Available online at http://www.advancedscientificjournal.com
http://www.krishmapublication.com
IJMASRI, Vol. 2, issue 5, pp. 592-597, May -2021
10.53633/ijmasri.2022.2.5.010

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY ADVANCED SCIENTIFIC RESEARCH AND INNOVATION (IJMASRI)

ISSN: 2582-9130 **IBI IMPACT FACTOR 1.5** **DOI: 10.53633/IJMASRI**

RESEARCH ARTICLE

LOAN FINANCIAL RISK ANALYSIS AND VISUALIZATION USING PYTHON

Rahul Ratusaria[1], Tushar Baghel[2], Ayush Chander Vanshi[3] and Ms. Sudha Narang[4]

[1,2,3,] Department of Computer Science and Engineering, Maharaja Agrasen Institute of Technology, Rohini,
[4]Assistant professor, Department of Computer Science and Engineering, Maharaja Agrasen Institute of Technology, Rohini, Delhi

Abstract

This project is made to compare the three gradient boosting algorithms in the loan risk analysis. Reason for using gradient boosting algorithm is that we have studied previous research on this topic in which different algorithms are compared and out of which gradient boosting algorithm seems to be standing out of all. So, it is beneficial to compare different gradient boosting algorithms and check which performs better and can be used for analysis in this problem. For that we spent major time on cleaning and manipulating the data in the form that we needed. Than after implementing the three algorithms we compared them on the basis of precision, recall and F1 scores and as a result we can see XGBoost seems to be performing better than LightGBM and CatBoost with precision more than 80%..

Keywords: Gradient Boosting, XGBoost, LightGBM, CatBoost, Machine learning

Introduction

The main objective of this project was to build machine learning algorithms that would be able to identify potential defaulters and therefore reduce company loss. In this fast-pacing world, it has become very difficult to trust anyone due to which person in need don't get proper help he/she is seeking for and also in order to minimize the chances of fraud there is a need for the companies to have a proper methods, algorithms and data driven approach to find the fraudsters. Machine learning is the best possible way for companies in order to separate out these people and get proper insights whether it is beneficial for them to provide them credit support or not. Being computer science background, our main motive is to devise such ways and algorithms by which we can reduce human efforts for managing things and make increase the automation. In order to do so, we have to

check which method we are implementing is best suitable and has best performance.
Credit risk is associated with the possibility of a client failing to meet contractual obligations, such as mortgages, credit card debts, and other types of loans. Minimizing the risk of default is a major concern for financial institutions. For this reason, commercial and investment banks, venture capital funds, asset management companies and insurance firms, to name a few, are increasingly relying on technology to predict which clients are more prone to stop honouring their debts. By this project, we are trying to compare three gradient boosting ensemble machine learning algorithms and find which one can work better based on some performance matrices like Precision, Recall and F1 score.

Related Work

In the last three decades, various estimation methods have been developed to solve this problem. However, the issue of credit risk predicting of banks remains an open issue (Nehrebecka 2021). Any work performed on this issue has diverse strengths and weaknesses. Financial professionals gain valuable information and knowledge due to the nature of credit risk and research in this field (Husejinovic 2019). This knowledge could not be achieved in other research and areas of study (Husejinovic and Husejinovic 2019). As a result, great attention and desire exist for establishing financial model variation. In the process of predicting credit risk, a great deal of research has been done. To rate specific banking customers, a new method had been utilized based on fuzzy sets and particular rules in the bank (Blahun *et al.*, 2020). A method of IT2FS has been used for scoring from intervals between 0 and 100 (Sidik *et al*., 2013).

Abdou et al. presented In demand for the facility from 487 actual data (achieved through the Islamic Bank of England), of which 336 cases have been accepted and 151 cases have been rejected (Abdou 2014). An artificial multilayer perceptron neural network has been trained by using this data. Finally, the usage of the mentioned model has been achieved in essential and nonessential components of bank credit risk forecasting.

(Gozer *et al.*, 2014) Offered a hybrid model of artificial neural networks and supported vector machines to evaluate the incapability of business units to pay debts. As primary data, 62 business units were considered, half of which could not repay the debt. Different methods, including RBF networks, multilayer perceptron, and backup vector machines, have been applied to analyze the results. According to the results, the vector-based backup method has better outcomes than other methods.

(Keramati *et al.*, 2013) utilized data mining techniques for analyzing credit data (Keramati and Yousefi 2011). In this research, a comprehensive study of all the articles on this topic affords a review article for summarizing the methods offered in assessing banks' credit. Finally, a comparison of the existing methods is presented. In 2021, (Doko *et al.*, 2021) utilized different machine-learning models for generating precise models for credit risk valuation based on the North Macedonia Central Bank. They compared the results with five machine-learning models including decision tree, logistic regression, artificial neural network, and support vector machines for categorizing the credit risk data. The studied models were then evaluated by diverse machine-learning metrics, and then a detailed credit registry data-based model was presented for credit risk prediction from the population history credit. The results indicated that the best precision was obtained by utilizing the decision tree classifier with no need for scaling.

(Giri *et al.*, 2021) proposed an operative rule-based classification technique for credit risk forecasting based on a metaheuristic technique, called Biogeography Based Optimization (BBO) algorithm. They modified the presented BBO algorithm using rule mining and named it locally and globally tuned biogeography-based rule-miner (LGBBO-RuleMiner). The new algorithm was then employed for the optimal rule set discovery by considering an accuracy with a high value than the dataset including the continuous and categorical attributes. The proposed methods were then validated by comparing with some different rule-miners like Decision Table, PART, OneR (1R), Conjunctive Rule, JRip, Random Tree, and J48 based

on some different bio-inspired optimization algorithms based on assuming two credit risk datasets. The results showed that the proposed method outperforms the other compared algorithms in dissimilar cases. (Hozic and Saracevic 2021) progressed a credit risk model to the company clients for probability prediction of default (PD). They considered 151 different companies which were the clients of an Islamic bank in Bosnia and Herzegovina. The model was developed based on logistic regression. Because of the profit-loss sharing in Islamic banks, they require for evaluating successful joint investment or client financing probability. The results showed that by globalization and growing of the Islamic financing worldwide, better tools and creditworthiness predictions are required for Islamic banks.

Methodology

Concept Used

Gradient Boosting Algorithms (XGBoost, LightGBM, CatBoost)

XGBoost:

XGBoost (Extreme Gradient Boosting) is an open-source free library that help in providing an effective and efficient implementation of the gradient boosting ensemble algorithm. There were already existing efficient algorithms for gradient boosting before XGBoost, after releasing XGBoost power of this technique is unleashed and made everyone in the applied machine learning to generally consider for analysis and to use for efficiently predicting. GBM is taken as the base for XGBoost. GBM is improved to make XGBoost. Both works on same procedure.. The trees in XGBoost are built sequentially, trying to correct the errors of the previous trees. In order to trying to correct errors of previous trees, XGBoost trees are built sequentially

LightGBM:

As LightGBM is very speedy and efficient algorithm, so it is getting popular day by day. LightGBM can handle huge amounts of data easily. LightGBM is not highly recommended for small data sets because it doesn't perform well with small number of data points.

Training process in this algorithm is speeded up by using a histogram-based method to select the best split**.** For continuous variables, training data is divided into the buckets and bins rather than taking the individual values. This makes the training process faster and lowers memory usage.

CatBoost:

CatBoost is a boosting algorithm that can handle categorical variables in the data. Most machine learning algorithms cannot work with strings or categories in the data. Thus, converting categorical variables into numerical values is an essential preprocessing step.

CatBoost can internally handle categorical variables in the data. These variables are transformed to numerical ones using various statistics on combinations of features.

Reason why CatBoost is being widely used is that it works well with the default set of hyperparameters. Hence, as a user, we do not have to spend a lot of time tuning the hyperparameters.

We have taken the data set containing 45,000 rows and 43 columns. First, we will analyze the data and try to convert it into useful form so that we can perform further analysis over it. After examining it, we will implement three gradient boosting algorithms namely XGBoost, LightGBM and CatBoost to compare them and to check which one of these can perform better for this type of problem and dataset. Parameters on which we will be comparing these algorithms are precision and recall. Any further insights can be drawn on the basis of results that we will obtain after properly implementing the algorithms.

Procedure

We are working with a data set containing 43 features for 45,000 clients. `target_default` is a

True/False feature and is the target variable we are trying to predict. After exploring the data set, we found that some features had outliers and missing values. Other variables, that would add no value to the model, were removed.

The following histogram helps us visualize the distribution of the numerical features:

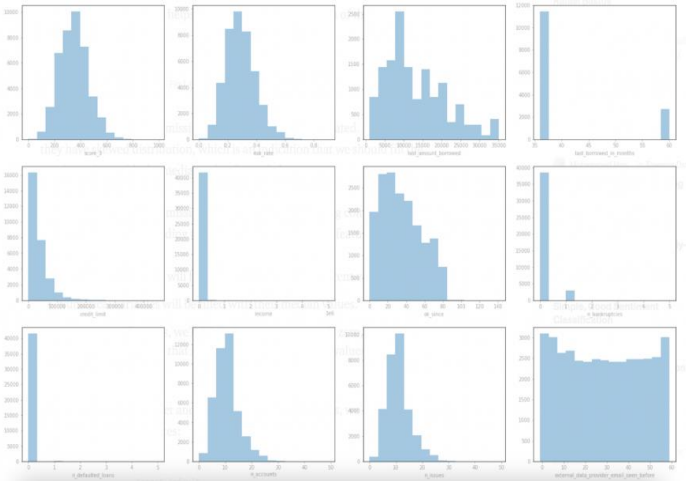


Fig. 1: Numerical features Histogram

For treating missing values, we can use the median value for each features as we can see from the skewed distribution. Other missing values can be filled with these assumptions:

- Categorical variables with the most recurrent value.
- Numerical variables with their median values.
- In some specific cases, missing values with zero.

Machine Learning Models:

Three gradient boosting algorithms to determine which yields better results are:

- XGBoost
- LightGBM
- CatBoost

At first we need to split training and testing data set. As data set is unbalanced we'll standardize and resample the data with `RandomUnderSampler` and `StandardScaler`

XGBoost:

For the XGBoost model, we'll tune the following hyperparameters

- n_estimators - Number of trees in model
- max_depth - Maximum depth of tree
- min_child_weight - Minimum sum of the instance weight needed in child
- gamma - Minimum loss reduction required to make further partition on leaf node of tree
- learning_rate - Step size shrinkage used in the update to prevents overfitting

LightGBM

- max_depth - Maximum depth of tree
- learning_rate – Rate of shrinkage
- num_leaves - Max number of leaves in a single tree
- min_data_in_leaf - Minimal number of data in the single leaf

CatBoost

- depth - Depth of tree
- learning_rate – Rate of shrinkage
- l2_leaf_reg - Coefficient at L2 regularization term of cost function

After tuning some hyperparameters, all the three models have displayed betters results. It is worth mentioning that the XGBoost presented a greater score increase, while LightGBM and CatBoost saw a little improvement.

Evaluation Metrics

Precision will give us proportion of the positive identifications that were correct.

Precision = (True Positives) / (True Positives + False Positives)

Recall will determine proportion of the real positives that were correctly identified

Recall = (True Positives) / (True Positives + False Negatives)

F1 Score is a metric that is useful when we need to seek a balance between precision and recall.

F1 = 2 * (Precision * Recall) / (Precision + Recall)

Experimental Result

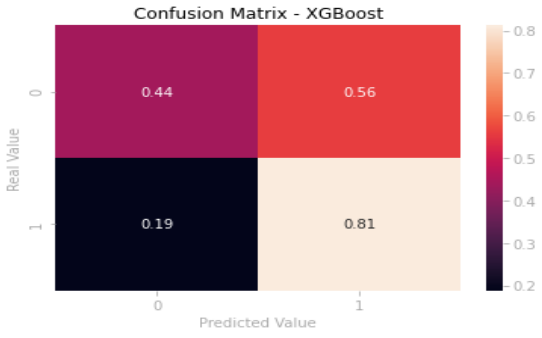
After completing and testing project we have reached to following results for each of the three gradient boosting algorithms:
XGBoost:
Best recall rate: 0.81

LightGBM:
Best recall rate: 0.69

CatBoost
Best recall rate: 0.65
Now, we can check how these models perform on the **test set.** To help us visualize the results, we are plotting a **confusion matrix** for each one of them.

XGBoost

	precision	recall	f1-score	support
0	0.92	0.44	0.60	8771
1	0.22	0.81	0.34	1665
accuracy			0.50	10436
macro avg	0.57	0.63	0.47	10436
weighted avg	0.81	0.50	0.56	10436



loss, we decided to give more emphasis on reducing false positives, searching for the best hyperparameters that could increase the recall rate.

Fig. 2: XGBoost Confusion Matrix

LightGBM

	precision	recall	f1-score	support
0	0.91	0.62	0.74	8771
1	0.25	0.68	0.37	1665
accuracy			0.63	10436
macro avg	0.58	0.65	0.55	10436
weighted avg	0.81	0.63	0.68	10436

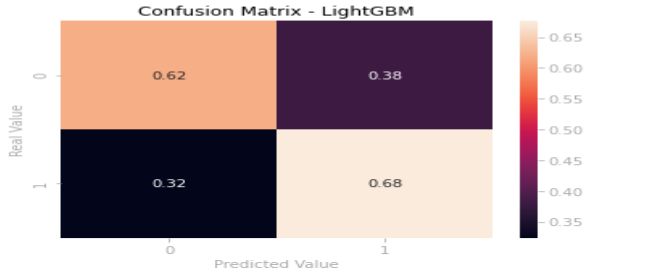
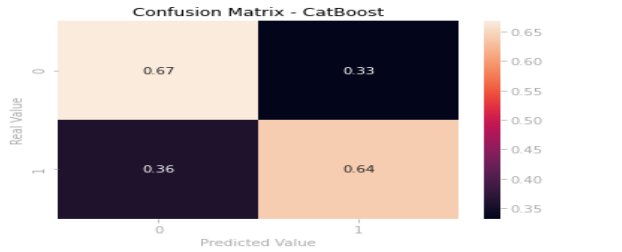


Fig. 3: LightGBM confusion Matrix

CatBoost

	precision	recall	f1-score	support
0	0.91	0.67	0.77	8771
1	0.27	0.64	0.38	1665
accuracy			0.66	10436
macro avg	0.59	0.65	0.57	10436
weighted avg	0.81	0.66	0.71	10436



Conclusion

The best model possible would be the one that could minimize false negatives, identifying all defaulters among the client base, while also minimizing false positives, preventing clients to be wrongly classified as defaulters.

Meeting these requirements can be quite tricky as there is a tradeoff between precision and recall, meaning that increasing the value of one of these metrics often decreases the value of the other. Considering the importance of minimizing company

Among the three **Gradient Boosting Algorithms** tested, **XGBoost** yielded the best results, with a recall rate of 81%, although it delivered an

undesired 56% of false positives. On the other hand, **LightGBM** and **CatBoost** delivered a better count of false positives, with 38% and 33% respectively, but their false negatives were substantially higher than that of XGBoost, resulting in a weaker recall rate.

This Major project presents a classic evaluation metrics dilemma. In this case, it would be up to the company's decision-makers to analyze the big picture, with the aid of the machine learning algorithms, and decide the best plan to follow.

References

1. Nehrebecka, N. (2021) "Internal credit risk models and digital transformation: what to prepare for? an application to Poland," *European Research Studies Journal*, vol. XXIV, no. 3, pp. 719–736, 2021. View at: Publisher Site | Google Scholar.
2. Husejinovic, A and Husejinovic, M. (2019). "Adoption of internet banking in Bosnia and Herzegovina," Available at SSRN 3501455. View at: Google ScholarHusejinovic, A and Husejinovic, M. (2019). "Adoption of internet banking in Bosnia and Herzegovina," Available at SSRN 3501455. View at: Google Scholar.
3. Husejinovic, A. (2019). "Efficiency of commercial banks operating in federation of Bosnia and Herzegovina using DEA method," *Sustainable Engineering and Innovation*, vol. 1, no. 2, pp. 2712–0562. View at: Google Scholar
4. Blahun, I.S., Blahun, I. I and Blahun, S.I. (2020). "Assessing the stability of the banking system based on fuzzy logic methods," *Banks and Bank Systems*, vol. 15, no. 3, pp. 171–183. View at: Publisher Site | Google Scholar
5. Sidik, G.K., Djatna, T and Buono, A. (2013). "An It2fs model for sharia credit scoring: analysis & design," *Jurnal Sistem Informasi*, vol. 9, pp. 58–66.View at: Publisher Site | Google Scholar
6. Abdou, H., Alam, S and Mulkeen, J. (2014). "Would credit scoring work for islamic finance? a neural," *Journal of the Operational Research Society*, vol. 54, pp. 822–832, 2014. View at: Publisher Site | Google Scholar
7. Gozer, I.C., leite de Albuquerque, A.R.P. Isotani, S. Gimenes, R. Márcio, R and Toesca, A. (2014). "Evaluation of insolvency in mutual credit unions by the models of artificial neural networks and support vector machines," *African Journal of Agricultural Research*, vol. 9, pp. 1227–1237. View at: Publisher Site | Google Scholar
8. Keramati, A and Yousefi, N. (2011). "A proposed classification of data mining techniques in credit scoring," in *Proceedings of the 2011 International Conference of Industrial Engineering and Operations Management*, pp. 22–24, Kuala Lumpur, Malaysia. View at: Google Scholar
9. Doko, F., Kalajdziski, S and Mishkovski, I. (2021). "Credit risk model based on central bank credit registry data," *Journal of Risk and Financial Management*, vol. 14, no. 3, p. 138. View at: Publisher Site | Google Scholar
10. Giri, P.K., De, S.S. Dehuri, S and Cho, S.B. (2021). "Biogeography based optimization for mining rules to assess credit risk," *Intelligent Systems in Accounting, Finance and Management*, vol. 28, no. 1, pp. 35–51. View at: Publisher Site | Google Scholar
11. Hozic, M and Saracevic, N. (2020). "Credit risk assessment for an islamic bank in Bosnia and Herzegovina," in *Islamic Finance Practices*, pp. 131–147, Springer, Berlin, Germany. View at: Google Scholar
12. A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting. View at: Google

******