Fraud Detection in E-Commerce Transaction

Xi Qian, M.S.

The George Washington University

Table of Contents

INTRODUCTION

Fraud is billion-dollar business and it is increasing every year. The PwC global economic crime survey of 2018 found that half of the 7200 companies they surveyed had experienced fraud of some kind. Fraud often consists many instances or incidents involving repeated transgressions using the same method. Fraud instances can be similar in content and appearance but usually are not identical. Traditional methods of data analysis have long been used to detect fraud for a long time. Techniques for fraud detection fall into two primary classes: statistical techniques and artificial intelligence. Fraud detection process involves complex and time-consuming investigations that deal with different domains of knowledge like financial, economics, business practices and law. By looking at history data, people may improve the efficacy of fraudulent transaction alerts for millions of people around the world, helping hundreds of thousands of businesses reduce their fraud loss and increase their revenue.

Although fraud detection techniques have been used in industries such as telephone companies, insurance companies and banks for a long time, there are two main criticisms of data mining based fraud detection research: the dearth of publicly available real data to perform experiments on; and the lack of published well research methods and techniques. Especially, most of these limited examples and papers focused on utilizing white box models for fraud detection due to regulations in those industries. As the development of the algorithm and technology, people who work in these industries do realize the limitation of using these traditional methodologies to detect fraud. Now, they are also interested in using more black box models to help them detect fraud.

Hence, the primary objective of this project is to measure and compare the results of using black box machine learning models on a history large-scale dataset from a real-life company for fraud

detection, as well as existing challenges in using these black box models. It also tries to highlight promising new directions from related adversarial data mining fields. In all, the result of this study can be used as reference for people who are interested in applying higher level methodologies in fraud detection, and also can be served as a reference for people who want to combine traditional methods and other machine learning algorithms to help industries detect fraud in the future.

LITERATURE REVIEW

Research scholars have conducted studies exploring fraud detection using different statistical and machine learning techniques. Traditional statistical classification methods (Hand, 1981; McLachlan, 1992), such as linear discriminant analysis and logistic discrimination, have proved to be effective tools for many applications, but more powerful tools (Ripley, 1996; Hand, 1997; Webb, 1999), especially neural networks, have also been extensively applied. Researchers who have used neural networks for supervised credit card fraud detection include Ghosh and Reilly (1994), Aleskerov (1997), Dorronsoro (1997), and Brause (1999). Rule-based methods are supervised learning algorithms that produce classifiers using rules of the form If, Then. Tree-based algorithms such - CART produce classifiers of a similar form. In all, supervised methods to detect fraudulent transactions can be used to discriminate between those accounts or transactions known it be fraudulent and those known (or at least presumed) to be legitimate. However, new algorithms such as LightGBM and XGBoost are not so common to be used in fraud detection yet. Hence, it would be worthwhile to try and see how they work, what's their advantage and disadvantage.

With regard to the use of supervised learning methods in fraud detection, fraudsters adapt to new prevention and detection measures, so fraud detection needs to be adaptive and evolve over time.

In this case, unsupervised methods are useful in applications where there is no prior knowledge as to the particular class of observations in data set (Richard, 2001; David, 2001). For example, we may not be able to know for sure which transactions in a database are fraudulent and which are legitimate. In this situations, unsupervised methods can be used to find groups or find outliers in the data. Essentially, we collect data to provide a summary of the system, we can identify those observations that do not fit in with this behavior, i.e. anomalous observations. In all, models generated should be either be updated at fixed time points or continuously over time if they are generated from supervised algorithm; unsupervised models may help people to identify new fraudulent behavior or that have not been captured by supervised model.

METHODOLOGY

Data Collection

Both Transaction and Identity datasets comes are provided Vesta Corporation, which is the forerunner in guaranteed e-commerce payment solutions. It is available on Kaggle website.

Dataset Description

This project uses data comes from Vesta's real-world e-commerce transactions and a wide range of features from device type to product features.

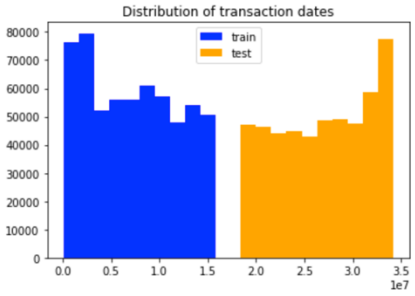
The data is broken into two files: identity and transaction, which are joined by TransactionID. Not all transactions have corresponding identity information.

Further, for training purposes, Vesta also divided their entire dataset into training and testing set. They include 560k and 500k of transaction records respectively. There are 433 feature variables in the dataset, with 1 target variable called "Fraud or Not". Among 433 feature variables, there are about 340 variables with their true meaning masked, they are Vesta engineered rich features.

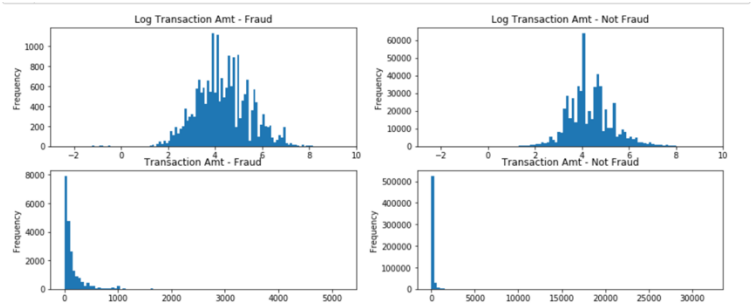
Exploratory Data Analysis

The distribution of the target variable indicates that this is an imbalanced dataset. With only 3.5% of fraud cases in all records.

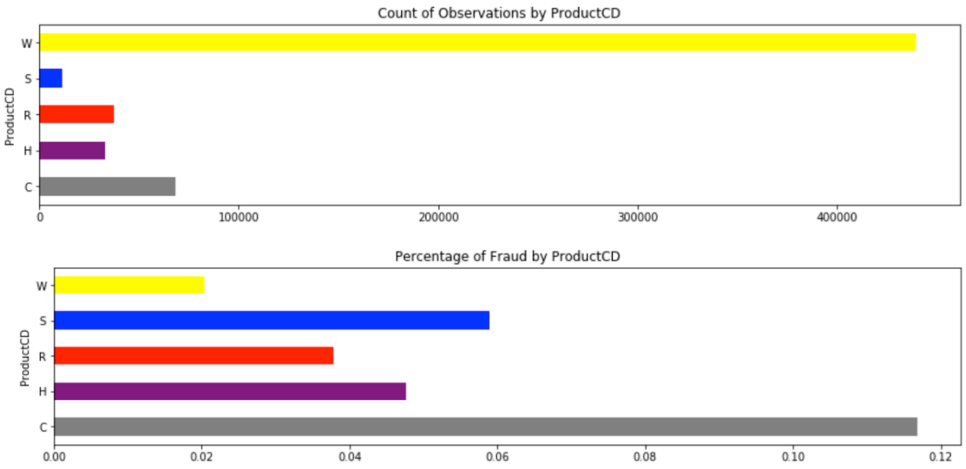
The transactionDT variable is a timedelta from a given reference datetime (not an actual timestamp). One discovery about the data is that the train and test appear to be split by time. There is also a gap in between, the training set was from an earlier period of time and the testing set was from a later period of time. This will impact which cross validation technique should be used.



Log transformation was performed on the variable called TransactionAMT (transaction amount). Otherwise, very few large transactions skew the distribution. After calculation, it shows that fraudulent charges appear to have higher average transaction amount.

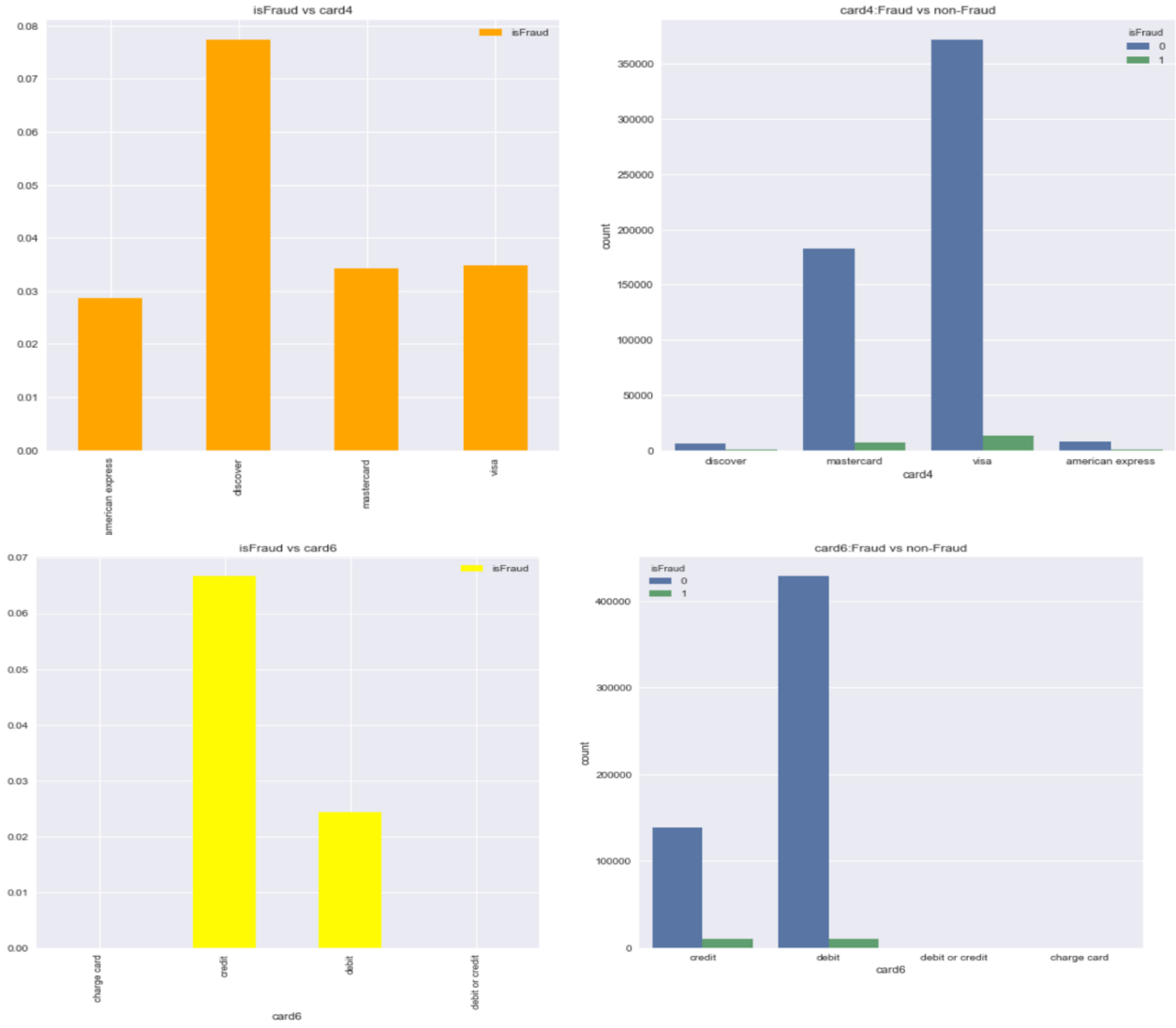


There is a variable called ProductCategory, but the exact category name with respect to each category is masked. The distribution plot shows that Category W has most number of records associated with it, but has smallest percentage of fraud cases. Category C has the small number of records associated with it, but with largest percentage of fraud cases.



There is a variable called Card4, it refers to the company the card associated with. The distribution plot shows that transactions associated with Discover has highest ratio of fraud cases. Ratio of fraud in transactions associated with Mastercard and Visa follow behind Discover.

There is a variable called Card6, it refers to the card type: credit/debit/charging. Although the amount of fraud cases in both credit and debit card types are about the same, but the fraud ratio in credit type of card are higher.

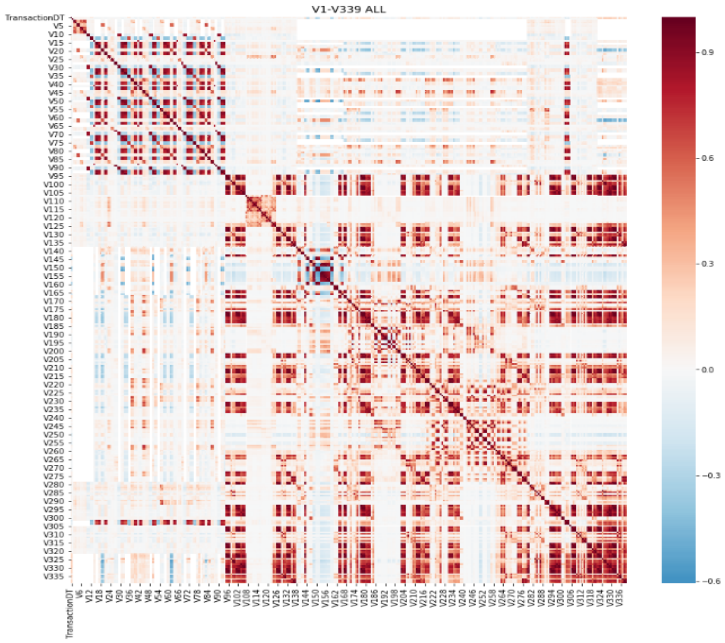


Data Preprocessing & Feature Engineering

Here is a correlation heatmap for 335 Vesta engineered rich features. The correlation heatmap obviously revealed the strong correlation in groups of V1-V95 and V96-V335. The way those rich features were handled in this project mainly involved in using NAN processing and Dimensionality Reduction using PCA (Principle Component Analysis).

First of all, a function that be used to calculate the percentage missingness in those variables were wrote. In addition, features that have the same percentage of missingness were grouped into the

same group. From those generated subgroups, only one of the features among groups are selected and used in modeling process in this project before performing PCA on it. After the reduced features are selected from each subgroup, PCA was performed on them and 90% of the explained variance was set as the threshold. At the end, 50 principal components were generated and be put into the modeling process.



Other feature engineering process were also performed, such as creating dummy variables, regrouping variables based on the distribution/frequency of values.

Modeling

In the modeling process, several main issues are handled.

First of all, because the train and the test datasets are split by time, time series split were used for model validation. In other words, for example, the num of splits were set to 5 in the training dataset, which means every time the model is trained on the first n fold and validated on the later 5-n folds.

Secondly, specific machine learning algorithms were selected for training the model. They are LightGBM and XGBoost. There are two reasons of using them: (1) The number of variables is very large in this dataset, even after dimensionality reduction. Both these two algorithms have advantage in processing dataset with large number of variables. (2) The goal of this project is to measure and compare the results – accuracy. In other words, the goal of this project is not to interpret how the model works, what rules can be derived from the model. Hence, these two algorithms are selected for training.

Thirdly, hyperparameter tuning were performed to find the optimal parameter values that could give the better result and higher accuracy in recognizing fraud in transactions. Parameters such as num_leaves, min_data_in_leaf, max_depth, learning_rate, reg_alpha, reg_lambda in LightGBM were tuned. Parameters such as max_depth, reg_alpha, reg_lamba, num_leaves, min_child_samples in XGBoost were tuned.

RESULT & ANALYSIS

LightGBM result	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Training Set	0.95745	0.997002	0.999465	0.991157	0.988608
Validation Set	0.872766	0.911545	0.924285	0.922122	0.917767

XGBoost result	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Training Set	0.97643	0.984565	0.992463	0.993552	0.987252
Validation Set	0.95353	0.964523	0.974672	0.975562	0.963622

	Best Accuracy (Testing Set)
LightGBM	0.9656
XGBoost	0.9794

Above are the results.

The top one is the result from LightGBM model using 5 folds for validation. Since the fraud rate is 3.5%, as long as the model result is at least as 96.5%, then the model works well overall. The accuracy increases in both training and validation set as the number of folds increases in the

training set, and the result makes sense. For example, when there is only one fold of data that was used for training and four folds of data that were used for validation, obviously the information the algorithm that can capture from the only fold is limited, so the accuracy of the validation set in Fold 1 is the lowest among all folds. Overall, the highest accuracy is up to 99% in the training set. However, the accuracy in validation set is not so high. One possible reason behind it is that new fraudulent behavior might have occurred in the last time period, but the old model was not able to learn that from history data. Similar result also occurs in result table of XGBoost.

Overall, the performance of XGBoost is better than that of LightGBM in this project using this specific dataset. The final accuracy of XGBoost model is a little higher than that of LightGBM, which indicates XGBoost is a more robust model. One thing to be noted that, the model that was trained on the history records might not work well on new observations, it might only be able to capture fraudulent behavior that has already occurred and captured in history records that exist in training set. This might explain why the accuracy on the test set for both models is not as high as it is in training and validation set.

CONCLUSION

In this project, several key findings are the following: (1) The dataset is a highly imbalanced dataset, with fraud cases only takes up to 3.5% among all transactions. (2) Average transaction amount of fraud cases is little higher than that of not-fraud cases. (3) Product Category that appear most often in transactions usually associated with lower fraud rate. Product Category that has smaller number of transactions usually associated with higher fraud rate. (4) Fraudulent activities have higher percentage of possibility of being associated with Visa and American Express than any other card issuer. However, that does not mean there is any correlation between fraudulent activity and card issuer. (5) Credit card has higher possibility of being related with

fraud than debit card. (6) XGBoost performs better than LightGBM in capturing fraudulent transactions in this dataset, the highest accuracy is 98.0%. However, when it comes to new observations that might involve with new fraudulent behavior pattern, both models have the possible weakness of not being able to classify those new observations correctly.

Obviously, there are also limitations with this project. The interpretation of the model is not known. Compared to those white box model such as decision tree, which rules can be derived, both models that were used in this project could only provide accuracy. Hence, these two models can only be used when the goal of is not to interpret how the model classify records.

With regard to the improvements that could be made in the future, there are two aspects can be worked on: (1) To handle imbalanced dataset, oversampling and undersampling are both feasible techniques can be used before training the model. Instead of using accuracy and confusion matrix, training algorithms can learn more from the updated datasets, so the model performance would get closer to reality too. (2) Besides supervised algorithms, unsupervised algorithms can also serve to detect anomaly in data. Since supervised algorithms can only capture fraudulent activities that have occurred in history dataset, using unsupervised algorithms might help to capture undiscovered or new fraudulent behavior such that helping business avoid loss in the future.

REFERENCE

Hand D.J. and Henley W.E. (1997) Statistical classification methods in consumer credit scoring: a review. Journal of the Royal Statistical Society, Series A, 160, 523-541.

Ghosh, S. and Reilly, D. L. (1994). Credit card fraud detection with a neural network. In Proceedings of the 27th Hawaii International Conference on System Sciences (J. F. Nunamaker and R. H. Sprague, eds.) 3 621-630. IEEE Computer Society Press, Los Alamitos, CA.

Aleskerov, E., Freisleben, B. and Rao, B. (1997). CARD-WATCH: A neural network based database mining system for credit card fraud detection. In Computational Intelligence for Financial Engineering. Proceedings of the IEEE/IAFE 220- 226. IEEE, Piscataway, NJ.

Dorronsoro, J. R., Ginel, F., Sanchez, C. and Cruz, C. S. (1997). Neural fraud detection in credit card operations. IEEE Transactions on Neural Networks 8 827-834.

Brause, R., Langsdorf, T. and Hepp, M. (1999). Neural data mining for credit card fraud detection. In Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence 103-106. IEEE Computer Society Press, Silver Spring, MD.

Bolton, Richard J.; Hand, David J. Statistical Fraud Detection: A Review. Statist. Sci. 17 (2002), no. 3, 235--255. doi:10.1214/ss/1042727940.

McLachlan, G. J. (1992). Discriminant Analysis and Statistical Pattern Recognition. Wiley, New York.

Ripley, B. D. (1996). Pattern Recognition and Neural Networks. Cambridge Univ. Press.

Webb, A. R. (1999). Statistical Pattern Recognition. Arnold, London.