Company Bankruptcy Prediction and Analysis

By

NIMIT AGRAWAL **19BCE1537**

MEHAK GUPTA **19BCE1652**

RADHIKA KOTECHA **19BCE1807**

A project report submitted to

Dr. Pattabiraman V

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

in partial fulfilment of the requirements for the course of

CSE3020 Data Visualization

in

B. Tech. COMPUTER SCIENCE AND ENGINEERING

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Vandalur – Kelambakkam Road

Chennai – 600127

APRIL 2022

INDEX

SI No.	CONTENT	PAGE No.
1.	ACKNOWLEDGMENT	3
2.	ABSTRACT	4
3.	INTRODUCTION	5
4.	LITERATURE SURVEY	6-7
5.	METHODOLOGY	8-10
6.	RESULTS	11
7.	CONCLUSION & GIIHUB LINK	12
8.	REFERENCES	13

ACKNOWLEDGMENT

Primarily, we would like to thank the almighty for all the blessings he showered over us to complete this project without any flaws.

The success and final outcome of this assignment required a lot of guidance and assistance from many people and we are extremely fortunate to have got this all along with the completion of our project. Whatever we have done is only due to such guidance and assistance by our faculty, Dr. Pattabiraman V, to whom we are really thankful for giving us an opportunity to do this project.

Last but not the least, we are grateful to all our fellow classmates and our friends for the suggestions and support given to us throughout the completion of our project.

ABSTRACT

For many economists, predicting company bankruptcy has been a major topic of discussion. The purpose of generating and predicting a company's financial distress is to create a predictive model that can be used to forecast a company's financial condition by combining many econometric factors of interest to the researcher. Based on Taiwanese listed firms, the study developed a complete study model for forecasting insolvency. This data comes from the Taiwan Economic Journal for the years 1999–2009, and it represents firm bankruptcy as defined by the Taiwan Stock Exchange's business standards.

INTRODUCTION

Bankruptcy prediction is a method of predicting and projecting financial problems in both public and private companies. Predicting bankruptcy is critical in evaluating a company's financial situation and future prospects.

In economics, predicting corporate insolvency is a critical occurrence. The many players and participants in the business cycle place a high value on a company's financial stability. Policymakers, investors, banks, internal management, and the general public, referred to as consumers, are among the players and interested parties.

The ability to accurately estimate a company's financial performance is critical for many stakeholders when making key and consequential decisions about their relationship and engagement with the company.

Financial distress is a worldwide problem that impacts businesses in many industries.

Investors, as well as suppliers and retailers to the business, need to know if the company will go bankrupt. To avoid major losses for banks and other credit lenders, credit lenders and investors must assess a company's financial insolvency risk before making an investment or extending credit.

Suppliers and retailers of a company always perform credit transactions with it, and they must be completely aware of the firm's financial situation in order to make credit choices.

The ability to accurately foresee a company's financial problems is of tremendous importance to the company's numerous stakeholders. Problems with bankruptcy have forced the creation of research to identify various stressors to businesses in order to assist investors in making prudent investment decisions.

LITRATURE SURVEY

[1] *Explainability of Machine Learning Models for Bankruptcy Prediction*, by Min Sue Park, Hwijae Son, Chongseok Hyun and Hyung Ju Hwang

In this study by leveraging the advantage of LIME, the authors have proposed a novel, highly accurate, and instance-wise interpretable bankruptcy prediction model. The proposed model meets the two aforementioned requirements of high accuracy and interpretability. The experiment results show that the instance-wise interpretation of a LightGBM (or XGBoost) based bankruptcy prediction model is mostly consistent with the model-wise interpretation, which implies that the instance-wise interpretation is reliable. They also empirically show that instance-wise feature importance is more robust along with the predicted probability when equipped with the LightGBM-based model than with the XGBoost-based approach. Moreover, the experiments show that the important feature distribution is similar in the training and testing data, which implies that the instance-wise interpretation is robust to a random splitting of the data.

[2] *A Bankruptcy Prediction Model Using Random Forest*, by Shreya Joshi, Rachana Ramesh and Shagufta Tahsildar

In this paper, the dataset used includes financial ratios as attributes that are derived from the financial statements of various companies. The most influencing ratios that are required for predicting bankruptcy are selected on the basis of the Genetic Algorithm which filters out the most important ones from different existing bankruptcy models. These ratios of different companies are fed as an input to train the model being implemented in R. The prediction algorithm used is Random Forest, which will enable to differentiate between bankrupt and non-bankrupt companies.

[3] *Bankruptcy prediction using machine learning and an application to the case of the COVID-19 recession*, by Aditya Narvekar and Debashis Guha

In this paper, the authors find that two different Machine Learning algorithms, Random Forest (RF) and Extreme Gradient Boosting (XGBoost) produce

accurate predictions of whether a firm will go bankrupt within the next 30, 90, or 180 days, using financial ratios as input features. The XGBoost based models perform exceptionally well, with 99% out-of-sample accuracy. The training dataset uses a large database of public US firms over a period of 49 years, 1970–2019, and 57 financial ratios. This study has used a substantially larger training dataset as compared to previous studies.

An application of their best performing XGBoost model to Q2-2020 financial data for a sample of both private and public U.S. firms shows that the bankruptcy rate will climb substantially higher in 2020 than in the expansion years of 2011–2019. However, their model suggests that the rate will be only marginally higher than in 2010.

[4] ***Comparative Study of Bankruptcy Prediction Models*****, by Isye Arieshanti, Yudhi Purwananto, Ariestia Ramadhani, Mohamat Ulin Nuha and Nurissaidah Ulinnuha**

In this paper, the authors perform a comparative study of several machine leaning methods for Bankruptcy prediction. According to the comparative study, the performance of several models that based on machine learning methods (k-NN, fuzzy k-NN, SVM, Bagging Nearest Neighbour SVM, Multilayer Perceptron (MLP), Hybrid of MLP + Multiple Linear Regression), it can be concluded that fuzzy k-NN method achieve the best performance with accuracy 77.5%. The result suggests that the enhanced development of bankruptcy prediction model could use the improvement or modification of fuzzy k-NN.

METHODOLOGY

A. Dataset Description:

```
1  df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6819 entries, 0 to 6818
Data columns (total 96 columns):
 #   Column                                                  Non-Null Count  Dtype
---  ------                                                  --------------  -----
 0   Bankrupt?                                               6819 non-null   int64
 1    ROA(C) before interest and depreciation before interest 6819 non-null   float64
 2    ROA(A) before interest and % after tax                  6819 non-null   float64
 3    ROA(B) before interest and depreciation after tax       6819 non-null   float64
 4    Operating Gross Margin                                  6819 non-null   float64
 5    Realized Sales Gross Margin                             6819 non-null   float64
 6    Operating Profit Rate                                   6819 non-null   float64
 7    Pre-tax net Interest Rate                               6819 non-null   float64
 8    After-tax net Interest Rate                             6819 non-null   float64
 9    Non-industry income and expenditure/revenue             6819 non-null   float64
 10   Continuous interest rate (after tax)                    6819 non-null   float64
 11   Operating Expense Rate                                  6819 non-null   float64
 12   Research and development expense rate                   6819 non-null   float64
 13   Cash flow rate                                          6819 non-null   float64
 14   Interest-bearing debt interest rate                     6819 non-null   float64
 15   Tax rate (A)                                            6819 non-null   float64
 16   Net Value Per Share (B)                                 6819 non-null   float64
 17   Net Value Per Share (A)                                 6819 non-null   float64
 18   Net Value Per Share (C)                                 6819 non-null   float64
 19   Persistent EPS in the Last Four Seasons                 6819 non-null   float64
show more (open the raw output data in a text editor) ...
 93   Interest Coverage Ratio (Interest expense to EBIT)      6819 non-null   float64
 94   Net Income Flag                                         6819 non-null   int64
 95   Equity to Liability                                     6819 non-null   float64
dtypes: float64(93), int64(3)
memory usage: 5.0 MB
```

The resulting panel is strongly informative for us, and it shows how:
The dataset is composed of a combination of 6819 observations per each of our 96 features. All of the features are numerical (int64 or float64). There are no missing values (Nan) among the data. Considering that all our features are numeric, we can easily calculate their descriptive statistics: a further source of information. As we can see this dataset is highly imbalanced because the total no of companies bankrupted are only 220 which is very less compared to 6599.

B. Data Pre-processing:

Missing data: Inaccurate data might be observed because missing information produces gaps that may be crucial to the final analysis. Missing data frequently occurs as a result of a difficulty in the data gathering phase, such as a system outage caused by a malfunction, data entry errors, or biometrics challenges, among other things.

```
bank_data.isnull().sum.any()
```

```
[]
```

```
bank_data.duplicated().sum()
```

```
0
```

C. Outliner Removal:

DBScan: It is a clustering algorithm used in machine learning to distinguish high-density clusters from low-density clusters.

LOF: The Local Outlier Factor (LOF) technique is an unsupervised anomaly detection method that calculates a data point's local density deviation from its neighbours. It considers samples with a significantly lower density than their neighbours to be outliers.

COF: COF is a variant of the outlier factor algorithm that calculates the amount of data point distraction given a dataset. One of the drawbacks of the density-based (LOF) factor algorithm (for a point) is that it is totally dependent on the density of the points around it. As a result, it crashes when the outlier's density is similar to that of neighbouring data points. The connectivity-based approach is created to deal with such situations.

Isolation Forest: It is a machine learning algorithm for detecting anomalies. It uses the Decision Tree algorithm as its foundation. It separates outliers by selecting a feature at random from a set of features and then selecting a split value between the feature's max and min values at random.

METHOD	ACCURACY
DBScan	-
LOF	73%
COF	68%
Isolation Forest	73%

LOF is completely dependent on the density of the neighbouring points. Hence the Isolation Forest Outliner Detection is well suited for this model.

D. Data Modelling:

RANDOM FOREST: A random forest is a machine learning technique for solving classification and regression problems. It makes use of ensemble learning, which is a technique for solving complicated problems by combining several classifiers. Many decision trees make up a random forest algorithm. Bagging or bootstrap aggregation are used to train the 'forest' formed by the random forest method. Bagging is a meta-algorithm that increases the accuracy of machine learning methods by grouping them together.

GRADIENT BOOST: Gradient boosting classifiers are a set of machine learning algorithms that integrate a number of weak learning models to generate a powerful predictive model. When doing gradient boosting, decision trees are commonly employed. Gradient boosting models are gaining popularity as a result of their ability to classify difficult information.

KNN: It assumes that the new case/data and existing cases are similar and places the new case in the category that is most similar to the existing categories.

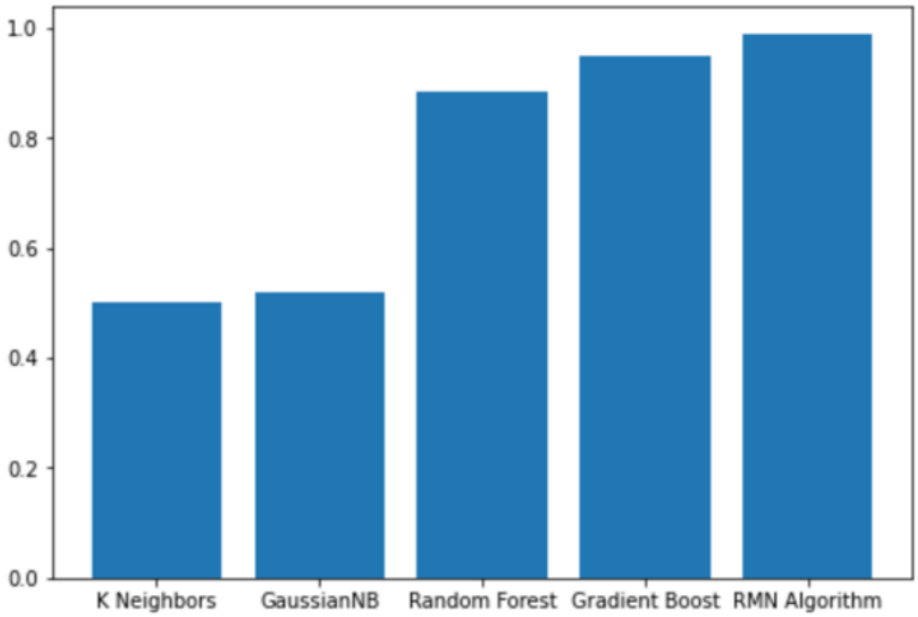
GAUSSIAN NAIVE BAYES: This algorithm is a special type of Naïve Bayes algorithm. It's specifically used when the features have continuous values. It's also assumed that all the features are following a Gaussian distribution i.e., normal distribution.

RMN ALGORITHM: We used the method of stacking for forming this hybrid model using the KNN and Gaussian Naïve Bayes Model. Stacking is a general procedure in which a learner is taught to combine multiple learners. Individual learners are referred to as first-level learners, whereas the combiner is referred to as a second-level learner, or meta-learner.

RESULTS

Comparative Analysis:

MODEL	ACCURACY
KNN	50%
Gaussian NB	52%
Random Forest	88%
Gradient Boost	92%
RMN Algorithm	99%




```
In [36]: accuracy_score
Out[36]: [0.5,
          0.5207564575645757,
          0.8842250922509225,
          0.9511070110701108,
          0.9910242968041134]
```

From the above table and graph, it is clear that the RMN model gives the highest accuracy. Alone the K Neighbours model and Gaussian NB give accuracy of 50% but after stacking they form a hybrid model giving accuracy of 99% which is even more than the random forest and Gradient Boost models.

CONCLUSION

We have developed a model which is used to predict the financial condition of a company using different economic variables of interest like operating gross margin, profit rate, sales, etc. This model is compared with other models and gives an accuracy more than them.

Github link for source code:

https://github.com/mac25-git/Company-Bankruptcy-Prediction- and-Analysis.git

REFERENCES

[1] Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of accounting research*, 71-111.

[2] Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, *23*(4), 589-609.

[3] Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, 109-131.

[4] Kubat, M., & Matwin, S. (1997, July). Addressing the curse of imbalanced training sets: one-sided selection. In *Icml* (Vol. 97, No. 1, p. 179).

[5] Singh, A., & Purohit, A. (2015). A survey on methods for solving data imbalance problem for classification. *International Journal of Computer Applications*, *127*(15), 37-41.

[6] Bruynseels, L., & Willekens, M. (2012). The effect of strategic and operating turnaround initiatives on audit reporting for distressed companies. *Accounting, Organizations and Society*, *37*(4), 223-241.

[7] Sun, J., Li, H., Huang, Q. H., & He, K. Y. (2014). Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowledge-Based Systems*, *57*, 41-56.

[8] Shi, Y., & Li, X. (2019). An overview of bankruptcy prediction models for corporate firms: A systematic literature review. *Intangible Capital*, *15*(2), 114-127.

[9] Zopounidis, C., & Doumpos, M. (1999). Business failure prediction using the UTADIS multicriteria analysis method. *Journal of the Operational research Society*, *50*(11), 1138-1148.