REPORT ON PROJECT

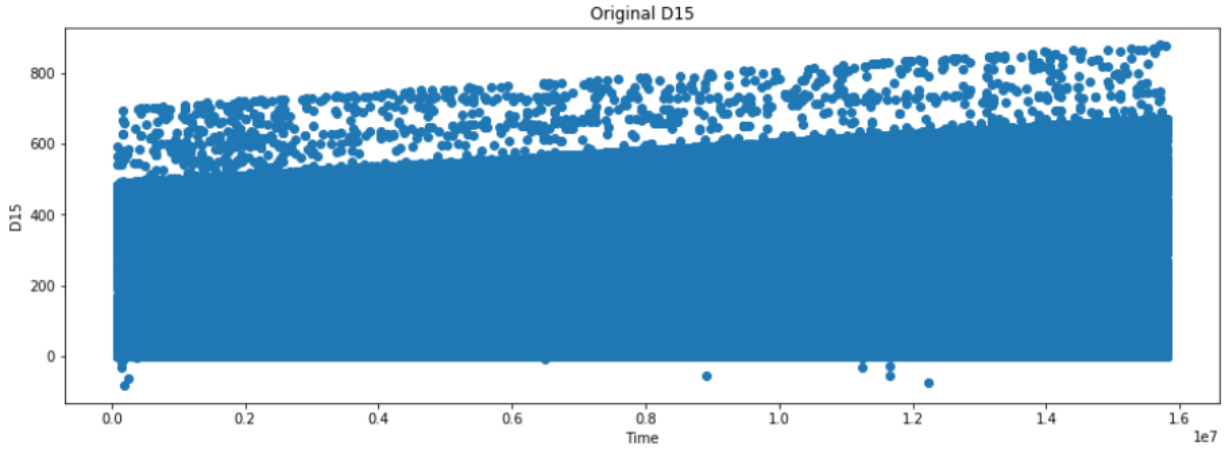
ABSTRACT:

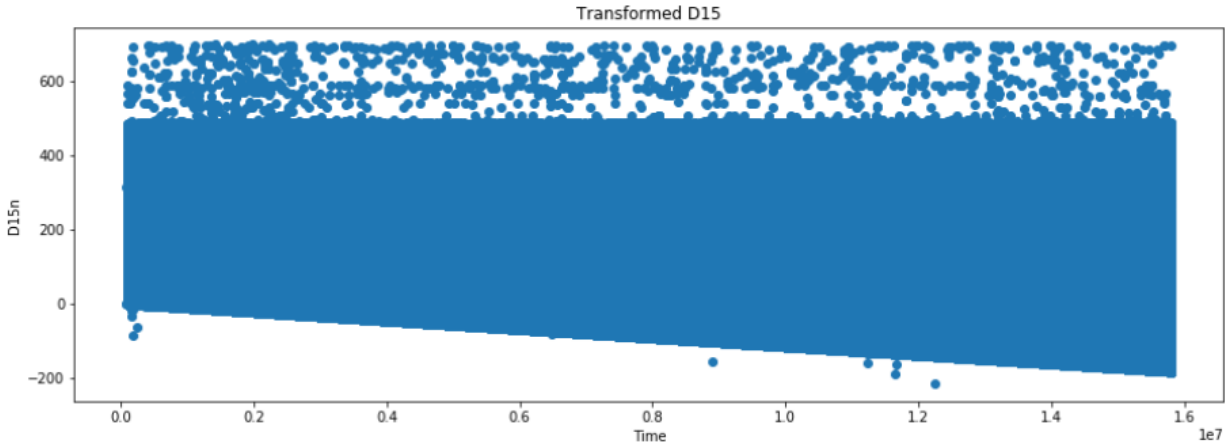
Fraud detection is a task of determining fraudulent transaction from the transaction data of a credit card using machine learning. It is a binary classification task(1=Fraud, 0=Not Fraud). Here in this competition vista company provided the dataset on customer transactions and their ids. The task was to predict if a card or new customer transaction is fraudulent from the model that is trained. There were total 590540 transactions in the training set and 506691 transactions in the test set. The probability of a card being fraud is to be submitted to be scored. The metric of evaluation used is area under the Roc curve. Six months of training data is used to train the model.

METHODOLOGY:

To find the redundant columns of V, correlation analysis was done. The V columns with similar nan structure or similar nan count were grouped together to check the correlation between each of the column. Within the similar nan structure there were group that has a correlation>0.75 which were replaced by a single column in the group. The reduced columns retained most of the information.

The D columns were normalized. They represented time deltas from some point in the past. So they were converted to the point in the point in the past to stop them from increasing. Formula used : D15n = Transaction_Day - D15 and Transaction_Day = TransactionDT/(24*60*60)





The seven D columns('D6','D7','D8','D9','D12','D13','D14') were removed as they were mostly nan.

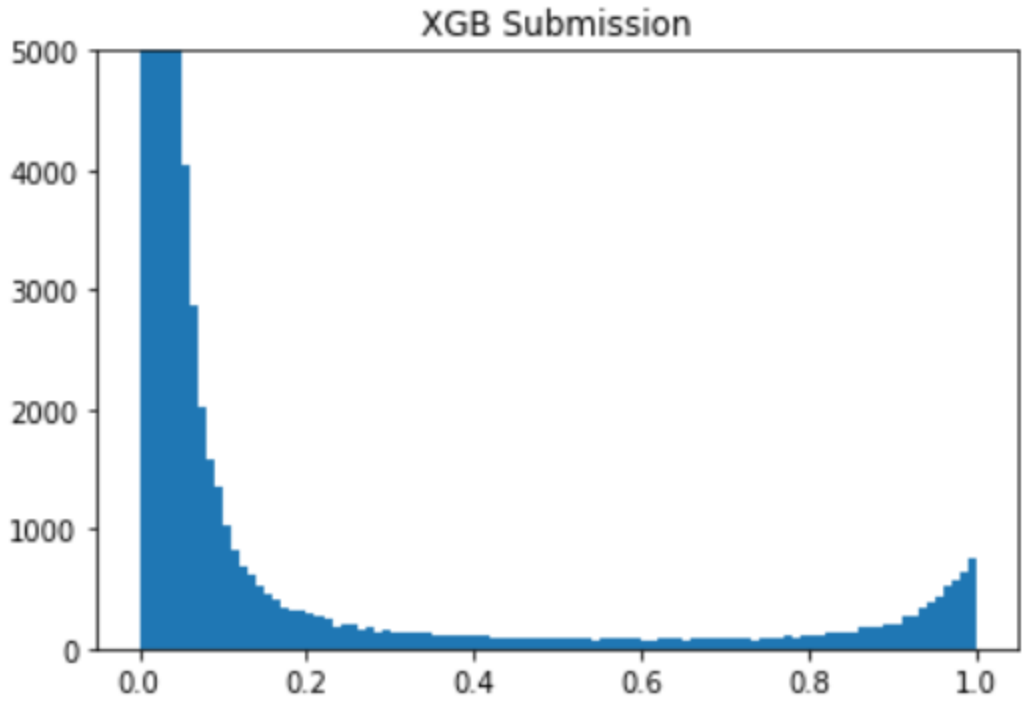
The id columns having more than 500000 nans were dropped along with the column V332, V325, V335, V338.

Label encoding was done on the categorical columns and all the nans were filled by -1.

Six months of training data was provided, that is why the data was grouped according to the months then model was trained using groupkfold. The predictions were given forward in time using the test data. XGBoost was used to train the model.

RESULT ANALYSIS:

The final test accuracy was 94.5338. The probabilities of the card being Fraud was submitted. Histogram is plotted to show the probability counts:



CONCLUSION:

The probability of a card being fraudulent is low seen in the above graph which makes sense as most of the cards are not fraudulent and are used by their respective owners in real life. The model built gives a pretty accurate prediction of a card being fraudulent or not. There are room for improvement using different feature engineering techniques and ensembling using different models which weren't tried.