International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.119

Performance Evaluation of Different Machine Learning Models for Bankruptcy Prediction

Joe RishwanthDemel J
Data Science and Business systems
SRM institute of science and technology
Chennai, India jd1594@srmist.edu.in

Mrs. S. Sindhu
Data Science and Business systems
SRM institute of science and technology
Chennai, India sindhus2@srmist.edu.in

***Abstract*—Bankruptcy of a company usually is a huge issue for companies. The bad influence of bankruptcy could lead to loss for components of the business such as the owners, the investors, the employees, and the consumers. Bankruptcy prediction is the practice of forecasting the financial distress and potential bankruptcy of a public firm. This area of research involves analyzing various financial ratios and other data to identify indicators of financial risk. We can prevent bankruptcy by predicting the likelihood of acompany getting bankrupt based on a company's financial ratios and data. With the advent of new data-intensive techniques, suchas machine learning, researchers have developed increasinglysophisticated methods for predicting bankruptcy. However, it is important to exercise caution when interpreting the results of such models, as they can suffer from biases and other limitations.Despite these challenges, bankruptcy prediction remains a critical area of study for investors and creditors seeking to assess the financial health of a company The issue of this whole processis managing the imbalance in class caused by the rare event of bankruptcy in the real economy. Advancements in Artificial Intelligence (AI) have helped these companies by applying those models to predict bankruptcy. This bankruptcy problem wassolved in this work by comparing various Machine Learning methods such as SVM, KNN, Ensemble Learning, Decision Tree classification was applied which achieved 60%, 84%, 93%, 94% accuracy.**

***Index Terms*—Finance, Bankruptcy, XGboost, KNN, Decision Tree, Classification, Ensemble Learning, SVM**

I. Introduction

For any company, insolvency is an undesirable phenomenon. It heavily affects the company owner and other responsible stakeholders. Companies getting bankrupt is increasing day by day. Therefore, bankruptcy prediction is of such high value and has been researched extensively over the past decades. There are a number of causes for business financial issues, and it's a difficult one to define, but some factors can be revealed that could signal impending bankruptcy. Determining your pre-bankruptcy status is critical to timely averting a difficult situation. Nowadays, highly advanced prediction- based computing techniques have been widely used by researchers to solve error prediction problems and are also discussed withsystematic literature review technique. In this study, we will focus on holistic methods to get the best possible results.

Bankruptcy is a legal process in which judges and court administrators investigate the assets and debts of individuals, partnerships, and corporations who believe they have too many debts to pay. The court decides to settle the debt. "Discharge" means that the person in debt is not legally obligated to pay. A court can also dismiss a lawsuit if it determines that the individual or entity has sufficient assets to pay the bills. Bankruptcy laws were written to give people a chance to start over when their finances collapsed.

Forecasting bankruptcy and assessing financial distress of public companies is an important area of research in finance and accounting. Creditors and investors are interested in predicting the likelihood of a firm going bankrupt, which has led to extensive research in this field. The abundance of data available for both bankrupt and non-bankrupt public firms, including various accounting ratios and other explanatory variables, has further spurred the development of sophisticatedforecasting techniques that rely on large datasets. The evolution of bankruptcy prediction involves the use of various statistical tools that have become increasingly available over time and requires a growing understanding of the limitations. Despite these advancements, some published research still fails to address long-known pitfalls in the field. The accuracy of bankruptcy prediction has improved with the development of more advanced machine learning models. This paper introduces recently developed methods for predicting bankruptcy using real-world data.

II. Literature Survey

Research done by [1] did model development for companies getting bankrupt for the given firms using ML algorithms. They have used 21 variables which are used to select the variables for model. The 1st technique used is based on conditional likelihood and the 2nd technique about conditional correlation between variables. Totally 3 models were used for predicting a company's bankruptcy and is created using NB model and all three are evaluated. First NB model showed90 % and the second showed 93 % and logistic regression gave an accuracy of 90 %. Finally, it was concluded that it is possible to predict financial bankruptcy-based prediction using Bayesian models.

In [2] applied the algorithm called as Partial Least Squares Regression which in general helps us in implementing a huge number of financial ratios in the given algorithm. Apart from that, it also solves the issue based on correlation, and also takes into account the missing data. They also showed demo about how the application of the Partial Least Squares approach to the companies usuallyprovides improved results. The application of this technique consists of two classes which are of healthy and the failing companies, allows the members to possibly gain high significant results and also to propose a model which is better than a model obtained by a parametric approach. This research helps the bankers and the investors by providing a detailed explanation about the indicators which indicate bankruptcy. It allows companies to do diagnostics of their models to predict bankruptcy prediction. Logit models which were implemented by [3], resulted in a stepwise selection process, which correctly predicted 84% and 91% of bankruptcies 1 and 2 respectively. The Implied

DOI: rp-9788770040723.119
estimation models that a hotel business is more likely to get bankrupt if it has more operating cash flow and increased total liabilities when compared to none. Models suggest that a conservative sales strategy coupled with a tighter operating cost control and some less debt financing can help improve a company's ability to meet its financial obligations and thus reducing the risk of bankruptcy.

A study conducted by [4] evaluated the financial data of more than 400 companies – 52 of which went bankrupt and 348 were "healthy". The results show the set of factors involved for financial bankruptcy prediction and neural network relevance. They applied a total of 17 factors that characterize liquidity, profitability, sustainability, efficiency, and innovation. The total predictive power of the model created in our study is close to 98%, which is extremely superior an efficient when compared to other models

[5] applied multiple data analyzing tools to bankruptcy data, with the sole aim of comparing the accuracy of model. For these data, decision trees were more accurate than support vector machines and deep neural networks.

In a survey done by [6] analyzed the design and application of many ML models for different process involving default events: (a) Estimate the probability of survival over a period of time period (b)Predicting default bankruptcy using the time series based on accounting data of varying length. Finally, they seriously talk about the most interesting metrics and also suggested it for future studies.

A exploration by [7] evaluated the motivations and trends of business failure in Lithuania in the period 2006-2010. The probability of bankruptcy was assessed in five companies that are currently active and two that have gone bankrupt using Springate, Zavgren, the Altman and Chesser models. After testing the usability and the applications of bankruptcy financial prediction model in 7 companies, the results show that the linear discriminant model most accurately reflects the financial position of the company.

Statistical methods were used to select the most appropriate indicators by [8] and after filtering the data, the indicators were more convincing. Second, unlike former study methods, they use the same set of samples to do the experiment. Finally, the result can prove the worthiness of the machine learning method, with an accuracy score of 95.9%.

Analysis conducted by [9] solved the issue of unbalanced data with subsampling and (SMOTE). Machine learning techniques involving random forests resulted in 99% accuracy, while Decision Trees, Logistic Model Trees (LMTs), Support Vector Machines (SVMs), and random forests (RFs) resulted accuracy 92%, 92.3%, 93.8% and99%, respectively.

Support vector machine (SVM) to improve bankruptcy prediction problem with the aim of proposing a brand new algorithm was done by [10] with way better explaining capacity and increased stability. They used 5-fold cross-validation to find the efficient value of the SVM kernel function parameters.Apart from that, to predict the accuracy of SVM, they compare SVM performance with the performance of logistic regression analysis, multiple discriminant analysis (MDA), and 3-way back-propagation neural network. fully connected layer (BPN). Test results show that SVM is superior to other methods.

Research [11] which proposed a method by applying XG- Boost to handle class imbalance in datasets. In this work, Randomized Search CV optimization helped find the parameters. To increase model efficiency, XGBoost applies data sampling techniques. The experiment was based on two real credit card records. The study showed that the combination of data sampling doesn't have much effect on efficiency of XGBoost. The proposed approach finally ended up with very high accuracy

S.S. Panigrahi et al made an effort to use discrete input variables on top of which decision tree was applied. Decision trees for NSE and BSE are generated. It is constructed differently and its set side by side with a decision tree that is directly generated over the same period. Empirical studies prove high effectiveness of the given model by outperforming other decision trees [12]

[13] used data from company based in Ghana. Regression analysis was fitted to data from 648 clients. They proposed that microfinance institutions bear the risk of default. A model was implemented for determining the effect of problem, as it is relatively effective. It was recommended that training is re- quired to improve their skills. Regulators should also consider enacting legislation to make sure that this is guaranteed.

Research focused on cancer classification by [14]. The distribution shows that the dataset is high Unbalanced and biased decision tree-like learning algorithms. A benign observation, leading to poor prediction performance malicious observation. Adaptive boosting is used here in this study. The performance of the models is analyzed, and they came to know that Adaboost algorithm performed better than decision trees with accuracy of 95.1% over decision tree which had 89In their study, [15] proposed a novel similarity-based approach totext classification, utilizing a KNN model-based classifier that combines both KNN and Rocchio techniques. The researchers developed a classification prototype using the KNN model and conducted experiments on two well-known document corpora, the ModApte version of the Reuters 21578 collection and 20 newsgroups dataset. The results of their experiments indicated that the KNN model-based classifier performed favorably compared to traditional KNN and Rocchio classifiers. The researchers concluded that their proposed approach could be a viable alternative to KNN and Rocchio in various domains.

III. Proposed Methodology

Classification algorithm is used to predict the target class asthe target class which are categorical in nature. We are going touse SVM, KNN, XGBoost and Decision Tree Classifier. Let see in detail about the following algorithms and the results they produce

A. Logistic Regression

Logistic Regression is a machine learning technique frequently used to predict binary outcomes, such as "Yes" or

DOI: rp-9788770040723.119
”No,” by analyzing the relationship between one or more independent variables and a dependent variable in a dataset.

B. SVM

Support Vector Machines (SVM) are popular supervised learning algorithms used for classification and regression problems. SVM creates optimal decision boundaries, or hyper- planes, to divide n-dimensional space into classes so that new data points can be categorized correctly. The SVM algorithm chooses support vectors, or extreme cases, to help create hyperplanes. SVM achieves an accuracy of 60%.

C. XGBoost classifier

XGBoost is an advanced gradient boosting library known for its efficiency, flexibility, and portability. It uses machine learning algorithms as part of gradient enhancement and offers Parallel Tree Boost, which solves many statistical problems quickly and accurately. The code runs on major distributed environments like Hadoop, SGE, MPI, and can solve problems beyond billions of examples. XGBoost achieves an accuracy of 92%.

D. KNN

K-Nearest Neighbor (KNN) is a simple machine learning algorithm based on supervised learning techniques. KNN calculates the space between a point and all points in the data, chooses the K number of specific examples closest to the point, and votes for the most common label (in classification) or averages the labels (in regression). KNN achieves an accuracy of 84%.

E. Decision Tree classifier:

The Decision Tree classifier is a popular model that creates a tree called a decision tree. Each node in the decision tree tests a column, and each branch corresponds to one of the possible values for that variable. The Decision Tree classifier achieves an accuracy of 94%.

IV. System Architecture

The dataset used here in this work is based on company bankruptcy which has 6820 rows and 95 columns. It was acquired from the Taiwan journal. The output class in this dataset has information about whether the company is bankrupt or not. Basic EDA techniques were applied to explore the data. Feature selection was initially done using Random Forest method which helped select the right features for the model. The data of the target class was initially evaluated to check the class imbalance and as expected a huge amount of class imbalance was found with the data where a company is bankrupt being the minority class. Imbalanced data are values in which the observed frequencies are very different across the different possible values of a categorical variable. It’s like there are many rows of some type and very few of another type. Correlation methods were also used to find out how various variables are correlated with each other. Figure 1 depicts the overall system architecture for evaluating the performance comparison of various ML models

The data is next being readied for model where we can apply the algorithm. The data imbalance of target class was

Financially stable: 96.52 % of the dataset
Financially unstable: 3.48 % of the dataset

As our data’s target class has a huge imbalance of data, we will use SMOTE sampling method to generate synthetic samples from the minority class.

SMOTE is an oversampling technique that produces synthetic samples of minority classes. SMOTE algorithm will help solving the problem of overfitting due to oversampling. We focus on the features and create new instances using interpolation between consecutive positive instances.

Then model was split into train and test and then predicted using various classification algorithms. The results were evaluated, and the best model was used for bankruptcy prediction

V. Experimental Evaluation

The performance of various Machine learning models are examined using the following metrics.

a) ACCURACY: Accuracy is the one of the predefined models which is used to find the balanced and unbalanced data in the models and how its performing in the given predicted models. A common way to calculate accuracy in machine learning is by dividing the total number of correct predictions by the total number of predictions made.

Accuracy = True positive + True Negative / True positive+ False positive + True Negative + False Negative

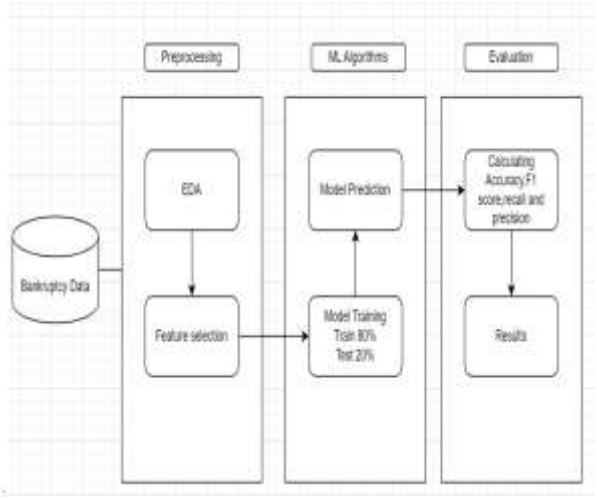

Fig. 1. Depicts the architecture of this work

b) PRECISION: It is calculated as the ratio of true Positives to all the positives predicted by the model. A higher precision indicates a lower number of false positives predicted by the model. In this data we have used precision for finding the perfectness of the model to know the true positive value, negative value. By the precision equation we can find the denominator for true positive of values.

DOI: rp-9788770040723.119

Precision = True Positive / True Positive + False Positive

c) RECALL: Recall, also known as Sensitivity, is another metric used to evaluate the performance of a machine learning model. It is calculated as the ratio of True Positives to all the positives in the dataset. A lower recall value indicates that the model is predicting more false negatives.

Recall = True Positive / True Positive + False Negative

(d) F1 SCORE

F1 score is a metric used to evaluate the overall performance of a machine learning model. It is calculated as the harmonic mean of precision and recall, taking into account the contribution of both metrics.

F1 Score = 2 *Precision*Recall / Precision+Recall

The models have been applied and the following results have been achieved

TABLE I: ACCURACY COMPARISON FOR VARIOUS ML MODELS

S.No	Model	Accuracy
1	SVM	60%
2	KNN	84%
3	XGBoost	92%
4	Decision Tree Classifier	94%

Table I shows accuracy of various ML models.We can see that Decision tree has more accuracy compared to other algorithms

TABLE II: PRECISION COMPARISON FOR VARIOUS ML MODELS

S.No	Model	Accuracy
1	SVM	64%
2	KNN	85%
3	XGBoost	93%
4	Decision Tree Classifier	94%

Table II shows precision of various ML models.We can see that Decision tree has more precision compared to other algorithms.

TABLE III: RECALL VALUE FOR VARIOUS ML MODELS

S.No	Model	Accuracy
1	SVM	61%
2	KNN	85%
3	XGBoost	92%
4	Decision Tree Classifier	94%

Table III shows recall of various ML models

TABLE IV: F1-Score for Various ML Models

S.No	Model	Accuracy
1	SVM	59%
2	KNN	85%
3	XGBoost	92%
4	Decision Tree Classifier	94%

Table IV shows F1-score of various ML models. F1 score is the highest for decsion tree as per the above table

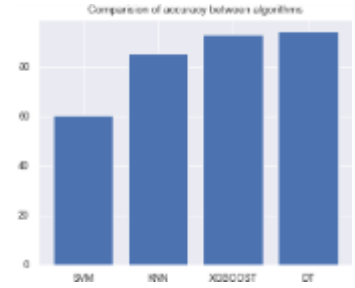

Fig. 2. Accuracy of various ML algorithms

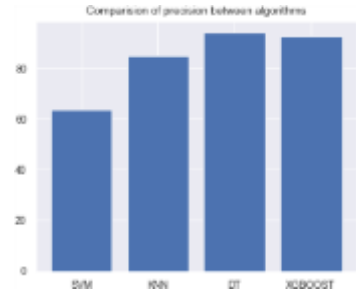

Fig. 3. Precision of various ML algorithms

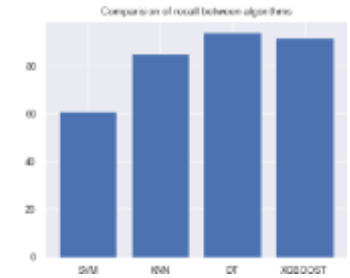

Fig. 4. Recall of various ML algorithms

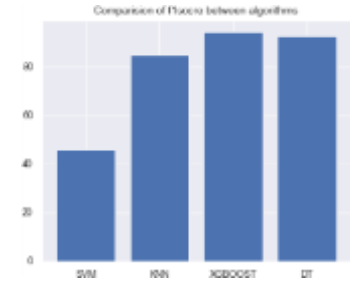

Fig. 4. Recall of various ML algorithms

Fig 2, Fig 3, Fig 4, Fig 5 shows the performance metrics of various ML algorithms

CONCLUSION

This study fully focused on applying and implementing the methods and techniques which actually help companies avoid getting bankrupt by analyzing their financials and thereby using a Machine Learning approach. Based on the results of the research study, the findings suggest that these models can offer improved accuracy and effectiveness

compared to traditional statistical methods. The study demonstrated the potential of sampling techniques such as SMOTE and also applied advanced models such as SVM, logistic regression, decision tree, and XGBoost which helped predicting bankruptcy ofpublic firms. These findings have implications for investors and creditors seeking to evaluate the financial health of a company and make informed decisions based on accurate risk assessments. The results of this study also highlight the importance of exploring a variety of models and techniques to develop robust and accurate predictions, as well as the need to continue refining and improving these models through further research and experimentation. Overall, the use of machine learning models for bankruptcy prediction holds promise as a valuable tool for improving financial risk assessment in the future. In conclusion, our study highlights the potential of machine learning models to predict bankruptcy, using a dataset of financial ratios from public companies. By using a range of algorithms, we were able to demonstrate the effectiveness of these models in identifying potential financial distress and predicting bankruptcy. We believe that future research could build on this work by applying more advanced algorithms, testing additional financial ratios, and exploring other potential features of interest. The proposed methods were successful in predicting a company's bankruptcy thereby saving company from getting bankrupt.

REFERENCES

[1] Aghaie, Arezoo, and Ali Saeedi, "Using bayesian networks for bankruptcy prediction: Empirical evidence from iranian companies," 2009 International Conference on Information Management and Engineering, IEEE, 2009

[2] Jabeur, and Sami Ben, "Bankruptcy prediction using partial least squares logistic regression," Journal of Retailing and Consumer Services, vol. 36, pp. 197-202, 2017.

[3] Kim, Hyunjoon, and ZhengGu, "A logistic regression analysis for pre- dicting bankruptcy in the hospitality industry," The Journal of Hospitality Financial Management, vol.14.1, pp. 17-34, 2006.

[4] Ivan, and Lobeev, "Bankruptcy Prediction for Innovative Companies," vol. 15.4, pp. 36-55, 2021.

[5] Olson, L. David, DursunDelen, and YanyanMeng, "Comparative analysis of data mining methods for bankruptcy prediction," Decision Support Systems, vol. 52.2, pp. 464 - 473, 2012.

[6] Lombardo, Gianfranco, et al., "Machine Learning for Bankruptcy Pre- diction in the American Stock Market: Dataset and Benchmarks," Future Internet, vol. 14.8, p. 244, 2022.

[7] Kiyak, Deimena, and DaivaLabanauskaite˙, "Assessment of the practical application of corporate bankruptcy prediction models," Ekonomikairvadyba, vol. 17, pp. 895-905, 2012.

[8] Li, Yachao, and Yufa Wang, "Machine learning methods of bankruptcy prediction using accounting ratios," Open Journal of Business and Management, vol. 6.1, pp. 1-20, 2017.

[9] Alam, TalhaMahboob, et al., "Corporate bankruptcy prediction: An approach towards better corporate world," The Computer Journal, vol. 64.11, pp. 1731-1746, 2021.

[10] H. Min, Jae, and Young-Chan Lee, "Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters,"

Expert systems with applications, vol. 28.4, pp. 603-614, 2005.

[11] C. V. Priscilla, and D.P. Prabha, "Influence of Optimizing XGBoost to handle Class Imbalance in Credit Card Fraud Detection," 2020 Third In- ternational Conference on Smart Systems and Inventive Technology (IC-

SSIT), 2020, pp. 1309-1315, doi: 10.1109/ICSSIT48917.2020.9214206

[12] Moshika, A., Thirumaran, M., Natarajan, B., Andal, K., Sambasivam, G., &Manoharan, R. (2021).Vulnerability assessment in heterogeneous web environment using probabilistic arithmetic automata. IEEE Access, 9, 74659-74673.

[13] Agbemava, Edinam, et al., "Logistic regression analysis of predictors of loan defaults by customers of non-traditional banks in Ghana," European Scientific Journal, vol. 12.1, 2016.

[14] Assegie, TsehayAdmassu, R. Lakshmi Tulasi, and N. Komal Kumar. "Breast cancer prediction model with decision tree and adaptive boost- ing." IAES International Journal of Artificial Intelligence, vol. 10.1, p. 184, 2021.

[15] Rajesh, M., &Sitharthan, R. (2022). Image fusion and enhancement based on energy of the pixel using Deep Convolutional Neural Network. Multimedia Tools and Applications, 81(1), 873-885.