# Datamining Project proposal

## Fraud detection analysis on online transactions and credit cards

*Team:*

Monica Dommaraju
Sri Sruthi Chilukuri
Vijaylaxmi Nagurkar

# Fraud detection analysis on online transactions and credit cards

## *Abstract:*

As the financial world moves from traditional currency exchange to online transactions, credit card and online usage has become very popular and are widely in use now. With this increase in transaction volumes and the amount of money involved, fraudsters are finding numerous ways to exploit the system. Detecting these fraudulent transactions is challenging for reasons such as transactions data being heavily skewed and the volatility of the normal and fraudulent behavioral profiles. Features selection, sampling techniques and choosing the right algorithms play a crucial role in identifying these transactions. We intend to use naïve bayes, k-nearest neighbor and logistic regression based approaches to find the better performing algorithm to detect the fraudulent transactions. Factors such as precision, accuracy, sensitivity, AUC are considered to evaluate the performance of the above mentioned algorithms.

## *Motivation:*

In 2018 for United States alone, there were 41 billion credit card transactions based on cards from four major networks – Visa, Mastercard, AmericanExpress, and Discover that accounted for about $3.8 trillion dollars [1]. Online retail sales for Feb 2019 has surpassed general merchandise stores [2]. These two indicate the potential for fraudulent activities in the online and the credit card transactions.

To mitigate credit card and online frauds, e-commerce companies and banks have been implementing different fraud detection mechanisms and as a preventive maintenance taking actions such as blocking credit cards, blacklisting and placing a temporary hold on accounts etc. These mechanisms may help financial institutions and banks save costs that may incur by these fraudulent transactions. However, they are not adequate to completely eradicate these frauds as fraudsters are quickly coming up with new innovative ways.

Identifying the threshold to tag the transaction as fraud is sensitive and require good predictive modelling algorithms. Incorrectly tagging the transaction as a fraud may result in losing customers loyalty and trust. For instance, consider a customer planning to book an air travel reservation from a country different than that issued a credit card. Though the transaction is genuine, credit card company may temporarily place a hold on the account until the customer authorize that transaction. This scenario may be considered reasonable for an infrequent flyer. However, certain frequent flyers may find this inconvenient to get assistance from customer service to unblock their credit cards. This is an example of treating the same parameter in different ways based on the customer types and their behavioural profile. There can be numerous different scenarios that can happen while performing online transactions too.

Considering the volatility of usage patterns and behaviours, identifying fraudulent transactions is a difficult problem and is considered a suitable candidate to solve using data mining and machine learning techniques.

## *Literature Survey:*

Early work on credit card transactions by dataset with highly skewed data was mostly done by undersampling the no-fraud transactions to match with fraudulent transactions close to 1:1 ratio. As the fraud transactions in the dataset provided by ULB Machine learning group has only 492 fraud transactions out of 590K, undersampling the no-fraud transactions data would significantly eliminate the valuable data. To overcome this, many approaches handled the problem by running multiple iterations with randomly selected sub-sampled datasets.

An IEEE conference paper [5] on credit card fraud detection analysis approached this problem using a hybrid technique of under-sampling and oversampling. A paper [7] on SMOTE also showed that a combination of over-sampling the minority class and under-sampling the majority class can achieve better classifier performance. Paper[5] also discussed about the best practices in selecting the features and how to categorize them. But these techniques weren't implemented on this dataset as the feature selection and dimensionality reduction were already performed on the dataset using Principal Component Analysis (PCA) to protect the data for confidentiality reasons.

Another conference paper on credit card fraud detection using personalized or aggregated model [6], proposed building a personalized model for credit card users to identify fraud. This personalized model is built for each user by taking online questionnaire and sets are built using random forest and naïve bayes. Though this approach is good, this model lacks accuracy because personalized model is generally worse than aggregated model.

The project we are targeting focuses on the practical implementation of different classification algorithms on the European credit card transactions dataset to learn the Predictive Data Analysis. We are also targeting the exploratory data analysis on online fraud transactions dataset which has 393 attributes and try to find the correlation between features and class prediction using different visualization techniques.

## *Methodology:*

For this project, we are planning to work on two different datasets. Online fraud transactions dataset can be obtained from Vesta corporation, a payment service company. This dataset was posted by IEEE for a Kaggle competition and it contains 590K transactions and associated user identities [3]. Credit card transactions dataset can also be obtained from Kaggle. This data set contains European credit card transactions for two days from September 2013 [4]. Unfortunately for this dataset, all the features except the transaction time intervals and transaction amounts were transformed using PCA (Dimensionality reduction technique) due to confidentiality reasons and it was hard to get other financial datasets. However, we can still use this dataset to train the classifier and predict if we can classify the transaction as fraud or not.

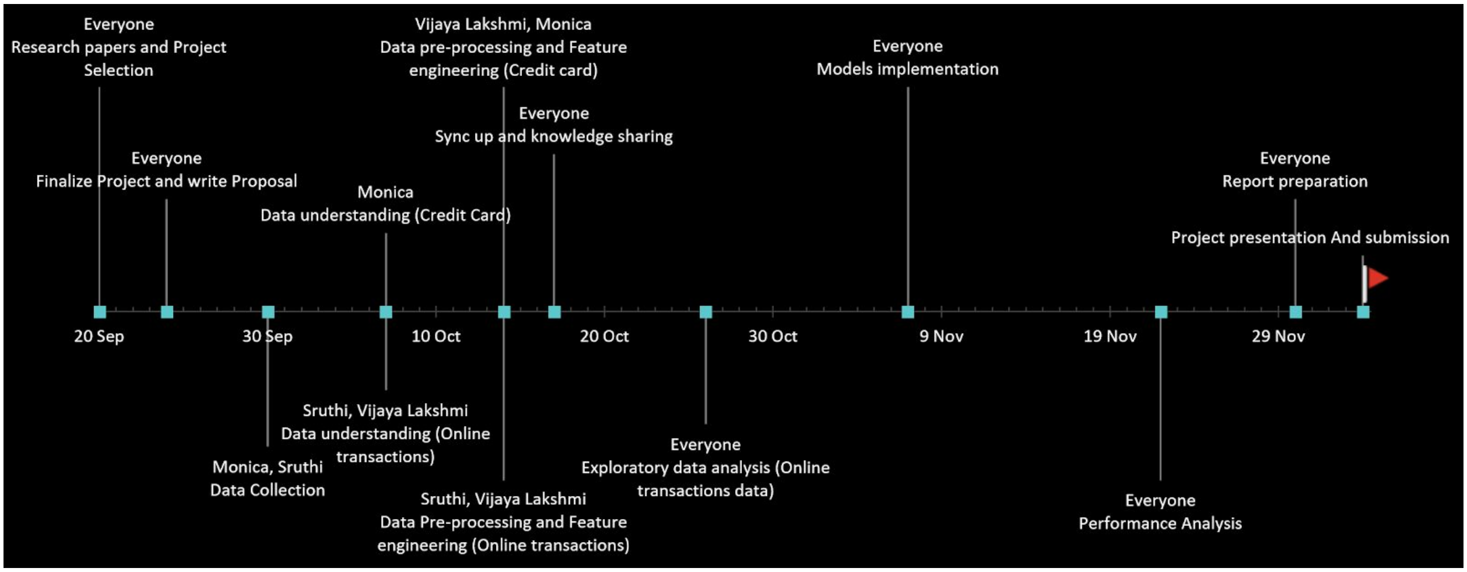
Both these datasets are highly imbalanced. In this project we will use different sampling techniques (oversampling, undersampling and hybrid) randomly to have different ratios of fraud/non-fraud transactions. We will then run this sampled training dataset on different classification algorithms such as naïve bayes, k-nearest neighbor and logistic regression, and choose the best performing one. As the data is highly skewed on both datasets, confusion matrix might not be the best one to evaluate the performance of the model. However we want to evaluate the performance using confusion matrix, AUROC and AUPRC techniques and understand the differences between them. Finally, we also want to run a few exploratory data analysis techniques on credit card related features of Vesta Corporation online transactions dataset. This dataset has numerous attributes, and at the same time some of these attributes have missing values, which is common in the real world. So, we would like to run pre-processing techniques and do feature engineering on it to choose the right attributes. We plan to use Python for all these implementations.

## *Deliverables:*

- Comparative performance analysis of classification algorithms such as naïve bayes, k-nearest neighbor and logistic regression.
- Visualizations such as boxplots, heat maps, etc if needed.
- Python Notebooks
- Write up a technical paper submission if results are satisfactory and time permits (Extended goal)

## *Team members and Roles:*

- Monica Dommaraju
- Sruthi Chilukuri
- Vijaya Lakshmi



| *Date* | *Milestone* | *Assigned To* |
|---|---|---|
| 09-20-2019 | Research papers and Project Selection | Everyone |
| 09-24-2019 | Finalize Project and write Proposal | Everyone |
| 09-30-2019 | Data Collection | Monica, Sruthi |
| 10-07-2019 | Data understanding (Credit Card) | Monica |
| 10-07-2019 | Data understanding (Online transactions) | Sruthi, Vijaylaxmi |
| 10-14-2019 | Data pre-processing and Feature engineering (Credit card) | Vijaylaxmi, Monica |
| 10-14-2019 | Data Pre-processing and Feature engineering (Online transactions) | Sruthi, Vijaylaxmi |
| 10-17-2019 | Sync up and knowledge sharing | Everyone |
| 10-26-2019 | Exploratory data analysis (Online transactions data) | Everyone |
| 11-07-2019 | Models implementation | Everyone |
| 11-22-2019 | Performance Analysis | Everyone |
| 11-30-2019 | Report preparation | Everyone |
| 12-04-2019 | Project presentation And submission | Everyone |

## *Sources:*

1. https://www.creditcards.com/credit-card-news/market-share-statistics.php
2. https://www.cnbc.com/2019/04/02/online-shopping-officially-overtakes-brick-and-mortar-retail-for-the-first-time-ever.html
3. https://www.kaggle.com/c/ieee-fraud-detection/data
4. https://www.kaggle.com/mlg-ulb/creditcardfraud
5. J. O. Awoyemi, A. O. Adetunmbi and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis," *2017 International Conference on Computing Networking and Informatics (ICCNI)*, Lagos, 2017, pp. 1-9.
   doi:10.1109/ICCNI.2017.8123782;
   http://ieeexplore.ieee.org.libaccess.sjlibrary.org/stamp/stamp.jsp?tp=&arnumber=8123782&isnumber=8123766
6. M. I. Alowais and L. Soon, "Credit Card Fraud Detection: Personalized or Aggregated Model," *2012 Third FTRA International Conference on Mobile, Ubiquitous, and Intelligent Computing*, Vancouver, BC, 2012, pp. 114-119.
   doi: 10.1109/MUSIC.2012.27
   http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6305834&isnumber=6305804
7. SMOTE: Synthetic Minority Over-sampling Technique
   https://doi.org/10.1613/jair.953