Project : Credit Card Fraud

04.02.2019

Joshua Owusu Ankomah
Your Company
123 Your Street
Your City, ST 12345

Overview

1. We are trying to create a model that detects fraud
2. Each year Millions of pounds are lost

Goals :

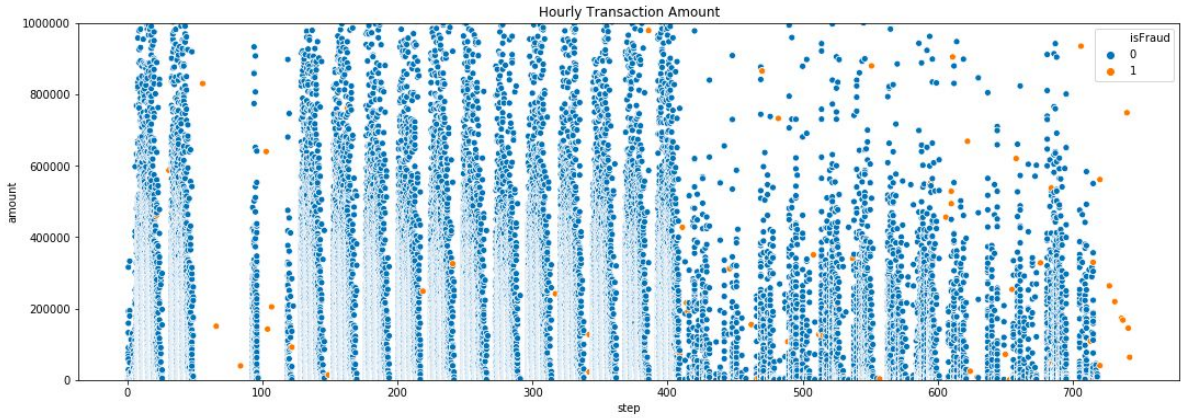
Methodology :

Project Steps

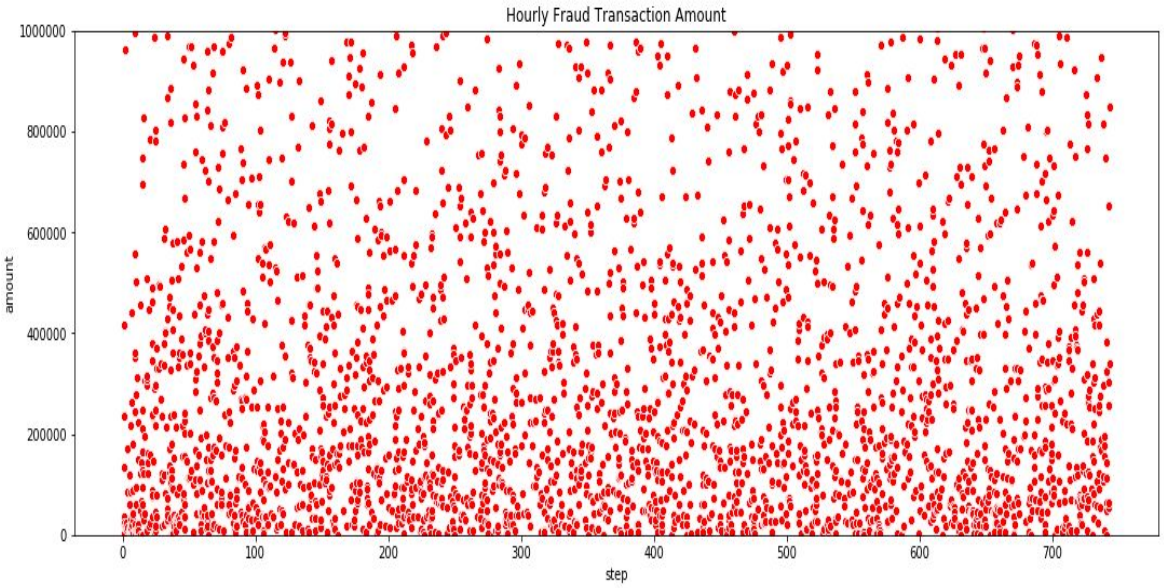
- 1.Loading Data and EDA
- 2.Feature Engineering
- 3.Machine Learning
    - 3.1. Baseline Models
    - 3.2. Grid Search for Best Hyper-parameter
    - 3.3. Dealing with Unbalanced Data
    - 3.3.1. Balancing Data via Resambling with SMOTE
    - 3.3.2. Subsampling Data from the Original Dataset
    - 3.3.3 Performing SMOTE on the New Data
    - 3.3.3a HyperParameter Optimization And Visualization of ROC_Curve of Models
    - - RandomForestModel and XGBoost Model
- 4.Machine Learning Pipeline
- 5.Feature Importance
- 6.Conclusion
- 7.Future Works

Further Findings: EDA Findings

Safe Transactions



Fraud Transactions



- Eventhough safe transactions slows down in 3rd and 4th day and after 16th day of the month, fraud transactions happens at a steady pace. Especially in the second half of the month there are much less safe transactions but number of fraud transactions does not decrease at all.

- Fraud proportion over all transactions is 0.01% while the fraud amount proportion is 0.1%

- There is some sort of seasonality in the number of transaction every 24 hours.Fraud transactions does not show that significant pattern. They happen every hour almost in the same frequency.

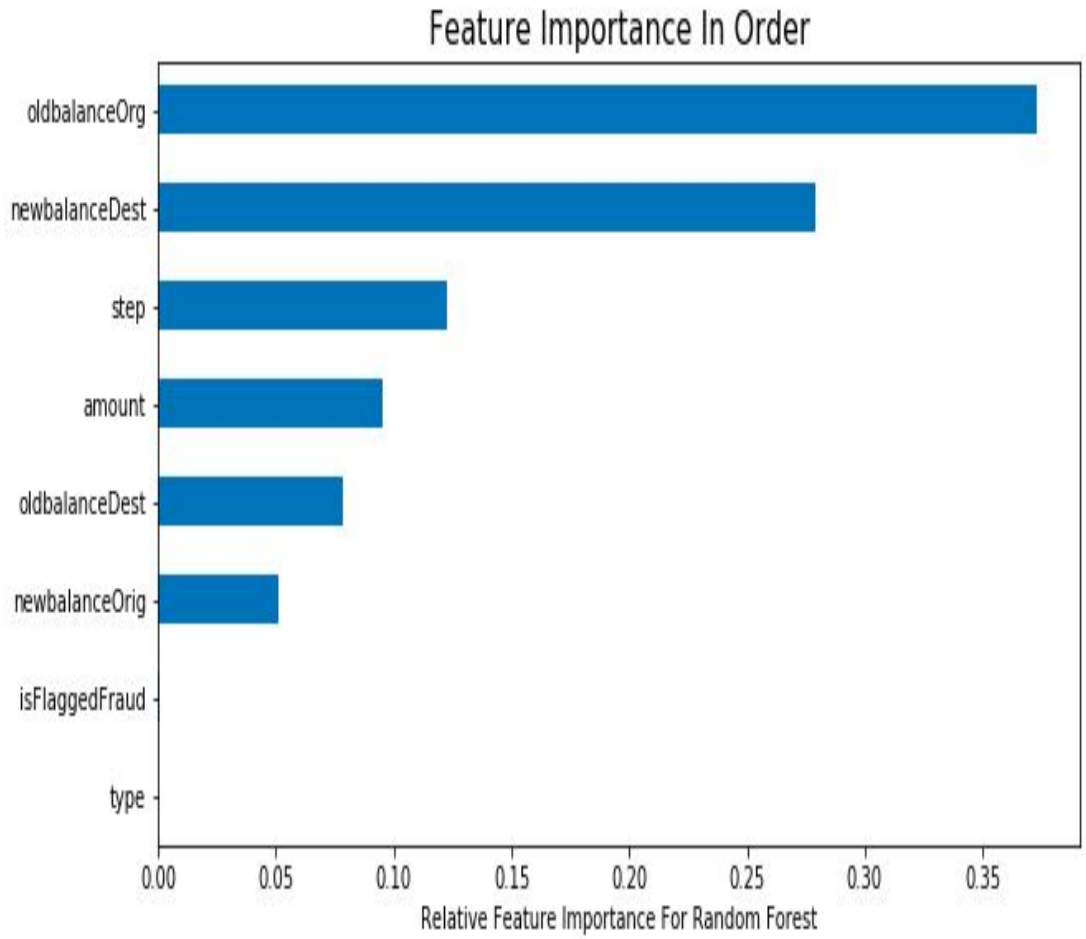
- There are more fraud transactions in low amounts and less in high amount. This distribution does not change much.

- Fraud transaction happens in a large range such as $119 to $10M . Most of the fraud transactions are of Lesser amount. But in 1M there is an interesting increase similar to safe transactions.

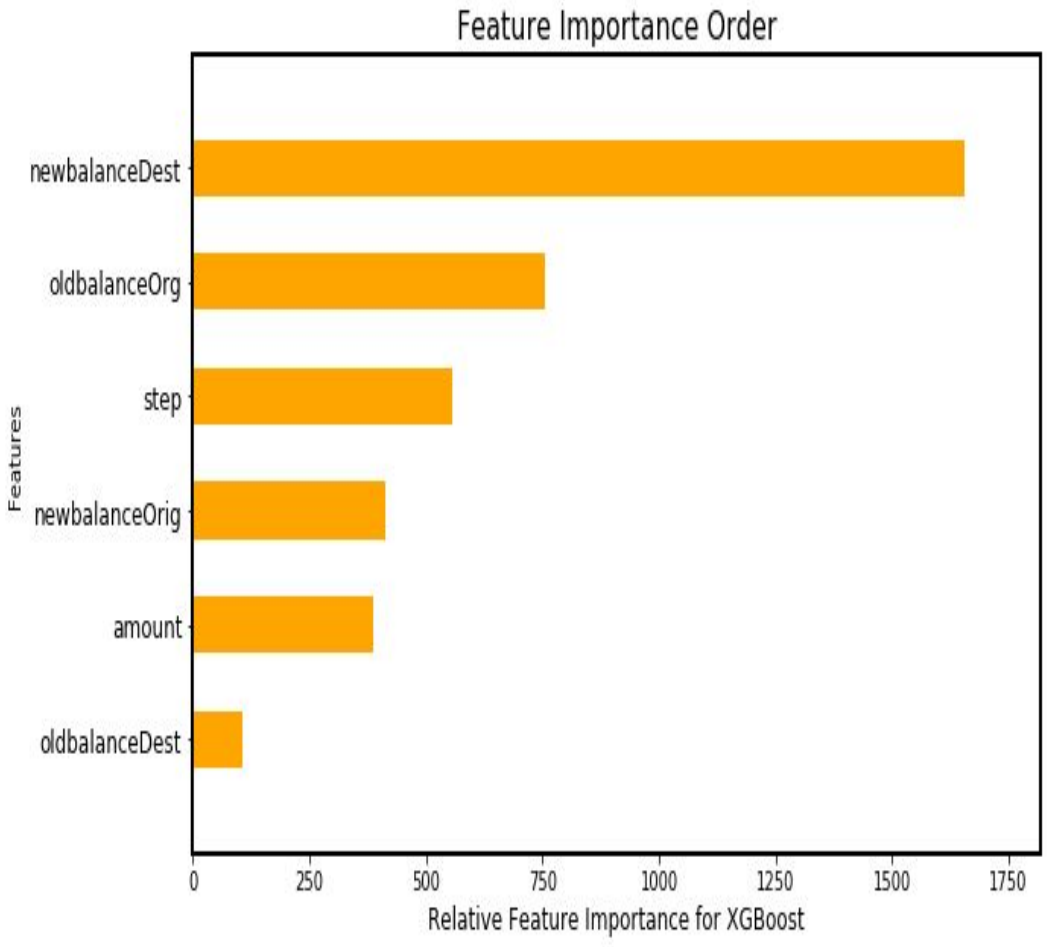
- There are 16 fake fraud cases with '0' amount.

- Fraud activities only happens with TRANSFER and CASH_OUT transactions. DEBIT usage is very safe.

Feature Importance For Random Forest:



Feature Importance For XGBoost:



Conclusion:

1. Random Sample

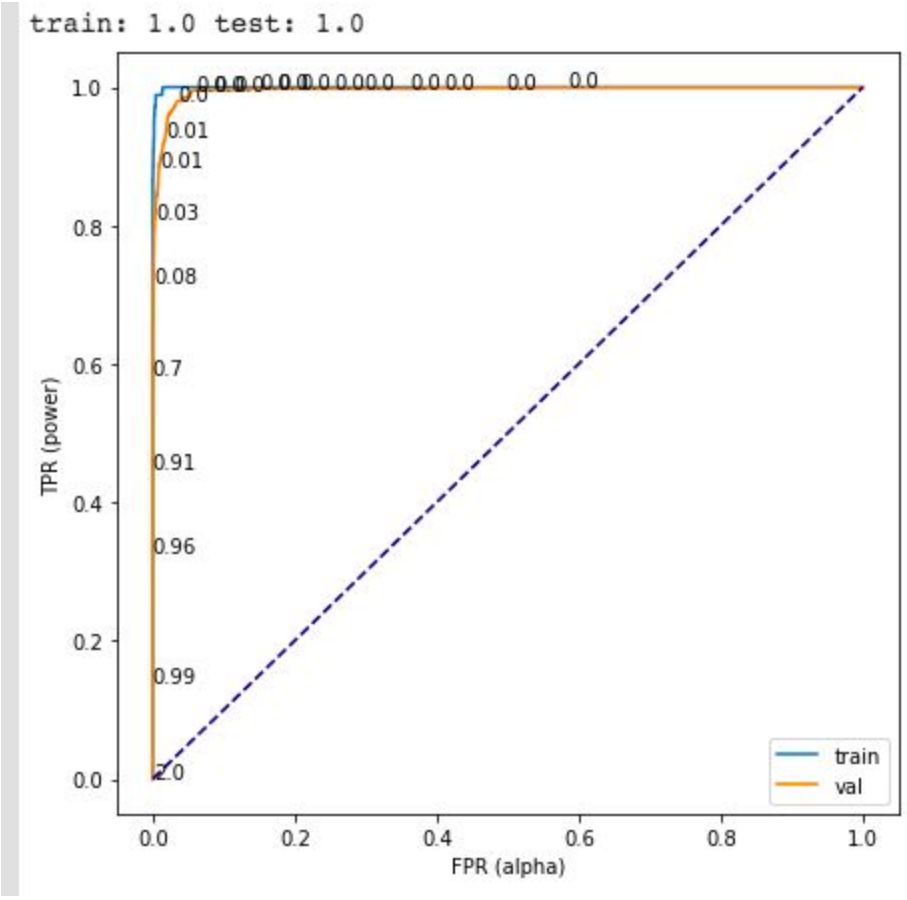
▾ Accuracy results after iterations

2. I created a model that can predict fraud transactions. I used XGBoost and RandomForest classifiers in this model.

(Data & Parameters)	(Accuracy)	XGBoost	RandomForest
Iteration 1			
- Random Sample & default parameters		100%	83%
- Random Sample & best parameters		85.5%	84.3%
Iteration 2			
Iteration 3			
- Balanced data with SMOTE & best parameters		99.4%	98.7%
Iteration 4			
- Random Safe trans. data and all Fraud data & best parameters		98.8%	99.6%
Iteration 5			
- New data balanced with SMOTE & best parameters		99%	92.1%
- New Balanced data with SMOTE & best HpyerParametr Tunning .		100%	100%

New Balanced data with SMOTE for hyperparameter selection

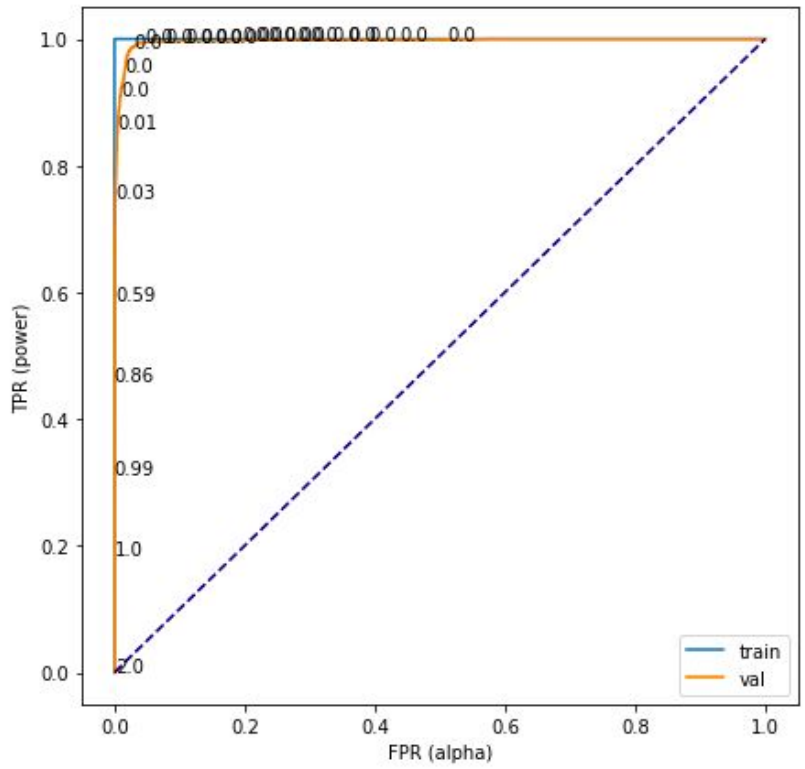
RandomForest:



New Balanced data with SMOTE for hyperparameter selection

XGBoost

```
train: 1.0 test: 1.0
```



Questions: