IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

Benchmarking Predictive System for Detection of Bankruptcy using Artificial Intelligence

Jaiprakash Prajapati[1] and Kumar Abhishek[2]
Manager, Data Scientist, Mumbai, India[1]
Sr. Data Scientist, Bihar, India [2]

Abstract: *Bankruptcy assessment is very important for creditors and investors who are predicting bankruptcy. In the future, it will be a principal area for research. In the past several years, machine learning methods and artificial intelligence has accomplished a big success into the prediction of bankruptcy. "Class imbalance Deep learning for Bankruptcy Prediction - 2020 First International Conference on Power, Control and Computing Technologies (I.C.P.C.2.T) – Shanmukha vellamcheti, Pradeep Singh" [32] research paper was examined in this study and used as a criterion to improvise the AUC, Accuracy, and other metrics for predicting future bankruptcy using the LightGBM algorithm, a fast and high-performance open source distributed gradient boosting system. The results obtained are promising and show a greater AUC than earlier published findings. The cause of the exceptional predictive performance is examined, and it seems that the Feature Engineering element is critical. It observed that predictive models with good performance could achieved with feature engineering of the data. The average AUC obtained with this methodology for predicting bankruptcy is 0.949.*

Keywords: Artificial Intelligence, Benchmarking, Bankruptcy, Light GBM, Machine Learning

I. INTRODUCTION

Bankruptcy forecasting has a large influence on management, workers, investors, and the national economy, it is an important and thoroughly researched topic in finance and accounting. Accuracy is a crucial function because of the financial ramifications. The Bankruptcy prediction has been improved by several statistical approaches, including integral analysis, Calculus, Bayes Theorem, Linear optimization techniques, Logistical models, and Probit models. [1].

In making economic decisions, the probability of a business collapse is paramount. The organization's commercial position impacts not just communities, key stakeholders, and customers, but also policymakers as well as financial system. Because corporate bankruptcy has serious economic and social consequences, scholars have been compelled to gain a better understanding of the causes before predicting future bankruptcy problems. [2].

Bankruptcy speculation is a well-studied issue as in financial and strategy management literature (Polemis & Gonopoulos, 2012). Before the advent of modern financial ratio analysis, early researchers (such as Ramsar & Foster (1931), Fitzpatrick (1932), and Vinakor & Smith (1935) concentrated on contrasting insolvent and non-insolvent company ratios and came to the conclusion that the ratio of insolvent firms is weaker (Ugurlu & A20). Multiple discriminant analysis was used by Altman (1968) to forecast the bankruptcy of firms.

The basic strategy for forecasting bankruptcy was based solely on analysis. Regression analysis approaches that rely on logistic regression have been more popular since 1980s (Virag & Christoph, 2005). Ohlson (1980) was the first to employ logistic regression to forecast bankruptcy. The neural network method to bankruptcy prediction has recently been used by a number of academics due to promising outcomes in predictive modelling (Ugurlu and Aksoy, 2006). Odom and Sharda (1990) pioneered the neural network when they used it to predict bankruptcy.

II. OBJECTIVE

Our goal is to develop effective modelling strategies for bankruptcy prevention, focusing on the salient topics mentioned below:

1. Experts in the domain provide economic criteria to describe the firm status, though it's uncertain how to include them in a viable model.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

2. Statistical results used in the learning model tend to be influenced by disproportionate data, as there are generally more successful organizations than unsuccessful ones. Consequently, even if some firms are in danger, a trained model seems to forecast their success (majority class).

The primary accomplishments of work are listed as follows:

1. The emphasis is on improving predictive power through the Feature Engineering/Selection aspect, followed by the gradient boosting framework LightGBM Classifier, which combines several less accurate models to create more accurate models.
2. Addressing specific challenges through the application of appropriate pre-processing methods:
   a. Outliers: omission approach
   b. Missing value: mean strategy
   c. Unbalanced data Sets: The oversampling Method (SMOTE)
   d. Feature Engineering
   e. Modelling

III. LITERATURE SURVEY

Using historical financial data to envision future bankruptcies is an intriguing subject. Bankruptcy Prediction has been the subject of several studies. [3]. Logit analysis and Discriminant analysis are broadly utilized statistical models for Bankruptcy Prediction [4]. Altman Z-score [5] is majorly utilized in this discriminant analysis. A Survey of various studies on Bankruptcy Prediction is depicted in the table.

TABLE I: A survey of numerous research of bankruptcy FORECASTING [6]

Reference	Classifiers	Datasets	Evaluation methods				
			CRD	Accuracy	Type I/II error	F-Score	Kappa
7	MLP	Australian	No	Yes	No	No	No
8	MLP	Australian/German	No	Yes	No	No	No
9	MLP	US	No	Yes	No	No	No
10	MLP+LDA	Taiwan	No	Yes	Yes	No	No
11	MLP	Taiwan/US	No	Yes	No	No	No
12	MLP	Korea	No	Yes	No	No	No
13	MLP	Korea	No	Yes	No	No	No
14	MLP ensembles	Australian/German	No	Yes	No	No	No
15	MLP	Taiwan	No	Yes	Yes	No	No
16	MLP ensembles	Australian/ German	No	Yes	Yes	No	No
17	GA	Korea	No	Yes	No	No	No
18	GA	Korea	Yes	Yes	Yes	No	Yes
15	GA+SVM	Korea	No	Yes	No	No	No
19	LDA	Spanish	No	Yes	No	No	No
20	LDA	South African	No	Yes	Yes	No	No
21	PLS-DA	USA	No	Yes	Yes	Yes	No
22	LDA	USABD/JPNBD	No	Yes	No	No	No
23	B-CDT	Australian/ German/Japanese	No	Yes	No	No	No
24	WNN	Turkish/Spanish/US	No	Yes	No	No	No
25	SVM	Canada	No	Yes	No	No	No

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

26	DT, LR, SVM	US	No	Yes	No	No	No
27	MLP, DT, SVM, LR	Australian/German Japanese/Bankruptcy Data/Ucc	No	Yes	No	No	No
28	MDA, LR, CRT, and ANNs	Interfax SPARK database	No	Yes	Yes	Yes	No
29	DA, LR, RF, MLP, SVM, SOM	U.S. commercial banks	No	Yes	Yes	Yes	No
30	EXGB with other methods	Polish companies	No	Yes	No	No	No

IV. PROPOSED METHODOLOGY

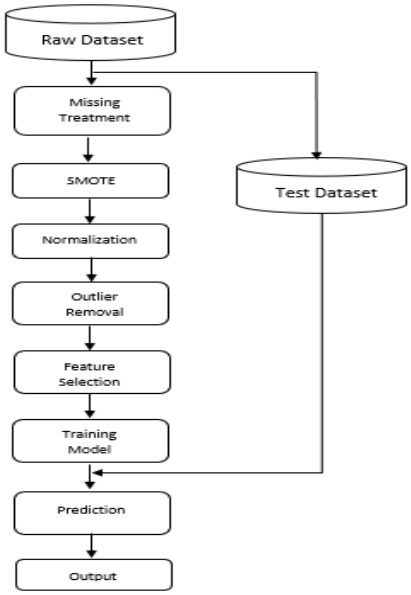

Fig. 1 Proposed Methodology

V. DATASET & PRE-PROCESSING

5.1 Dataset

The UCI database of Polish bankruptcies was taken into consideration while trying to solve the problem of bankruptcy. Polish firms' chances of bankruptcy are examined in this dataset. Between year 2000 and 2012, troubled businesses and from 2007 to 2013 successful businesses were studied. A large number of samples from Polish companies were evaluated throughout the course of five different time periods.

TABLE III: Summary of the Dataset

Year	Bankrupt Instances	Non-bankrupt Instances	Total Instances
1	271	6756	7027
2	400	9773	10173
3	495	10008	10503

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

4	515	9277	9792
5	410	5500	5910

5.2 Missing Data

TABLE IIIII: Missing Values

Year	Total Instances	Instances with missing values	Instances after removing missing value rows	Data loss if missing values were dropped
1	7027	3833	3194	54.54 %
2	10173	6085	4088	59.81 %
3	10503	5618	4885	53.48 %
4	9792	5023	4769	51.29 %
5	5910	2879	3031	48.71%

There are a lot of blanks in this dataset. Table II displays the amount of missing's for each data set into 3rd column, Instances after removing missing value in the 4th column; 5th column displays the proportion of data that has been lost if all rows with missing values have been disregarded. NaN values could not be eliminated since they lead to a considerable decrease in the data representation, which is more than 50 percent in a fixed data collection. Some cases of missing values of around 50% were ignored during the initial processing of missing data. The technique described in the following stages handles NaN values.

The MAE score was compared to a different approach to missing value imputation. The 4th approach in the below table, imputing missing values with mean achieved minimum error.

TABLE IVV: Imputation of Missing Values

Sr. No	Missing Imputation	MAE
1	Dropping columns with missing values	0.045
2	Iterative imputation	0.031
3	KNN imputation	0.039
4	Mean Imputation	0.028

5.3 Imbalance dataset

TABLE V: Imbalance dataset summary

Year	Total Instances	Before using SMOTE			After using SMOTE		
		Bankruptcy Instances	Non-Bankruptcy Instances	% Minority class	Bankruptcy Instances	Non-Bankruptcy Instances	% Minority class
1	7027	271	6756	3.85%	5838	5838	50%
2	10173	400	9773	3.93%	8177	8177	50%
3	10503	495	10008	4.71%	8504	8504	50%
4	9792	515	9277	5.25%	7787	7787	50%
5	5910	410	5500	6.93%	4776	4776	50%

Data imbalance usually represents an invalid category in a dataset: If a dataset contains two groups, the balanced dataset should comprise 50 per cent points for each group. The table above presents a class label population description for every dataset. Because of the dataset's imbalance, input samples from a minority class may be insufficient for the model if not correctly handled. Because of this, it is possible to have over-fitting circumstances. An imbalanced dataset is dealt by using the synthetic minority oversampling method (SMOTE), which equalizes data classes and builds matching datasets. Oversampling is applied when the quantity of data is inadequate so that the unique sample is balanced by expanding scale.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

5.4 Normalization and Outliers

The purpose of standardization is to shift the quantitative attribute value in the dataset to a single dimension, without distorting values. Data was normalized by transforming it into the z-score. Then for each variable if any value was above 3 standard deviations in any of the direction, it was marked as a potential outlier and was removed. We discovered that feature X4 is highly skewed, and that feature X3 contains significant outliers.

5.5 Feature Selection

It is necessary to pick a group of features that will provide the best prediction in modelling. Reducing the number of unnecessary and redundant features reduces the time and complexity of a model significantly. We used the wrapper method – Recursive feature elimination technique, which is a greedy optimization algorithm, to find the best performing feature subset. It recursively creates models and keeps aside the best or the worst performing feature at each iteration. It builds the next model using the left features until all the features are used up. The features are then ranked in the order in which they were eliminated. In total 56 features were selected, below Fig. 2 shows that the highest accuracy is achieved with 56 features.

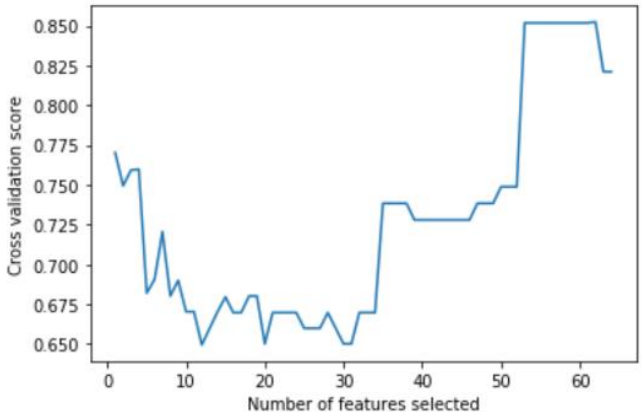

Fig. 2 Features Cross Validation Score

5.6 Why use a Nonlinear Model?

We used t-distributed stochastic neighbour embedding (t-SNE), a statistical method for visualizing high-dimensional data, to visualize the raw data. In the below figure we observed the non-linearity separation between the two-classes. Therefore, non-linear classifiers will perform better than linear classifiers have been concluded.

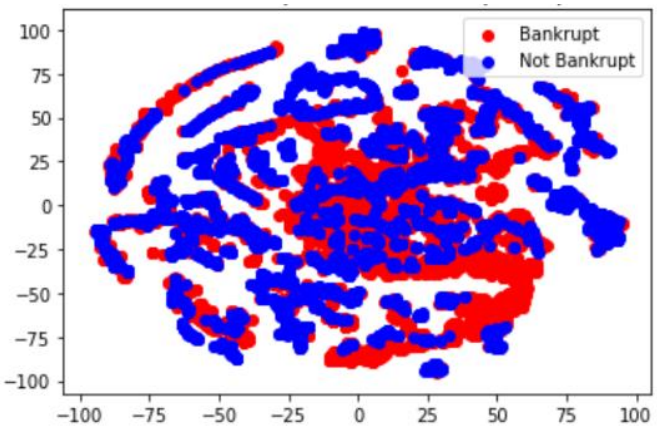

Fig. 3 t-SNE visualization of Bankrupt and Non-Bankrupt samples

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

VI. RESULTS

The Research paper "Class imbalance Deep learning for Bankruptcy Prediction - 2020 First International Conference on Power, Control and Computing Technologies (I.C.P.C.2.T) – Shanmukha vellamcheti,Pradeep Singh" average AUC score on 5 years data after SMOTE is 0.926

TABLE VI

Dataset	MLP Model AUC Observed in paper	Proposed Model AUC
Year-1	0.942	0.984
Year-2	0.912	0.935
Year-3	0.919	0.928
Year-4	0.917	0.94
Year-5	0.938	0.959

By comparing the results, the proposed model's average AUC score on 5-year data after SMOTE is 0.949. In the year-wise comparison between the benchmark results and the proposed model, the proposed model outperforms the benchmark model on each of the 5-year datasets.

VII. CONCLUSION

Bankruptcy is a major global problem. Its prediction is critical in the business world, and it must be done to reduce financial distress. The dataset used in the proposed model had a lot of missing data, and the data that was in it was also unbalanced. In that the companies that went bankrupt were very few in comparison to the companies that did not go bankrupt. In order to deal with missing values and unequal data spread, the mean imputation technique and Synthetic Minority Oversampling Technique (SMOTE) were used. After pre-processing and feature selection was done, model was trained, and it gave an AUC of 0.949 which is higher than the benchmark model. Having a better model and being able to predict a company's or organization's financial position can aid in averting global financial problems.

REFERENCES

[1]. Kim, M.J. & Kang, D.K. 2010. J. Expert Systems with Applications 3373–3379
[2]. Y. Zhang, S. Wang and G. Ji, A rule-based model for bankruptcy prediction based on an improved genetic ant colony algorithm, Mathematical Problems in Engineering, Hindawi.
[3]. Sartori, A. Mazzucchelli, and A. Di Gregorio, ―Bankruptcy forecasting using case-based reasoning: The CRePERIEapproach,‖ Expert Syst. Appl., vol. 64, pp. 400–411, 2016.
[4]. M. Tseng and Y. C. Hu, ―Comparing four bankruptcy prediction models: Logit, quadratic interval logit, neural and fuzzy neural networks,‖ Expert Syst. Appl., vol. 37, no. 3, pp. 1846–1853, 2010. 3
[5]. I.-P. Muhammad, ―Business bankruptcy prediction models: A significant study of the Altman's Z-score model,‖ Asian J. Manag. Res., vol. 3, no. 1, pp. 212–219, 2012.
[6]. S. Anand Christy, R.Arunkumar - Machine Learning Based Classification Models for Financial Crisis Prediction, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-4, November 2019
[7]. Fan and M. Palaniswami, Selecting bankruptcy predictors using a support vector machine approach,‖ 2000. IJCNN 2000, Proc. no.table 2, pp. 354–359, 2000.
[8]. West, ―Neural network credit scoring models,‖Comput. Oper. Res., vol. 27, no. 11–12, pp. 1131–1152, 2000.
[9]. F. Atiya and S. Member, ―Bankruptcy Prediction for Credit Risk Using Neural Networks : A Survey and New Results,‖ vol. 12, no. 4, pp.929–935, 2001.
[10]. T. Lee, C. Chiu, and C. Lu, ― Credit scoring using the hybrid neural discriminant technique.....file danneggiato,guardarenellacartellaorigine,‖ Expert Syst. Appl., vol. 23, pp. 245–254, 2002.
[11]. Z. Huang, H. Chen, C.-J. Hsu, W.-H. Chen, and S. Wu,―Credit rating analysis with support vector machines and neural networks: a market comparative study,‖ Decis. Support Syst., vol. 37, no.4, pp. 543–558, 2004.
[12]. J. H. Min and Y. C. Lee, ―Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters,‖ Expert Syst. Appl., vol. 28, no. 4, pp. 603–614, 2005.

[13]. K.-S. Shin, T. S. Lee, and H. Kim, ―An application of support vector machines in bankruptcy prediction model,‖ Expert Syst. Appl., vol. 28, no. 1, pp. 127–135, 2005.

[14]. West, S. Dellana, and J. Qian, ―Neural network ensemble strategies for financial decision applications,‖Comput. Oper. Res., vol. 32, no. 10, pp. 2543–2559,2005.

[15]. S.-H. Min, J. Lee, and I. Han, ―Hybrid genetic algorithms and support vector machines for bankruptcy prediction,‖ Expert Syst.Appl., vol. 31, no. 3, pp. 652–660, 2006.

[16]. F. Tsai, ―Financial decision support using neural networks and support vector machines,‖Expert Syst., vol. 25, no. 4, pp. 380–393, 2008.

[17]. K.-S. Shin and Y.-J. Lee, ―A genetic algorithm application in bankruptcy predictionmodeling,‖ Expert Syst. Appl., vol. 23, no. 3, pp. 321–328, 2002.

[18]. M. J. Kim and I. Han, ―The discovery of experts‘ decision rules fromqualitative bankruptcy data using genetic algorithms,‖ Expert Syst. Appl., vol. 25, no. 4, pp. 637–646, 2003.

[19]. Alfaro, N. García, M.Gámez, and D. Elizondo, ―Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks,‖ Decis. Support Syst., vol. 45,no. 1, pp. 110–122, 2008.

[20]. Altman, G. Marco, and F. Varetto, ―Corporate distress diagnosis: Comparisons using linear discriminantanalysis and neural networks (the Italian experience),‖ J. Bank. Financ., vol. 18, no. 3, pp. 505–529, 1994. Tsai and J. Wu, ―Using neuralnetwork ensembles for bankruptcy prediction and credit scoring,‖ Expert Syst. Appl., vol. 34, no. 4, pp. 2639–2649, 2008.

[21]. Tsai and J. Wu, ―Using neural network ensembles for bankruptcy prediction and credit scoring,‖ Expert Syst. Appl., vol. 34, no. 4, pp. 2639–2649, 2008

[22]. L. Zhou,―Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods,‖ Knowledge-Based Syst., vol.41, pp. 16–25, 2013.

[23]. J. Abellán and C. J. Mantas, ―Improving experimental studies about ensembles of classifiers for bankruptcyprediction and credit scoring,‖ Expert Syst. Appl., vol. 41, no. 8, pp. 3825–3830, 2014.

[24]. N. Chauhan, V. Ravi, and D. Karthik Chandra,―Differential evolution trained wavelet neural networks: Application to bankruptcy prediction in banks,‖ Expert Syst. Appl., vol. 36, no.4, pp. 7659–7665, 2009.

[25]. T. Xiong, S. Wang, A. Mayers, and E. Monga, ―Personal bankruptcy prediction by mining credit card data,‖ExperSyst. Appl., vol. 40, no. 2, pp. 665–676, 2013.

[26]. L. Olson, D. Delen, and Y. Meng, ―Comparative analysis of data mining methods forbankruptcy prediction,‖Decis. Support Syst., vol. 52, no. 2, pp. 464–473, 2012.

[27]. C. F. Tsai and K. C. Cheng, ―Simple instance selection for bankruptcy prediction,‖ Knowledge-Based Syst., vol. 27, pp. 333–342, 2012.

[28]. Fedorova, E. Gilenko, and S. Dovzhenko, ―Expert Systemswith Applications Bankruptcy prediction for Russian companies : Application of combined classifiers,‖ vol. 40, pp. 7285–7293, 2013.

[29]. J.L. Iturriaga and I. P. Sanz, ―Expert Systems with Applications Bankruptcy visualization and prediction using neural networks : A study ofU . S . commercial banks,‖ vol. 42, pp. 2857–2869, 2015.

[30]. M. Zi, S. K. Tomczak, and J. M. Tomczak, ―Ensemble boosted trees withsynthetic features generation in application to bankruptcy prediction,‖ vol. 58, pp. 93–101, 2016.

[31]. GuolinKe, Qi Meng, Thomas Finely, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu, ―LightGBM: A Highly Efficient Gradient Boosting Decision Tree,‖Advances in Neural Information Processing Systems 30 (NIP 2017) | December 2017.

[32]. Class imbalance Deep learning for Bankruptcy Prediction - 2020 First International Conference on Power, Control and Computing Technologies (I.C.P.C.2.T) – Shanmukha vellamcheti,Pradeep Singh

BIOGRAPHY

- Jaiprakash Prajapati possesses an engineering degree in computer and an MBA from Northeast Virginia along with several certifications from prestigious institutions and corporates such as IIT Kharagpur, IIT Roorkee, AWS and IBM having more than a decade experience in data science. A multidisciplinary data scientist with an extensive experience in Customer Level Personalization, Pricing and Promotion Analytics, Marketing Mix and

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

Multi-touch Attribution, Risk Prediction & Predictive Maintenance. I can be reached on jaiprakash.prajapati@outlook.com.

- Kumar Abhishek holds an engineering degree in Computer Science with a decade of experience in Data Science. A seasoned Data Scientist with acquaintance in Risk Analytics, AML Analytics, Fraud Analytics, Personalization, Pricing and Promotion Analytics, Marketing Mix and Multi-touch Attribution, and Predictive Maintenance. I am reachable at abhishek465@gmail.com.