# IEEE-CIS Fraud Detection

## Abstract

Since the first day of innovation, fraud detection has always been one of the most important duties of ML algorithms. In this project, the important part is data manipulation because of the number of columns and understanding better of situation each of them. Data manipulation is a job that each person does specifically and there no right or wrong exists in it, but some points can make it better,

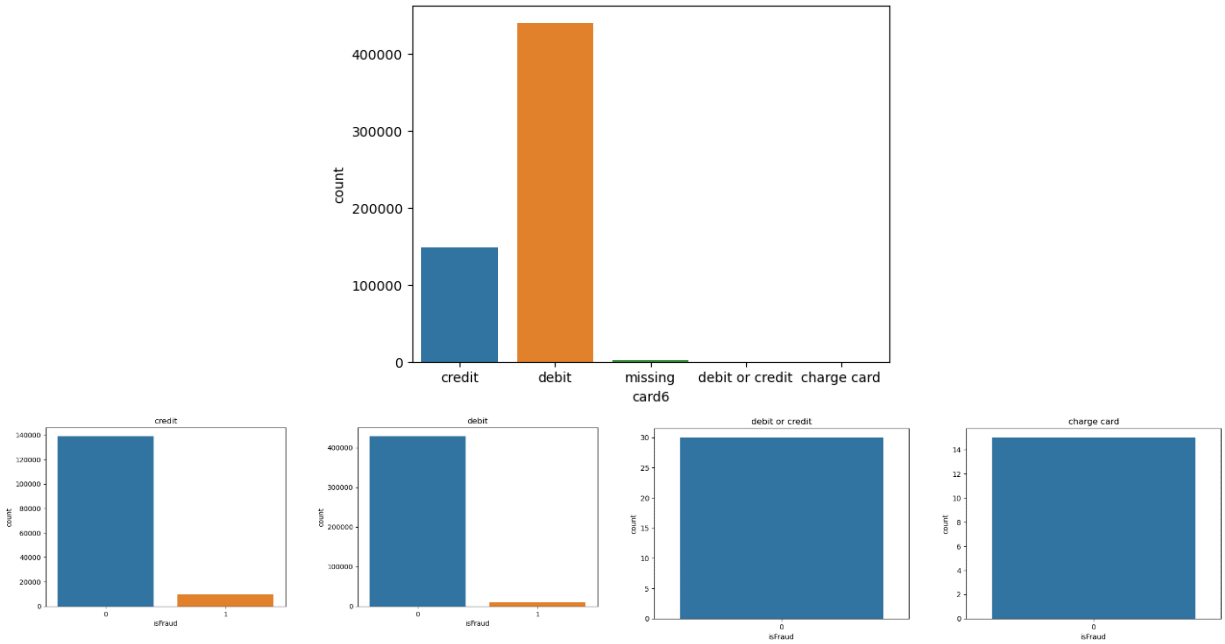
- Finding the distribution of the features (whole or with grouping by another feature)
  it helps to use the better solution to fill empty or invalid fields
- Visualizer data and find the related features
  remove the most related feature to avoid bias is one of activity that exists in the data cleaning part but separating these two activities almost be difficult

After checking the features, we need to focus on the target values especially when we talk about Fraud detection systems (the sample of the fraud is very less) and checking objective features to sure all values have a sample of fraud or not, it is very helpful because we can decrease the complicity in our dataset and help to the machine to learn more smoothly.

The unbalanced dataset's evaluation value can be defined as the f1-score because accuracy can't be an acceptable value. But in this project, as you will see, the author chooses accuracy as the main evaluation value.

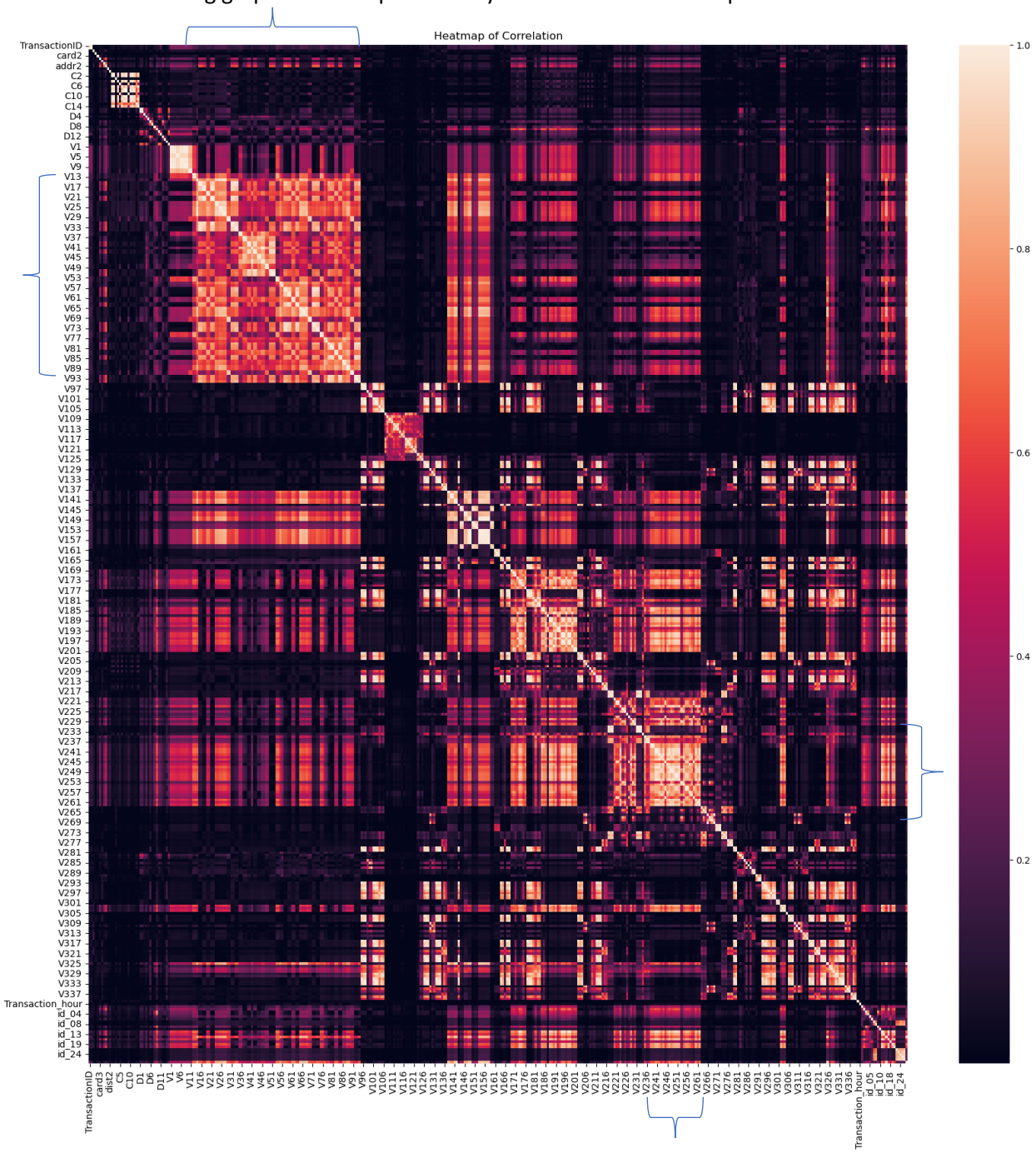
## Data Analyze

The analysis and visualization of this project are its strong point of it and based on the above explanation it is something completely in the author's hands.




In the specific feature, two of the four possible values don't affect the target value, so we will drop the related sample of them.

One of the interesting graphs that are provided by the author is a heatmap



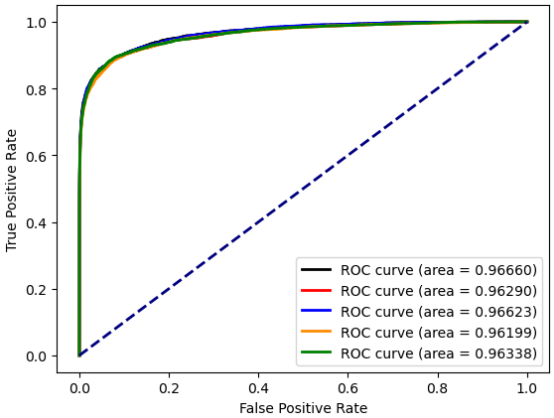

In two different spots, we can see the strong relationship between the features.

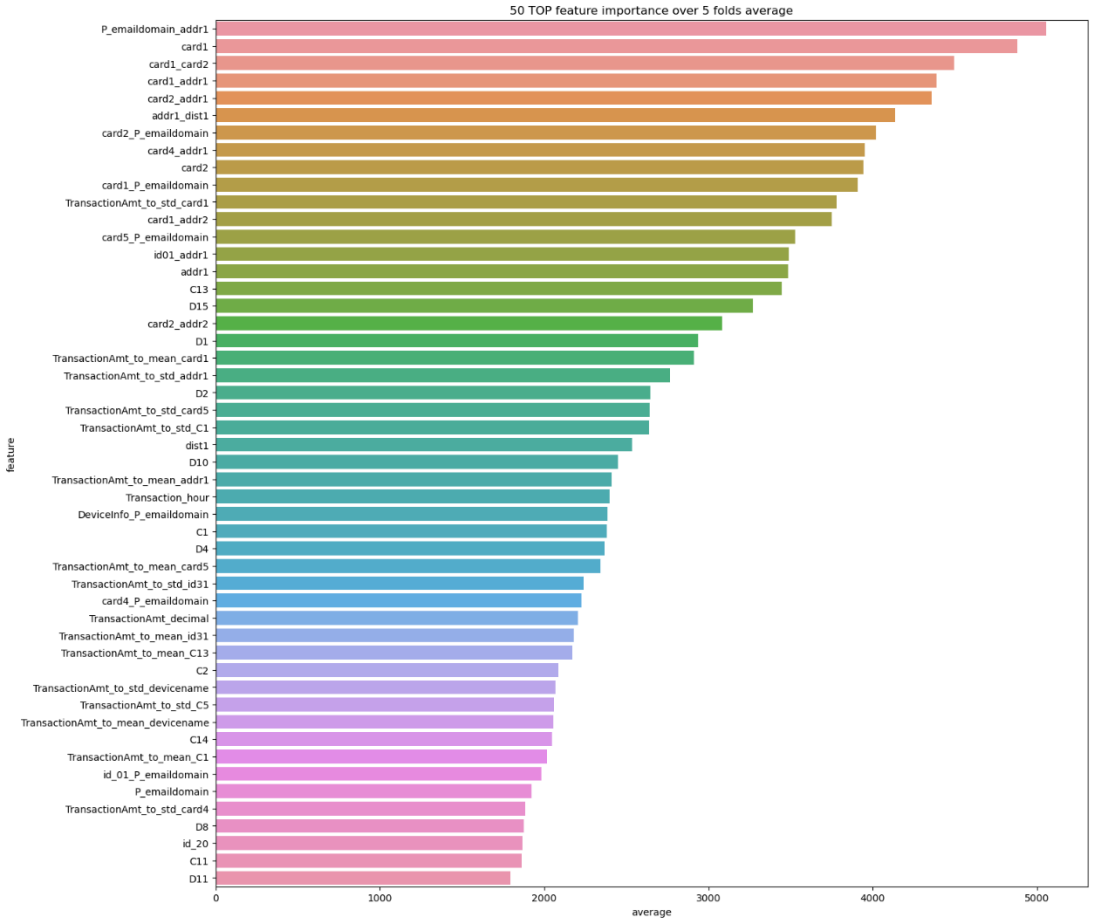
As checking all features has a null value, in this project author prefers not to spend much time to fill them perfectly just uses the simple method, filling the null value with a dummy value that is separate from other values.
For categorical: ‘missing’
For numerical: ‘-1’

For data preparation, one important point and task is not done by the author and it uses a pipeline to avoid any manual mistake in changing the value of the train and test. This pipeline will use on the production side when the new value is coming.

## ML algorithms

The author used the 'LGBMClassifier', modifies its hyperparameter and run it 5 different times, and got results to show in the ROC graph



## My exploration

I've followed several other machine-learning algorithms in binary classification and I've gotten below results for default hyperparameters value. It needed more time to spend on the normal machine to run and tunned parameters, I'd written the program but I didn't run it.

```
Classification report RandomForest:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99     57049
           1       0.95      0.48      0.63      2005

    accuracy                           0.98     59054
   macro avg       0.96      0.74      0.81     59054
weighted avg       0.98      0.98      0.98     59054


Classification report LogisticRegression:
              precision    recall  f1-score   support

           0       0.97      1.00      0.98     57049
           1       0.16      0.00      0.00      2005

    accuracy                           0.97     59054
   macro avg       0.56      0.50      0.49     59054
weighted avg       0.94      0.97      0.95     59054
```

## Summary

In general, to improve the result of this project, some important points must be followed,

- Data visualization of features based on each target value
- Use another technic to separate train and validation datasets to keep them with the same distribution
- Use the different functions in evaluation in the train time and find the best of them

Other jobs can do but it needs time and have the complete results of the above tries and errors.