An Xgboost based system for financial fraud detection

Shimin LEI[1,a], Ke XU[2,b], YiZhe HUANG[3,c], Xinye SHA[4,d]

[1]University of Virginia, Virginia, United States
[2]Washington University in Saint Louis, Saint Louis, United States
[3]Fudan University, Shanghai, China
[4]University of Cambridge, Shanghai, China

Abstract—Credit card fraud leads to billions of losses in online transaction. Many corporations like Alibaba, Amazon and Paypal invest billions of dollars to build a safe transaction system. There are some studies in this area having tried to use machine learning or data mining to solve these problems. This paper proposed our fraud detection system for e- commerce merchant. Unlike many other works, this system combines manual and automatic classifications. This paper can inspire researchers and engineers to design and deploy online transaction systems.

1 Introduction

Due to the outbreak of online transactions, online frauds increase rapidly per year. Risk assessment is a key technique to avoid frauds and hence adopted by many banks and companies. They have invested billions of dollars to prevent frauds and protect consumers' profit. However, it is impossible to risk every transaction manually. Actually, most of these transactions are processed by machine without any human interfere. In past years, data mining is applied to transaction frauds tasks [1-5]. Data mining is a process to extract interesting patterns from metadata which can be fitted to machine learning models like logistic regression [6], Naïve Bayes [7] and support vector machine [11] etc. Over the task, we extracted risk patterns via data mining and did inference via Xgboost [9] predictor. However, automatic detection and decision may be problematic in some cases and some normal transactions would be rejected and some fraud transactions would be passed. Hence, it is necessary to combine automatic detection with manual review.

1.1 Related Work

The increasing computation resources and data mining tools promotes development of financial systems. Many intelligent algorithms and data mining techniques are applied in fraud detection system. Bhusari and Patil have proposed a Hidden Markov Model [3] helps to obtain a high fraud coverage combined with a low false alarm rate. In paper [4], the authors proposed a CNN-based fraud detection framework, to capture the intrinsic patterns of fraud behaviors learned from labeled data. detection algorithm to detect fraud click from millions of click actions. The study [5] mainly focused on the feature engineering part, which is a most important part in data mining. The researchers expanded the transaction aggregation strategy, and proposed to create a new set of features based on analyzing the periodic behavior of the time of a transaction using the von Mises distribution. However, most of them lack the manual classification parts and dataset is not large enough. Manual classification can cover the shortage of automatic classification. Hence, the integrated fraud detection system performs better in this task. The richness of data will remarkably improve the accuracy of fraud detection models and avoid overfitting problems.

1.2 Our Contribution

This paper proposes an Xgboost based system for financial fraud detection. It can be divided into two parts, automatic part and manual part. The automatic part used a large dataset which contains millions of transactions to train our model. This dataset is the IEEE-CIS data which can be downloaded in Kaggle platform. With this data, the model can conduct a better result and avoid overfitting. The model for detection is Xgboost, which 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. In the other part, this paper introduce manual review to monitor the transaction. The transactions of high risk will be reviewed by human. The final decision will be made by combining machine scoring and manual feedback.

The following parts are organized as follows. Our system and data mining techniques are introduced in Section II. Section III presents the Xgboost based model and related parameters. In Section IV, experiments of the Xgboost based model and other classical models are carried out. In the end, final section draws a conclusion of this paper.

[a]sl2kd@virginia.edu; [b]ke.xu@wustl.edu; [c]13761536034@163.com; [d]xinyesha98@outlook.com

2 Xgboost Based system

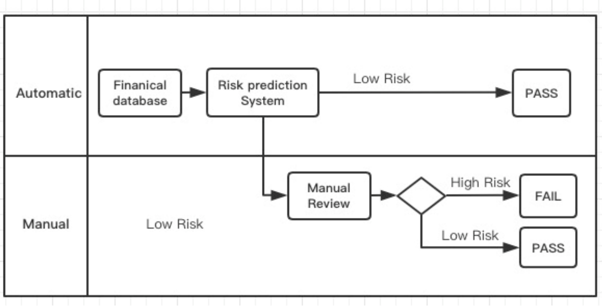

Figure 1. Our proposed system

The Figure 1 shows our proposed system. The first part is automatic, it gets data from financial database and put the data to risk prediction system. Risk prediction system is the intelligent system built by Xgboost based model. If a transaction is detected as low risk, then no extra actions will be needed and the transaction will be automatic approved by system. In manual part, high risk transaction will be reviewed manually. We can set a threshold like 0.8 and the transaction with a score more than this value can be viewed as high risk transaction. High risk transactions will be reviewed by some experts. They will search historical transactions of this identity or make a call to corresponding sender. If the experts assess the transaction as a fraud transaction, then the transaction will be failed. In practical, many companies or banks will call their users to guarantee effectiveness of a transaction.

Some data mining techniques like feature engineeering and feature selection are applied in this task. The IEEE-CIS [8] dataset has two kinds of tables transaction tables and identification tables. Therefore, firstly did data mining on them and then combined them together to a final data.

2.1 A. Transaction table

The transaction tables have 22 categorical features and 372 numeric features. The detail of these features is listed in table 1.

TABLE 1 TRANSACTION TABLE

Name	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

2.2 Identification table

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

The identification table have 11 categorical features and 29 numerical features.

Our data mining method mainly includes data cleaning, missing value filling, label encoding and feature elimination. The data cleaning part removed the columns with large percentage of Nan values (missing values). For example, if more than 90 percent of values in a feature column is Nan, this column will be deleted. Then the next process is filling Nan values with a specific value which hasn't appeared in data like -999. In addition, for categorical features, they need be transformed to numerical data by using label encoder technique. Finally, in feature elimination part, the high related features are removed.

3 Xgboost Based model

In this section, the Xgboost based model is used in automatic system and it will be introducted in this section. Xgboost is a highly scalable end-to-end tree boosting system. It integrates many advantageous designs. The justified weighted quantile sketch is used for efficient proposal calculation. The sparsity-aware algorithm is developed for parallel tree learning. An effective cache-aware block structure is implemented for out-of-core tree learning. Due to these pros, Xgboost outperforms most of other machine learning algorithms in both speed and accuracy. It is also easy for researchers to deploy it on distributed system and accelerating via GPUs. Due to its superiority, we applied Xgboost predictor in this task.

TABLE 3 XGBOOST PARAMETERS

Parameter	Parameter Description	Value
n_estimators	number of estimators	5000
learning_rate	learning rate	0.01
subsample	sample rate of data	0.8

max_depth	max depth of trees	15
colsample_bytree	sample rate of features	0.8
tree_method	boosting tree method	gpu_hist

Table 3 lists the parameters of our model. The other parameters are default. By selecting different para-meters, we can generate a different model. For example, because the n_estimators determine the iteration steps of the Xgboost, a large amount of steps and small learning rate usually result in a more accurate model. In addition, we can also change parameters to accelerate the training speed. Setting the tree method to gpu_hist means that the Xgboost will be trained on GPUs which is at least 50 times faster than CPUs.

4 Experiments

The training of Xgboost based model can be done offline or online. At this time, we utilized the online computer resources with 4 Cpu cores and a Gpu of P100 provided by Kaggle. Due to the acceleration of Gpu, the training time is compressed to a few hours. To show the superiority of our model with use accuracy and Auc-Roc score as metrics to compare it with other models. The roc curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. In Figure 2, it is the roc curve of our model, in which x axis is false positive rate and y axis is true positive rate. This score is the area under roc curve. The roc score of our model is 0.942 which can also be found in Table 4. In addition to Auc-Roc score, we also provided the accuracy score of different models. Table 4 compared the model used with other three models and recorded the auc-Roc score as well as accuracy.

TABLE 4 Performance of different models

Models	**Auc Roc Score**	**Accurac y**
Naïve Bayes	0.782	0.824
Logistic Regression	0.865	0.882
GBDT	0.905	0.935
Xgboost	0.942	0.976

From table 4, the Xgboost based model outperforms the other three models (Naïve Bayes, Logistic Regression, GBDT [10]) on both Auc-Roc score and accuracy, which means Xgboost can detect fraud more accurately than other three models.

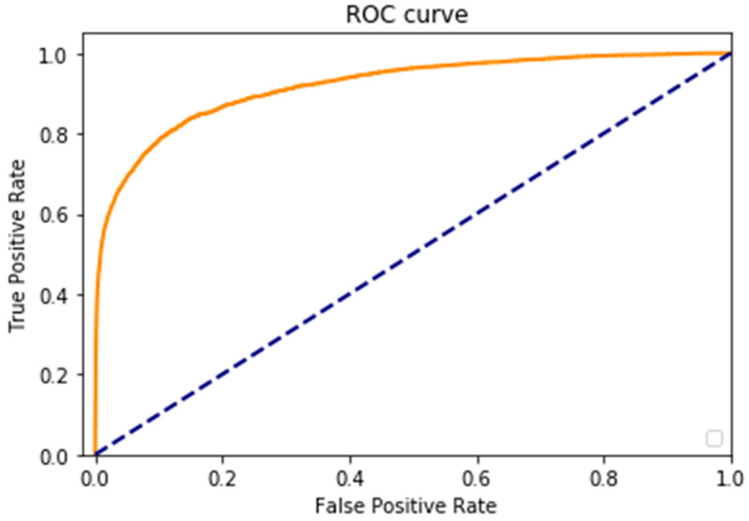

Figure 2. Roc-Auc Curve for proposed method

5 Conclusions

This paper presents an Xgboost-based financial system to detect transaction fraud. The framework of this system and the superiority of Xgboost is illustrated in this paper. Specif-ically,in Section II, we firstly put forward our system, which can be divided into two parts, automatic part and manual part. Then the data mining techniques like data cleaning, feature engineering and feature elimination are introduced. The details of IEEE-CIS dataset are described in this section. Se-ction III discussed the pros of Xgboost model and share the details of its parameters including learning rate, number of estimators, max depth and so on. In experiment part of the paper, the deploy environment is introduced and results of our model and other classical models are shown in Figure 2 and Table 4.

Acknowledgement

The author Shimin LEI thanks Xu KE for help in carrying out experiments on classical models, YiZhe HUANG for idea contribution on feature engineering and data cleaning and Xinye SHA for meaningful online discussions. Last but not least, we thank Kaggle platform to provide dataset and online computation resources.

References

1. 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.
2. Minastireanu, E. A., & Mesnita, G. (2019). Light gbm machine learning algorithm to online click fraud detection. J. Inform. Assur. Cyberse-cur, 2019.
3. 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.
4. Fang, Y., Zhang, Y., & Huang, C. Credit Card Fraud Detection Based on Machine Learning. Fu, K., Cheng, D., Tu, Y., & Zhang, L. (2016, October). Credit card fraud detection using convolutional neural networks. In International Conference on Neural Information Processing (pp. 483-490). Springer, Cham.
5. 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.
6. Tolles, J., & Meurer, W. J. (2016). Logistic regression: relating patient characteristics to outcomes. Jama, 316(5), 533-534.
7. Murphy, K. P. (2006). Naive bayes classifiers. University of British Columbia, 18, 60.
8. https://www.kaggle.com/c/ieee-fraud-detection/data
9. 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).
10. Friedman, J. H. (2002). Stochastic gradient boosting. Computational statistics & data analysis, 38(4), 367-378.
11. Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. Neural processing letters, 9(3), 293-300.