheses
- We learnt that even with a small number of attributes we were able to accurately predict the outcome of a company.

The results from this report are quite different from what we expected before performing the experiments. Originally, we were under the assumption that the random forest model would be the most accurate, however, linear regression turned out to be the best. During our experiments we predicted KNN would out-perform the other models due to the low positive rate of bankruptcy in our dataset. The experiments helped us to determine important attributes of a dataset and apply them to a model which could be used by existing companies to predict their bankruptcy.

Even though we ended up with a different result from our expectation, our experiment is still meaningful that our prediction model could detect the company that will not go bankrupt, as proved by the high accuracy rate. This could still be a useful model for the individuals who would like to search for the stable companies to invest money. Also during the experiment, we could still figure out the top factors affecting the company bankruptcy so that other researchers could still use these features for constructing a company bankruptcy prediction model with the different dataset for their research. Our company bankruptcy prediction model could be used to check if any startup company has a potential of going bankrupt or not to help the investors to choose their investment strategy.

6 MEETING ATTENDANCE

Meetings related to the project were recorded in this section along with the attendees.

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

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 m000

8 GITHUB LINK

https://github.com/mark7588/Company_Bankruptcy_Prediction.git