JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Enhanced Credit Risk Analysis Using Light-GBM and SMOTE: A Comparative Study with Traditional Classifiers

Vansh Choubey, Prerna Gupta, Aditya Singh, Abhishek Ambadwar, Himanshu Taiwade

Student[1,2,3,4], Professor

Priyadarshini College of Engineering

Abstract

Credit risk analysis is becoming more and more essential in this day and age as bank credit risk is a significant challenge in modern financial transactions and the ability to identify qualified credit card holders among a large number of applicants. In the past it was done by screening each applicants credit history which in turn increased manual labour and time consuming. Although credit scoring models were build they often struggled with complex and non-linear relationships and imbalanced datasets leading to inaccurate predictions. In this study we used the public dataset (name of the dataset) available on Kaggle to explore the application of Light Gradient Boosting Machine (Light-GBM) and Synthetic Minority Oversampling Technique (SMOTE) for credit Risk Analysis and also perform a comparison between various other Machine learning methods. These models are trained and compared based on the Accuracy, Precision, Recall, F1-score, and AUC-ROC to find out the effectiveness of Light-GBM. Results shows that Light-GBM combined with Smote significantly enhances the accuracy of the model and reducing the bias present in the minority class and outperforming the base line models. The experiments thus demonstrate benefits of combining gradient bosting with oversampling techniques for improved Credit Risk Analysis.

Keywords: Credit Risk Analysis, LightGBM, SMOTE, Machine Learning, Class Imbalance, Ensemble Learning, Financial Risk Prediction, AUC-ROC, Predictive Modeling.

Introduction

Credit risk an important concern for banks caused by the debtor's inability to full fill the obligations as per the contract. This can cause the bank to have economic losses and hinder the bank's lending operations. For this creating a credit risk prediction model is important so that banks can know if a customer will default or not. Traditional risk assessment methods are not able to capture the complex relationships present in financial data which then lead to inaccurate predictions.

However, class imbalance is also one of the major challenges in credit risk datasets as the machine learning models tends to get bias towards the majority class which leads to poor predictive performance. To overcome this issue, we use SMOTE to balance the dataset by generating synthetic samples for the minority class.

In this paper we investigate the efficacy of Light-GBM with SMOTE and benchmark the performance against various other traditional models such as Random Forest, Logistic Regression and also compare it with other

gradient boosting methods such as XG-Boost, Cat-Boost, etc. This would help us understand how it performs as compared to the other models.

Literature Review

To successfully implement this, we reviewed a vast amount of relevant literature and summarized the works of previous literature.

This research by Chang Yu et al. (2024) explores credit risk prediction with a data set of more than 40,000 cases from a commercial bank. Authors compare some dimension reduction methods like PCA and T-SNE to pre-process high-dimensional data sets. Authors also try out LightGBM, XGBoost, and TabNet models, optimizing the models with distributed processing and hyperparameter search. The key observation is that the combination of LightGBM, PCA, and SMOTEENN greatly enhances the accuracy of credit risk prediction compared to other models, especially in predicting high-quality borrowers. The research emphasizes the significant role that oversampling methods like SMOTEENN play in handling class-imbalanced datasets and improving the predictive power.[1]

Aruleba and Sun (2024) investigated the importance of ensemble learning techniques, i.e., LightGBM, XGBoost, AdaBoost, and Random Forest, in predicting credit risk. Their study employed SMOTE-ENN as a class imbalance reduction method, and SHAP (Shapley Additive Explanations) to improve model interpretability. With both the German Credit Dataset and Australian Credit Approval Dataset, the result indicated that XGBoost provided the highest recall rate (0.930) and specificity (0.846) for the German dataset while Random Forest provided better performance for the Australian dataset with recall of 0.907 and specificity of 0.922. The findings indicate that ensemble methods enhance prediction accuracy, while SMOTE-ENN effectively balances class distribution, thereby minimizing bias in model predictions..[2]

With the increase in number of fraudulent transactions in mobile and online payment systems has prompted the application of advanced machine learning algorithms such as XGBoost and LightGBM for fraud detection. The study develops fraud prediction and also apply SMOTE (Synthetic Minority Over Sampling Technique) to overcome the issue of class imbalance and feature selection techniques to increase the model’s accuracy and precision. The models' performance was compared with Random Forest, Neural Networks, and Logistic Regression with various performance metrics such as Precision, Recall, and F1 Score. The experiment indicates that combination of SMOTE with XGBoost and LightGBM enhanced the detection accuracy by 6% in comparison to the standard models and 5% compared to each of the individual classifiers. Additionally, a Local Ensemble Model combining XGBoost and LightGBM shows greater accuracy and stability.[3]

Naik's 2021 research highlights the increasing need for credit risk prediction models due to the rise in cashless transactions and credit card usage. The study tackles the issue of uneven data distribution in credit default records by using SMOTE to enhance predictions. It compares seven machine learning models and finds that LightGBM performs better than others because it is efficient, learns quickly, and handles large datasets well. The study suggests that using LightGBM in banks can help predict credit defaults, which aids financial institutions in assessing risks and making informed decisions.[4]

Methodology

This section describes the step-by-step methodology employed to perform credit risk analysis using LightGBM and SMOTE to achieve an objective and reproducible research process. The methodology involves four major stages: data preprocessing, exploratory data analysis, baseline model comparison (without SMOTE), and model evaluation after applying SMOTE.

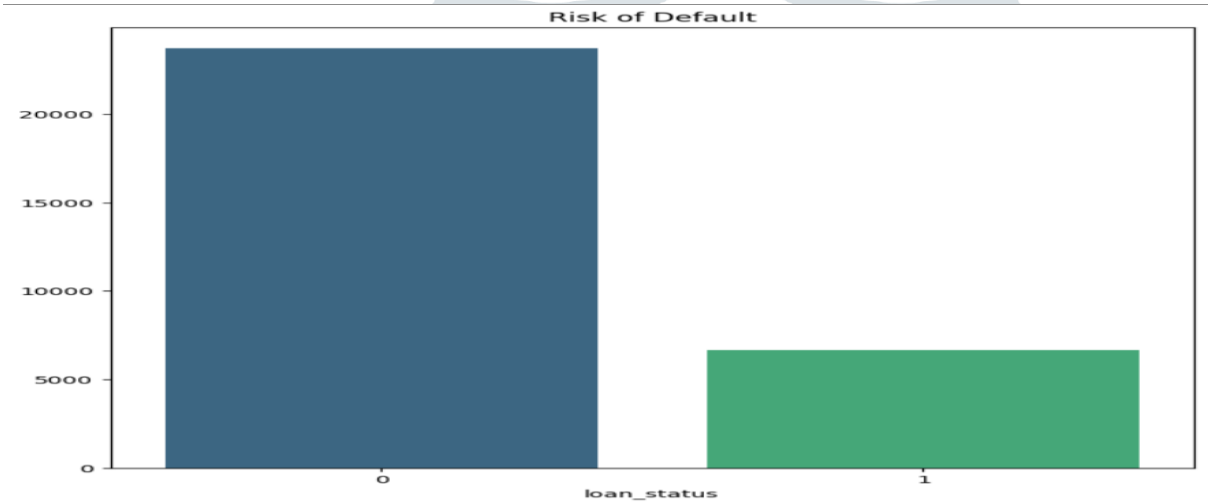
1.Data Pre-Processing:

The dataset contains features related to borrower characteristics, credit history, and loan repayment behaviour. The first phase of this was to perform the data preprocessing in which we rid the data of any sort of discrepancies that might be present in it to ensure high quality data. To do this we performed the following steps,

- Dealing with Missing Values: Median imputation for numeric attributes and mode imputation for categories.
- Encoding Categorical Variables: One-hot encoding and label encoding for non-numeric features.
- Feature Scaling: Min-Max Scaling for standardizing numerical features.
- Outlier Detection & Removal: Z-score technique to detect and eliminate extreme outliers.

2.Exploratory Data Analysis:

In the next stage we perform the Exploratory Data Analysis to identify data distribution patterns, feature correlations, and class imbalance, which tells us that the non-default cases significantly outnumber the default cases as shown on the figure given below,



where, 0 represents the non-default cases and 1 represents the default cases.

3.Baseline Model Comparison:

Next, we perform a baseline model evaluation without SMOTE to establish a benchmark and for various classifiers including Logistic regression, Random Forest, SVM, XGBoost and LightGBM all models are trained on the original imbalanced dataset. Models are assessed using Accuracy, Precision, Recall, F1-score, and AUC-ROC, with LightGBM showing strong performance but lower recall due to class imbalance.

4.Application of Smote:

At last, we apply SMOTE to the imbalanced dataset to balance it by synthetically generating minority class samples. LightGBM is then again applied and the performance metrics are compared with and without SMOTE. Results show notable increase in the recall and AUC-ROC, making the model more effective in identifying the high-risk borrowers while maintaining the accuracy.

Therefore, to conclude this methodology that LightGBM enhanced with SMOTE significantly improves the credit-risk prediction of the model.

Results

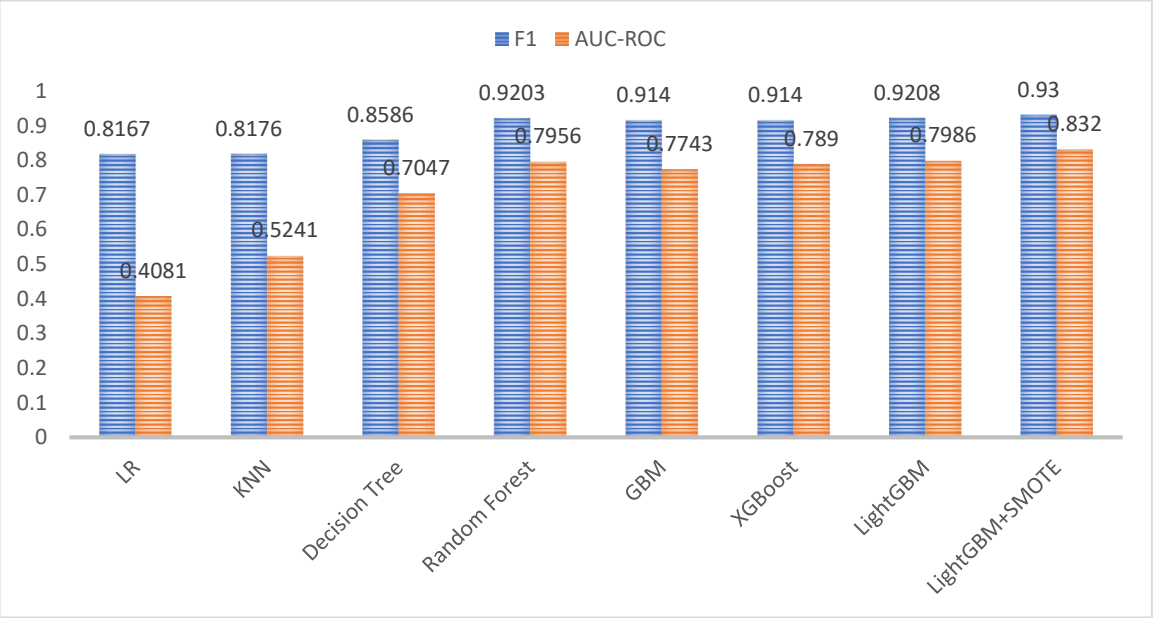
The machine learning model analysis for credit risk prediction showed that dealing with class imbalance is critical to the improvement of the prediction accuracy. In the current research, various classifiers were compared before and after applying SMOTE, with the LightGBM model performing best.

Model performance without SMOTE

When trained on imbalanced dataset, LightGBM shows the highest AUC-ROC scores and highest accuracy outperforming other models such as Logistic Regression, Random Forest, SVM and XGBoost.

LightGBM performance with SMOTE

Using SMOTE enhanced the capability of LightGBM to identify high-risk borrowers with greater recall and AUC-ROC values and no loss of precision. In particular, LightGBM's accuracy improved from 0.92 to 0.93 following the use of SMOTE, showing the utility of improving classification balance. This is consistent with the finding that data balancing effectively mitigates majority class bias and improves the overall performance of the model.



This chart compares machine learning models based on F1-score (blue) and AUC-ROC (orange). LightGBM with SMOTE performed the best, achieving the highest accuracy (0.93) and AUC-ROC (0.832), showing the advantage of SMOTE in handling class imbalance. Random Forest, GBM, and XGBoost also performed well, while KNN and Logistic Regression had the lowest AUC-ROC, indicating weaker performance.

Conclusion

This research assessed LightGBM for credit risk prediction and confirmed its superiority over conventional models such as Logistic Regression and Random Forest. SMOTE use enhanced class balance, leading to enhanced recall and overall model performance. Findings indicated LightGBM providing enhanced AUC-ROC and F1 scores, making it a viable tool in financial risk evaluation. The findings confirm the efficacy of the use of boosting algorithms in combination with oversampling techniques. Future studies can optimize feature selection, hyperparameters, and investigate other ensemble methods to further improve predictive accuracy in credit risk evaluation.

Refrences

[1] C. Yu, Y. Jin, Q. Xing, Y. Zhang, S. Guo, and S. Meng, "Advanced User Credit Risk Prediction Model using LightGBM, XGBoost and Tabnet with SMOTEENN," Nov. 13, 2024, *arXiv*: arXiv:2408.03497. doi: 10.48550/arXiv.2408.03497.

[2] I. Aruleba and Y. Sun, "Effective Credit Risk Prediction Using Ensemble Classifiers With Model Explanation," *IEEE Access*, vol. 12, pp. 115015–115025, 2024, doi: 10.1109/ACCESS.2024.3445308.

[3] Q. Zheng, C. Yu, J. Cao, Y. Xu, Q. Xing, and Y. Jin, "Advanced Payment Security System: XGBoost, LightGBM and SMOTE Integrated," in *2024 IEEE International Conference on Metaverse Computing,*

Networking, and Applications (MetaCom), Hong Kong, China: IEEE, Aug. 2024, pp. 336–342. doi: 10.1109/MetaCom62920.2024.00063.

[4] K. S. Naik, "Predicting Credit Risk for Unsecured Lending: A Machine Learning Approach," Oct. 05, 2021, *arXiv*: arXiv:2110.02206. doi: 10.48550/arXiv.2110.02206.