# Comparative Study of Machine Learning Algorithms for Bankruptcy Prediction

Shuvam Sanyal
Correspondence: Symbiosis International (Deemed University), Pune, India, Phone Number: 8777780597, E-mail: 19070243015@sig.ac.in

## ABSTRACT

Since Lehman Brother's bankruptcy in USA during 2008 global financial crisis, advanced bankruptcy prediction has been of huge importance in the industry. The aim of this paper is to show , how modern-day boosting algorithms can bring great accuracies in bankruptcy prediction specially in challenging times like corona virus related economic depression, when applied with different oversampling and dimensionality reduction techniques , alongside traditional algorithms. This comparative study is completely an original and dedicated research work. Results of this study may be of interest for financial institutions and for academics. further research steps include deep learning algorithms-based bankruptcy prediction, natural language processing-based analysis of the relation between a firm's bankruptcy chances and its recent social media activities.

## CONTACT

Shuvam Sanyal
Symbiosis International University, Pune,India
Email: 19070243015@sig.ac.in
Phone: 8777780597
Website: www.siu.edu.in

## INTRODUCTION

Corporate bankruptcy is some kind of legal way of handling those businesses which are unable to repay their outstanding debts. The organization's solvency is calculated using the probability value of the scenario where the company is not able to pay its debt. The main frameworks of corporate bankruptcy are: - (1) the smaller the failure probability, larger the reservoir and net liquid asset operation flow. (2) the greater the failure probability, greater the debt amount and operations fund expanses. Business insolvency can occur two ways: Receivership and Liquidity. Few financial ratios that influence business insolvencies are: - 1) Profit Scenario 2) Costs Incurred 3) Capital turnover 4) Liquidity Ratio 5) Asset 6) Capital 7) Growth Rate 8) Tobin's Q.

In 2008, Lehman Brothers filed for the biggest bankruptcy event in the history of America with more than 600 billion in debts. Just before becoming bankrupt, it was the fourth-largest U.S. investment bank. This firm went bankrupt due to its huge investment to the mortgage and real estate markets of America. Following the trend of most of the IB banks, they were mostly dependent on short-term markets in order to make profit in billions each day. Ultimately, it was a failure in securing funding that caused this sudden bankruptcy. Currently due to this ongoing global coronavirus pandemic, all leading experts predicted that the economic depression, we are going to face in coming few years, will be much higher on scale than what we faced in 2008. In this study, different machine learning techniques are employed to predict bankruptcy. Further based on the performance of the classifiers, the best model is chosen for development of a probabilistic decision modelling in Python programming language. The modelling can be utilized by stock holders and investors to predict the performance of a company based on their financial history.

## METHODS AND MATERIALS

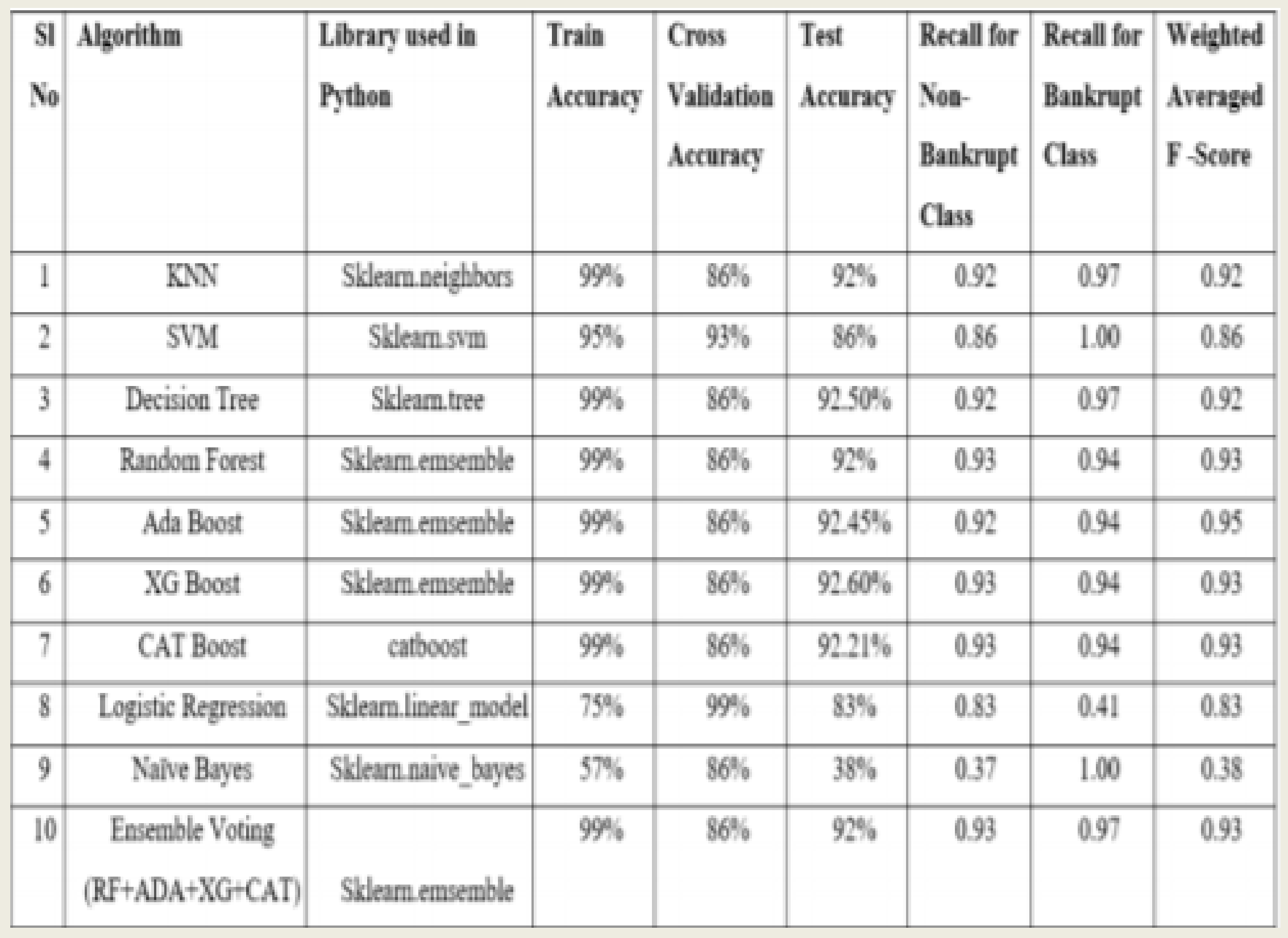
This paper uses a sample dataset of US Corporate firms comprising thirteen financial ratios as feature variables and after data preprocessing, an analysis based on nine different machine learning techniques((Logistic Regression, KNN, SVM, Naïve Bayes, Decision Tree, Random Forest, AdaBoost, XgBoost, CatBoost) on the dataset is done. I have used Accuracy Scores, ROC-AUC Curve, Confusion Matrix, Precision, Recall as well as F Score and Cumulative Gain Chart. The best models came out to be Random Forest, XgBoost, AdaBoost, CatBoost. After applying an Ensemble Voting Method on top 5 algorithms, it votes for Random Forest and the boosting algorithms to be the best two predictors on both training and testing data of bankruptcy cases, thus reducing over fitting problem. Final model is able to correctly predict all 8033 bankrupt firms correctly and 17208 non-bankrupt firms correctly out of 17217 in the test dataset. Only 9 corporations have been misclassified as bankrupt when they have no chance of going bankrupt. Furthermore, this model allocates significance to discrete components related to assets, liquidities.

## RESULTS

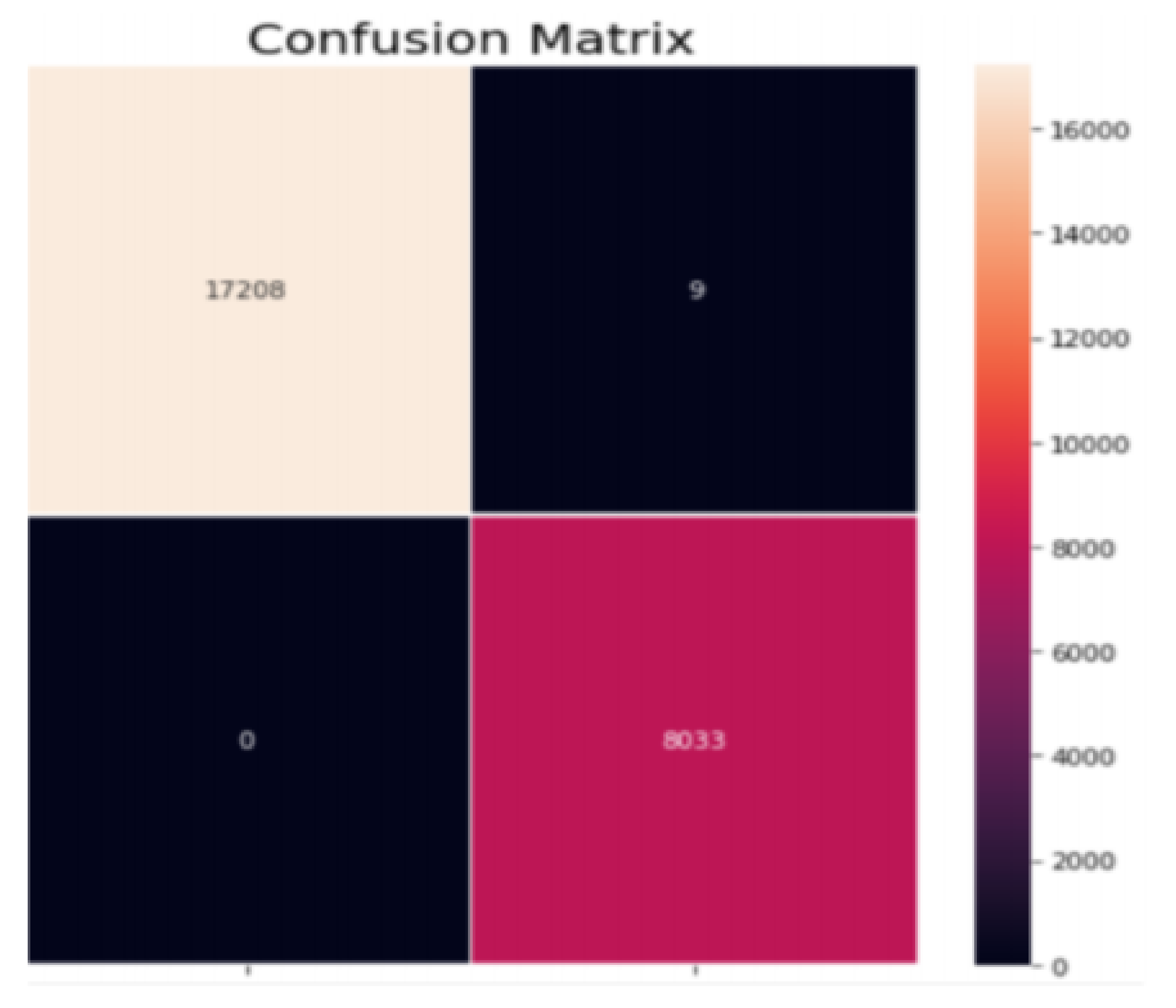
Our research indicates that after doing cross validation, Random Forest, Boost, AdaBoost & Cat Boost are the four top performing models, obtaining test accuracies of 92.5%, 92.6%, 92.45% and 92.21% respectively while achieving a ~ 99% training accuracy across dataset. Overall, these models obtain an AUC value of more than 0.80 without PCA and close to one after doing PCA. Since, this was initially a problem of asymmetric target class distribution, so F score will be a great measure of accuracy and our overall F score is 0.9994. Lastly, we did perform an ensemble-based majority voting classifier and found that maximum votes have gone to Random Forest and Gradient Boosting algorithms.

Additionally, our overall model confusion matrix result indicates, only 9 corporations out of all, have been misclassified as bankrupt when they are actually not with false negative rate being as low as close to 0.005 and false positive rate being zero after doing PCA and SMOTE. Logistic Regression accuracy is 75% and thus unacceptable. Naïve Bayes gives very less training and testing accuracies with ~57% and ~37% respectively. KNN and SVM are memory intensive, expensive to tune due to not being trained during training phase and thus don't perform well on huge datasets despite producing some good accuracy scores. .

Also, before doing PCA , our cumulative gain chart has shown only 20% positive results was achieved when we were considering 25% percentage of population with high probabilities to target according to the model. In contrast to that, when we applied PCA to our dataset , 60% positive results are achieved finally when we considered only 20% percentage of population with high probabilities to target according to the model. That is huge improvement in the model's predictive capability.



| Sl No | Algorithm | Library used in Python | Train Accuracy | Cross Validation Accuracy | Test Accuracy | Recall for Non-Bankrupt Class | Recall for Bankrupt Class | Weighted Averaged F -Score |
|---|---|---|---|---|---|---|---|---|
| 1 | KNN | Sklearn.neighbors | 99% | 86% | 92% | 0.92 | 0.97 | 0.92 |
| 2 | SVM | Sklearn.svm | 95% | 93% | 86% | 0.86 | 1.00 | 0.86 |
| 3 | Decision Tree | Sklearn.tree | 99% | 86% | 92.50% | 0.92 | 0.97 | 0.92 |
| 4 | Random Forest | Sklearn.emsemble | 99% | 86% | 92% | 0.93 | 0.94 | 0.93 |
| 5 | Ada Boost | Sklearn.emsemble | 99% | 86% | 92.45% | 0.92 | 0.94 | 0.95 |
| 6 | XG Boost | Sklearn.emsemble | 99% | 86% | 92.60% | 0.93 | 0.94 | 0.93 |
| 7 | CAT Boost | catboost | 99% | 86% | 92.21% | 0.93 | 0.94 | 0.93 |
| 8 | Logistic Regression | Sklearn.linear_model | 75% | 99% | 83% | 0.83 | 0.41 | 0.83 |
| 9 | Naïve Bayes | Sklearn.naive_bayes | 57% | 86% | 38% | 0.37 | 1.00 | 0.38 |
| 10 | Ensemble Voting (RF+ADA+XG+CAT) | Sklearn.emsemble | 99% | 86% | 92% | 0.93 | 0.97 | 0.93 |



## CONCLUSION AND FUTURE SCOPE

The results suggest if we compare it with the literature review then our study indicates due to advancement of modern machine learning and artificial intelligence-based technologies, much better accuracy in bankruptcy prediction results can be achieved with gradient boosting algorithms like Xg Boost, Ada Boost, Cat Boost alongside traditional random forest algorithm. Earlier the techniques like principal component analysis or oversampling to achieve more accurate results were not common, but in our research work, we have introduced the application of these modern techniques to handle dimensionality and data imbalance problem with ease. The predictive model helps to produce high test accuracy in predicting bankruptcy. In future to be continued with the research, I want to create multi-year bankruptcy prediction model which, can predict potential bankruptcies several years prior in advance. Also, we want to see if there is any underlined relation between what the firm's social media activities and odds of bankruptcy using Natural Language Processing based sentiment analysis. Additionally, I want to apply recent deep learning algorithms to predict bankruptcy

## REFERENCES

[1] Bredart.X (2014), Bankruptcy prediction model using neural networks. Accounting and Finance Research; vol 3 No 2, pp. 124-128

[2] Chen Ming, J (2019), "Models for Predicting Business Bankruptcies and Their Application to Banking and to Financial Regulation", available at: https://dx.doi.org/10.2139/ssrn.3329147 (accessed 20 May 2018).

[3] Koro, T (2017), Evaluation of the factors influencing business bankruptcy risk in Poland, Financial Internet Quarterly, e-Finanse; vol.13 No 2, pp.72-80.

[4] Mattsson, B and Steinert, O (2017), Corporate bankruptcy prediction using machine learning techniques, working paper, Department of Economics, University of Gothenburg School of Business, Economics and Law, Gothenburg, Sweden.

[5] Nagaraj K and Sridhar A (2015), A predictive system for detection of bankruptcy using machine learning techniques, International Journal of Data Mining & Knowledge Management Process; Vo l.5 No.1, pp. 29-40

[6] Sundal, H.K & Hatlestad, K (2015), Factors Affecting the Probability of Bankruptcy, working paper, Norwegian School of Economics.

[7] (S. Kotsiantis1, D et al., 2005), Efficiency of Machine Learning Techniques in Bankruptcy Prediction, Tampakas, V, ICESAcc 2005: 2 nd International Conference on Enterprise Systems & Accounting, Thessaloniki, Greece.

[8] Wagenmans, F and Feelders, A.J (2017), Machine Learning in Bankruptcy Prediction, working paper, Universiteit Utrecht, Utrecht, Netherlands.

[9] Cat Boost vs . Light GBM vs XG Boost(Mar 14, 2018) by Alvira Swalin