Bankruptcy detection over various machine learning models for unbalanced data

Kommareddy Tarun[#1],Manideep Mokurala [#2], Vasudeva Reddy Satti[#3],Koyya Sreelatha[#4]

[#] *Department of Electronics and Computer engineering, sreenidhi institute of science and technology Hyderabad*

[1]tarunkommareddy@gmail.com

[2]manideep.7219@gmail.com

[3]sattivasudev Gmail

[4]sreelathaGmail

Abstract: **Machine learning has come a way long from past decades and the use of machine learning and machine learning models has been rapidly increasing in various fields. Today machine learning is playing a critical role in many streams especially finance. Bankruptcy is a technical term used to describe whether a company can run its operations in the future because of its losses or bad debts.**

The paper focuses on delivering a machine learning algorithmic approach to detect a company's state of bankruptcy based on the available data. The project implements a study on a dataset of companies in a specific period and By applying various machine learning algorithms like logistic regression, random forest, and artificial neural networks.

Keywords: **Machine learning, logistic regression, random forest, artificial neural networks**

I. INTRODUCTION

Bankruptcy is a state where a company is not able to run its business or continue its operation in the future because of its financial condition or debts. Any firm or investor needs to know the state of bankruptcy before they start investing in the company. Having a standardized approach looks like a generous idea in predicting any state but is very unlikely that the method will give satisfying results because of the real-world situations established.

The goal of the research is to make an analysis of the state of bankruptcy based on various machine learning models and also to Identify how likely a machine learning algorithm will be able to predict the accuracy. we have used supervised learning algorithms to identify the bankruptcy. instead of using a single machine learning approach, we have used multiple models like Logistic regression, Random Forest classifier, Xg-boost, Cat-boost, Artificial neural network

II. METHODOLOGY

The methodology is simple, first, we are taking the input data and splitting the data into training data and testing data, and perform pre-processing and pre-processing is removing NAN values, Duplicate values. exploratory data analysis is a graphical representation of the data and it usually consists of heat maps box plots which help to find the outliers etc and simultaneously the data is filtered by removing outliers. Now the training data is ready to train with various machine learning models but there is a small catch. The data we are using is the Taiwanese bankruptcy dataset and it is an imbalanced dataset. To avoid the false results SMOTE(Synthetic Minority Oversampling) is applied to oversample the minority values in the dataset. This is done while applying the model through the pipeline.

Various models will be performed and looked for accuracy, precision, recall, and f1 on all the trained models. from the analysis we got from the training data models we tend to validate the results of the testing data. A brief flowchart of methodology is shown in Fig 2.1

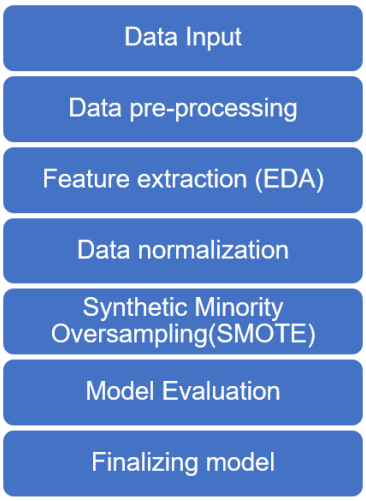
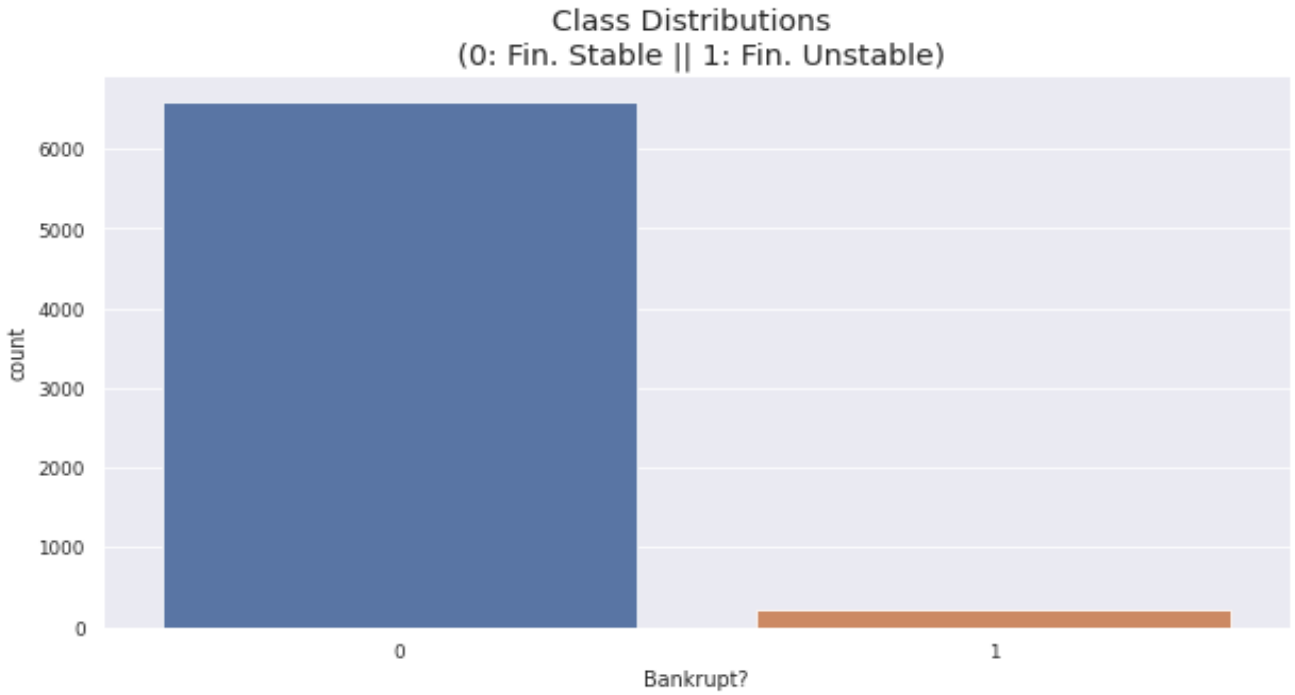


Fig 2.1

III. ALGORITHMS USED

Before applying the processed data to the machine learning algorithms the data s performed with EDA and the outliers are removed. we can observe that the data is imbalanced which means the ratio between the companies which went bankrupt and the companies which didn't go bankrupt is very high. As shown in Fig 3.1 there is a lot of difference. To balance the imbalanced data we are using SMOTE.



Smote is a technique that stands for synthetic minority oversampling oversamples the minority data and tries to balance the dataset to get the results more precisely.

There were been several attempts made to handle imbalanced data and most of the methods were not at all solving the major problem that is balancing the imbalanced data. Undersampling the majority seems to be like a nice idea to balance the data but that comes with a risk that the machine might not receive a proper amount of data to train, so anything that should be proceed should be only with the minority variable and oversampling to the perfect fit balances the data and makes much more easier to perform analysis and thankfully this oversampling is taken care by the machine with a simple function.

Logistic regression: is a basic machine learning algorithm used for classification problems. a confusion matric is drawn over the validation data after training and this can be observed in Fig 3.2

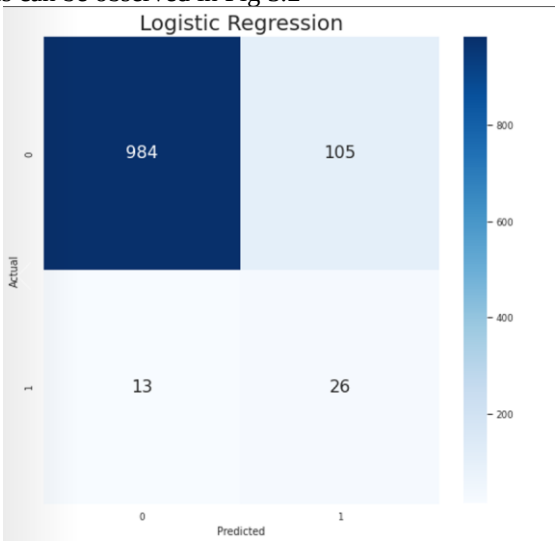


Fig 3.2

Random Forest Classifier: A Random Forest Classifier is a decision tree-based classifier that takes an average of the data and tries to improve the accuracy. The Confusion matrix of the plotted for the random forest can be observed in Fig 3.3

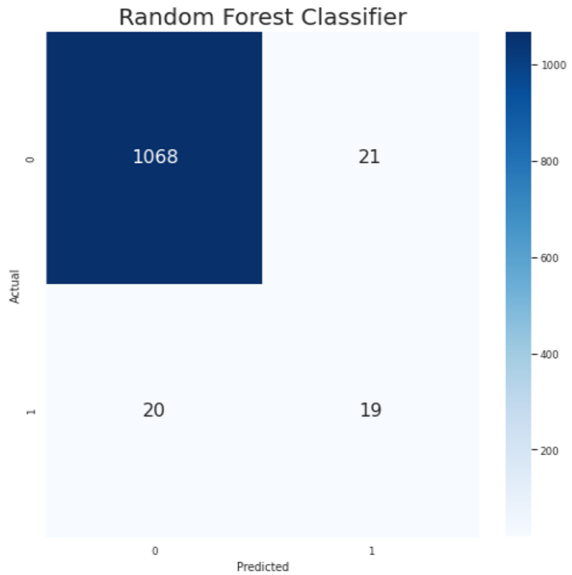


Fig 3.3

XG Boost: it is an implementation of gradient boosting which pushes the computational power of the algorithm. The Confusion matrix of the plotted for the XG Boost can be observed in Fig 3.4.

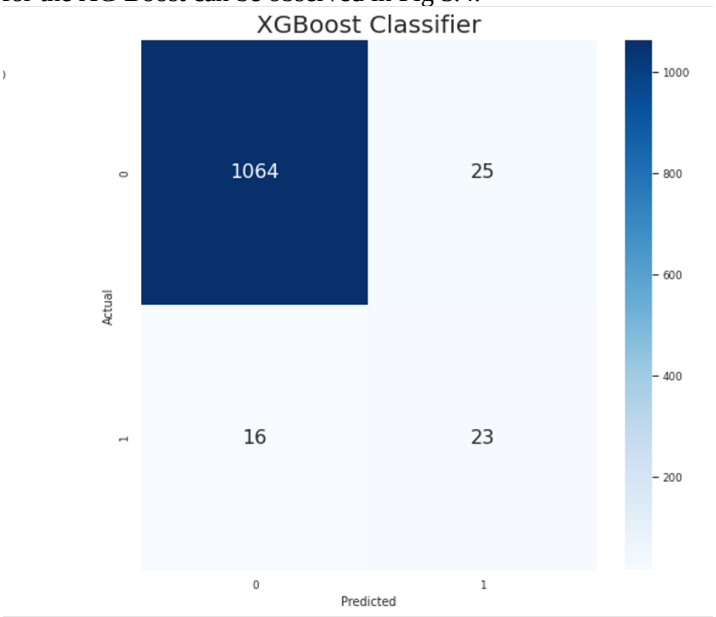


Fig 3.4

Cat Boost: categorical boosting is also an ensemble method of boosting technique typically used for classification problems. The Confusion matrix of the plotted for the cat boost can be observed in Fig 3.5

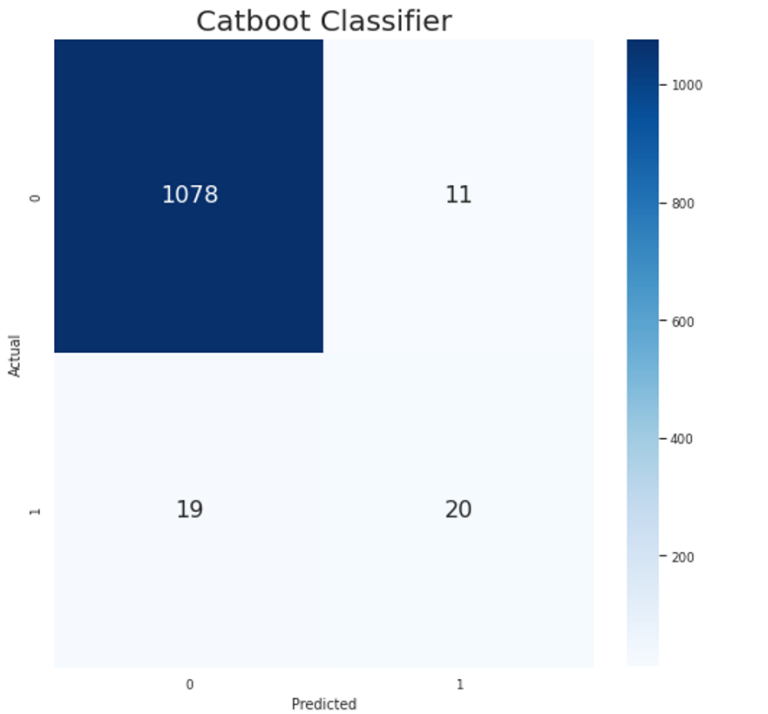


Fig 3.5

Artificial neural network: The ANN used contains 3 hidden layers with 3 nodes each helps to perform analysis. The Confusion matrix of the plotted for the ANN can be observed in Fig 3.5

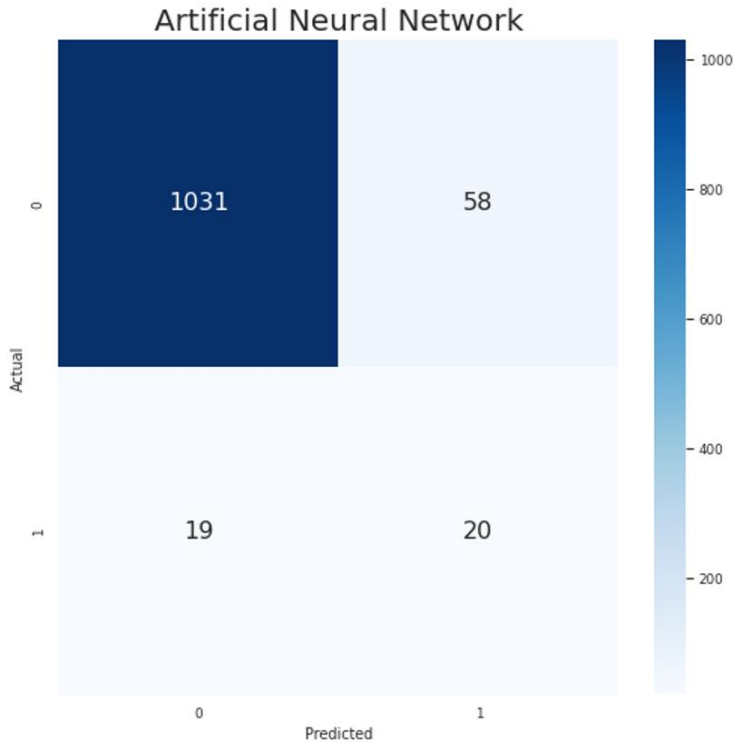


Fig 3.6

TABLE I
COMPARISON OF ALL MODELS WITH THEIR SCORES ON THE VALIDATION DATA

Model name	Accuracy	Precision	Recall	F1
Logistic regression	0.8872	0.2087	0.7895	0.3279
Random forest	1.0	1.0	1.0	1.0
XG Boost	1.0	1.0	1.0	1.0
Catboost	0.9951	0.8895	0.9935	0.9224

Artificial Neural Network	0.9627	0.4879	0.9556	0.6441

After the analysis of all the measures, we are plotting a bar graph featuring accuracy and this can be observed in Fig 3.7

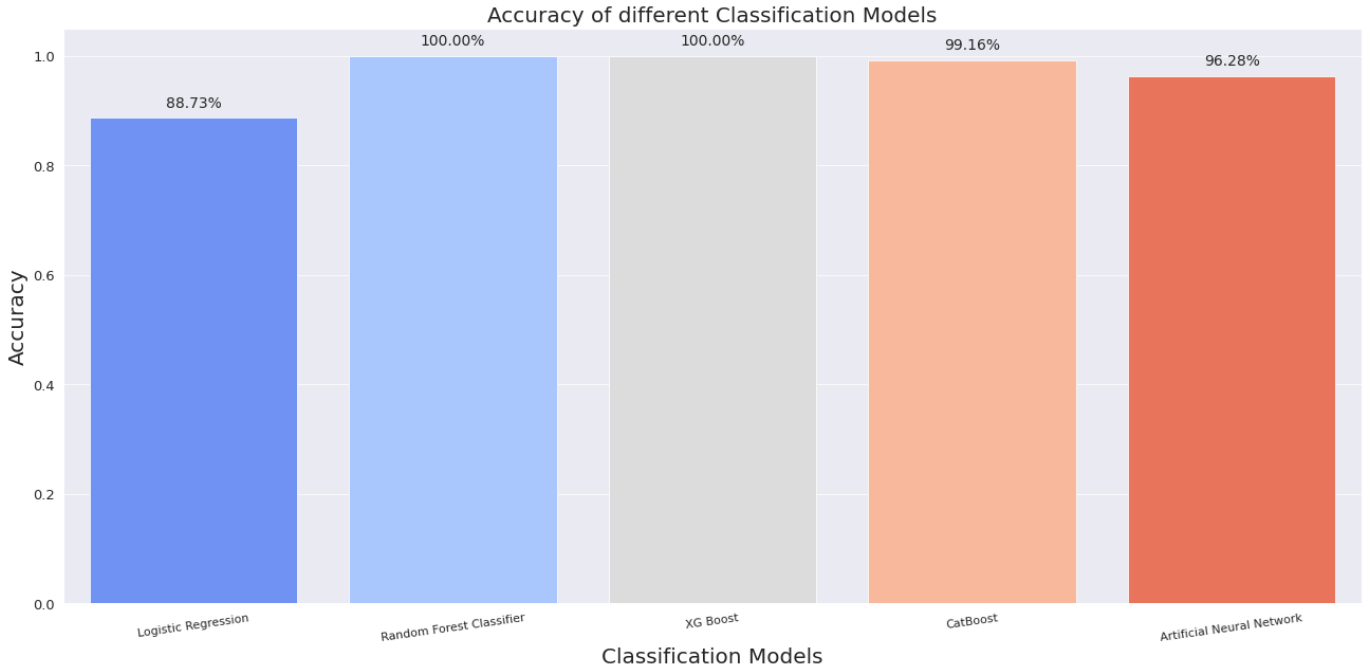


Fig 3.7

Although all the models performed very well on the validation data we can observe that random forest and Xg Boost gave an accuracy of 100% and that cannot be justified when we observe the confusion matrix of those two corresponding models. the ratio of true positives and false negatives is very irrelevant and lead to giving a high accuracy

IV. CONCLUSION

as we have observed in the methodology all the algorithms were applied with the validation dataset and also a ROC curve has been plotted based on the values acquired from each model and they are observed in Fig 4.1

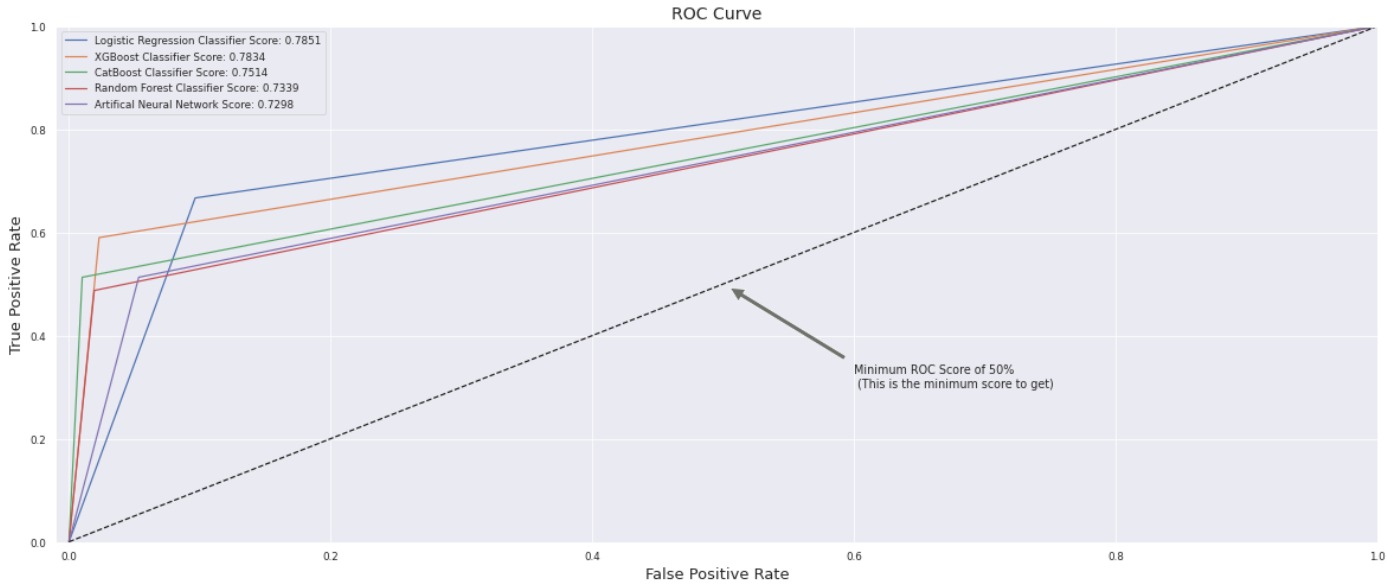


Fig 4.1

From the ROC curve listed above, we can observe that the ratio between True Positive Rate and False Positive Rate is higher for the Logistic Regression and cat boost. Though the XG boost has a better ROC score we are not considering it just because there is a high False positive rate.

Now the confusion matrix for the testing data is plotted for the selected algorithms (logistic regression and boost) can theFig4.2

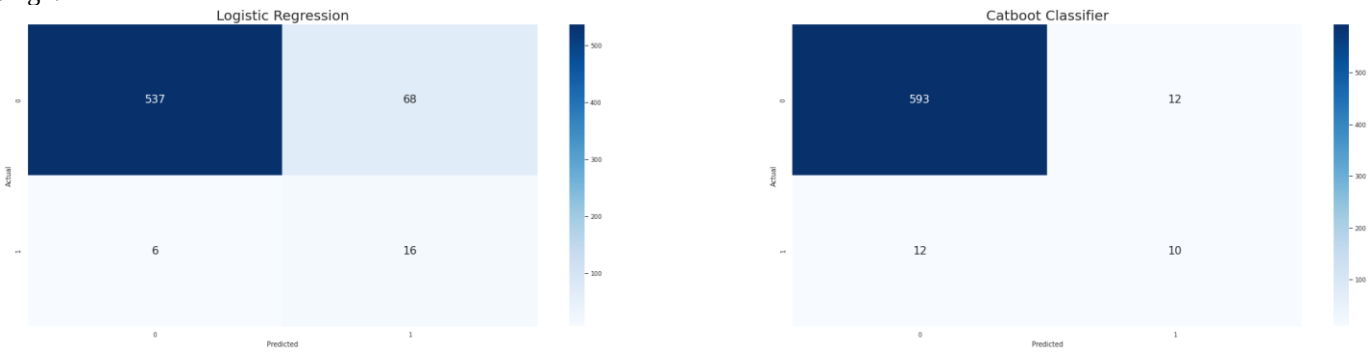


Fig 4.2

REFERENCES

[1] M. S. Keya, H. Akter, M. A. Rahman, M. M. Rahman, M. U. Emon and M. S. Zulfiker, "Comparison of Different Machine Learning Algorithms for Detecting Bankruptcy," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 705-712, doi: 10.1109/ICICT50816.2021.9358587.

[2] Devi, S. S., Radhika, Y. (2018). A survey on machine learning and statistical techniques in bankruptcy prediction. International Journal of Machine Learning and Computing, 8(2), 133-139

[3] P. Gnip and P. Drotár, "Ensemble methods for strongly imbalanced data: bankruptcy prediction," 2019 IEEE 17th International Symposium on Intelligent Systems and Informatics (SISY), 2019, pp. 155-160, doi: 10.1109/SISY47553.2019.9111557.R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sept. 16, 1997.

[4] Nitesh Chawla, et al. (2002) titled "SMOTE: Synthetic Minority Over-sampling Technique.*FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.

[5] Company Bankruptcy Prediction, https://archive.ics.uci.edu/ml/datasets/Taiwanese+Bankruptcy+Prediction

[6] Deron Liang, Chia-Chi Lu, Chih-Fong Tsai, Guan-An Shih, Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study

[7] Erdogan, B. E. (2013). Prediction of bankruptcy using support vector machines: an application to bank bankruptcy. Journal of Statistical Computation and Simulation, 83(8), 1543-1555

[8] F. Barboza, H. Kimura, and E. Altman, "Machine learning models and bankruptcy prediction," Expert Systems with Applications, vol. 83, pp. 405-417, 2017