LightGBM Model for Credit Card Fraud Discovery

Appala Srinuvasu Muttipati, Sangeeta Viswanadham, Radha Dharavathu, and Jayalakshmi Nema

Abstract The exponential growth of e-commerce and online-based payment options has created an empirical universe of financial fraud, with credit card fraud being the most prevalent. For several years, many researchers have developed a variety of data mining-based methods to address this issue. To detect credit card fraud, there has recently been a lot of interest in using machine learning algorithms instead of data mining techniques. In the digital space of financial transactions, on-going work is being conducted to put in a conceptual difference between fraud identification and predicting likely fraudulent opportunities. This paper extends the fraud detection technique and proposes a LightGBM-based detection algorithm. The dataset is a credit card dataset for credit card transactions in Europe. Our approach outperformed other traditional approaches such as random forest, AdaBoost, and XGBoost in this experiment. Furthermore, it demonstrates the value of feature engineering in terms of feature selection and performance tuning.

Keywords Credit card fraud · Machine learning · Feature selection · LightGBM

A. S. Muttipati (✉)
Department of Computer Science and Applications, KL Deemed to be University, Vaddeswaram, Guntur, Andhra Pradesh, India
e-mail: srinuvasu.mutti@gmail.com

S. Viswanadham
Department of Computer Science and Engineering, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh, India

R. Dharavathu · J. Nema
Department of Computer Science and Engineering, Raghu Engineering College, Visakhapatnam, Andhra Pradesh, India

J. Nema
Department of Computer Science and Engineering, Raghu Institute of Technology, Visakhapatnam, Andhra Pradesh, India

V. V. S. S. Chakravarthy et al. (eds.), *Advances in Micro-Electronics, Embedded Systems and IoT*, Lecture Notes in Electrical Engineering 838,
https://doi.org/10.1007/978-981-16-8550-7_6

1 Introduction

Nowadays, society is growing globally in all areas, and one of the areas is e-commerce. Due to the increase in e-commerce possibilities in making online payments and as they are easier to use, e-commerce business gained user confidence. This confidence leads to increase in number of users. The online transactions have given a drastic rise in revenue generation. Increase in the user's revenue generation has paved a path to be vulnerable to fraudulent behavior. Credit card fraud is one of the acclaimed problems in the present world. In 2016, there happened to be a benchmark increase in credit card fraud up to 92% compared to the 2012 count. The credit card may happen in one of the following ways: (1) application fraud, (2) stolen or lost cards, (3) account taken over, (4) card counterfeit. The stolen or lost card and account takeover are major problems and are named as card not present (CNP) fraud. In CNP, the cardholder is cheated by stealing the card's sensitive information like CVV, card No and using it remotely. It leads to the transfer of a large amount or the purchase of costly items before the cardholder discovers. As the availability of Internet is increasing in the world, people are showing interest in purchasing things online rather than offline. Due to this, the growth of e-commerce sites is increasing, and thereby the chance of credit card fraud. To solve credit card fraud, we have to find out algorithms that may either avoid or reduce credit card fraud.

1.1 Related Work

Reference [1] have suggested some ensemble models for detecting credit card fraud. Models like random forest, logistic regression, CatBoost have shown better results. The results when compared, random forest and CatBoost have outperformed and could create ROC curve and area under curve. References [2–4] have done performance comparison of naive Bayes, K-nearest neighbor, and logistic regression models in the binary classification of imbalanced credit card fraud. KNN has outperformed the competition based on all of the evaluation metrics. To identify fraudulent transactions in European credit card data, traditional algorithms such as decision tree, support vector machine (SVM) [5], least square regression, naive Bayes classifier, K-nearest neighbors (KNN), and gradient boosting (GB) have proven useful. KNN and outlier detection approaches were suggested [6] and are effective in fraud detection. They can help reduce false alarm rates and improve fraud detection rates. In an experiment, the author has tested and compared the KNN algorithm with other classical algorithms, and KNN performed well [7]. Random forest uses random tree-based and CART-based methods to train the behavioral features of standard and non-standard transactions [8–10]. Despite the fact that random forest obtained results on a small dataset, it faces the issue of imbalanced data. The focus of future work will be on resolving datasets that are imbalanced.

1.2 Our Contribution

This paper suggested a LightGBM-based credit card fraud detection algorithm. The dataset is organized based on the sequential transactions executed using credit cards by European credit cardholders. The dataset encloses a total of 284,315 transactions and is a complex dataset containing 30 variables like the difference between transaction times, transaction amount. In our work, data preprocessing to eradicate some irregular data is of the first importance. It is of great significance since some irregular data can lead to worst performance. LightGBM is executed as our twofold order. LightGBM is one of the tree-boosting framework models utilized by many data scientists to chronicle cutting-edge results to solve many machine learning issues, likewise executed other traditional models in this work like random forest, AdaBoost, and XGBoost. Experiment shows LightGBM performs better compared to other models.

2 Proposed Methodology

The proposed approach uses a three-step procedure which is stated below:

Step 1: Attaining the dataset from repository. The dataset is organized based on the sequential transactions executed using credit card by European credit cardholder. The dataset encloses a total of 284,315 transactions and is a complex dataset which containing 30 variables like difference between transaction times, transaction amount. It also contains 28 other attributes which are kept anonymous in order to protect the identity of the customer. It also contains a column with binary values ‘0’ directs non-fraudulent transaction and ‘1’ directs fraudulent transactions. One thing we can observe in the dataset is it is highly skewed. It is because the dataset is sway toward the genuine class. We can observe this as out of the 284315 transactions, only 492 are not genuine. So, only 0.172% fraudulent transactions are present when compared to whole number of transactions.

Step 2: Dataset splitting. The dataset is divided into two sets, (1) training and test set and (2) training and validation set using cross-validation. Cross-validation is a technique for evaluating a machine learning model and testing its performance. It helps in comparing and selecting an appropriate model for the precise extrapolative modeling problem. The dataset splitting can be carried out by the following steps:

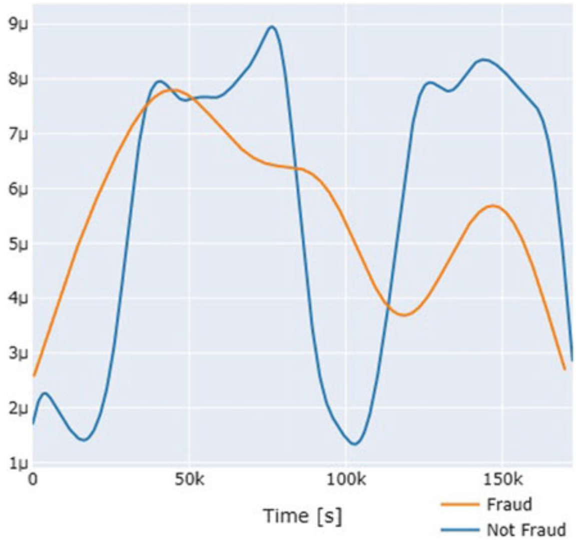
1. To split the dataset into two segments: one segment for training set and other segment for testing
2. To train the model on the training set
3. To validate the model on the test set
4. Repeat Steps 1–3 until k-fold has assisted as the test set.

Step 3: The Creation of Machine Learning Models. Machine learning is categorized into four: supervised, unsupervised, semi-supervised, and reinforcement learning. The deliberated machine learning algorithms are ensemble models and gradient boosting algorithms.

3 LightGBM-Based Fraud Detection Model

This section will momentarily present our model and offers the parameters of our model. Compared with XGBoost and other traditional models, LightGBM embraces numerous enhancements like gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB). Utilizing GOSS keeps all the instances with large gradients and performs arbitrary sampling on the occurrence with small gradients. In order to compensate the influence to the data distribution, when computing the information

Fig. 1 Credit card transactions time density plot



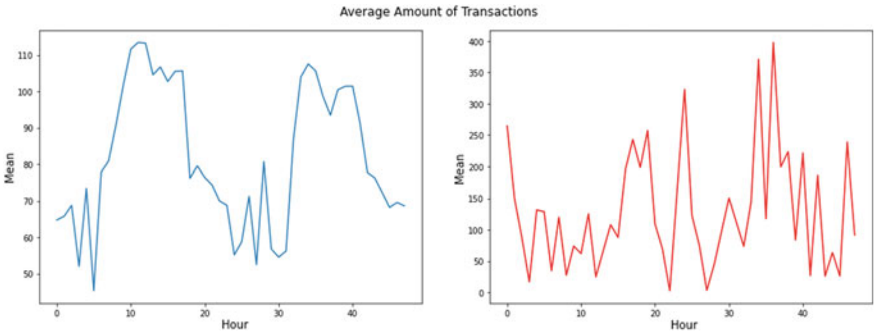


Fig. 2 Average amount of transaction over hour

Table 1 LightGBM parameters

Parameter	Parameter decryption	Values
n_estimater	Number of estimaters	2000
learning_rate	Rate of learning	0.05
Num_leaves	Number of leaves < 2	7
Max_depth	Tree maximum depth	4
Min_child_samples	Minimum number of data needs in a chils	100
Max_bin	Number of bucketed bin for feature values	100
Subsample	Subsample ratio of the training instance	0.9
Boosting_type	Type of boosting	gbdt

gain, GOSS introduces a constant multiplier for the data instances with small gradients [11]. With EFB, the model's special features are to reduce the number of features and subsequently improve forecast speed. Through these optimizations, LightGBM beats the large portion of other machine learning algorithms in speed and accuracy. In view of the limits of LightGBM, we applied this model to our exploratory work (Figs. 1 and 2).

To accomplish a superior value of our model, we utilized framework search to tune the parameters of our models. Practically speaking, it is helpful in improving the score around 1 or 2%. We implemented it to the some key parameters like learning rate, completed as of not long ago. Important features for implementation are further selected using feature selection process. To give better detail, Table 1 runs down the parameters of our model, and different parameters which do not show in this table are default parameters.

4 Experimental Analysis

In this session, the experiment was performed on Windows 7 operating system and the open-source software environment. The Jupyter notebook environment is used to develop and run our model. Various libraries are utilized such as NumPy, Pandas, Matplotlib, Seaborn, Sklearn, and imblearm.

Here, AUC-ROC score proves to be the better model. This score value is actually is the area under ROC curve, which is also known as receiver operating characteristic curve value. The curve is plotted by using true positive rate (TPR) against the false positive rate (FPR) at different threshold settings. The formula of TPR and FPR are defined as follows:

Table 2 Performance of various models

Model	AUC value	Accuracy values
Random forest	0.96	0.99
AdaBoost	0.87	0.99
XGBoost	0.90	0.99
LightGBM	0.94	0.99

Table 3 Fivefold cross-validation of LightGBM model

Five folds	Training_AUC_value	Valid_AUC_value
Fold 1	0.967	0.994
Fold 2	0.977	0.962
Fold 3	0.981	0.948
Fold 4	0.970	0.987
Fold 5	0.972	0.993

$$\text{True Positive Rate} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$\text{False Positive Rate} = \frac{\text{False Positive}}{\text{False Positive} + \text{True Negative}}$$

In addition to AUC-ROC value, we also provide the accuracy value of different models. In Table 2, it compared our model with other three models.

Form Table 2, it is easy to find out that our LightGBM-based model outperforms the other models on both AUC-ROC value (Table 3).

Tree-based algorithms like LightGBM or XGBoost are not difficult to yield the feature significance of each feature. In Figs. 3 and 4, it shows the significant features in diminishing request. The feature significance charts give us direction on the most proficient method to implement. We can pick portions of significant features as indicated by the diagram.

5 Conclusion

This paper presents a LightGBM model to recognize fraudulent transactions. Here, we utilized both train-validation set split and cross-validation to calculate the model efficiency to forecast 'class' value (i.e., discovering if a transaction was fraudulent or not). In this preliminary work, comparison of various machine learning models based on metrics is presented along with identification of significant features.

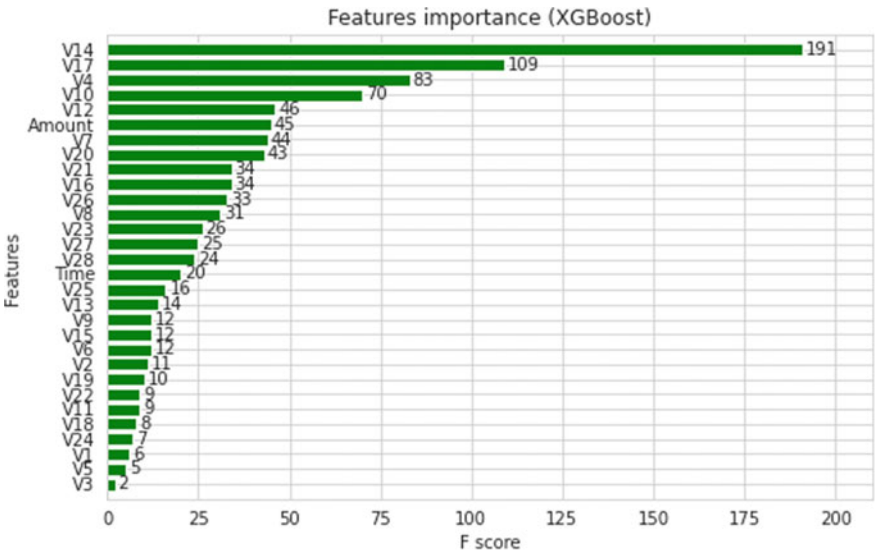


Fig. 3 Top important features of XGBoost

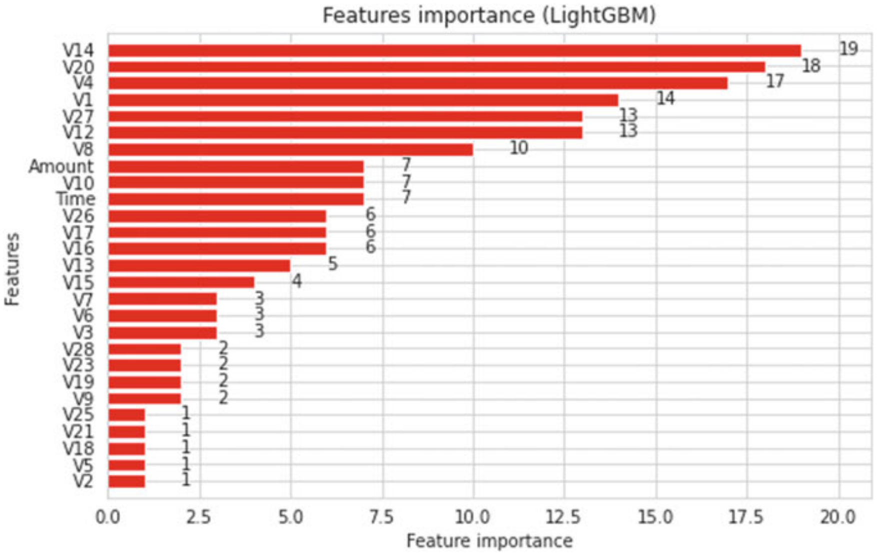


Fig. 4 Top important features of LightGBM

References

1. Awoyemi JO, Adetunmbi AO, Oluwadare SA (2017) Credit card fraud detection using machine learning techniques: a comparative analysis. In: 2017 international conference on computing networking and informatics (ICCNI)
2. Dhankhad S, Mohammed E, Far B (2018) Supervised machine learning algorithms for credit card fraudulent transaction detection: a comparative study. In: 2018 IEEE international conference on information reuse and integration (IRI)
3. Dornadula VN, Geetha S (2019) Credit card fraud detection using machine learning algorithms. Procedia Comput Sci 165
4. Godi B, Viswanadham S, Muttipati AS, Prakash Samantray O, Gadiraju SR (2020) E-healthcare monitoring system using IoT with machine learning approaches. In: 2020 international conference on computer science, engineering and applications (ICCSEA)
5. Hema G, Muttipati AS (2021) Machine learning methods for discovering credit card fraud. Int Res J Comput Sci 8(1):1–6
6. Kaithekuzhical LK, Jeet Ch (2019) Detection and prediction of credit card fraud transactions using machine learning. Int J Eng Sci Res Technol 8(3):199–208
7. Malini N, Pushpa M (2017) Analysis on credit card fraud identification techniques based on KNN and outlier detection. In: 2017 third international conference on advances in electrical, electronics, information, communication and bio-informatics (AEEICB)
8. Sailusha R, Gnaneswar V, Ramesh R, Rao GR (2020) Credit card fraud detection using machine learning. In: 2020 4th international conference on intelligent computing and control systems (ICICCS)
9. Varmedja D, Karanovic M, Sladojevic S, Arsenovic M, Anderla A (2019) Credit card fraud detection—machine learning methods. In: 2019 18th international symposium INFOTEH-JAHORINA (INFOTEH)
10. Muttipati AS, Sangeeta V, Radhika S, Brahmajirao KN (2021) Recognizing credit card fraud using machine learning methods. Turk J Comput Math Educ 12(12):3271–3278
11. Ge D, Gu J, Chang S, Cai J (2020) Credit card fraud detection using Lightgbm model. In: 2020 international conference on E-commerce and internet technology (ECIT)