Customer Transaction Fraud Detection Using Xgboost Model

Yixuan Zhang
Beijing Jiaotong University
Beijing, China
yixuanzhang025@gmail.com

Jialiang Tong
Zhejiang University
Hangzhou, China
tongjialiang@zju.edu.cn

Ziyi Wang
Beijing University of Technology
Beijing, China
wangziyi901@126.com

Fengqiang Gao
Beijing Jiaotong University
Beijing, China
Fqgao_wy@163.com.

Abstract—**Customer transaction fraud detection is an imp-ortant application for both the public and banks and it is becoming a heated topic in research and industries. Many data mining techniques have been utilized in financial sys-tem to save consumers millions of dollars per year. In this study, we presented a Xgboost-based transaction fraud detection model with some feature engineering and visuali-zation. The dataset is from IEEE-CIS Fraud Detection Compet-ition on Kaggle, which is a well-informed data science organi-zation. The study indicated that xgboost based model outperformed the other three methods including Support Vector Machine, Random Forest and Logistic Regression. As to two feature selection methods, Xgboost performed better. Our best model achieved 95.2% roc auc score on leader-board and defeated other 98 percent parti-cipants.**

Index Terms—Fraud detection, Xgboost, Binary classification, Data Mining

I. Introduction

Fraud is a billion-dollar business and it keeps increa-sing per year. Credit card is a nice target among many frauds since it is easy for attackers to steal the information of consumers and commit the fraud in a short time. Hence, many financial banks and insu-rance companies devoted millions of dollars to building a transaction detection system to prevent high risk transactions. The key component of the system is the detection algorithms, which mainly can be divided into two directions: clustering based methods [1, 2] and classification based methods [3 - 7]. Apart from the powerful model, it is equally important to do data mining [8, 9]. Some techniques like feature selections, data cleaning and data visualization contribute to a good prediction. In this work, we used Xgboost [10] as the classification model and take some data clea-ning strategies to build up a fraud system.

A. Related Work

The term “data” mining is a process to find insights which are statistically reliable, unknown previously and valuable from data. In this task, the goal of it is to extract fraud patterns and knowledge from ordinary data. Then these fraud patterns can be used in further detection via clustering methods or classification methods. Over the decades, several data-mining algo-rithms were proposed for financial fraud detection task[7,8,11]. In [7], the authors presented a data mining algorithm on UCI public the German data set[12]. This set contains 1000 sample, each sample has 20 variables including 7 numerical variables and 13 qualitative variables. However, the dataset size is too small so that it is hard to apply this algorithm to real scenarios. The paper[8] explored the combination of manual and automatic classification based on data mining. In automatic part, the writer developed a risk scoring model to estimate the risk score of every e-tail orders ranging from 0 to 1. In manual part, orders with scores surpass upper threshold would be revised by support teams and they would manually call cust-omers to ensure the validity of orders. The work[11] presented a hybrid detection model which benefited from combining the financial data like quick ratio, sales growth rate and rate of return on total assets .etc and non-financial data like board of supervisors and LSHR. The descriptive statistics related to these vari-ables also calculated such as Min, Max, Mean and Std values.

However, these methods suffer from lack of data or inadequate feature engineering. In [7], the size of data is limited so that it is hard to build a robust and accurate fraud detection system. In [8], it only con-sidered the financial data which may be not enough for a good data mining algorithms. In [11], the number of feature variables isn’t enough and it ignore to pre-process the skew distributed data.

B. Our Contribution

This paper puts forward a Xgboost-based fraud dete-ction algorithm fitting on IEEE-CIS dataset[13].

It contains more than 1 million samples and each sample consists of more than 400 feature variables. These variables involved financial features and non- financial features. It is complicated to deal with such size dataset and it requires processing features and selecting features carefully. In our work, we firstly did data cleaning for purpose of putting out with some anomalies. It is vital and necessary to remove them since they are common to happen in big data set. In addition,

978-1-7281-5904-1/20/$31.00 ©2020 IEEE
DOI 10.1109/ICCEA50009.2020.00122

to solve unbalanced distribution problem of y label, the binary classification output, the SMOTE (Synthetic Minority Oversampling Technique) is used to oversample the minority class. In feature enginee-ring period, some descriptive statistical data like min, max, mean, std of is also generated for some num-erical features. While for categorical features, label encoding algorithm is used to encode categorical data. Finally, high effective and widely used GBDT method Xgboost is implemented as our binary classification. Xgboost is a scalable end to end tree boosting system, used by many data scientists to achieve state-of-the-art results on many machine learning challenges. To show its superiority, we compared it with other class-ical machine learning models like SVM, logistic Regression, Random Forest and so on. Experiments showed that Xgboost model achieve a better score in both accuracy and roc auc score.

The remainder of this paper is organized as follows. The feature engineering on both financial and non-financial data are introduced in Section II. Section III presents the Xgboost based model and corresponding parameters in this task. Section IV compares the performance among the proposed algorithm with other classical machine learning model via metrics like accuracy and Auc Roc score. Finally, Section V draws a conclusion of this paper.

II. Feature Engineering

IEEE CIS Fraud Dataset is provided by Vesta Corporation which is a forerunner in guaranteed e-commerce payment solutions. In this set, we can extract four tables including train / validation dataset of transaction and identity tables. These two kinds of tables can be processed independently and be joined together by the key Transaction ID. For simplicity, we can explore these two tables separately.

A. Transaction tables

Transaction tables has 394 feature variables including 22 categorical features and 372 numeric features. Most of the numeric features are anonymous with a fixed prefix such as V100. To give a specific and clear description, we summarize these variables in Table 1.

Table 1. Transaction Table

Variables	**Variable Description**	**Type**
TransactionID	ID of transaction	ID
isFraud	binary target	categorical
TransactionDT	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

Among the anonymous numerical features like V1-V339, it is sufficient to choose part of them via correlation analysis. For instance, in our work, the features with a correlation coefficient more than 0.95 are removed to prevent overfitting and save computation and space resources. TransactionDT refers to transaction datetime and can be parsed to exact time information like year, month, day, week etc.

This step is vital for later steps to generate time statistical features. Besides, we also generated descriptive statistical features and combined features. For example, TransactionAmt_to_mean_card1 stands for the mean transaction amount of card1. These kinds of features show their importance in following experiments.

As it is illustrated in Figure.1, 0 represents normal transactions and 1 represents fraud transactions. It is easy to find that the target class distribution is unbalanced. In fraud detection task, that majority of transactions are not fraud transactions.

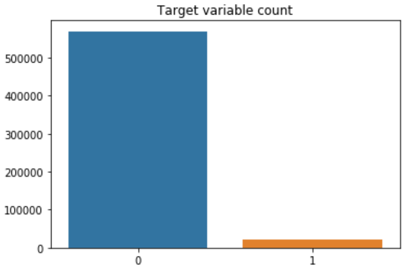

Figure 1. Unbalanced target class

To overcome it, SMOTE algorithms oversample the minority class so that ratio of positive and negative samples closes to 1:1. The Figure 2 shows the difference of target distribution before and after using SMOTE.

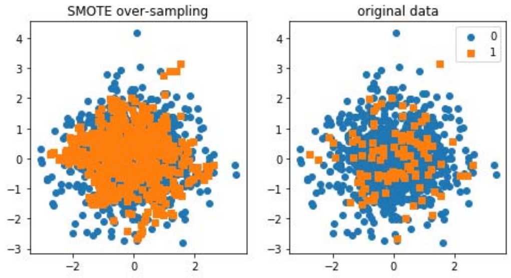

Figure 2. Over-sampling target data

B. Identification tables

Table 2. Identification Table

Variables	**Variable 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

Identification table contains 41 features including ID, categorical and numerical features. DeviceType can indicate device type like desktop, mobile or unknown. DeviceInfo represents the device information like MacOs, SAMSUNG, Moto and Windows etc. We adopt similar data processing methods like previous processing methods on identification tables.

After processing the two tables separately, they should be joined on TransactionID key to generate a new table. However, there are some null values remain in this table which requiring filling out some values like -999 to indicate they are missing values. The whole feature engineering is finished until now and the next step is to do feature selection to select important features from above features. For classification with high dimensionality, feature selection is important to prevent overfit problems and improve classification results. RFECV[16] (recursive feature elimination with cross validation) is often applied in data science competition to recursively choose important features. In this paper, we utilized a 5 fold RFECV method on Xgboost models to find top 100 features.

III. Xgboost Based Fraud detection model

In this section, we will briefly introduce several classification models like Logistic regression, Random forest, SVM and Xgboost model and then display the details of training process.

Logistic regression is a classifier which uses a logistic function to model a binary dependent variable. Supposed the x represented the features, $g(z)$ represents the activation function and $h_\theta(x)$ stands for the hypothesis function. Then we can define logistic model as:

$$h_\theta(x) = g\left(\theta^T x\right), \tag{1}$$

$$g(z) = \frac{1}{1+e^{-z}}, \tag{2}$$

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}, \tag{3}$$

Based on this function, we can output a possibility of a fraud.

Support vector machine [18] is a statistical learning model which is suitable for binary classification task like credit fraud detection. It takes positive samples and negative samples as input and outputs the hyperplane to best separate the task labels. In transaction detection if a test instance lies within the learned zone, it is recognized as normal, otherwise it is regarded as fraud. In this case, linear SVM is not suitable to learn nonlinear relationship between inputs and outputs. Hence, (radial basis function) RBF kernel is used to learn complex regions.

Random Forest classifier[19] consists of an ensemble of decision trees where each tree is generated using a random vector sampled independently, and each tree casts a vote to find the most popular class to classify the input. In this case, the classifier can choose randomly selected features or a combination of features to generate a tree from top to down. Bagging technique is a key component in random forest. It can generate a training set by randomly choose examples in a certain possibility so that the random forest would be diversiform. Since random forests are an ensemble of decision trees, it is essential to introduce how to generate decision trees. Each node of a decision tree is an attribute and there exist many methods to choose a quality attribute. The most frequently attribute selection methods are Gain Ratio criterion and Gini index method. Given a training set T, select one example which belongs to some class C_i, the formula of Gini index can be defined as followed:

$$\sum\sum_{j\neq i}\left(\frac{\int\left(C_i,\ T\right)}{|T|}\right)\left(\frac{\int\left(C_j,\ T\right)}{|T|}\right), \tag{4}$$

Where $\frac{\int(C_i,\ T)}{|T|}$ is the probability that the selected case belongs to class C_i.

Xgboost which is also called eXtreme Gradient Boosting, is an excellent tree boosting system. According to Tianqi Chen, Xgboost optimize the loss function by adding regularization to handle the sparse data and weighted quantile sketch for tree learning. Besides, they provide some insights that helps to build a fast and scalable tree boosting system. These insights contains data compression, sharding and cache access patterns. Through these mechanism and insights, Xgboost outperforms most of other machine learning algorithms in both speed and accuracy. It is also convenient for data scientists or engineers to employ Xgboost systems in distributed system or GPU machine. Considering the advantages of Xgboost, we applied this model in our work.

To best score of our model, it is useful to tune para-meters of Xgboost according to your task. However, it doesn't mean that every parameters should be used or tuned. Actually, for most task, people only need to some key parameters like learning rate, number of estimator and tree methods etc. To give a better detail, the Table 3 lists the parameters of our model and the other parameters don't show in this table are default parameters.

Table 3. Xgboost parameters table

Parameter	Parameter Description	Value
n_estimators	number of estimators	5000
learning_rate	learning rate	0.02
subsample	sample rate of rows	0.8
max_depth	max depth of each tree	12
colsample_bytree	sample rate of columns	0.4
tree_method	method for boosting tree	gpu_hist

The above table shows the parameters of xgboost. For further optimization, it is easy to use some search algorithms like grid search and Bayes optimization search methods. Given the range of each parameters and the score function of model, these methods would find the best parameters automatically. In most case, Bayes search would find a solution with a shorter time since it uses the prior knowledge in searching.

IV. Experiments

In this section, we applied our proposed algorithm on IEEE-CIS datasets and compared our algorithm with other algorithms. The environments of experiments are on ubuntu 18.04 with a GPU Nvidia Rtx 2080Ti and with the processor i9 9900K. Cuda 7.5 com-putation package was also installed in the computers so that we can quicken the pace of model training.

To compare with different models, the IEEE-CIS competitions used Auc-Roc score to judge the performance of models. This score actually is the area under roc curve, which is also known as receiver operating characteristic curve. The roc curve is created by ploting the true positive rate(TPR) against the false positive rate(FPR) at various threshold setting. The formula of TPR and FPR are defined as follows:

$$TPR = \frac{TP}{TP+FN}, \quad (5)$$

$$FPR = \frac{FP}{FP+TN}, \quad (6)$$

Where TP means true positive prediction, FN means false negative prediction, FP stands for false positive prediction and TN means true negative prediction.

In addition to Auc-Roc score, we also provided the accuracy score of different models. In table 4, four models logistic regression, SVM, random forest and Xgboost with their scores are shown.

Table 4. Performance of different models

Models	**Auc Roc Score**	**Accuracy**
Logistic Regression	0.862	0.931
SVM	0.907	0.958
Random Forest	0.925	0.965
Xgboost	**0.952**	**0.981**

From table 4, our Xgboost algorithm outperforms the other four algorithms on both Auc- Roc score and accuracy.

Another advantage of Xgboost is that it can output the feature importance of each features. In figure 3, it shows the importance of some part of features in a decreasing order.

Figure 3 shows that card features, address features together with transaction amounts is relative more important than other features. However, some features like M4, id_33, M6 and D15 share less significance in combinations of features. In practice, to make a trade off between accuracy performance and computation speed, we can choose parts of features according to feature importance. To make a better prediction, more features should be selected while for a faster detection speed fewer features should be selected.

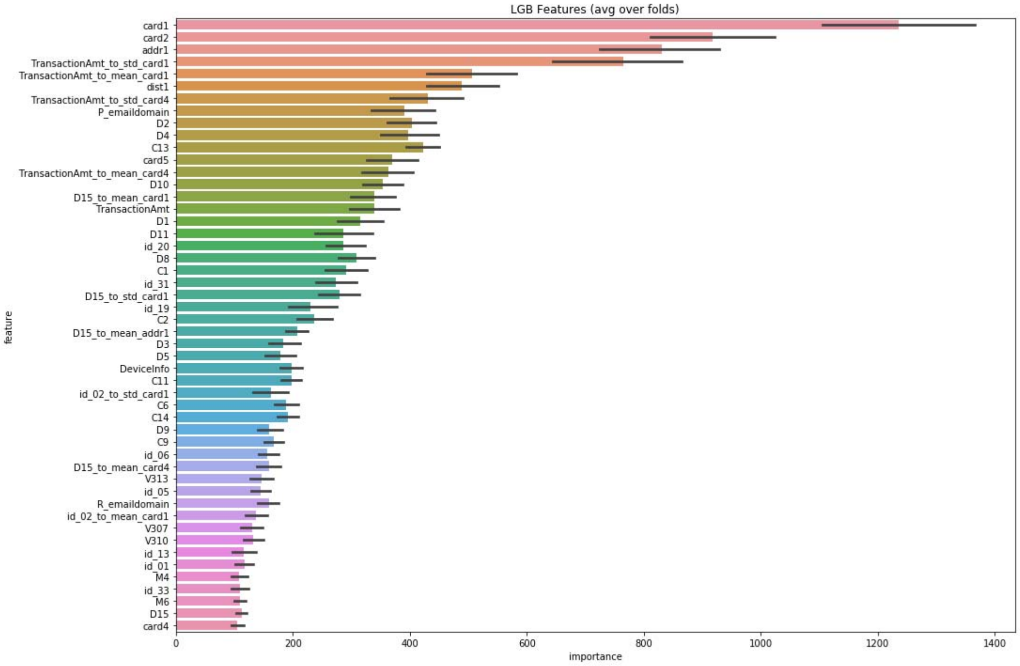

Figure 3. importance of features

V. Conclusions

In our work, we present a Xgboost-based model to detect customer transaction fraud on IEEE-CIS dataset. Some data mining techniques like data cleaning, feature selection and feature engineering are

Discussed in section II. In section III, we introduce some machine learning models including logistic regression, support vector machine, random forest and Xgboost. In fourth part of the paper, experiments show the performance of the four models. According to the result, we can draw the conclusion that Xgboost outperforms other machine learning models in both Auc Roc score and accuracy score. In figure3, features are ranked according to their importance. This information is valuable for feature selection to reduce feature dimensions.

Acknowledgement

The author Yixuan Zhang thanks Jialiang Tong for help in experiments, Ziyi Wang for idea contribution and discussions and Fengqiang Gao for fruitful discussions

References

[1] Majhi, S. K. (2019). Fuzzy clustering algorithm based on modified whale optimization algorithm for automobile insurance fraud detection. Evolutionary Intelligence, 1-12.

[2] Carneiro, E. M., Dias, L. A. V., da Cunha, A. M., & Mialaret, L. F. S. (2015, April). Cluster analysis and artificial neural networks: A case study in credit card fraud detection. In 2015 12th International Conference on Information Technology-New Generations (pp. 122-126). IEEE.

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

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

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

[6] 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).

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

[8] 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.

[9] West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: a comprehensive review. Computers & security, 57, 47-66.

[10] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[11] Kamesh, V., Karthick, M., Kavin, K., Velusamy, M., & Vidhya, R. (2019). Real-Time Fraud Anamaly Detection in E-banking Using Data Mining Algorithm.

[12] UCI(1994). Statlog (German credit data) data set. https://archive.ics.uci.edu/ml/machine-learning-databases/statlog /german.

[13] https://www.kaggle.com/c/ieee-fraud-detection/data

[14] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

[15] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[16] Chen, X. W., & Jeong, J. C. (2007, December). Enhanced recursive feature elimination. In Sixth International Conference on Machine Learning and Applications (ICMLA 2007) (pp. 429-435). IEEE.

[17] Tolles, J., & Meurer, W. J. (2016). Logistic regression: relating patient characteristics to outcomes. Jama, 316(5), 533-534.

[18] Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. Neural processing letters, 9(3), 293-300.

[19] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.