Customer Transaction Fraud Detection Using Random Forest

Du Shaohui[1],
Saint-Petersburg State University,
Saint-Petersburg State, Russia,
1641449599@qq.com,

GuanWen Qiu[1],
University of California, Irvine,
California, United State,
huafengmai@email.arizona.edu,

Huafeng Mai[2],
University of arizona,
Arizona, United State,
huafengmai@email.arizona.edu,

Hongjun Yu[4],
Beijing Foreign Studies University,
Beijing, China,
yuhongjun_apply@163.com.

Cai Heng*
London School of Economics and Political Science
London, United Kindom
H.Cai4@lse.ac.uk

Abstract— In the evolution of the electronic money system, frequent transaction fraud has been a shadow behind the prosperity. It not only endangers the property security of users, but also hinders the development of digital finance in the world.With the development of data mining and machine learning, some mature technologies are gradually applied to the detection of transaction fraud. This paper proposes a transaction fraud detection model based on random forest. The experimental results of IEEE CIS fraud dataset show that the method of this model is better than the benchmark model, such as logistic regression, support vector machine.Finally, the accuracy of our model reached 97.4%，and the AUC ROC score was 92.7%.

Index Terms—Fraud detection, Data mining, Online Transaction

I. Introduction

In recent years, the number of e-commerce users is increasing day by day, and the scale of online transactions is also expanding. Fraud criminals often use various channels to steal card information and transfer a large amount of money in the shortest time, causing a lot of property losses to users and banks.Fraudulent transactions are usually small probability events hidden in a large amount of data, and the transaction form is extremely flexible. So tMachine learning and data mining are useful to build detection system of fraudulent transactions. The core detection algorithms of the system are mainly based on classification [1-5]. In order to face a large amount of data, data mining related processing methods are introduced into transaction fraud task [6-9]. In this paper, we establish a fraud detection system based on the classification model of random forest and the data processing related to feature engineering.

A. Related Work

Data mining is a process of exploring information hidden in a large number of data through algorithms. Through the cleaning, correction, extraction, selection and summary of a large number of data features, the hidden knowledge behind the data is obtained. For our problem, it is to extract the difference information between the behavior patterns of real users and fraud behavior patterns in the data, so as to help the subsequent classification model better achieve the purpose of detecting fraudulent transactions. In [3], Fang et al. Proposed a framework for fraud detection based on CNN to capture the inherent patterns of fraud learned from the marked data. Wang [5] proposed a data mining algorithm based on UCI public dataset. Bhusari and Patil proposed a hidden Markov model [10] to help obtain high fraud coverage and low false alarm rate. The research of work [11] mainly focuses on the most important feature engineering part of data mining. Researchers extend the transaction aggregation strategy, and propose to create a new set of features by using von Mises distribution on the basis of analyzing the periodic behavior of transaction time.

However, these methods have some defects, such as insufficient data or limited feature processing. In reference [5], due to the limited amount of data, it is difficult to establish a robust and accurate fraud detection system for practical scenarios. In reference [12], it only considers financial data, which may not be enough for a good data mining algorithm.

B. Our Contribution

Referring to the fraud transaction model based on xgboost [13], this paper proposes a fraud detection algorithm suitable for IEEE-CIS fraud data set based on random forest [13]. The data set contains more than 1 million samples, each sample contains more than 400 characteristic variables, including financial characteristics and non-financial characteristics. Sufficient amount of data and data type will provide a reliable basis for the subsequent classification model. Of course, the processing process is more complex, so we need to clean the data carefully and select the most valuable features carefully. In our work, we first carry out the truth of the data to eliminate some outliers and excessive missing data. In the further feature engineering cycle, the data will be transformed and the statistical data such as maximum, mean and standard deviation will be extracted. Then, Recursive feature elimination (RFECV) is used to eliminate some unimportant features.

Finally, we implement a binary classifier based on random forest according to the data and features. Random forest is a classifier with multiple decision trees. It has flexible model and fast training. It can solve the classification error caused by the extremely unbalanced data of fraud transaction detection. In order to show its superiority, we compare it with support vector machine and logistic regression. The experimental results show that the random forest model has achieved good results in accuracy and ROC AUC score.

The rest of this paper is arranged as follows. The second section introduces the characteristic engineering of financial data and non-financial data. The third section introduces the model based on random forest. In the fourth part, the performance of this algorithm is compared with other classical machine learning models by accuracy and AUC ROC score. Finally, the fifth part summarizes this paper.

II. Feature Engineering

The IEEE CIS fraud dataset is divided into two groups of tables: transaction table and identity table. These two kinds of tables are connected by key transaction ID, but not all transactions have identity information. The data has been labeled as two categories, i.e. isfraud = 0 or 1, but other characteristic information is still messy. Therefore, we first conduct data mining on them, and then combine them to form the final training data. In order to better handle data, we can study the two tables separately.

A. Transaction table

The transaction table has 394 characteristic variables, including 22 classification features and 372 numerical features. Most digital features are anonymous with fixed prefixes. To give a specific and clear description, we summarize these variables in Table 1. TransactionDT refers to the transaction date and time, which can be parsed into precise time information, such as year, month, day, week, etc. TransactionAmt refers to the amount of transaction payment in U.S. dollars. A small part of the amount with irregular decimal, may represent the transaction for remittance calculation.

B. Identification table

The identification table contains 41 features, including identity information, network connection information associated with transactions (IP, ISP, agent, etc.) and behavior information. The field names are masked and no paired dictionaries will be provided to protect privacy and sign contracts. When the two tables are then processed separately, they are joined on the transactionid key to generate a new table.It also records behavioral fingerprints, such as account login time & login failure time, account duration staying on the page, and so on.

TABLE 1 Transaction Table

Name	Description	Type
Transaction ID	ID of transaction	ID
isFraud	binary target	categorical
Transaction DT	transaction date	time
TransactionAmt	transaction amount	numerical
card1-card6	card	categorical
addr1-addr2	address	categorical
M1-M9	anonymous features	categorical
P_email domain	purchaser email domain	categorical
R_email domain	receiver email domain	categorical
dist1-dist2	country distance	numerical
C1-C14	anonymous features	numerical
D1-D15	anonymous features	numerical
V1-V339	anonymous features	numerical

TABLE 2 Identification Table

Name	Description	Type
TransactionID	ID of transaction	ID
DeviceType	device type	categorical
DeviceInfo	Device Information	categorical
id01-id11	Identification data	numerical
id12-id38	Identification data	categorical

Our data mining methods mainly include data cleaning, missing value filling, data transformation and feature extraction. In the data cleansing section, we deleted columns with a large percentage of Nan values (missing values). For example, if more than 90% of the value in a feature column is Nan, the column is deleted. For data that is not very defective, we'll fill it with - 1000 (a specific value that doesn't appear in the data). For the time code, we will also convert it into more accurate time information for better use. In addition, we also generate many descriptive statistical features, such as the mean and extreme value of numerical characteristics such as transaction amount, billing address and mailing address.Finally, a large part of the digital features have correlation. It can greatly improve the efficiency of data fitting and improve the performance of data classification. Therefore, in our work, recursive feature elimination with cross validation (RFECV) is used to eliminate each feature iteratively.

III. Random Forest Fraud detection model

In this part, we will introduce the Random Forest fraud detection model.

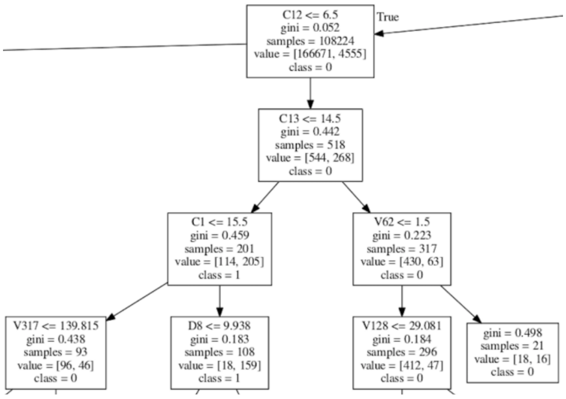

Figure 1. Part of a decision tree in a random forest

The random forest classifier is composed of a group of decision trees. Each tree is generated by independent sampling random vectors, and each tree votes to find the most popular category to classify the input. Random forest has both sample randomness and characteristic randomness, and its generalization performance is superior. At the same time, random forest has good processing ability for high-dimensional data sets, which is very suitable for IEEE CIS data sets. It can process a large number of inputs and determine the most important characteristics. Therefore, further feature mining is carried out on the data extracted by RFECV.

IV. Experiments

.In order to show the superiority of the model, we compared the accuracy and AUC ROC score with other models. AUC ROC score is actually the area under the receiver operating characteristic curve, which is created by drawing the relationship between true positive rate (TPR) and false positive rate (FPR) under different threshold settings. The formulas for TPR and FPR are defined as follows:

$$TPR = \frac{TP}{TP + FN} \quad (1)$$

$$FPR = \frac{FP}{FP + TN} \quad (2)$$

TP was the true positive prediction, FN was the false negative prediction, FP was the false positive prediction, and TN was the true negative prediction. The configuration matrix corresponding to the model can also be obtained (Fig. 2).As can be seen from table 3, the random forest model is superior to the other two models in terms of AUC ROC score and accuracy.

TABLE 3 Performance of different models

Models	Auc Roc Score	Accuracy
Logistic Regression	0.853	0.925
SVM	0.918	0.959
Random Forest	**0.927**	**0.974**

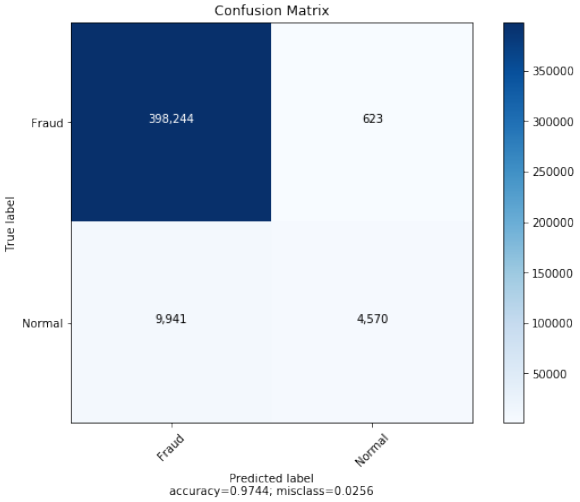

Figure 2. The confusion matrix of Random Forest Model

V. Conclusions

In this paper, we propose a method to detect fraud based on ieee-cis. The second section introduces the processing of data sets, including data cleaning, feature selection and feature engineering. The third section introduces the model based on random forest. In the fourth part, the performance of the algorithm is compared with other two classical machine learning models by accuracy and AUC ROC score.

ACKNOWLEDGEMENT

The author Du Shaohui and co-author GuanWen Qiu thanks Huafeng Mai for help in experiments, Hongjun Yu for idea contribution on feature engineering and Cai Heng for meaningful discussions.

REFERENCES

[1] Duan, L., Xu, L., Liu, Y., & Lee, J. (2009). Clusterbased outlier detection. Annals of Operations Research, 168(1), 151-168.

[2] Minastireanu, E. A., & Mesnita, G. (2019). Light gbm machine learning algorithm to online click fraud detection. J. Inform. Assur. Cybersecur, 2019.

[3] Fang, Y., Zhang, Y., & Huang, C. Credit Card Fraud Detection Based on Machine Learning.

[4] Maes, S., Tuyls, K., Vanschoenwinkel, B., & Manderick, B. (2002, January). Credit card fraud detection using Bayesian and neural networks. In Proceedings of the 1st international naiso congress on neuro fuzzy technologies (pp. 261-270).

[5] Wang, M., Yu, J., & Ji, Z. (2018). Credit Fraud Risk Detection Based on XGBoost-LR Hybrid Model.

[6] Dhingra, S. (2019). Comparative Analysis of algorithms for Credit Card Fraud Detection using Data Mining: A Review. Journal of Advanced Database Management & Systems, 6(2), 12-17.

[7] Minastireanu, E. A., & Mesnita, G. (2019). Light gbm machine learning algorithm to online click fraud detection. J. Inform. Assur. Cybersecur, 2019.

[8] Bhusari, V., & Patil, S. (2016). Study of hidden markov model in credit card fraudulent detection. In 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave) (pp. 1-4). IEEE.

[9] Bahnsen, A. C., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. Expert Systems with Applications, 51, 134-142.

[10] Carneiro, N., Figueira, G., & Costa, M. (2017). A data mining based system for credit-card fraud detection in e-tail. Decision Support Systems, 95, 91-101.

[11] Zhang, Y. , Tong, J. , Wang, Z. , & Gao, F. . (2020). Customer Transaction Fraud Detection Using Xgboost Model. 2020 International Conference on Computer Engineering and Application (ICCEA).