Machine learning for Bankruptcy Prediction Final Report

Abhay DA
SRN : PES1UG19CS011
Computer Science and Engineering
PES University
Bangalore, India
abhayda2001@gmail.com

Akshay SP
SRN : PES1UG19CS046
Computer Science and Engineering
PES University
Bangalore, India
Akshaysp2090@gmail.com

Akash S
SRN : PES1UG19CS042
Computer Science and Engineering
PES University
Bangalore, India
akashpari2013@gmail.com

Akshay SD
SRN : PES1UG19CS045
Computer Science and Engineering
PES University
Bangalore, India
akshaysdoddamani@gmail.com

***Abstract*—Bankruptcy prediction constitutes an important area of research. In recent years artificial intelligence and machine learning methods have achieved promising results in corporate bankruptcy prediction.**

In this report we aim to analyse how well various machine learning models fare in predicting bankruptcy of an organisation. We will compare various models and rank them based on their performance on the test data.

***Index Terms*—Financial Distress prediction, Machine learning, Bankruptcy prediction**

I. Introduction and Background

In today's world of business and economics the correct way of usage of funds to maximize the end results of the product being developed is very important and should also be in accordance with the trend of the market present, to achieve success for the product. When the allocated resources be it monetary or materials or manpower are not used in the proper way which result in the failure of the product along with the losses the company takes for the following. The company would have taken certain credit or funds to fulfill the requirements of the product and when it fails leads to owing of debts to the creditors and over a period of time the company may not be able to repay back the debts due to various factors like repeated failure of product, product not according to the current trend in market to name a few then bankruptcy comes in to the picture. Bankruptcy is a legal process wherein the company or any other entities who cannot repay the debts to their creditors, may seek relief for some or all of their debts, initiated by the debtor. The word bankruptcy usually means "broken bench" or "broken bank" here.

The dataset that we have chosen tells us which company have gone bankrupt depending on various factors or attributes, in the country of Taiwan. The dataset that we deal with also tells us which companies in the country are on the verge of bankruptcy by taking into consideration such as ROA(A), Debt Ratio %, Net Income to Total Assets, Current Liability to Assets, Working Capital to Total Assets, and Long-Term Liability to Current Assets. These attributes or factors are the ones that have the largest contribution of saying that the whether the company is bankrupt, going to be bankrupt in near future or is not bankrupt. This above process is called bankruptcy prediction. It is basically predicting the financial distress of public firms and is very important due to the fact that it enables the creditors and the investors to look at the financial situation of the firm before investing into the future of it.

This model helps us calculate various or numerous accounting ratios as mentioned above that might indicate danger along with various other explanatory variables. This area is well suited for testing increasingly sophisticated, data intensive forecasting approaches or models. With this help of this early signs of bankruptcy we might be able to reduce the economic losses that it entails both in qualitative and quantitative terms. At the end our model solves above problems and makes a detailed report of the financial situation of the firm in Taiwan from the year 1999-2009.

II. Previous Work

A. Brief Overview

We have taken into account the two most important predecessor work which guides us in the right direction. The first study is in fact the first study of forecast which uses the Economic index to predict. Further evolution of the study led scholars to come up with three main factors that really affect the bankruptcy of a company. The second one is a Z-score model that predicts if a company goes bankrupt in 2 years based on three important positions of a company. The

positions are safe, grey, and problem areas. Given the various features (information) related to that company like assets, fees, etc… we can calculate a Z-score corresponding to which we map three positions as described above. Depending on that position we can say we a company will go bankrupt in 2 years or in the near future or how similar it is to the companies that are similar to the one in context.

B. Limitations identified

We have seen rather an uncanny limitation that really concerns the composition of the dataset. We have identified that the data is largely imbalanced, meaning the number of companies that are bankrupt is minuscule when compared with the companies which are standing firm.

The challenge in working with an imbalanced dataset is the fact that most machine learning models tend to ignore the minority class and in turn will have a poor performance on the minority class as data imbalance makes is tough for the model to choose an appropriate decision boundary. But typically the performance in the minority class is more important when it comes to problem statements such as ours.

Since the number of non-bankrupt companies is extremely low, we cannot resort to undersampling methods as it will reduce the amount of data we will be left to work with. To tackle this problem we resort to some well known oversampling techniques. We have used accurate frameworks to achieve the same. This gives us a good perspective on the way the features contribute to the bankruptcy of a company. In this method, we try to balance the dataset composition so that the machine learning algorithms can infer more accurately than was possible in the later stage.

C. Assumptions made

We have assumed that the features in the dataset comprising of 96 robust columns have an unequal effect on the outcome of the bankruptcy, but in the meantime, we also assume that the given features are the only factors playing a role in the downfall of the company, meaning the features are alone responsible for the bankruptcy of the company excluding any external issues or calamities.

III. Proposed Solution

Pre-processing is the process of dropping of values in order to enhance the performance of them model to get more accurate and reliable data to make predictions based on it. We pre-process the data by dropping all single value data, null values. Exploratory data analysis is an approach of analysing datasets to summarise their main characteristics often using statistical graphs and other visualization methods. It often tells us how best to manipulate data in order to get the best possible result, discover patterns, anomalies, test hypothesis and to check our assumptions made based on the dataset.

EDA makes use of factors, attributes or columns which have a very high proportionality, or dependency or have a say in whether the company is going bankrupt or not. In order to determine the relationships between these factors or columns with bankruptcy we draw graphs and determine how much effect it has on the dataset.

In our model we make use of ROA(A) which is return of assets, debt ratio %, Quick ratio, current liability, working capital and long term liabilities to determine or to estimate the number of companies which are bankrupt or on verge of bankrupt. This allows the creditors and investors to know the financial situation of the company and also even tells the company about its own situation in terms of the cashflow generated and spent. At the end we make use of a heat map to determine the magnitude of effect each attribute or factor has its influence on the situation of the company. One major problem present in our dataset is the imbalanced dataset as majority of companies are not bankrupt. To solve this problem we make use of oversampling method particularly smote technique reducing imbalance in the data. We don't use undersampling method due to the fact that many important features will be lost will dealing with this technique. Smote stands for Synthetic Minority Oversampling Technique where synthetic samples are generated for minority classes. It mainly focuses on feature space to generate new instance by interpolation between positive instances that lie together. Afterwards we normalize the data using the standard scaling techniques.

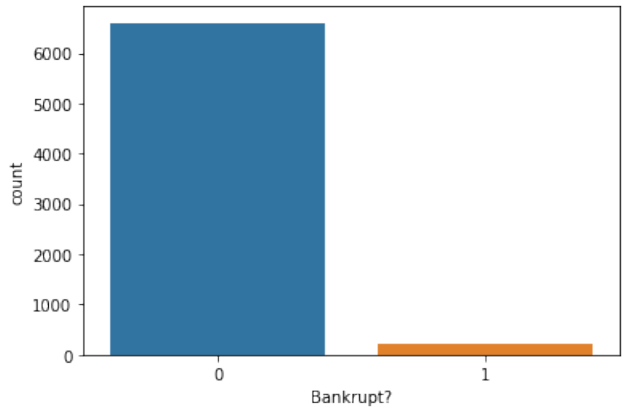
Next step, we made use of PCA (principal component analysis) which is a collection of points in a real co-ordinate space. It is the process of computing the principal components and using them to perform a change a basis of data to get the condensed form of components influencing the prediction mainly. It is an exploratory data analysis used for making predictive models. Then we later experimented with the various models to get the best prediction metrics and the data models used such as Support vector machines (SVM), Logistic regression (LR), k-nearest neighbours (knn), decision trees, Neural network, Random forest and gradient boosting for prediction. These models help predict the values and the factors leading to bankruptcy with great accuracy by identifying the pattern present between the factors.

The percentage prediction of each model after running through the dataset will be displayed without applying the PCA which are the original models. By solving the data imbalance present with the help of PCA we get reduced model which will be more accurate and more precise. The one with the highest prediction percentage is chosen to represent the the bankruptcy model to investors and creditors as they are more accurate and gives a better picture of the financial situation of the firm.

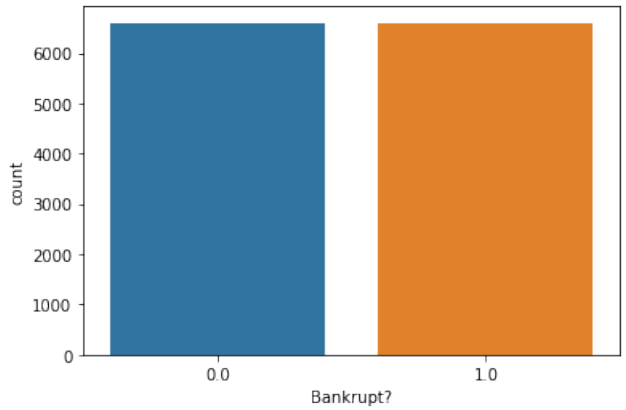
IV. Experimental results and insights gained

As described in the above section we have used various regression and classifier machine learning algorithms which is eight in total to be precise. Training all these algorithms and analyzing them for our dataset has given us insightful results. The main problem with our data as mentioned earlier was imbalanced data, which is typical in binary classification cases. As we have the majority of the companies as not being bankrupt and the minority being bankrupt we had to solve this issue beforehand to use any accuracy metrics on the data for that matter. This is because the metrics would have high bias and give inaccurate results. As mentioned in the above section this was solved by an oversampling method called SMOTE. Before using this method the non-bankrupt companies were a whopping 6599 and the bankrupt companies were a meager 220. After using this method to augment some of the instances in the minority class the dataset balanced to an equal composition of 6599 companies both bankrupt and non-bankrupt. Now we had an equal composition of binary classification dataset to move forward, analyze the data with models, and apply accuracy metrics.

Balance of data before performing oversampling :



Balance of data after performing oversampling :



After balancing the data there were some features that were not scaled. So the applying a StandardScaler resolved the issue to give us a completely scaled dataset to work on. After partitioning the dataset to give us train and test datasets we were ready to train our models. We trained eight models in total and analyzed the metrics to see which was performing well. The results which we got for each model is as follows :

”Logistic Regression”: 90.35,

”K-Nearest Neighbors”: 93.71%,

”Decision Tree”: 94.24%,

”Support Vector Machine (Linear Kernel)”: 90.45%,

”Support Vector Machine (RBF Kernel)”: 95.08%,

”Neural Network”: 98.48%,

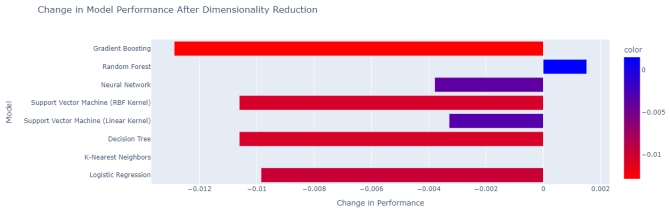
”Random Forest”: 97.60%,

”Gradient Boosting”: 94.90% .

As evident, we got the best results from Neural Network and Random Forest. Speculating that the model would be improved if we used PCA for selecting some of the features explicitly as there were 96 features for the model to learn from, we used the PCA dimensionality reduction to check for the features with the highest variance and affecting the decision factor the most. After analyzing the results from the PCA and selecting only the features that have high variance(We selected the top 50 features) and thus affect the results the most, we followed the same methods to train our models on the updated data. As the training had completed for all the eight models, we applied the same accuracy metrics on the models, but we were in for a surprise when the results rolled out or should we say computed out!

Even though we had selected the features which had high variance and affected the outcome the most, all the models including the ones that performed well had a major setback in terms of performance metrics, and the situation worsened as we decreased the number of components selected for training. As the models including neural networks do not perform well on the PCA data, we can say that the complete set of features is necessary for the models to predict accurately and that PCA is an overrated privilege not necessary in this case.

The following graph depicts the change in performance of the ML models before and after performing PCA. The horizontal bars falling on the negative side of the number line indicates a decrease in performance while the bars falling on the positive side indicates an increase in performance.



Inferencing on the results gained by experimentation we can come up with some insights which describe the model better. As we have used eight learning models we can select one best model which gives us the best performance, which is the neural network. The neural network has given us an accuracy of 98.46% and has stood as the best performing model in both before PCA and after PCA cases. So we can say for the given problem statement and the features of the dataset, the neural network model performs the best.

The model works really well when the data has a fair composition of bankrupt and not bankrupt companies, as it solves the problem of imbalanced data and gives more instances with necessary features, rather than augmenting the instance using distance metrics in the smote method. The models also work well when all the features of the data are retained. As seen in the PCA case, it is necessary for all the features to stay put for the modes to perform better. The model also works well when the data has scaled values which makes the training easier.

As mentioned earlier the model does not perform well even if some of the necessary features required for the prediction are missing or ignored the performance can be affected slightly in a negative manner. And another thing that may affect the model's performance is the gap between the majority data instances and minority data instances. Even though we have taken care of it through oversampling methods, it is better to keep the gap as small as possible because if the gap is insanely large the augmented instances may lead to underfitting due to the lack of accurate instances.

V. Conclusions

The aim of this paper is to analyse how we can use machine learning models to aid us in predicting the financial health of a company. We saw how to handle imbalanced data which is the main issue of many real world classification problems. From simple financial ratios used by earlier researchers to complex machine learning models - we showed that bankruptcy prediction models continue to evolve with time and there is no perfect model that can be classified as generally better than others. We can conclude that the model accuracy and metrics depends on tweaking the algorithm based on the data being considered and its properties rather than relying on predefined models based on previous studies.

Further studies may consider analysing more characteristics with more emphasis on hybrid machine learning models.

VI. Acknowledgements

We would like to acknowledge our Data Analytics Course Professor Dr. Gowri Srinivasa for providing constant guidance during each phase of our project. We would also like to acknowledge our assistant professors who have prepared the course content and also the teaching assistants who have been constantly providing resources to practice the learnt concepts.

References

[1] Sun J, Li H, Huang QH, He KY. Predicting Financial Distress and Corporate Failure: A Review from the State-of-the-Art Definitions, Modeling, Sampling, and Featuring Approaches. Knowledge-Based Systems. 2014;57:41–56.

[2] Shi Y, Li X. An Overview of Bankruptcy Prediction Models for Corporate Firms: A Systematic Literature Review. Intangible Capital. 2019;15(2):114–127.

[3] Beaver WH. Financial Ratios As Predictors of Failure. Journal of Accounting Research. 1966;4:71.

[4] Altman EI. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. The Journal of Finance. 1968;23(4):589–609.

[5] Gissel, J. (2007). A Review of Bankruptcy Prediction Studies:1930-Present, Gissel, Don Giacomino, Michael D. Akers. Journal of Financial Education, 33(Winter 2007), 1-42. The author of this document, Jodi L. Gissel, published under th ename Jodi L. Bellovary at the time of publication.

[6] Altman, E. I. (2018) 'Applications of Distress Prediction Models: What Have We Learned After 50 Years from the Z-Score Models?', International Journal of Financial Studies. MDPI AG, 6(3), p. 70. doi:10.3390/ijfs6030070.

[7] Antunes, F., Ribeiro, B. and Pereira, F. (2017) 'Probabilistic modeling and visualization for bankruptcy prediction', Applied Soft Computing Journal. Elsevier B.V., 60, pp. 831–843. doi:10.1016/j.asoc.2017.06.043.

VII. Contributions of team members

Abhay DA (PES1UG19CS011) - Studied various research papers as part of the literature review and worked on the literature review report. Worked on the making the final video. Made both the reports in IEEE format.

Akash S (PES1UG19CS042) - Performed exploratory data analysis on the dataset and gathered relevant information from it. Worked on the final report and video.

Akshay SP (PES1UG19CS046) - Worked on building models required for the problem statement and used PCA to derive inferential results using python and scikit learn(for coding) along with another teammate (Akshay SD). Worked on writing the final report.

Akshay SD (PES1UG19CS045) - Worked on building models required for the problem statement and used PCA to derive inferential results using python and scikit learn(for coding) along with another teammate (Akshay SP).