Springboard - Data Science Track
Jacqueline (Xuan) Guo

First Capstone Project – Milestone Report

May 27, 2020

Title:

Improve the efficacy of fraudulent transaction alerts

Introduction:

When doing some transactions with a debit card or credit card, we often face a situation that the card cannot work normally. Imagine there is a bunch of people waiting behind you and you tried your card once, twice, three times, but it's still not working. This is a very embarrassing situation when fraud detection went wrong. Therefore, improving the accuracy of fraudulent transaction alerts would help to avoid this embarrassing situation. Also, reduce the risk when the detection undiscovered leading people to lose money.

Companies such as JP Morgan Chase, Bank of America, and Credit Card Processing would love to use the improved model to predict fraud activities. If this project is successful, it will help hundreds of thousands of businesses reduce their fraud losses and increase their revenue. Customers may also want to choose banks or financial companies with high accuracy fraud detection.

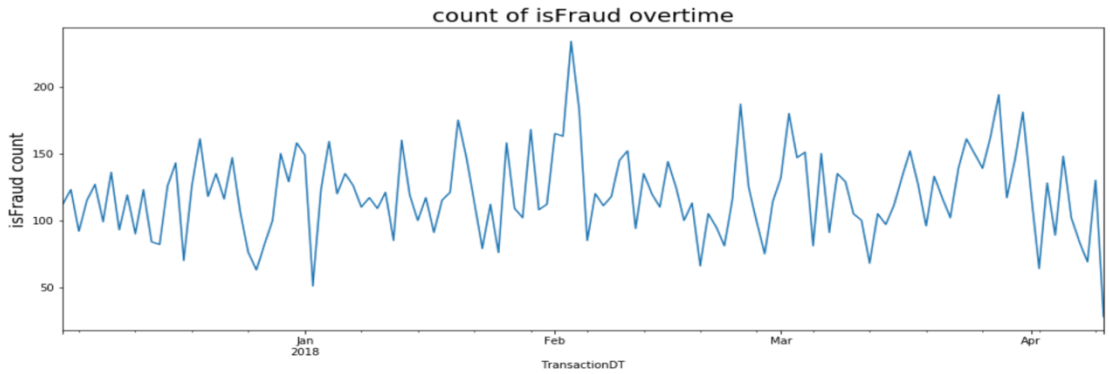
Dataset:

The data comes from Kaggle Competition - IEEE-CIS Fraud Detection**,** Kaggle's data comes from Vesta's real-world e-commerce transactions and contains a wide range of features from device type to product features. The dataset contains 20 kinds of features in total such as *DeviceType* (transaction device type), *addr* (where this transaction was made), and *card* (what kind of card the customer used). Then using those features to analyze the relationship between features and labeled fraud and build machine learning models to improve better accuracy fraud prediction.

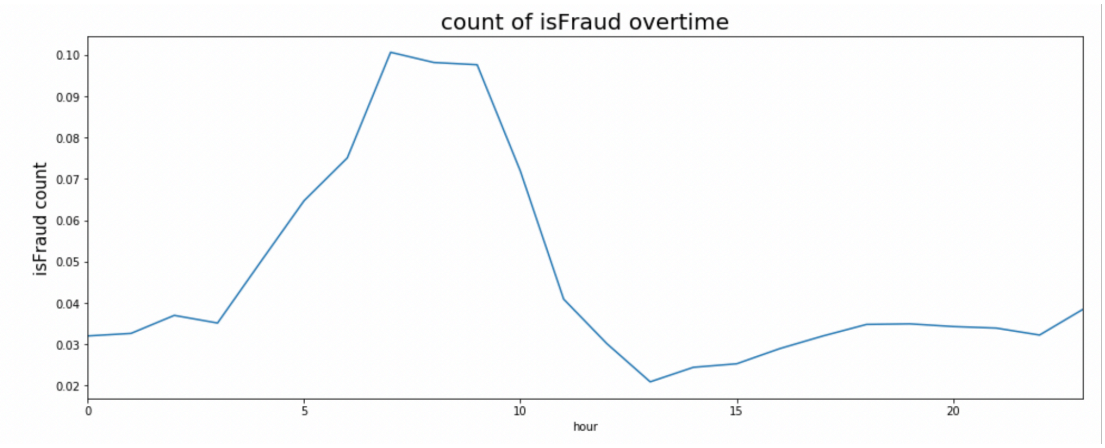
Preprocessing

Since there are many available features in our samples, I decide to do feature selection first to see if there are some features have strong correlations to fraud results.

I. I've noticed that there are features about the date and time in the transaction data set. I want to measure if the fraud action is somehow related to the time of transaction. Therefore, I draw a fraud count graph with daytime and try to find if there's any obvious trends:

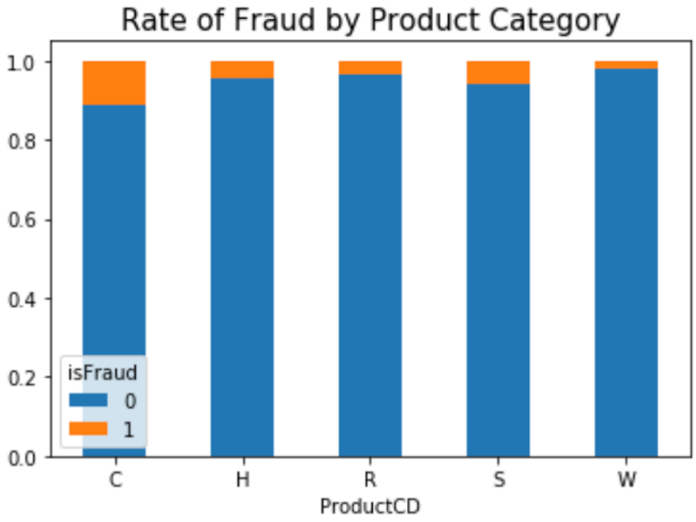


However, according to the figure above: no significant pattern visually based on days. Thus, I changed the unit to hours:



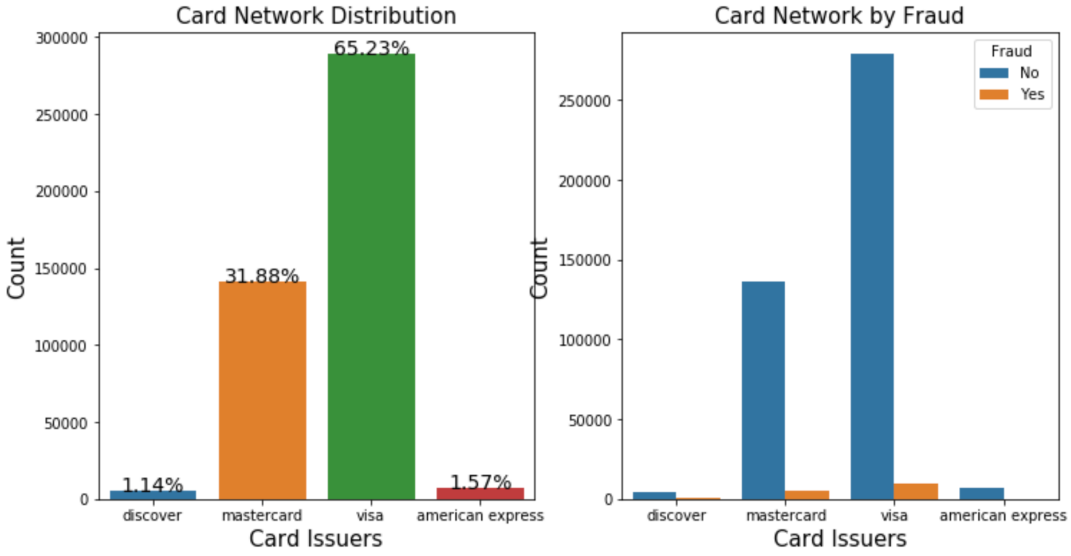
Then we can notice that there's a sharp increase between 5 am to 10 am. Thus we could conclude that the fraud action seems to peak during early morning.

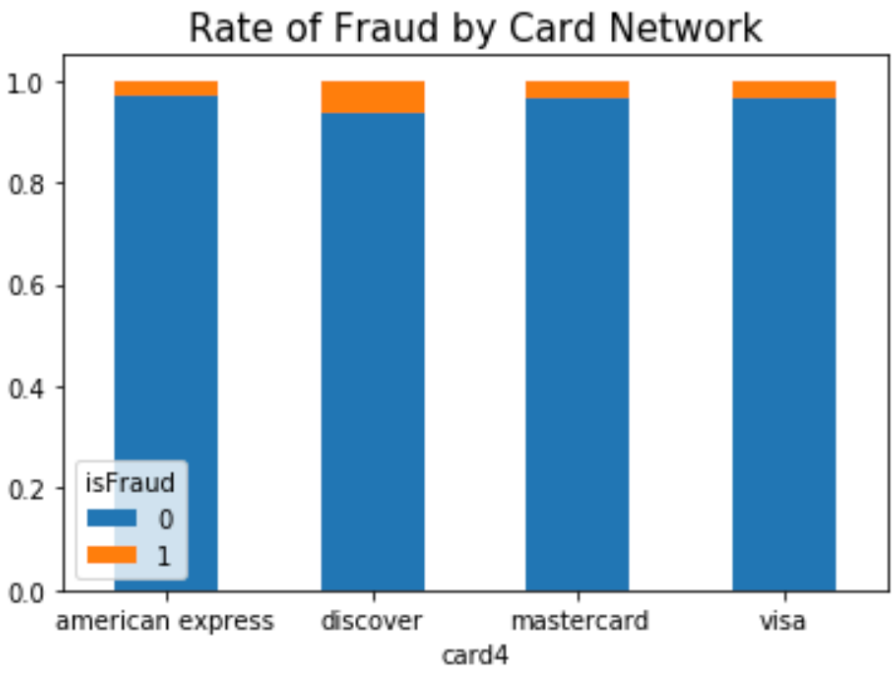
II. In addition, I've noticed that different product type may cause more or less frauds as the following figure shows:



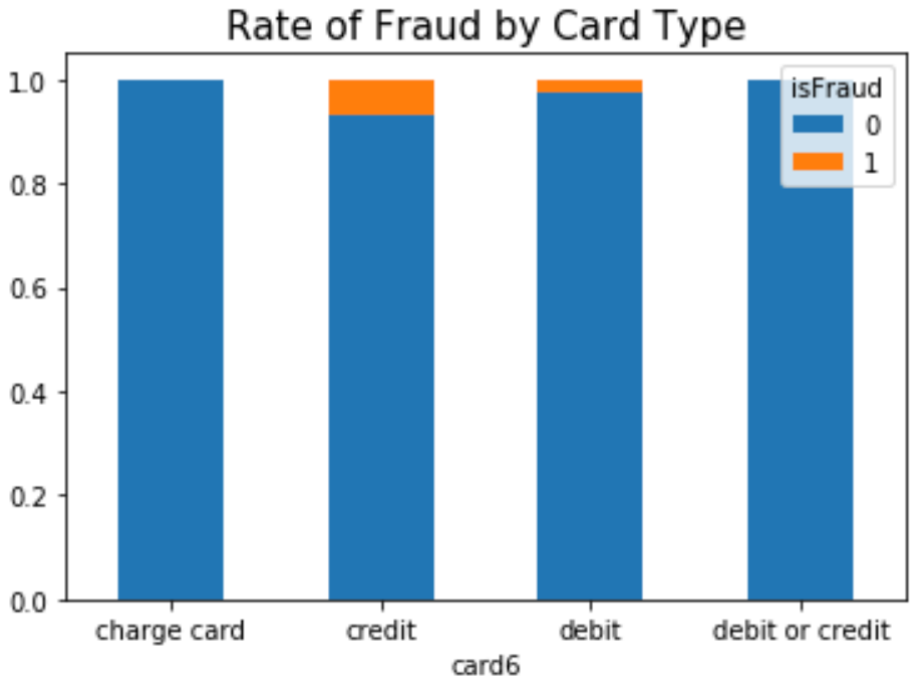
As the figure shows, product C takes up 67.5% of fraud cases for transactions that have identity. It also has highest rate of fraud -- 12%, more than double any other class of product.

III. I've reduced some of the card features and only keep card issuers and card types as the elements of our analysis. I used data visualization to figure out the relationship between card issuers, card types and fraud actions as the following figures show:





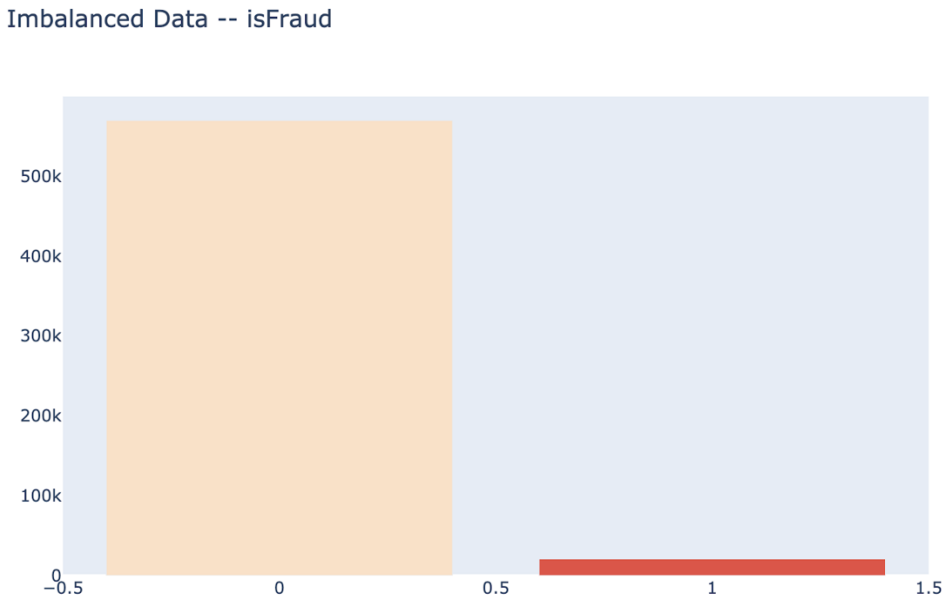
As we can see, although visa has the highest counts of "isfraud", this is because visa is most used type of card. If we divide it by its population then we can see that American Express have a lower fraud rate.



The above figure indicates that credit card tend to have a higher fraud rate.

Limitations:

I. I checked if the labeled data is reasonable for us to do data analysis. However, I found the labeled data "isFraud" are imbalance data as the following graph shown:



II. I found some of the features is hard to analyze because we could not know the feature meaning due to security reasons. For example, the card information, 4 of them are numeric data and we cannot know what's the meaning of them. What I did is drop them because I think it's probably some personal information such as card expiration date or card number. Thus, those kinds of information would not influence our analysis. I also did the Pearson correlation coefficient between those features to check if it's necessary to be included in our analysis.

III. There are some outliers in the data set, but this data is from the real-world collection. It's hard to determine if the outliers are due to real world phenomenon or some human mistakes. Therefore, it's hard to count outliers into our analysis. For instance, the distance between billing address and shipping address of some credit card transactions are super far, then we could not say it's an outlier or something else. Since it's possibly that the consumer traveling out of their country and made the purchase. Thus, how to deal with the outliers is hard to determine.

Models:

Deliverables:

1. Code notebooks
2. Report on the capstone project
3. Presentation on the capstone project