International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

Effective Appeal of Fraud Detection in Credit CardData Using Machine Learning Techniques

K. Geetha Rani[1], Chippe Ashok[2], Haalini Mandava[3], R. Sai Pranav[4], Santati Uttej[5]

[2345]Final Year B.Tech Student, Department of CSE, Jain University, Bengaluru, Karnataka, India.
[1]Associate Professor, Department of CSE, Jain University, Bengaluru, Karnataka, India.

Email:[1]AssistantProfessor,geetha.rani@jainuniversity.ac.in,[2]ashokc8500@gmail.com,[3]haalinimandava@gmail.com,[4]saipranavyadav0707@gmail.com, [5]santatiuttej@gmail.com

Abstract

Banks and other financial organizations are extremely concerned about credit card theft on a global scale.More and more, real-time fraud detection is being done using machine learning algorithms. In this study, two wellknown machine learning techniques—Light Gradient Boosting Machine and Isolation Forest—are used to investigate the detection of credit card fraud. These algorithms were utilised in the study to analyse transactional data and look for patterns of creditcard fraud. The results showed that both algorithms had high rates of success in identifying fraudulent behaviour. The Isolation Forest approach outperformed the Light Gradient Boosting Machine technique in problems involving classification, regression, and ranking. The authors conclude that utilising machine learning techniques can significantlyimprove by using machine learning techniques.

Key Words: Light Gradient Boosting Method, Isolation Forest**.**

I. INTRODUCTION

Mainly we use two machine learning techniques to build a model that identifies fraudulent activity in the dataset. Moreon that below. Microsoft was the first company to develop Light GBM, sometimes referred to as a distributed gradient boosting machine learning framework known as Light Gradient Boosting Machine. For problems including classification, ranking, and other types of machine learning, use decision tree algorithms. Performance and scalability are the main concerns of development. The reasons for using this strategy are: The lack of unique weights in GBDT data instances is exploited by gradient-based one-sided sampling (GOSS) techniques. Data instances with higher gradients provide greater contributions to the calculation of information gain because they perform distinct roles depending on their gradients. This algorithm has an accuracy of 91.60 for identifying fraud within the dataset (including AUC, Precision, and Fi-Score). Isolation Forest (IF): Just like Random Forest, IF is built using decision trees. Because there are no predetermined labels, this model is also unsupervised. Isolation Forests were built using anomalies, or "few and different" data points.

II. LITERATURE REVIEW

[1]. The application of this method is justified by processing randomly subsampled data in an isolated forest with a treestructure based on randomly selected quality. More cuttings were needed to isolate specimens that had moved further up the tree, making them less likely to be antiques. Occasional specimens landing on short branches can be a strange sign. These use a modified version of the gradient boosting algorithm for classification. The model parameters were tuned using the hyperparameter tuning method.

[2] Investigating imbalance problems caused by credit card fraud data being so imbalanced that we have no choice but to rely on anomaly detection techniques.

[3] They used isolation forests and local outlier coefficients to detect anomalies. Works well with unlabelled records. This algorithm avoids partial detection tasks.

[4] A hybrid method that combines unsupervised and supervised approaches to increase the accuracy of fraud detection. Outliers are calculated at multiple granularity levels, outperforming existing schemes.

[5] To find credit card fraud, we applied machine learning methods like decision trees, logistic regression, and random forests. Logistic regression, decision trees, and random forests all have accuracy rates of 90.0%, 94.3%, and 95.5%, respectively.

[6] Three different machine learning techniques focused on outlier detection. The techniques used in this article are isolation forests, local outlier coefficients, and support vector machines. The isolated forest accuracy is 98.65%, the local outlier is 98.58%, and the SVM is 69.87%.

III. PROPOSED MODEL

We are developing new technologies to detect and prevent such fraud in credit card data transactions. The technique weuse to detect fraud is the use of an optimized light gradient boosting algorithm and Isolation Forest. The following stepsare the various processes that take place in our fraud detection system.

STEP 1 : DATA PREPROCESSINGSTEP 2 : FEATURE SELECTION STEP 3 : DATA SPLITTING

STEP 4 : ENSEMBLE METHODSTEP 5 : FEATURE SCALING

STEP 6 : K – FOLD CROSS VALIDATION

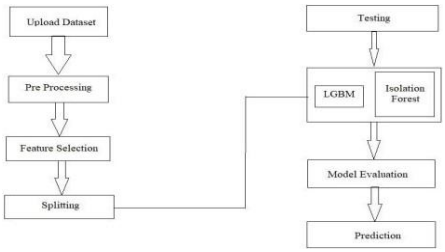

Fig 1: Flow diagram of methods

STEP 1: DATA PREPROCESSING:

Data preparation includes any sort of processing done to raw data to get it ready for another data processing procedure. Data pre-processing converts data into a format that data mining can handle more quickly and effectively. The followingtools and techniques are used for data pre-processing:

1.Sampling;2. Transformation;3. Imputation;4. Normalization and;5. Feature extraction.STEP 2: FEATURE SELECTION

A technique for reducing the input variables of a model by using only relevant data and eliminating irrelevant data. Themethod of automatically choosing characteristics for a machine learning model that are appropriate given the type of issue it is attempting to tackle.

STEP 3:DATA SPLITTING

Data partitioning is the process of dividing a dataset into two or more parts. It is frequently necessary to test or evaluatedata on one portion of the split while training a model on the other. Especially when creating models from data, data slicing is a crucial component of data science. By employing this technique, the data model's construction and use are made more accurate.

STEP 4: ENSEMBLE METHOD

Ensemble techniques attempt to improve the accuracy of model results by combining multiple models rather than relyingon just one model. The integrated model greatly improves the accuracy of the results. As a result, ensemble approaches to machine learning are growing in popularity.

Primary Ensemble Methods:

1. **Sagging or Bagging:** Classification and regression frequently employ bootstrap aggregation, also known as bagging. By using decision trees, model accuracy is increased and variability is greatly decreased. Overfitting is a problem with many predictive models that can be fixed by lowering variability and raising precision. Bagging falls under the conceptsof aggregation and bootstrapping.

1.1 Bootstrapping is a sampling technique that selects samples from the entire population (set) using the exchange method. Surrogate sampling aids in ensuring selection is made randomly. An algorithm for baseline learning is then applied to the samples to complete the process.

1.2 Aggregation: Aggregation is used in bagging to incorporate all possible outcomes of a prediction and to randomisethe outcomes. Forecasts without aggregation are inaccurate because not every scenario is taken into account. As a result, either all predictions from predictive models or probabilistic bootstrapping techniques are used for aggregate. Bagging is beneficial since it produces a single stable strong learner as opposed to a single unstable base learner. Additionally, variance is removed, and model overfitting is diminished. One of the limitations of bagging is the computational expense. Therefore, disregarding proper bagging strategies can result in significant model distortions.

2. **Boosting:** Boosting is an ensemble method that uses a combination of weak learners to improve the accuracy ofa model. The idea is to train many weak models sequentially, where each model tries to correct the errors of the previous model. This way, the final model becomes a strong learner that can make accurate predictions. Gradient Boosting, AdaBoost, and XGBoost are some of the most popular boosting algorithms. AdaBoost, for example, trains decision trees as weak learners with equal weights on observations. The algorithm adjusts the weights of themisclassified observations to enhance the performance of the subsequent decision trees.

2.1 Gradient boosting: gradually improves the predictors of an ensemble, allowing early predictors to correct later predictors, improving model accuracy. New predictors are adjusted to account for the error results of previous models. Gradient enhancers can identify and address learner prediction problems thanks to descent gradients.

2.2 XGBoost: Boosted gradient decision trees are used in XGBoost to provide faster performance. This is highly dependent on the computational efficiency and effectiveness of the target model. Model training should be sequential, so gradient-enhanced machines should be implemented slowly.

2.3 Stacking: An ensemble approach is stacking, also referred to as stacked generalisation. This technique operatesby enabling ensemble of predictions from multiple learning algorithms similar to the training algorithm. Regression, density estimation, distance learning, and classification all make good use of stacking. It can also be used to estimate how often errors occur when bagging.

3. **Deviation reduction or variance reduction:** The ensemble technique is the most effective tactic for reducing model variation and raising forecast accuracy. Variance is eliminated when numerous models are combined to provide a singleforecast that is selected from a range of potential outcomes. An ensemble of models integrates numerous models to guarantee that the final forecast is the best one that can be made while taking into account all predictions.

STEP5: FEATURE SCALING

The numbers are shifted and rescaled using a scaling method called normalization shifts such that they lie between 0and 1. Likewise known as min-max scaling.

Xstd = x – mean(x)/Standard deviation(x)

Normalisation is yet another scaling technique. This utilises a unit standard deviation and centres the data on the mean. As a result, the attribute has a mean of 0 and a standard deviation of 1 for the distribution it produces.

Normalisation Formula:

Xnormalised = (X – Xmin)/(Xmin – Xmin)

Metric	Formula	Interpretation
Accuracy	$\frac{TP + TN}{TP+TN+FP+FN}$	Overall performance of model
Precision	$\frac{TP}{TP+FP}$	How accurate the positive predictions are
Recall Sensitivity	$\frac{TP}{TP+FP}$	Coverage of actualnegative sample
Specificity	$\frac{TN}{TN+FP}$	Coverage of actualnegative sample
F1-Score	$\frac{2TP}{2TP+FP+FN}$	Hybrid metric useful for unbalanced classes

STEP6: K – FOLD CROSS VALIDATION

By training a model on a subset of the original dataset and then testing the model using a complementary subset of the dataset, this technique is called cross-validation.

IV. IMPLEMENTATION

Isolation Forest: With the aid of the outlier identification method known as Isolation Forest, anomalies rather than typical observations are found. It is constructed using an ensemble of binary (isolation) trees, much like Random Forest. It can be expanded to deal with huge, multidimensional datasets. The isolation method is used in this case to identify uncommon patterns since it focuses on something unique and out of the ordinary, hence the name anomaly. Hence, we finally give it the name Anomaly Detection algorithm. Anomaly Detection is now widely used in all sectors of the economy (banking, finance, healthcare, manufacturing, and networking). This operates similarly to the Decision Tree algorithm, which begins at the node root and proceeds to additional spaces. If, for example, we tried to use our eyes to find a mole in a similar data set, we would be unable to do it. Thus, we have employed such techniques to discover andenhance credit card fraud detection. This method aids in isolating anomalous data from the entire dataset when many ofthe data in a large dataset are the same and one of them differs from the others. The main advantage of this method is the potential for using sampling techniques on a dimension that is prohibited by methods based on profiles. While it follows a somewhat shorter path than profiling regular data, the Isolation Forest approach concentrates on anomalies data. It mostly aids in the development of quick algorithms that consume very little memory for detecting anomalies. The unusual portion can be located using an algorithm that is provided. Create a profile for the normal data, observe andimport the complete csv file, report everything that cannot be kept as normal, and use the procedure to determine the anomalies score.

s (a, b) = 2^−$E(h(a))/c(b)$ b=quantity of data points

In a binary search tree, c(b) is the average path length for unsuccessful searches. Every observation is given anomaly score,

and based on that score, the following choice can be made: A score of less than 0.5 indicates normal observations,

whereas a value of close to 1 indicates anomalies. If all scores are near to 0.5, then there do not appear to be any obvious anomalies over the entire sample.

LGBM: Model performance is enhanced while utilising less memory thanks to LightGBM, a gradient boosting framework based on decision trees. Gradient-based one side sampling and exclusive feature bundling (EFB) are two novel strategies used to solve the limitations of the histogram-based approach, which is mainly used in all GBDT (Gradient Boosting Decision Tree) frameworks. The two GOSS and EFB techniques that are described here constitute the foundation of the LightGBM Algorithm's attributes. They provide the model an edge over rival GBDT frameworksand allow it to perform well as a whole.

DATA SPLITTING: It is called data splitting when the data are split into two or more components. Each half is applied to testing or analysing data, while the second part is used for training this model in a common two part split. Data splitting is a necessity for data science, in particular when it comes to creating models based on the data. This approach makes itfeasible to ensure the precision of activities like machine learning and data model construction.

The training is followed by the use of the testing data set. To ensure that the final model performs as anticipated, the training and test sets of data are compared. Data is often divided into three or more sets for machine learning. The settingsfor the learning process are changed on the dev set, which is one more set with three sets. A predictive model's number of predictors or the size of the initial data pool may determine how the data should be divided; there is no defined way or measure for doing so. Organisations and data modellers may decide to classify data into the three groups listed below depending on data sampling methodologies:

1. Random sampling: is a common technique used during the data modelling process to prevent bias towards specific data features. In this technique, a random subset of the available data is selected to train the model. However, when the data is unevenly distributed, random splitting may cause problems by producing training and testing sets that do not accurately reflect the distribution of the original data

2. Stratified random sampling: This approach uses pre-set settings to randomly select data samples. Ensures that data is evenly distributed across test and training sets.

3. Non-random sampling: Data modelers often employ this strategy when they need up-to-date data as a test set.

V. RESULTS AND DISCUSSION

Table 2: The comparison between the proposed EL and the other researcher's work on the CCF dataset

Work	Approach	Base Learner	Accuracy	Performance Evaluation TPR	Performance Evaluation TNR
Our Approach	EnsembleLearning	LGBM + IF	0.906	0.818	0.995
	Bagging	RF	0.816	0.371	___
	Multiple Classifiers	MLP + RBF + NB	____	0.534	0.831

	precision	recall	f1-score	support
0	1.00	0.90	0.95	56864
1	0.01	0.91	0.03	98
accuracy			0.90	56962
macro avg	0.51	0.90	0.49	56962
weighted avg	1.00	0.90	0.94	56962

Accuracy: 0.8964221761876339

Fig2: Result of accuracy achieved with precision, recall &f1 score.

Fig3: Prediction window (front-end part)

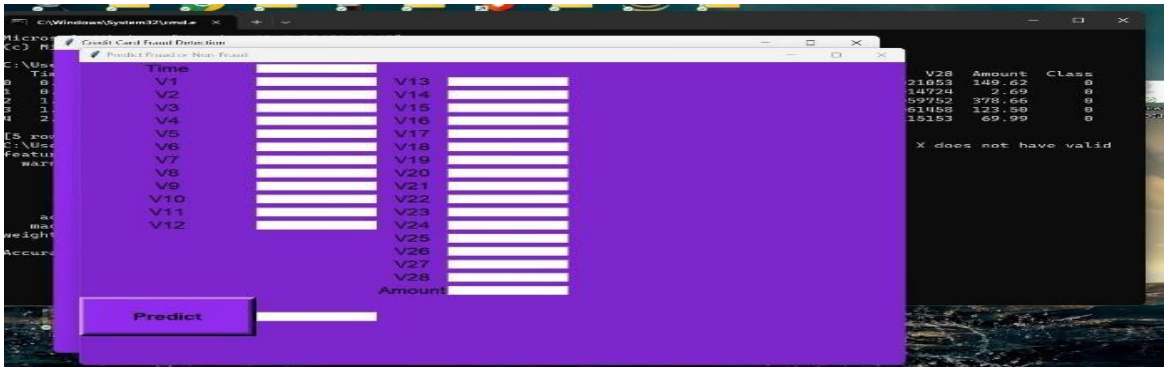

Fig4: Restriction values input window.

VI. CONCLUSION & FUTURE SCOPE:

Since there are wide variety of credit card transaction datasets which include bot attacks, malicious traffic, content moderation and loan risks issues, effective solution is needed to face such issues. Using LGBM and IF algorithms is the best combination technique to trace out the fraudulent activities and for handling such vast datasets. As the accuracy of these both algorithms is much better than the other existing machine learning techniques, it is efficient enough to trace out the fraudulent activities in credit card transactions of both existing and real-time datasets. Researchers who have used LGBM algorithms with other type of techniques concluded that in future work if this algorithm is used and developed with alternative combinations efficient accuracy can be achieved as it mainly focuses on the development of performance and scalability. Isolation forest among the random forest and other outlier methods, it is considered to be the best outlier method to detect the accurate anomalies. In this project we are mainly developing an effective Appeal for fraud detection using combination of two effective machine learning techniques for the detection of fraudulent activities in two vast datasets of credit card. We are working and experimenting on both existing and real time datasetsto extract the best accuracy from these algorithms which may become the best technique in future when the better accuracy is obtained with such combination.

There is still a lot of room for further research in this area, but recent applications of machine learning techniques in credit card fraud detection have yielded encouraging results. Many areas for improvement include:

Ensemble methods: Ensemble approaches, which combine different machine learning algorithms, can be used to improve the accuracy of fraud detection systems.

Deep learning: Deep learning techniques such as convolutional neural networks and recurrent neural networks can be used to identify complex patterns in credit card transactions.

Explainable AI: Explainable AI techniques help businesses understand and address potential problems by revealing the reasons why specific transactions are marked as fraudulent.

Big Data: With the volume of credit card transactions rising, there is a need to develop more efficient algorithms that can handle large volumes of data in real time.

Internet of Things (IoT): IoT devices can be used to gather and evaluate data in real-time, which can help detect fraudulent activities more accurately and quickly.

The future of credit card fraud detection using machine learning techniques looks promising, and further research in thisarea can help financial institutions prevent fraudulent activities and minimize financial losses.

REFERENCES

[1] Taha, Altyeb & Malebary, Sharaf. (2020). An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine. IEEE Access. 8. 25579-25587.
[2] Assaghir, Zainab & Taher, Yehia & Haque, Rafiqul &Hacid, Mohand-Said &Zeineddine, Hassan. (2019). An Experimental Study With Imbalanced Classification Approaches for Credit Card Fraud Detection. IEEE Access.
[3] L. Meneghetti, M. Terzi, S. Del Favero, G. A Susto, C. Cobelli, "DataDriven Anomaly Recognition for Unsupervised Model-Free Fault Detection in Artificial Pancreas", IEEE Transactions On Control Systems Technology, (2018) pp.1-15.
[4] F. Carcillo, Y.-A. Le Borgne and O. Caelen et al., "Combining unsupervised and supervised learning in credit cardfraud detection", Information Sciences, Elsevier (2019), pp. 1-15.
[5] Lakshmi S V S S` Selvani Deepthi Kavila, "Machine Learning for Credit Card Fraud Detection System". International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 24 (2018) PP.16819-16824.
[6] Franklin Ore-Areche, Kelyn Nataly Munoz-Aleja, Cledi Puma-Condori, "ANALYSIS AND DETECTIONOF FRAUD IN CREDIT CARD USING SILOF" July 2022 – Sep 2022, volume 10 Issue 3, ISSN: 2322-2394.
[7] SURAYA NURAIN KALID, KENG-HOONG NG, GEE-KOK TONG AND KOK-CHIN KHOR,"A Multiple Classifiers System For Anomaly Detection in Credit card Data With Unbalanced And overlapped classes". Feb 2020, IEEE Access,2020,2972009.
[8] KULDEEP RANDHAWA, CHU KIONG LOO, (SENIOR MEMBE, Feb2018.IEEEAccess,2018,2806420.
[9] Andrea Dal Pozzolo, Giacomo Boracchi, Olivier Caelen, Cesare Alippi and Gianluca Bontempi. "Credit Card Fraud Detection: A Realistic modelling and a novel learning strategy".2017 Sep 14, IEEE Access.20736643. [10] Kartik Madkaikar, Manthan Nagvekar, PreityParab, Riya Raikar, Supriya Patil."Credit Card Fraud Detection System", International Journal Of Recent Technology and Engineering(ISRTE) July 2021,volume 10 Issue 2 ,2277-3878.
[10] Vaishnavi Nath Dornadula, Geetha S. "Credit Card Fraud Detection Using Machine Learning Algorithms". ICRTAC2019.
[11] Emmanuel Ilberi, Yanxia Sun, and Zeghui Wang. "A machine learning based credit card fraud detection using theGA algorithm for feature selection". Journal Of Big Data (2022).S40537-022-00573-8.