Lab 1 - Credit Card Fraud Data Wrangling and EDA, Stat 415, Spring 2023

Wing Yan Yau

13th April 2023

1 Introduction

This report aims to investigate the key factors that contribute to credit card fraud. The focus of this report is on data preparation and analysis, while the findings can be applied to construct predictive models for detecting credit card fraud. Consequently, we will be considering the fraud status as the dependent variable throughout this report, with the awareness that it would function as the response variable in a predictive model developed in the future.

2 The Data

2.1 Data Cleaning

There is one pair of duplicated columns i n the data set, namely the 'accountNumber' column and the 'customerId' column, where both of them are the identifying number for the customer. Hence one of them is deleted from the data set (here we chose to delete the 'customerId' column).

A total of 6 columns are with entirely missing data, which include 'echoBuffer', 'merchantCity', 'merchantState', 'merchantZip', 'posOnPremises' and 'recurringAuthInd', and they are deleted due to an obvious reason thata we cannot get any information from them.

Besides, there are also some missing data for individual rows that is different from the above scenario. The columns with this problem are 'acqCountry', 'merchantCountryCode', 'posEntryMode', 'posConditionCode', 'transactionType', where the maximum amount of rows with missing data is less than 5000, and that of those with also 'isFraud' as True is 269.

On the numerical variables, note that 'creditLimit' and 'availableMoney' actually implies that value of 'currentBalance' (more precisely, 'currentBalance' is 'creditLimit' - 'availableMoney'). This nice linear-combinational relationship among them supports our decision to get rid of the 'currentBalance' column from our data set.

In terms of eliminating the outliers among the numerical variables {'creditLimit', 'availableMoney', 'transactionAmount', 'currentBalance'} in the data set, we examined the magnitudes of the numerical data individually and cross-variables-wise. In the former perspective, there is no outliers as all of them are bounded within 0 to 50000, which is the possible range of credit limit. When comparing any 2 of them, there is no clear 'cluster v.s. outliers' pattern either. See figure 1.

2.2 Data Preparation

Also, for consistency, we unified the data type used to represent the time variables { 'currentExpDate', 'transactionDateTime', 'accountOpenDate', 'dateOfLastAddressChange'}, which is the class of 'pandas._libs.tslibs.timestamps.Timestamp'.

Lastly, to better capture the information provided by columns 'cardCVV' and 'enteredCVV', we added a new column 'disCVV' to the data set which represents the difference between the corresponding 'cardCVV' and 'enteredCVV' value in a binary string of length 3. For the i^{th} digit of 'disCVV', it would be 1 if the i^{th} digits of 'cardCVV' and 'enteredCVV' are different, 0 otherwise. For example, if the 'cardCVV' is 123 and the 'enteredCVV' is 122, the 'disCVV' would be 001. It is believed that the position and number of wrongly entered digit of CVV may provided some useful information in determining if the transaction is an unauthorized use of an individual's accounts, i.e. is fraudulent.

3 Findings

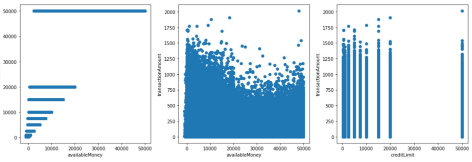


Figure 1: Scatter plot with pairwise comparison among numerical variables

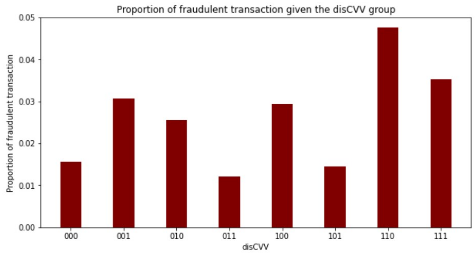


Figure 2: Refer back to session 2.2 to recall the definition of disCVV

In most of the cases, the fraud rate where CVV entered does not match with the correct CVV is higher than that where the entered CVV is correct. The difference in fraud rate is quite significant in terms of their ratio, where fraud rate of each subclass with wrong entered CVV is at least around twice as high as that of the class with correct entered CVV. In a more holistic picture where we only consider a binary split of classes, namely one with correct entered CVV and the other one with wrong entered CVV, the difference in fraud rate is 0.0157 and 0.0284 respectively, also implying a more possible situation to find fraudulent transaction when the entered CVV is wrong. This is a clear evidence that 'disCVV' could provide us with perhaps useful information when deciding if a transaction is fraudulent or not. One point to not that would be also the fraud rate ratio make them distinctive, the absolute difference is actually very small since the proportion of fraudulent transactions among all is small.

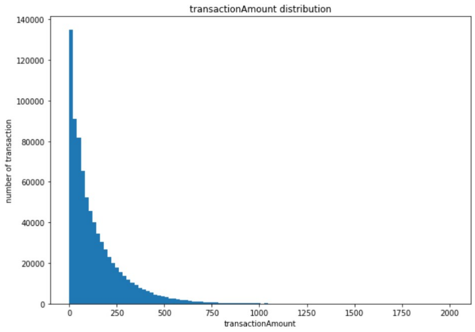


Figure 3:

The distribution of 'transactionAmount' looks like an exponential distribution. Such nice approximation give an easy way to formulate the descriptions on the supposed-to-be distribution of transaction conditioned on some predictors. A output distribution that deviate too much compared to the exponential shape might hint the existence of abnormality.

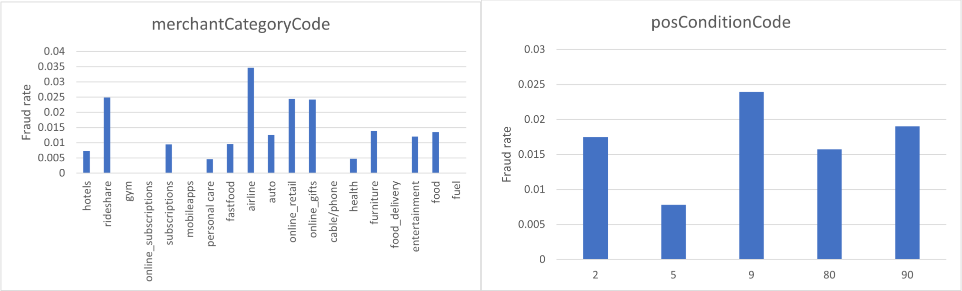


Figure 4:

The above figures provide a glimpse of how different categorical predictors might relates to the fraudulent rate. We can see that among classes of a categorical predictor, the fraudulent rate is not usually uniform, but with higher value for some of the classes and lower or even 0 for the others. This should tell us a useful predictive model might probably include a structure to distinguish the classes from one another among categorical variables and base on this information to predict how likely would a transaction be fraudulent.

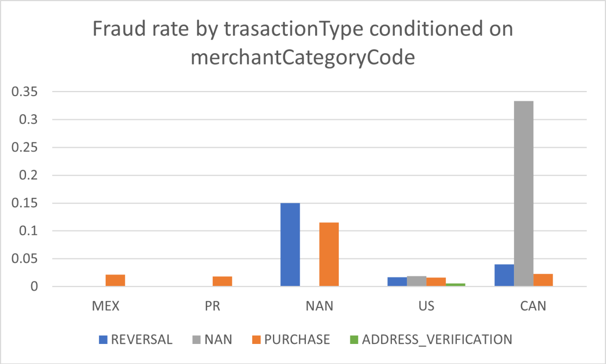
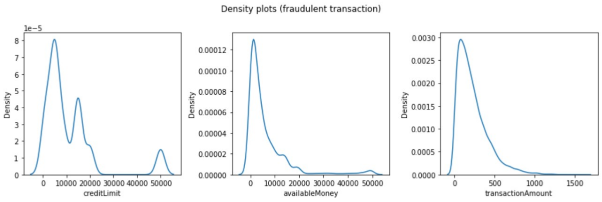
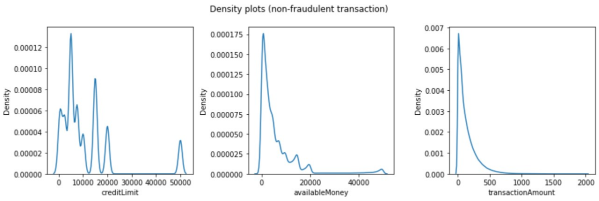


Figure 5: The NAN group corresponds to those with merchantCategoryType missing. The NAN bars correspond to those with transactionType missing.

Having a deeper look into the relationship between Fraudulent rate and 'transactionType' conditioned on 'merchantCategoryCode', we can see that for France and Mexico, transaction of purchase is the only significant transaction type to have fraudulent ones among. While both the case for US and Canada are distinctly different, and that only in US is address verification a non-negligible transaction type in terms of fraudulent proportion. This disparate patterns implies that the predictive model might need the ability to adjust its reaction when given a certain transaction type based on the 'merchantCategoryCode' variable.





For the distribution of fraudulent and non-fraudulent transaction on different numerical variables in the data set, our first observation is that they are very similar in the rough shape but things tend to get a lot more spiky for the density plots of non-fraudulent transactions, especially for the case of 'creditLimit'. This could be a result of the relatively small sample size of fraudulent transactions.

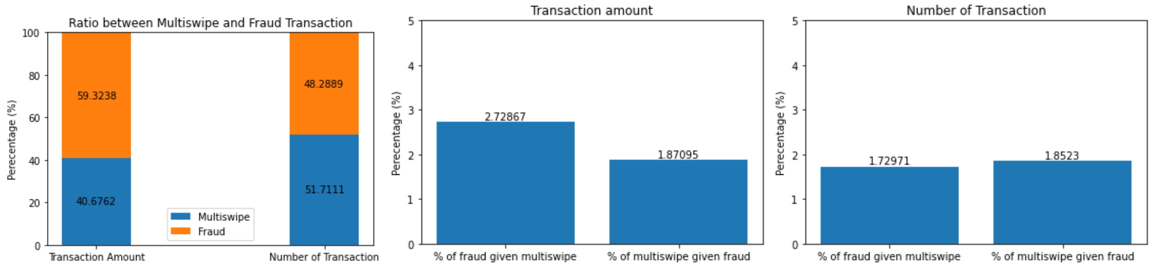


Figure 6:

Lastly, we observed that there is a disparity between the relationships among multi-swap transaction and fraudulent transaction between the number of transaction and the transaction amount. Before mentioning the finding, we first describe the definition of 'multi-swap transaction' : a transaction that made using the same account with transaction amount being the same as the previous transaction within 5 minutes. The different choices of time interval length may hold different interpretation of 'multi-swap', but since upon investigation, the fraudulent rate and multi-swap rate do not vary much when ranging the interval from 10 days to 5 minutes, we chose the one with the strictest definition to avoid outliers as best. It is clear that although the number of fraudulent transaction is less than that of multi-swap transaction, the transaction amounts of the two are in an opposite relation, and the proportion of fraudulent transaction among multi-swap transaction is higher in terms of transaction amount (2.73%) than that of number of transaction (1.73%). This finding might indicate possible relative severeness of fraud involving multi-swap and that we might want to increase the sensibility of our model towards this phenomenon.

4 Discussion

A potential problem with this data set in putting to train a machine learning model is class imbalance, where among around 780000 data points, only around 12000 (approximately 1.5%) of them are fraudulent. Yet if we consider real life data, according to the 7th report on card fraud released by European Central Bank, from 2015 to 2021, the value of fraud as a share of the value of transactions and of the volume of transactions are around 0.030%-0.050% and 0.020%-0.025%, respectively. The extent of class imbalance observed in the data set might be reflective of the real-world distribution of fraudulent transactions. Therefore, training a machine learning model on it could capture the general characteristics and patterns associated with fraud. Hence, dealing with under-sampling is not needed for this data set.