**Credit Card Fraud Detection**

Jyoti Sahu

Bellevue University

DSC680- Applied Data Science

June 28, 2022

**Business Problem**

My objective for this project is to assist in preventing, reducing, or avoiding credit card fraud. fraud is typically only discovered after it has already occurred, it is crucial for me to evaluate and visualize the Credit Card Transactions datasets. However, fraud detection is the best option for eliminating it from the environment and averting a recurrence in the event that they are unable to prevent it in time.

**Background/History**

The use of credit card to commit fraud is known as credit card fraud. Cybersecurity is becoming increasingly significant in our daily lives. When addressing digital life security, the main issue is identifying anomalous behavior. Many consumers prefer using credit cards when they make purchases or do online transactions. Occasionally, we can make purchases even when we don't have the cash on hand thanks to credit card credit limitations. However, cyber thieves take use of these advantages.

**Data Explanation (Data Prep/Data Dictionary/etc)**

I have used the below dataset in my project

<https://www.kaggle.com/adityakadiwal/credit-card-fraudulent-transactions>

Here are the columns and definition of the file

'DOMAIN', 'STATE', 'ZIPCODE', 'TIME1', 'TIME2', 'VIS1', 'VIS2', 'XRN1',

'XRN2', 'XRN3', 'XRN4', 'XRN5', 'VAR1', 'VAR2', 'VAR3', 'VAR4', 'VAR5',

'TRN\_AMT', 'TOTAL\_TRN\_AMT', 'TRN\_TYPE'

* *DOMAIN*: The domain name of the customer's email address that was used for the transaction (Masked)
* *STATE*: The state code of the customer's location.
* *ZIPCODE*: The zip code of the customer's location.
* *TIME1*: Hour feature #1 of the transaction.
* *TIME2*: Hour feature #2 of the transaction.
* *VIS1*: feature #1 for feature VIS.
* *VIS2*: feature #2 for feature VIS.
* *XRN1*: feature #1 for feature XRN.
* *XRN2*: feature #2 for feature XRN.
* *XRN3*: feature #3 for feature XRN.
* *XRN4*: feature #4 for feature XRN.
* *XRN5*: feature #5 for feature XRN.
* *VAR1*: feature #1 for feature VAR.
* *VAR2*: feature #2 for feature VAR.
* *VAR3*: feature #3 for feature VAR.
* *VAR4*: feature #4 for feature VAR.
* *VAR5*: feature #5 for feature VAR.
* *TRN\_AMT*: The transaction amount.
* *TOTAL\_TRN\_AMT*: The total transaction amount.
* *TRN\_TYPE*: type of transaction whether FRAUD or LEGIT

**Methods**

I am going to two execute exercises here.

**EDA-** In this part I am going to see the distribution of feature variables and derive the correlation matrix.

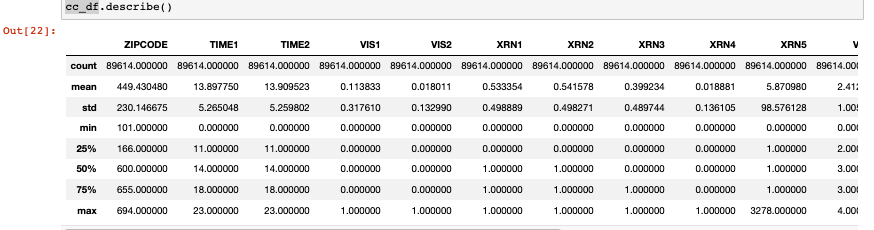
**Model development-** In this portion I will be decision tree model to determine the accuracy of test and train data. After this I might use K-means or PCA to predict the credit card fraud

**Analysis**

I examined the data. The total dataset's Total Transaction Amount count is 89614.

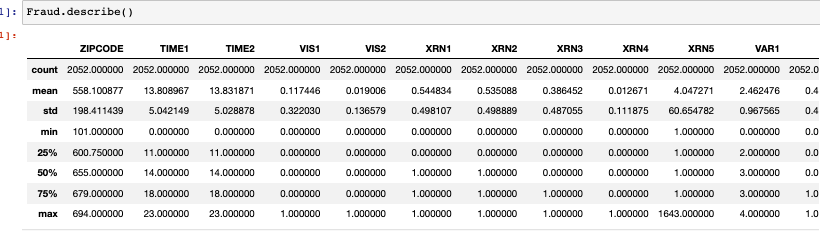
For the entire dataset, 25% of the Total Transaction Amount is equal to $12.95.

For the entire dataset, the total transaction amount's standard deviation is 14.17.



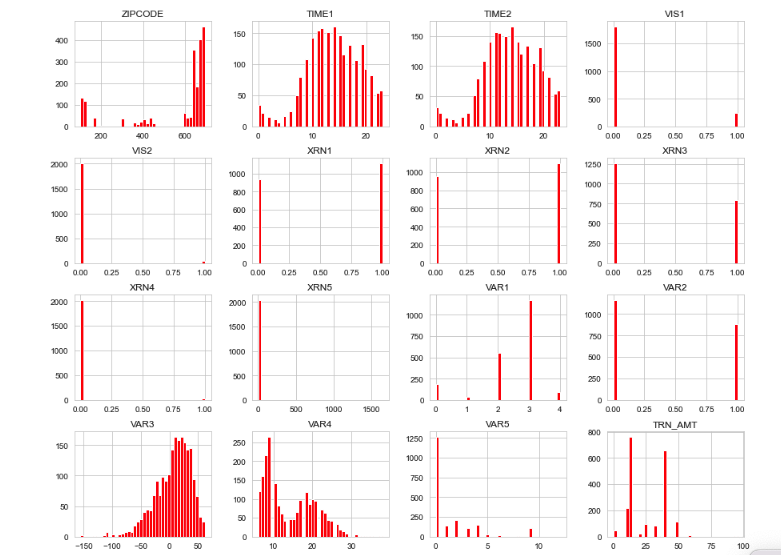
I have separated the fraud and legit data to check the distribution.

2052 out of the total transaction amount are fraudulent transactions. The amount of the whole transaction at the very least that is fraudulent is 0. The Total Transaction Amount's fraud-related standard deviation is 14.42.



There have been 87562 legal transactions. There were 2052 fraudulent transactions. I found 2.34 percent of transactions are fraudulent.

After plotting the distribution of the variable using the Kernel density plot it was observed that for legitimate and fraud transactions var 1 and var 2, time1 and time2 are the variables which can be considered to detect the fraud or legit transaction.

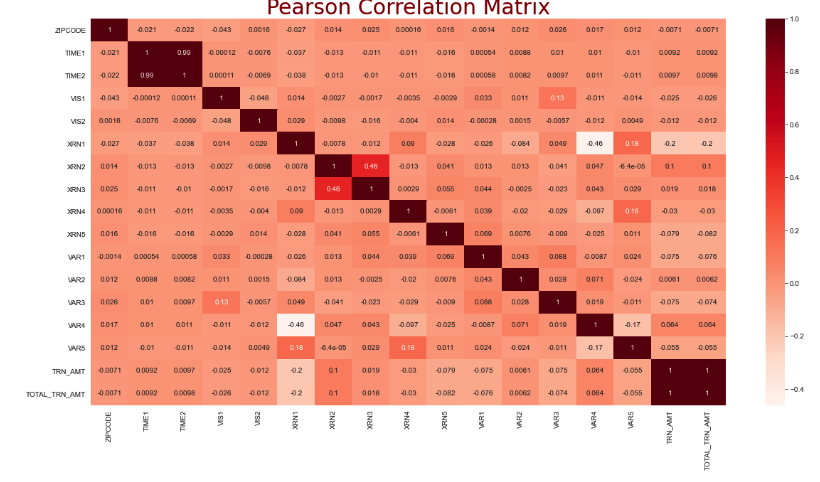


I did apply person coefficient matrix to determine the correlation between variables and my observations are

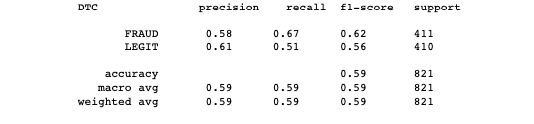
TIME1 and TIME2 have a strong correlation of 0.99, which is close to 1, indicating a significant positive link.

- The correlation between XRN2 and XRN3 is 0.46, which is quite close to 1, indicating a favorable link.

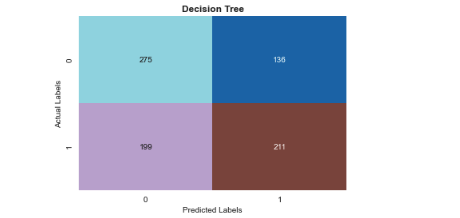
- The correlation between XRN1 and VAR4 is -0.46, which is close to zero and indicates that there is no association.



Moving to Model development, I have used decision tree model to detect the fraud. I have taken the fraud data into consideration and splited the data in 80-20 ratio test and train data. The accuracy of my model came as 60%



In the confusion matrix the first row stands for positive, while the second row for negative. I therefore have 275 real positives and 136 erroneous positives. Accordingly, out of 411, I have 275 transactions that are successfully classed as normal, and 136 transactions that were fraudulently classified as normal.



**Conclusion**

I have looked into the data, ensuring that it is balanced, visualizing the features, and understanding how they relate to one another. Our recent results for detecting credit card theft showed a 60% accuracy rate. Given that our data was weighted in favor of one class, this number shouldn't come as a surprise. Our model is not overfitted, which is a positive finding from the confusion matrix.

**Assumptions and Limitations**

Dataset is too small and contains duplicates. My model will be useful assuming after I do the EDA and massaging the data. A time series data can be helpful if the frauds are spiking more during holidays and what other factors involved. A considerable number of values being eliminated will also have an impact on our model and have drawbacks of its own. Our accuracy may be overly high as a result of the fact that we will be drastically reducing the quantity of training points. The only requirements are historical data and an appropriate algorithm that can better fit our data. Overall, I have demonstrated that it is possible to accurately detect credit card fraud even though the model may be prone to overfitting. We can alert customers in a short time for potential fraud with integration of this technology with transaction data. Fraud pattern may change faster than model

**Challenges**

I have faced multiple challenges including Duplicate data. The dataset I worked in not real time hence unavailability of real time dataset is a challenge to measure the accuracy of model. As we see in the EDA Data is highly skewed as only very minimal fraud transaction data so my model can be biased. I used jupyter notebook but it dosed not come up with all the packages hence faced many Pythons package issues. The data source We have used in this case study does not say how accurate the data is and also verification is not defined. Kaggle dataset may be too small to predict.

**Future Uses/Additional Applications**

Numerous legitimate use cases in the actual world exist for this concept. For instance, a bank may adopt a similar strategy and cut costs by automating the process of trying to find fraud. By providing an additional degree of security for lost cards and stolen products, our concept might also help the user save a significant amount of time and money.

I will probably test multiple models and test the accuracy of best fit model along with PCA /K-means.

**Recommendations**

Recommend to use the best fit model based on testing the accuracy of models.

**Implementation Plan**

The implementation plan will contain business problem, EDA, Model development, conclusion and recommendation.

**Ethical Assessment**

Credit card information is twisted and unbalanced. Due to GDPR and privacy concerns, financial institutions are not permitted to divulge their credit card information. Therefore, tests were conducted while taking this issue into consideration using federated learning techniques. With this technique, the participants' local data was trained. Credit card companies are aware of the security dangers associated with online transactions, as well as the available solutions that might reduce the amount of online fraud transactions. The issue is that credit card firms are under little pressure to improve their online security because victims of identity fraud caused by third party debt purchasers are bearing the losses.