Fraud Detection in Banking Data by Machine Learning Technique

N.Usha,Department o fCSE, Narayana Engineering College,Gudur.

P.Yesdhani khan,M.Tech,(Phd),AssociativeProfessor,DepartmentofCSE, Narayana Engineering College,Gudur.

ABSTRACT:

As technology advanced and e-commerce services expanded, credit cards became one of the most popular payment methods, resulting in an increase in the volume of banking transactions. Furthermore, the significantincreaseinfraudrequireshighbankingtransactioncosts.Asa result,detectingfraudulent activities has becomea fascinating topic. Inthis study, weconsider the useofclass weight-tuning hyper-parameters to control the weight of fraudulent and legitimate transactions. We use Bayesian optimization in particular to optimize the hyperparameters while preserving practical issues such as unbalanced data.We propose weight-tuning as a pre-process for unbalanced data, as well as CatBoost and XGBoost to improve the performance of the LightGBM method by accounting for the voting mechanism. Finally, in order to improve performance even further, we use deep learning to fine-tune the hyperparameters, particularly our proposed weight-tuning one. We perform some experiments on real-world data to test the proposed methods. To better cover unbalanced datasets, we use recall-precisionmetrics in addition to the standard ROC-AUC. CatBoost, LightGBM, and XGBoost are evaluated separately using a 5-fold cross-validation method.

INDEX TERMS :Bayesian optimization, data Mining, deep learning, ensemble learning, hyperparameter, unbalanced data , Machine learning.

I.INTRODUCTION

In recent years, there has been a significant increase in the volume of financial transactionsdue to the expansion of financial institutions and the popularity of web-based e-commerce. Along with credit card development, the pattern of credit card fraud has always been updated. Fraudsters do their best to makeitlooklegitimate,andcreditcardfraudhasalwaysbeenupdated.Fraudstersdotheirbestto

Fraud detection in banking is considered a binary classification problem in which data is classified as legitimate or fraudulent. Because banking data is large in volume and with datasets containing a large amount of transaction data, manually reviewing and finding patterns for fraudulent transactions iseither impossible or takes a long time. Therefore, machine learning-based algorithms play a pivotal role in fraud detection and prediction. Machine learning algorithms and high processing power increase the capability of handling large datasets and fraud detection in a more efficient manner. Machine learning algorithms and deep learning also provide fast and efficient solutions to real-time problems . In this paper, we propose an efficient approach for detecting credit card fraud that has been evaluated on publicly available data setsand has used optimized algorithms LightGBM,XGBoost, CatBoost,and logistic regression individually,as well asmajority voting combined methods,aswell as deeplearning and hyper parameter settings. An ideal fraud detection system should detect more fraudulent cases, and the precision of detecting fraudulent cases should be high, i.e., all results should be correctly detected, which will lead to the trust of customers in the bank, and on the other hand, the bank will not suffer losses due to incorrect detection. The main contributions of this paper are summarized as follows:

- We adopt Bayesian optimization for fraud detection and propose to use the weight-tuninghyperparametertosolvetheunbalanceddataissueasapre-processstep.Wealsosuggestusing CatBoost and XGBoost alongside LightGBM to improve performance.
- We use the XGBoost algorithm due to the high speed of training in big data as well as the regularizationterm,whichovercomes overfittingbymeasuringthecomplexityofthetree,andit does not require much time to set the hyperparameters.
- We also use the Catboost algorithm because there is no need to adjust hyperparameters for overfitting control, and it also obtains good results without changing hyperparameterscompared to other machine learning algorithms.
- We propose a majority-voting ensemble learning approach to combine CatBoost, XGBoost, and LightGBM and review the effect of the combined methods on the performance of fraud detectiononreal,unbalanceddata.Wealsoproposetousedeeplearningforadjustingandfine- tuning the hyperparameters.
- To evaluate the performance of the proposed methods, we perform extensive experiments on real-world data. Tobetter covertheunbalanceddatasets, weuserecallprecision inaddition

tothetypicallyusedROCAUC.WealsoevaluatetheperformanceusingF1_scoreandMCC metrics.

II.IMPLEMENTATION

MODULES:

Data

selectionDataprepr

ocessing Data

splitting

Classification

Prediction

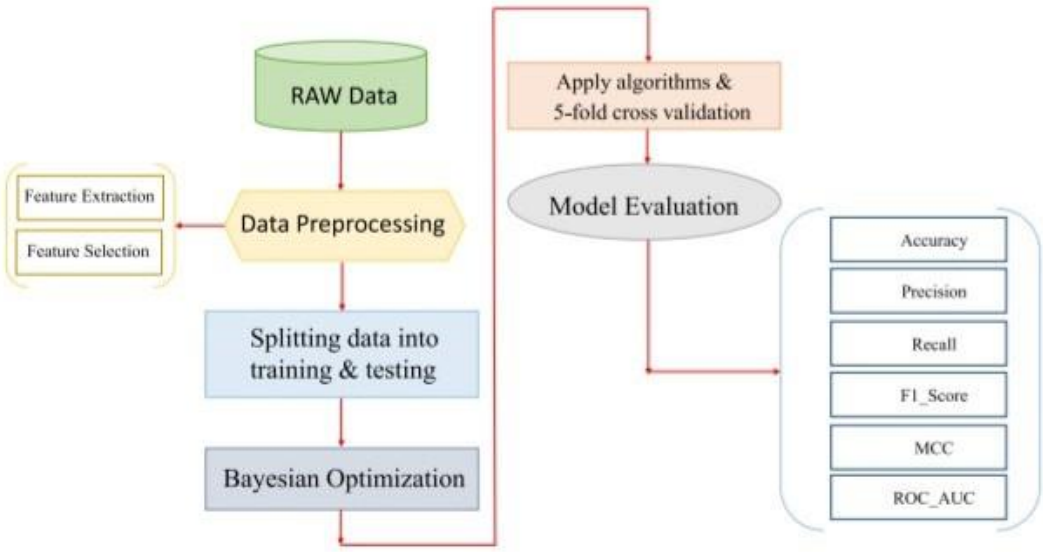
PerformanceMetrics

Graph Comparison

MODULESDESCRIPTION:

DATASELECTION:

- Theinput datawascollectedfromthedatasetrepositorylikeUCIRepository.
- Inthisprocess,theinputdatahavesomecolumnslikestep,type,amount,nameOrig, balanceOrig, nameDest, balanceDest, isFlaggedFraud, etc.
- Inourcollecteddatasetwasreadinthisprocessusingpandas



DATAPREPROCESSING:

- Datapre-processing is the process of removing the unwanted data from the dataset.
- Pre-processingdatatransformationoperationsareusedtotransformthedatasetintoa structure suitable for machine learning.
- This step also includes cleaning the dataset by removing irrelevant or corrupted data that can affect the accuracy of the dataset, which makes it more efficient.
- Missing data removal
- Missing data removal: In this process, the null values such as missing values and Nan values are replaced by 0.
- Missingandduplicatevalueswereremovedanddatawascleanedofanyabnormalities.
- LabelEncoding:Inthisprocess,thestringvaluesareconvertedintointegerformoreprediction.

DATASPLITTING:

- Duringthemachinelearningprocess,dataare neededsothatlearningcantake place.
- In addition to the data required for training, test data are needed to evaluate the performance ofthe algorithm but here we have training and testing dataset separately.

- Inourprocess,wehavetodivide as trainingandtesting.
- Data splittingis the act ofpartitioningavailable data intotwoportions,usuallyforcross-validator purposes.
- OnePortionofthedataisusedtodevelopapredictivemodelandtheothertoevaluatethe model's performance.

CLASSIFICATIONS

RandomForest Algorithm

- **Randomforest**isamachinelearningalgorithmforfraud detection.
- It'sanunsupervisedlearningalgorithmthatidentifiesfraudbyisolatingoutliersinthedata.
- RandomForestisbasedontheDecisionTreealgorithm.
- It isolates the outliers by randomly selecting a feature from the given set of features and thenrandomly selecting a split value between the max and min values of that feature.

KNN Algorithm:

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN is anon-parametric algorithm, which means it does not make any assumption on underlying data.

- It is also called alazy learner algorithmbecause it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

AdaBoostAlgorithm

- AdaBoostalso calledAdaptiveBoostingisatechniqueinMachineLearningusedasanEnsemble Method. The most common algorithm used with AdaBoost is decision trees with one level that means with Decision trees with only 1 split. These trees are also called Decision Stumps.
- AdaBoost is best used to boost the performance of decision trees on binary classification problems.
- AdaBoost was originally called AdaBoost.M1 by the authors of the technique Freund and Schapire. More recently it may be referred to as discrete AdaBoost because it is used for classification rather than regression.
- AdaBoost can be used to boost the performance of any machine learning algorithm. It is best used with weak learners. These are models that achieve accuracy just above random chance on a classification problem.
- The most suited and therefore most common algorithm used with AdaBoost are decision trees with one level. Because these trees are so short andonly containone decision for classification, they are often called decision stumps.

III. Methodology

Identifying fraudulent financial statement is critical for capital market regulation and is generally formulated as a classification problem. Feature selection in traditional machine learning methods does not consider correlation information among financial features which may influence performance of classifiers. To explore correlation information on conducting financial statement fraud detection (FSFD), we combine traditional features with knowledge graph models, and learn new representationsenriched

withfeatureembeddingofvariousfinancialcategories.Thesefeaturerelationsdefinedbycorrelation

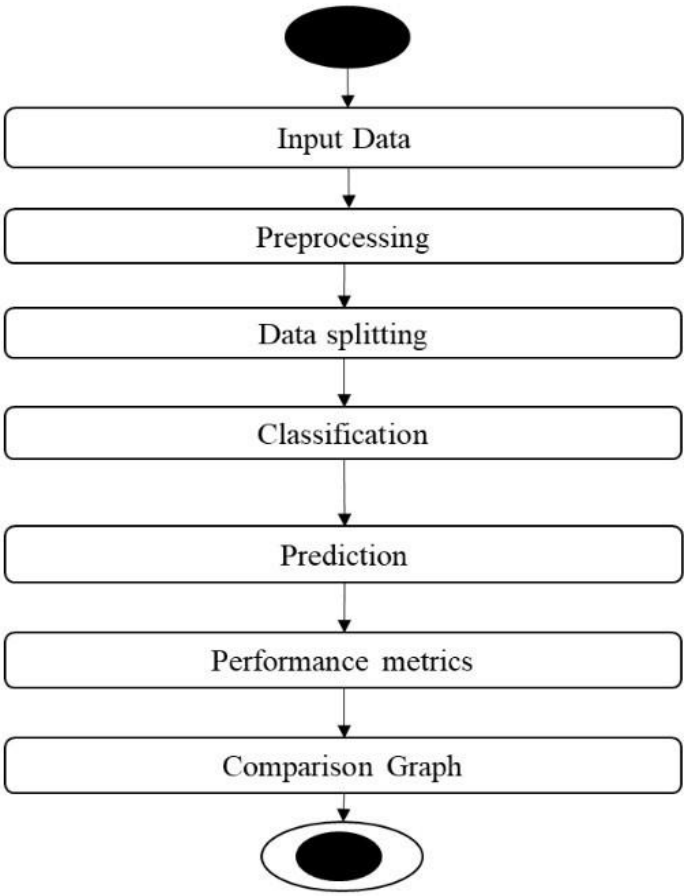
types may form knowledge graphs with features as nodes and correlation relations as edges. Experimentations demonstrate that financial feature representations with correlation information significantly improve classification performances for SVM and K-NN, marginally better than decision trees and logistic regression, but not outperforming naive Bayes (Kernel).

Advantages:

Changeofdetectingunknownattack. May
be more efficient.

Disadvantages:

Mustbeusedwithsignaturedetection.
Fraudimplies unusual activity.

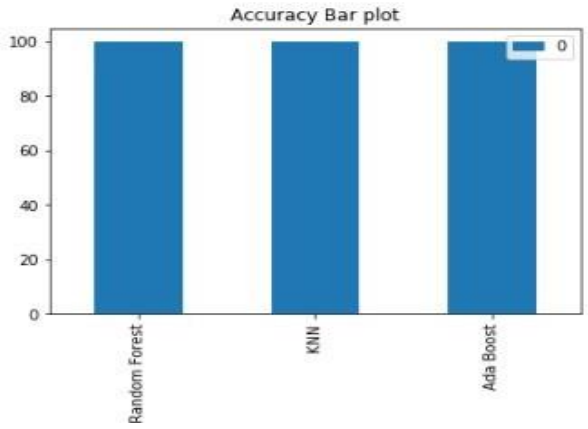


IV. EXPERIMENTALRESULTSANDDISCUSSION

We use the stratified 5-fold cross validation method and the boosting algorithms with the Bayesian optimization method to evaluate the performance of the proposed framework. We extract the hyperparameters and evaluate each algorithm individually before usingthemajority voting method. We examinethealgorithmsintripleanddoubleprecision.ThecomparisonresultsarepresentedinTable .Most studies in the literature rely on AUC diagrams to evaluate performance. However, as can be seen from the ROC-AUC curve in the value of AUC in severely unbalanced data is not a good evaluationmetric. It is influenced by the real positives and considers the negatives irrelevant. According to theROC-AUC Fig. 4, the logistic regression algorithm 0.9583 has the highest number of fraud detection, but it has the lowest value in other criteria. The precision-recall curve is illustrated in the system performance ina more precise mannercompared withthe ROC-AUC curve.However, the results cannot be cited because false negatives are far from the view of this diagram. As the highest value belongs to the combination of the CatBoost and LightGBM algorithms with a value of 0.7672, and the lowest value belongs to logistic regression and is 0.7361. Comparing the precision, recall, and F1-score as well as the MCC. The best performance is related to the combination of lightGBM and XGBoost algorithms, which have an MCC value of 0.79 and an F1-score of 0.79. In individual algorithms, XGBoost has the highest values. According to the digits has achieved better performance compared with individual algorithms and majority voting ensemble learning. The MCC and F1-score metrics have values of 0.8129 and0.8132, respectively. The area under the ROC curve in the deep learning method is illustrated .

	step	type	amount	...	newbalanceDest	isFraud	isFlaggedFraud
0	1	PAYMENT	NaN	...	0.00	0	0
1	1	PAYMENT	1864.28	...	0.00	0	0
2	1	TRANSFER	NaN	...	0.00	1	0
3	1	CASH_OUT	181.00	...	0.00	1	0
4	1	PAYMENT	11668.14	...	0.00	0	0
5	1	PAYMENT	7817.71	...	0.00	0	0
6	1	PAYMENT	7107.77	...	0.00	0	0
7	1	PAYMENT	7861.64	...	0.00	0	0
8	1	PAYMENT	4024.36	...	0.00	0	0
9	1	DEBIT	5337.77	...	40348.79	0	0
10	1	DEBIT	9644.94	...	157982.12	0	0
11	1	PAYMENT	3099.97	...	0.00	0	0
12	1	PAYMENT	2560.74	...	0.00	0	0
13	1	PAYMENT	11633.76	...	0.00	0	0
14	1	PAYMENT	4098.78	...	0.00	0	0
15	1	CASH_OUT	229133.94	...	51513.44	0	0
16	1	PAYMENT	1563.82	...	0.00	0	0
17	1	PAYMENT	1157.86	...	0.00	0	0
18	1	PAYMENT	671.64	...	0.00	0	0
19	1	TRANSFER	215310.30	...	0.00	0	0

```
#--------------------Camparision batween 3 Algorithm Accuracy---------------#
********************
```



V. CONCLUSIONANDFUTUREWORK

In this paper, we studied the credit card fraud detection problem in real unbalanced datasets. We proposed a machine learning approach to improve the performance of fraud detection. We used a publicly available ‘‘credit card’’ dataset with 28 features and 0.17 percent of the fraud data. We proposed two methods. In the proposed LightGBM, we used class weight tuning to choose the proper hyperparameters. We used the common evaluation metrics, including accuracy, precision, recall, F1-score, and AUC. Our experimental results showed that the proposed LightGBM method improved the frauddetection cases by50% andthe F1-scoreby20% comparedwiththe recently presentedmethodin [17]. We improve the performance of the algorithm with the help of the majority voting algorithm. We also improved the criteria by using the deep learning method. The assurance of the results of MCC for unbalanced data proved that, compared to other criteria of evaluation, it’s stronger. In this paper, by combining the LightGBM and XGBoost methods,we obtained 0.79 and 0.81 for the deep learning method.

REFERENCES

- J. Nanduri, Y.-W. Liu, K. Yang, and Y. Jia, ‘‘Ecommerce fraud detection through fraud islands and multi-layer machine learning model,’’ in Proc.Future Inf. Commun. Conf., in Advances in Information and Communication. San Francisco, CA, USA: Springer, 2020, pp. 556–570.
- I. Matloob, S. A. Khan, R. Rukaiya, M. A. K. Khattak, and A. Munir, ‘‘A sequence mining-based novel architecture for detecting fraudulent transactions in healthcare systems,’’
- H. Feng, ‘‘Ensemble learning in credit card fraud detection using boosting methods,’’ in Proc.2nd Int. Conf. Comput. Data Sci. (CDS), Jan. 2021,pp. 7–11.
- M. S. Delgosha, N. Hajiheydari, andS.M. Fahimi, ‘‘Elucidationof big data analytics inbanking: A four-stage delphi study,’’ J. Enterprise Inf.Manage., vol. 34, no. 6, pp. 1577–1596, Nov. 2021.
- M. Puh and L. Brkić, ‘‘Detecting credit card fraud using selected machine learning algorithms,’’inProc.42ndInt.Conv.Inf.Commun.Technol.,Electron.Microelectron.(MIPRO),May2019,pp. 1250–1255.
- K. Randhawa, C. K. Loo, M. Seera, C. P. Lim, and A. K. Nandi, ‘‘Credit card fraud detection using AdaBoost and majority voting,’’ IEEE Access,vol. 6, pp. 14277–14284, 2018.

- N. Kumaraswamy, M. K. Markey, T. Ekin, J. C. Barner, and K. Rascati, ‘‘Healthcare fraud data mining methods: A look back and look ahead,’’ Perspectives Health Inf.Manag.,vol. 19, no.1, p. 1, 2022.
- A. Malik, K.W. Khaw, B. Belaton, W. P. Wong, and X.Chew, ‘‘Credit card fraud detection using a new hybrid machine learning architecture,’’ Mathematics, vol. 10, no. 9, p. 1480, Apr. 2022.
- K. Gupta, K. Singh, G. V. Singh, M. Hassan, G. Himani, and U. Sharma, ‘‘Machine learning based credit card fraud detection—A review,’’ in Proc.Int. Conf. Appl. Artif. Intell. Comput. (ICAAIC), 2022, pp. 362–368.
- R. Almutairi, A. Godavarthi, A. R. Kotha, and E. Ceesay, ‘‘Analyzing credit card fraud detection based on machine learning models,’’ in Proc. IEEE Int. IoT, Electron.
- N. S. Halvaiee and M. K. Akbari, ‘‘A novel model for credit card fraud detection using artificial immune systems,’’ Appl. Soft Comput., vol. 24, pp. 40–49, Nov. 2014.
- A. C. Bahnsen, D. Aouada, A. Stojanovic, and B. Ottersten, ‘‘Feature engineering strategies forcredit card fraud detection,’’ Expert Syst. Appl., vol. 51, pp. 134–142, Jun. 2016.