Credit Card Fraud Detection

By Yiming Zhao

1. Problem Description

Credit card is the most popular payment tool in contemporary society. It is quicker and more convenient than other traditional payment methods, such as cash and checks. Although paying by credit card has significantly changed the payment structure, it has also brought a series of crime problems related to it. For example, credit card fraud, financial chaos, and bankruptcies are frequently reported issues in the spotlight. The serious issue among them is credit card fraud. It has been identified as a form of identity theft where criminals buy different products or obtain cash using other people's credit cards account.

According to related statistics, more than 459,297 credit card fraud offenses including credit card fraud and identity theft were reported in 2020. Cardholders lost more than $3.3 billion due to credit card fraud in 2020, an increase of $1.5 billion compared to 2019. Official data have revealed a sharp increase in credit card fraud, causing huge financial losses to consumers. Therefore, it is necessary for financial institutions to detect possible fraudulent transactions in time so that they can ensure financial safety for the cardholder. In this project, I developed a hybrid method for a credit card fraud detection to identify potential fraudulent transactions.

2. Project Procedure

2.1 Data Description and Feature Engineering

I used a simulated credit card transaction records from a financial institution as the dataset. Each record includes detailed information about a transaction for each credit. The independent variables of the task are the features of the transaction, such as time, amount, merchant, etc. The target dependent variable is whether a transaction is fraud, it is a binary classification task (1 or 0). There are more than one million credit card transaction records in the dataset, and the ratio of positive case to negative cases is 1:160 (shown in figure 1). For highly imbalanced datasets, it is necessary to sample the minority class before building the classification model to ensure that the model will not be affected by the distribution of the classes.

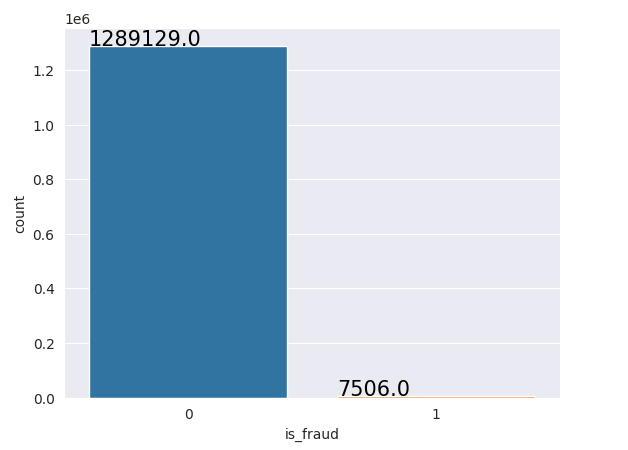


Figure 1

Many features in the dataset are not closely associated with the fraudulent transactions. For those features, I tried to synthesize them into meaningful features through feature engineering methods. Zhang et al. (2019) designed and tested a homogeneity-oriented behavior analysis (HOBA) model to process the features in credit card data. This method applies a novel framework to classify all features into four different groups, namely recency, frequency, monetary, and location. Recency reflects how recently a credit card user make a transaction and frequency reflects how often the user make transactions. These two groups of features focus on identifying transfer transactions that do not match customers' historical purchasing preferences and individual habits. Monetary considers the amount of a single transfer. For example, if the amount of the transaction is too large, it may be risky. The last feature is location which involves the real-time location of the transactions. Table 1 shows all the features I created as well as the feature descriptions.

Table 1

|  |  |
| --- | --- |
| Features | Description |
| Transfer interval | The time interval between the current transaction and the last transaction |
| Time period | Morning, afternoon, evening, and evening |
| Purchasing frequency | The purchasing frequency of current merchant category |
| Distance | Distance between purchase location and original location |
| Age | The interval between birthday and transaction time |

2.2 Exploratory Data Analysis

The next step is to conduct exploratory data analysis. This allows me to have a basic understanding of the dataset as well as detailed information on the columns. Figure 2 shows that groceries paid at POS machine and online shopping are the channels where fraudulent transactions occur frequently. Form figure 3 we can see that evening and night are the most frequent time period for fraudulent transactions.

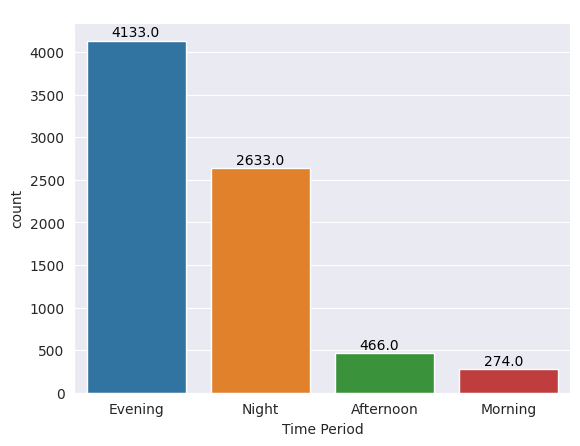
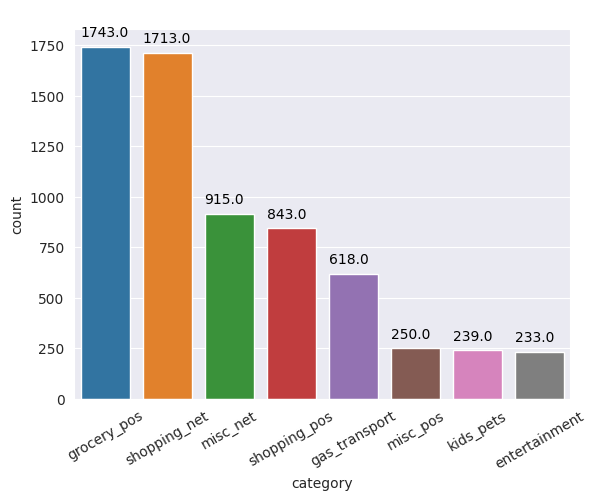


Figure 2 & Figure 3

A heat map is drawn to verify the correlation between features, from Figure 4 we can see that there are no highly correlated features.

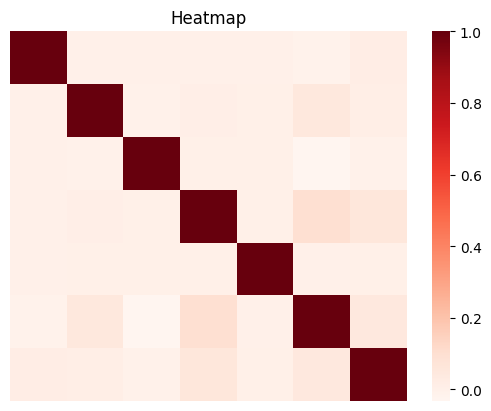


Figure 4

2.3 Data Sampling

Compared to normal transactions, fraudulent transactions are not common in real life. As a result, the credit card fraud dataset is often highly imbalanced. There are two different ways to solve the class imbalance problem. The first one is undersampling method which remains all data points in the minority class and samples data from the majority class. However, this method may lose critical information in the majority class. Another method is oversampling which augments the number of observations in the minority class. In this project, I applied SMOTEENN method that can do undersampling and oversampling simultaneously. After sampling the data, the ratio of positive cases to negative cases is 1.05 in the balanced dataset.

3 Related Work

3.1 Unsupervised Methods

Unsupervised models are widely used in the task of outlier detections (Fabrizio et al. 2019), such as one class SVM, k-means, isolation forest, DBSCAN, and Gaussian Mixture Model. Since Unsupervised models are not required to identify the type of class of each observation based on domain experience, unsupervised can help us quickly discover possible fraudulent transactions or new types of fraudulent transactions that have never been learned before.

3.2 Supervised Methods

Previous research also applied various supervised methods, such as Xgboost model, Random Forest model, SVM model, KNN model, ANN model, and Logistic Regression model to identify potential fraudulent transactions. These models are proven to be effective in previous research (Zhang et al. (2022), Acosta et al. (2017), Taha et al. (2020)). In this research I used them as bench mark models to evaluate the performance of the hybrid method. In addition, supervised models may have different sensitivities for different classes, which enables the model to identify the pattern of the fraudulent transaction effectively. For example, if a financial institution wants to ensure that all fraudulent transactions are found as much as possible to guarantee the security of the cardholder's account, then a model with a low false negative rate or high recall value should be applied. On the contrary, if a financial institution does not want to cause trouble to customers due to fake alarms, it is necessary to select a model with low false positive rate or high precision value.

Moreover, I also discuss the explainability among all the models. For credit card fraud issue, highly explainable models can be used to provide more insight into the features of the fraudulent transactions. Practitioners can develop targeted preventive measures based on these features.

4. A Hybrid Method

We know that kernel functions are able to project data into higher dimensions to deal with non-lineal datasets. Therefore, in this project, I designed a hybrid method to test whether the kernel function can further improve the performance of supervised models by striping outliers through projecting data into higher dimensions. First, I used One Class SVM to find the optimal kernel function that can identify outliers (which are also believed as the fraudulent transactions) effectively. Then, I applied this optimal kernel to transform the dataset and train different models. Finally, I compared the model trained on transformed dataset with other bench mark models trained on the original dataset to verify the effectiveness of the hybrid approach.

5. Experiments and Evaluations

Table 2 shows the performance of different kernel functions in detecting outliers in One Class SVM. We can find that the Gaussian RBF kernel performances best.

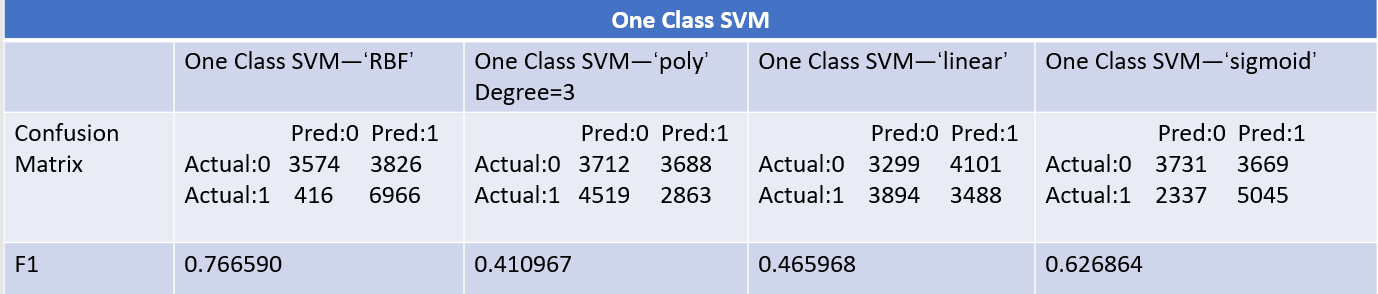


Table 2

I also compared the result of One Class SVM with RBF kernel with other unsupervised models used for anomaly detection, and RBF still has the best performance.

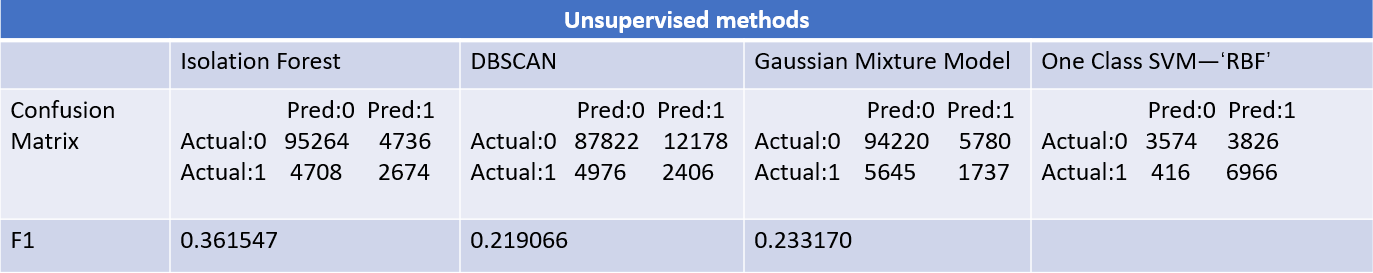


Table 3

In next step, I used models, such as XGboost, ANN, Random forest, and KNN, to train the transformed dataset. All models have gone through 5 folds cross-validation to find the optimal feature combination. The model comparison results are shown in Table 4.

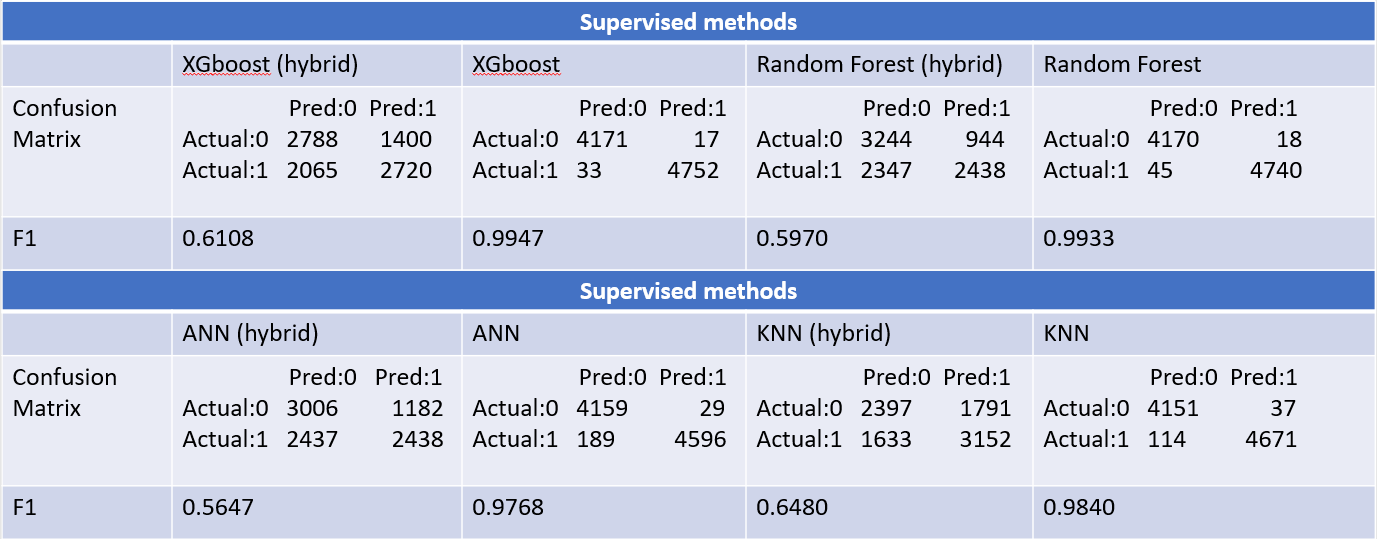


Table 4

The comparison results shows that the model fits extremely well in the original dataset and performs poorly in the transformed dataset. The reason for this result may be that the original dataset is linear separable and introduce too much noise during the kernelizing process.

In credit card fraud detection task, we also care about the important features. Financial institutions can develop measures to identify potential fraudulent transactions timely as well as develop measures to prevent credit card fraud. Table 5 lists all the features offer more information gain for the XGboost model. All the colored features are created by the feature engineering. The results prove that credit card fraud occurs at night most frequently.

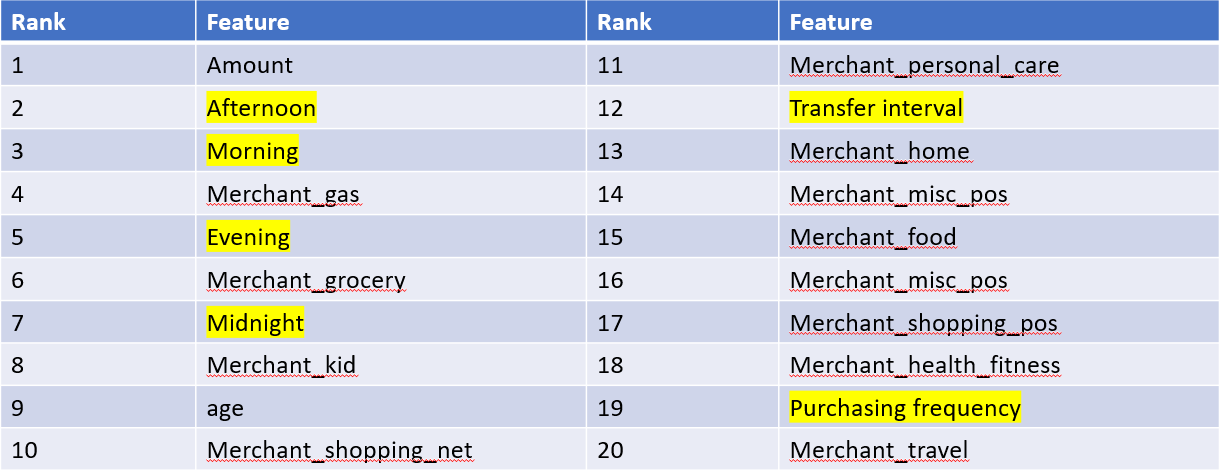


Table 5

6. An interesting finding

An interesting phenomenon is that when I did the sampling first and then the kernelized the dataset, I got a model with half of misclassifications. However, when I adjusted the order of sampling and kernelizing, I could get a smaller training dataset and a model that performed significantly better than the previous one (results shown in table 5).

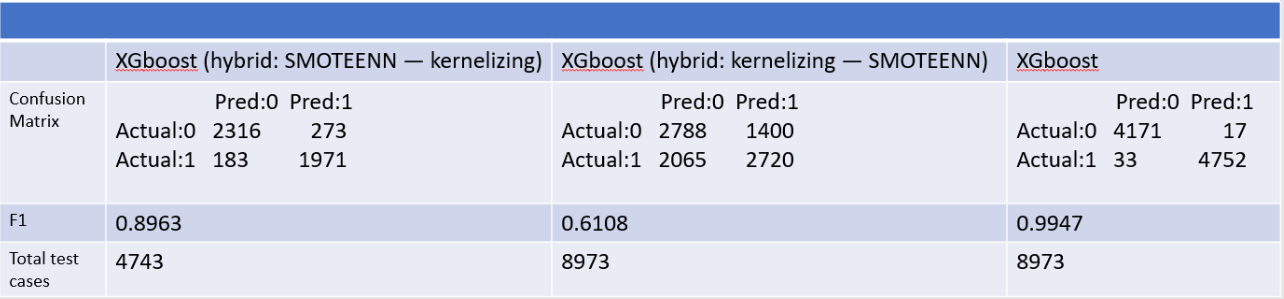


Table 6

I did some research on this finding. One possible explanation is that the ENN process in the SMOTEENN sampling method discards the noise introduced during the kernelization process and oversampling process. Therefore, the number of observations in the generated dataset decreases and the model performance improves.

7. Discussion

One purpose of this project is to prove that kernel trick is able to improve the performance of the credit card fraud detection model. Unfortunately, the hybrid method did not fit on the dataset quite well. kernel trick is a powerful tool to deal with linearly non-separable data. In addition, kernel trick can be not only used to transform numerical data but also on strings and graphs. I also want to discuss the importance of feature engineering. From table 5, we can see that one-third of the important features are created by feature engineering. Appling feature engineering in model implementation helps maximize the utilization of information in the data. Finally, it is necessary to pay attention to the sequence of the steps of model building process. In this project, sampling the kernelized data can filter out the noise introduced by kernelization to improve model performance.

Reference

1. Melo-Acosta, G. E., Duitama-Munoz, F., &amp; Arias-Londono, J. D. (2017). Fraud detection in big data using supervised and semi-supervised Learning Techniques. 2017 IEEE Colombian Conference on Communications and Computing (COLCOM).

2. Sisodia, D. S., Reddy, N. K., &amp; Bhandari, S. (2017). Performance evaluation of class balancing techniques for credit card fraud detection. 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI).

3. Zhang, X., Han, Y., Xu, W., &amp; Wang, Q. (2021). Hoba: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture. Information Sciences, 557, 302–316.

4. Correa Bahnsen, A., Aouada, D., Stojanovic, A., &amp; Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. Expert Systems with Applications, 51, 134–142.

5. Kewei, X., Peng, B., Jiang, Y., &amp; Lu, T. (2021). A hybrid deep learning model for online fraud detection. 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE).

6. Zhou, H., Sun, G., Fu, S., Wang, L., Hu, J., &amp; Gao, Y. (2021). Internet financial fraud detection based on a distributed big data approach with node2vec. IEEE Access, 9, 43378–43386.

7. Forough, J., &amp; Momtazi, S. (2021). Ensemble of deep sequential models for credit card fraud detection. Applied Soft Computing, 99, 106883.

8. Jing, R., Tian, H., Zhou, G., Zhang, X., Zheng, X., &amp; Zeng, D. D. (2021). A GNN-based few-shot learning model on the credit card fraud detection. 2021 IEEE 1st International Conference on Digital Twins and Parallel Intelligence (DTPI).

9. Ileberi, E., Sun, Y., &amp; Wang, Z. (2022). A machine learning based credit card fraud detection using the GA algorithm for Feature Selection. Journal of Big Data, 9(1).

10. Caroline Cynthia, P., &amp; Thomas George, S. (2020). An outlier detection approach on credit card fraud detection using Machine Learning: A Comparative Analysis on supervised and unsupervised learning. Intelligence in Big Data Technologies—Beyond the Hype, 125–135.