JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 553-558
https://publishoa.com
ISSN: 1309-3452

Gradient Boosting Techniques for Credit Card Fraud Detection

Kasarapu Ramani[1] , Irala Suneetha[2] , Nainaru Pushpalatha[3] , P. Harish[4]

[1]Professor, Dept. of Information Technology
Sree Vidyanikethan Engineering College, Tirupati
ramanidileep@yahoo.com
[2]Professor, Dept. of Electronics and communication Engineering
Annamacharya Institute of Technology and Science, Tirupati
iralasuneetha.aits@gmail.com
[3]Professor, Dept. of Electronics and communication Engineering
Annamacharya Institute of Technology and Science, Tirupati
pushpalatha825@gmail.com
[4]Associate Professor Dept. of Electronics and communication Engineering
Annamacharya Institute of Technology and Science, Tirupati
harishpasupulati@gmail.com

Received 2022 April 02**; Revised** 2022 May 20**; Accepted** 2022 June 18**.**

Abstract

Now-a -Days e-commerce has enabled increased online transactions, and hence growing serious credit card frauds. This malicious activities are affecting millions of people' identity theft and loss of money. The fraudsters are continuously adopting new ways to perform illegal activities. This paper gives a detailed analogy of different supervised and unsupervised machine learning techniques for detecting fraudulent activities. The new schemes Cat Boost and Light Gradient Boosting Machine (LGBM) are proposed for fraud discovery. The performance of these methods is compared with approaches of Auto Encoder (AE), Logistic Regression and K-Means clustering and Neural Network (NN) and found that Cat Boost and LGBM are giving high accuracy in fraud detection.

Keywords: CatBoost, Light Gradient Boosting Machine, Auto Encoder, Logistic Regression, Neural Network (NN) and K-Means clustering

1. Introduction

Unauthorized usage of funds in any transaction through credit card represents a Credit card fraud [1]. With digital transactions these credit card frauds increased rapidly especially during pandemic. The credit card frauds are of two categories: the first category is card-not-present fraud, where customer's card number along with its expiration date including card verification code (CVC) are comprised without presenting the card physically to the vendor, especially it happens through online transactions and the second one is card-present fraud, where the card information is stolen during its physical transactions through point of sale system [2]. Now, with chip based cards the card-present fraud can be combated. As per the Federal Trade Commission's (FTC) Annual Report 4,59,297 credit card frauds were reported in 2020. The instances of identity theft become the most common credit card fraud and is increased by 44.6% in 2020 compared to 2019 statistics [3]. From this it is evident that robust fraud detection methods are needed to avoid monitory losses. The credit frauds can happen in various forms, such as: Non- Mail receipt related card Fraud, Account Take Over, Electronically or Manually prepared credit card Imprints, Counterfeit-Card Fraud, Synthetic Theft, ID document Forgery, Formjacking redirection, Intercepting from mailed , Application Fraud, Merchant Collusion, Fraudulent credit applications, Location Spoofing, Phone number spoofing, Copying a Buyers behaviour, Lost and

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 553-558
https://publishoa.com
ISSN: 1309-3452

Stolen card fraud, False Merchant Sites, and credit card theft [4]. The classification algorithms play a major role in predicting the target class.

1.1 Background of the proposed work::

Machine Learning: It is best suited for Fraud detection as it provides quick prediction results. With large datasets the performance of the Machine Learning algorithms improve and decision making becomes accurate as it can learn from the past and predict from the future transactions[5].

Classifiers: Fraud detection models can be built with both Unsupervised Machine Learning and Supervised Machine Learning methods. The supervised ML algorithms mainly focus on classifying the transactions as fraud or not, whereas the unsupervised algorithms identify type of anomalies. Similarly, neural networks are also applicable for fraud detection, but their performance will depend on training data, that is equal amount of normal and abnormal data points [6].

Classifier Models: In this paper Cat Boost and Light Gradient Boosting Machine are used for credit card fraud detection.

1.2 Gradient Boosting:

An ensemble learning represents a group of techniques, which unite the predictions obtained from multiple number of weak learning models to achieve the finest predictive performance value. The ensemble learning techniques are categories as:

Bagging: This technique results in multiple parallel models using randomly chosen subsets and then aggregates the prediction value from all these predictors deterministically.

Boosting: This technique is an adaptive technique applied iteratively in sequential manner and, where each predictor fixes its predecessor model's error value.

Stacking: This technique is a meta-learning one, which combines predictors from multiple machine learning techniques, such as bagging and boosting.

Gradient boosting algorithms can work as a Regressor or a Classifier. This technique focuses on training the model based upon reducing the differential loss function using gradient descent optimization. There will be equal distribution of weights to all the learning models. A gradient boosting represents a series of decision trees, forming stage wise additive model. it reduces bias error of the model.

1.3 CatBoost: This belongs to a gradient boosting algorithm applied on decision tree. It is used in many applications such as search, personal assistant, recommendation systems, self-driving cars, and weather prediction. This algorithm works without parameter tuning, it gives good results with default parameters. It is best suited for categorical data. No need to convert categorical data into numerical values, as it can be applied directly on non-numerical data. The gradient boosting improves accuracy and reduces overfitting problem [7].

Symmetric Trees: The CatBoost algorithm constructs number of symmetric trees also known as balanced trees. In each and every step, the split is done at every leaf is based on the same condition. The feature split is done in such way that it will lead to lowest loss and same criteria is applicable to all levels of the nodes. With such balanced tree organization, it is possible to reduce prediction time, enhancement of CPU's efficiency, and also it controls overfitting problem as this structure works as a regularization process [8].

Ordered Boosting: Due to prediction shift the regular boosting algorithms may lead to overfitting issue on small and noisy datasets. Sometimes the gradient estimate calculation may consider the same data instances based on which the model was built, therefore the model can not experience the training based on remaining data. But, the CatBoost

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 553-558
https://publishoa.com
ISSN: 1309-3452

algorithm applies permutation based approach on every subset of data for training and on the other hand computes the residuals on different data subsets, which prevents the target leakage and overfitting problems [8].

Native feature support: In other classification models the non-numerical features need to be converted into numerical before applying classification, whereas the CatBoost algorithm supports numeric and categorical features therefore it avoids conversion time and effort required for such preprocessing [8].

1.4 Light Gradient Boosting Machine(LGBM): It is also a kind of gradient boosting method based on decision tree and is used to increase the efficiency of a given classification model and works with reduced memory usage. It is used in many Machine Learning application tasks such as ranking, classification, etc. It is based on two innovative techniques. The first one is called as Gradient based One Side Sampling (GOSS) and Second one is known as Exclusive Feature Bundling (EFB), developed to address the drawbacks of the histogram approach used in GBDT Gradient Boosting Decision Tree (GDBT) models. The characteristics of LGBM model are achieved by methodologies of EFB and GOSS [9].

2. Objectives

The objectives of this paper are as follows:

i). To implement Cat Boost and LGBM algorithm based fraud detection schemes for fraud detection in credit card data.

ii). To compare the performance of proposed algorithms with the existing schemes like Auto Encoder, Neural Networks, Logistic Regression and K-means clustering on a credit card dataset.

This paper compares various supervised, unsupervised methods and neural networks to explore most accurate prediction algorithm to determine fraudulent credit card fraud transaction.

3. Methods

In the existing system, classification models are built based on Auto Encoder (AE), Synthetic Minority Over Sampling (SMOTE) and Logistic Regression to estimate fraudulent and non-fraudulent transactions. As these techniques give low precision, recall scores and also lack the robustness because of higher computational time and not suitable for imbalanced data. Thus Credit Card based fraud detection scheme need some powerful Machine Learning techniques to prevent fraudulent transactions [10].

Proposed System:

The block diagram of Credit card Fraud detection System using Gradient Boosting Techniques is as shown in Fig.1

Implementation of CatBoost algorithm is as follows: .

Algorithm1:

Input: The UCI Machine Learning Repository of Credit card dataset with 30,000 instances.

Step 1: Remove the rows containing NaN values for the given target column.

Step 2: Convert non-numerical columns, if any into the category data type. Get column indices for this categorical data.

Step 3: *Training*: Wrap the training as well as testing datasets into a selected Catboost pool constructor. The Sci-Kit learns Grid Search CV, which is an in-built grid search method is used. Prepare a dictionary with hyperparameters which are used to adjust various performance parameters such as tree depth, learning rate, L2 leaf regularisation and number of iterations.

Step 4: The grid search divides the training data into an 80:20 split for training and testing respectively with a three fold cross validation mechanism.

Step 5: *Evaluation*: The model is tested with dynamic user transaction to predict whether it is fraudulent or legitimate transaction.

Step 6: The model's performance is evaluated using precision, recall and accuracy parameters.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 553-558
https://publishoa.com
ISSN: 1309-3452

Implementation of LightGBM algorithm is as follows:

Algorithm2:

Input: A Credit card dataset with 30,000 instances is taken from UCI ML Repository. The parameters: No. of iterations, loss function , Weak learner and sampling ratio are chosen.

The LightGBM model is optimized with the following steps:

Step 1: The number of estimators or boosted trees will influence the performance of the LGBM. Models with varying numbers of trees are constructed and evaluated to decide the optimal number of $n_{opt.}$

Step 2: In low and medium datasets the occurrence of overfitting is the most common problem. Therefore, the maximum depth D_{max} of trees should be limited.

Step 3: Set the number of tree leaves, $N_{leaves}=2^{Dmax}$ to get the same number of leaves for depth-wise trees. Appropriate value of this parameter is used to moderate the complexity level of the LGBM tree. If depth is unconstrained , it can induce overfitting, therefore the N_{leaves} should be smaller than2^{Dmax}.

Step 4: Build multiple number of LGBM models with varying D_{max}, and N_{leaves} parameters using 10-fold cross validation.

Step 5: Validate the model using dynamic Credit Transaction input and predict whether it is Fraudulent or legitimate transaction.

Step 6: The performance of this model is evaluated using precision, recall and accuracy parameters.

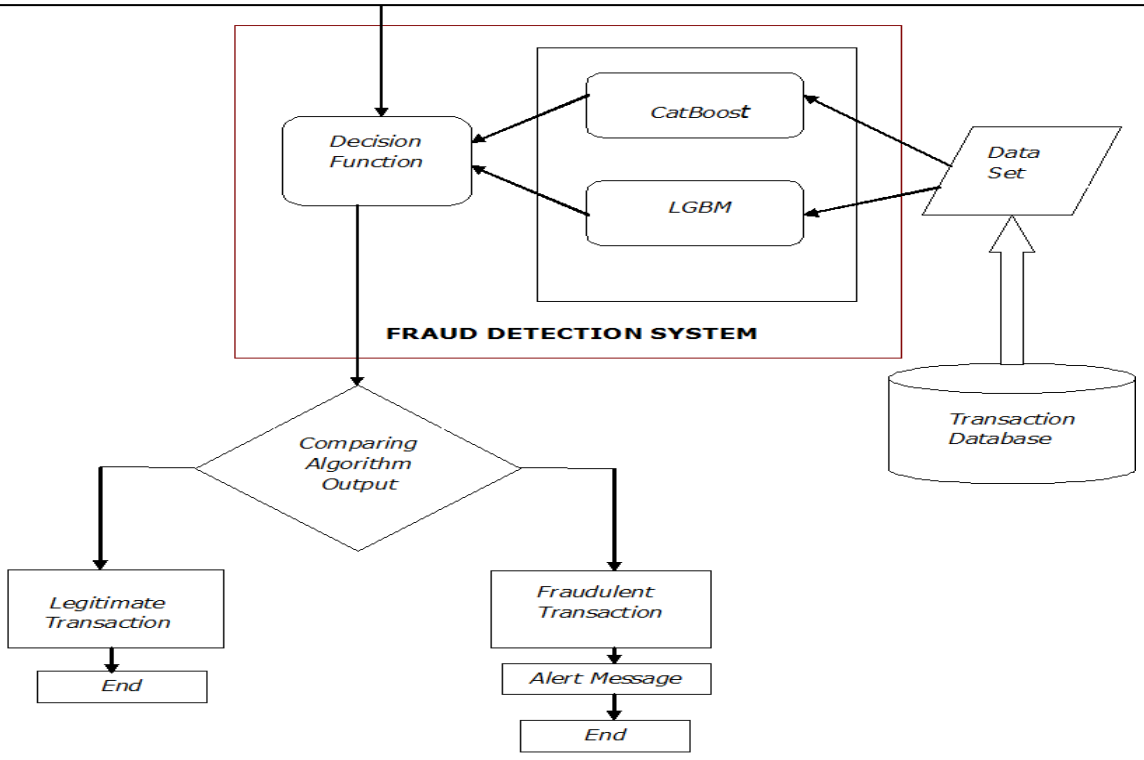

Fig. 1 Credit card Fraud detection System using Gradient Boosting Techniques

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 553-558
https://publishoa.com
ISSN: 1309-3452

4. Results

To verify the effectiveness of CatBoost and LGBM models for predicting the fraudulent credit card transactions they are compared with supervised, unsupervised techniques such as Logistic Regression, Auto Encoder and K-Means Clustering Models and neural networks respectively. The Accuracy, Precision and Recall obtained by these methods are shown in Fig. 2. The CatBoost Model is efficient in terms of accuracy and LGBM is efficient when large datasets are is used [12].

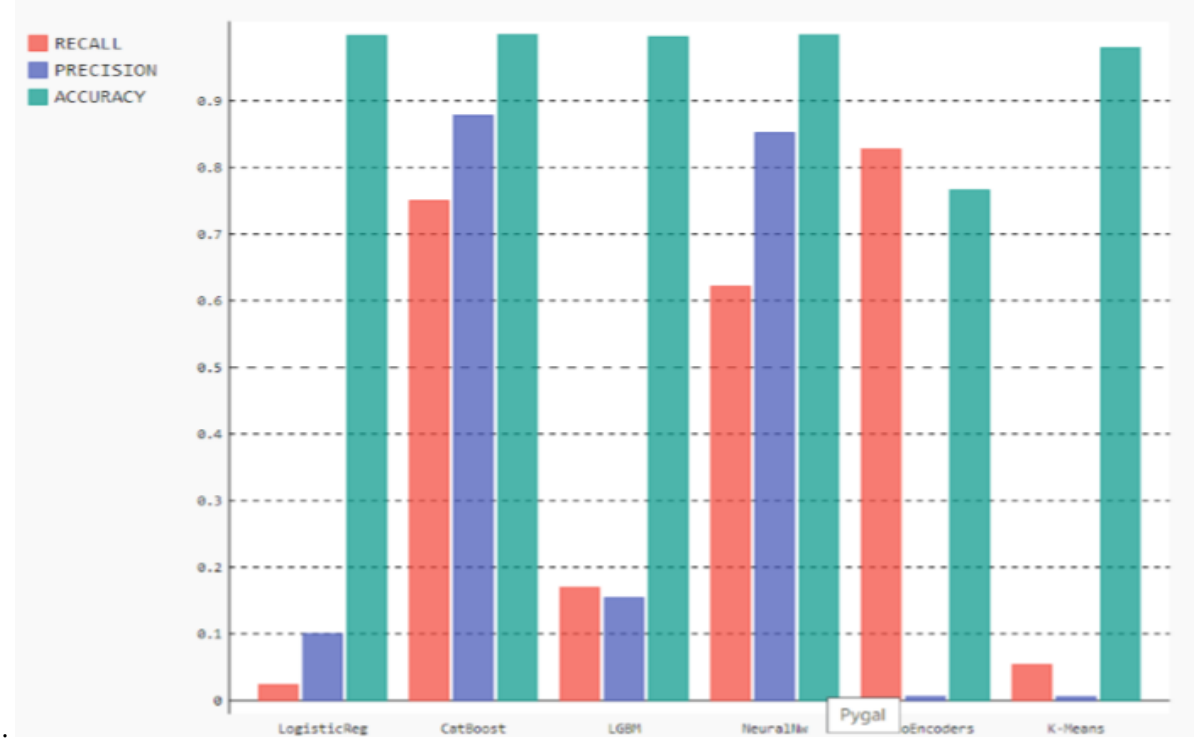

Fig. 2 Comparison of Machine learning Techniques applied on credit card data set

5. Discussion

The effective Machine Learning models - CatBoost and LGBM are used to determine whether the given credit card transaction is fraudulent or legitimate. LGBM outperforms Logistic Regression, Neural Networks, Auto Encoders, K-Means Clustering, Cat Boost, and LGBM with a 97% accuracy score. The accuracy of neural network-based technique has a 96% accuracy, while Logistic Regression, Auto Encoder, K-Means clustering, Cat Boost, and LGBM have accuracies of 77%, 96%, 93%, 98%, and 99% respectively. In future a hybrid model can be built to achieve the better performance.

Refrences

[1]. Jain R., Gour B., Dubey S., A hybrid approach for credit card fraud detection using rough set and decision tree technique, International Journal of Computer Applications, Vol.139, Issue.10, 2016.
[2]. Rishi Banerjee et.al, Comparative Analysis of Machine Learning Algorithms through Credit Card Fraud Detection, IEEE MIT Undergraduate Research Technology Conference (URTC), 2018.
[3]. Hala Z Alenzi, Nojood O Aljehane, Fraud Detection in Credit Cards using Logistic Regression, International Journal of Advanced Computer Science and Applications, Vol. 11, No. 12, 2020..
[4]. Aashi Maharjan, Partha Chuda, Comparative Analysis of Algorithms for Credit Card Fraud Detection, KEC Conference, 2019.

[5]. V Jyothsna, K Munivara Prasad, K Rajiv, G Ramesh Chandra, Flow based anomaly intrusion detection system using ensemble classifier with Feature Impact Scale, Cluster Computing, Vol. 24, Issue.3, 2021,pp. 2461-2478.
[6]. K.K. Baseer, V. Neerugatti, S. Tatekalva, A.A. Babu, Analysing various regression models for data processing, nternational Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-8 June, 2019.
[7]. https://catboost.ai/
[8]. https://neptune.ai/blog/when-to-choose-catboost-over-xgboost-or-lightgbm
[9]. Guolin Ke. et.al, LightGBM: A Highly Efficient Gradient Boosting Decision Tree, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.
[10]. Huang Tingfei, Cheng Guangquan, and Huang Kuihua "Using Variational Auto Encoding in Credit Card Fraud Detection," doi:10.1109/ACCESS.2020.3015600
[11]. https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-it-how-to-fine-tune-the-parameters-60347819b7fc.
[12]. https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/iet-its.2020.0396