**Preliminary Analysis Report**

Credit Card Fraud Detection

**Name: Maviya Shaikh**

**Student number: 100766785**

**Date: December 1st,2020**

Table of Contents

[1. Executive Summary 3](#_Toc57407372)

[2. Problem Statement and Analytical Statement 3](#_Toc57407373)

[3. Identification and Justification of Output Variable Class Structure 3](#_Toc57407374)

[4. Data Sources 3](#_Toc57407375)

[5. Data Requirements 3](#_Toc57407376)

[6. Exploratory Data Analysis (EDA) 4](#_Toc57407377)

[Summary of EDA 4](#_Toc57407378)

[Insights 4](#_Toc57407379)

[7. Feature Engineering 5](#_Toc57407380)

[8. Analytical Score Card 5](#_Toc57407381)

[9. Assumptions 6](#_Toc57407382)

[10. Constraints 6](#_Toc57407383)

[11. Limitation 6](#_Toc57407384)

[References 7](#_Toc57407385)

# Executive Summary

In this digital era where physical cash flow is reduced and we are wrapped around the global pandemic of covid19 that pushes us to use only electronic methods for transaction, there is always a risk of misuse of this method that leads to higher losses to the user. In response to deal with this, the project aims to develop a system to detect fraudulent credit card transaction to prevent such cases by analysing their historical pattern. Specifically, this project is tailored to identify the credit card frauds as use of credit card has increased now-a-days because it provides instantaneous money without having actual money in the pocket.

# Problem Statement and Analytical Statement

**Problem Statement**: The analysis will be conducted to identify whether transaction occurred is fraudulent or not.

**Analytical Statement**:The model will be developed to predict whether the transaction carried out is normal payment or a fraud*.*

# Identification and Justification of Output Variable Class Structure

* The class structure is *binomial* because the analysis is conducted to determine whether the transaction is fraudulent or not
* The naming convention that is used to represent the binomial problem is *No Fraud* (i.e. Class 0) and *Fraud* (i.e. Class 1)

# Data Sources

The dataset used for the analysis of this project is downloaded from Kaggle which is well known website in data science community to explore the open source datasets. Originally, the dataset has been collected and analysed during a research collaboration of Worldline and the Machine Learning Group of ULB (Université Libre de Bruxelles) on big data mining and fraud detection.

# Data Requirements

The datasets consist of 284,807 transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. There are 31 variables in the dataset which are explained below::

* Due to privacy reason, variables ‘**V1, V2, … V28’** are transformed with Principal Component Analysis (PCA)
* The variable '**Time'** contains the seconds elapsed between each transaction and the first transaction in the dataset.
* The variable '**Amount**' is the transaction amount
* The dependent variable '**Class**' is the response variable that has value 1 in case of fraud and 0 otherwise.

# Exploratory Data Analysis (EDA)

## Summary of EDA

1. **Analysis of Basic Statistics:**

* With the help of **sweetviz report**, it was found that dataset has total 284807 observation with 1081 duplicate rows. In addition, all the independent variables are numerical continuous values whereas dependent variable Class is Boolean i.e. 0 for No fraud and 1 for fraud. **Note**: The duplicate rows will not be removed because it seems to be duplicate records, but it is original data and that may include fraud cases
* Tukey method is used to detect the outliers that lies beyond **1.5** times the Interquartile Range. There are **138473 rows** detected as outliers which is almost half of the dataset. I believe removing outlier will cause the problem of information loss, therefore outliers are not removed from the dataset

* **Analysis of the graphs:**
  + With the help of **heatmap**, it is determined that variables are not correlated with each other
  + With the help of **histogram**, the time variable is plotted and found out that for both fraud and no fraud, distribution is skewed and there is not much difference in the time that can classify both the classes and for the amount variable, there is slightly larger value of amount for fraudulent transaction
  + With the help of **kernel density estimate (KDE) plot,** the distribution of all the variables with target variables is plotted that gives various insights such as Variables V4 and V11 have clearly separated distribution, V12, V14, V18 have partially separated distribution for class 0 (no fraud) and class 1 (fraud). However, V26 has exactly same distribution for class 0 and class 1 whereas V24, V25, V27 and V28 have nearly same distribution for both classes.

## Insights

From the Exploratory Data Analysis, the key insights developed that will help in model selection process are as follows:

1. **The dataset is not balanced:** This nature of the dataset will restrict the use of certain metrics in the machine learning models. So, while selecting the evaluation metrics, this insight should be considered. Moreover, there is a constraint that it is not possible to collect or add more observations into the dataset to balance it. However, we can use the Synthetic Minority Oversampling Technique (SMOTE) technique to Random Under-sampling technique to eliminate the bias in prediction
2. **The features of the dataset are not correlated with each other:** According to our assumption that all independent variables in the dataset are useful to predict the outcome variable, however because of less correlation we may have to use the feature importance to improve the performance of the model.

# Feature Engineering

Feature Engineering is an important step to validate that all the features used are important in the prediction of the model and it allows us to add or remove the features according to the business problem requirements. Here, two variables Time and Amount are not scaled whereas all other variables are scaled and transformed using Principal Component Analysis. Thus, we will scale the variables using Standard Scaler which is the technique to rescales the value of the data points in the range of mean 0 and standard deviation 1.

Time 🡪 Scaled Time

Amount 🡪 Scaled Amount

Furthermore, the original time and amount is dropped from the dataset and scaled time and scaled amount is added in the dataset. Thus, our dataset is ready for model prediction.

# Analytical Score Card

The analytical score card is updated from the previous version because the accuracy metrics will not be useful for this project as the target variable class is not balanced.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sr. No.** | **Metrics** | **Explanation** | **Its use in the project** | **Target** |
| 1 | Precision | It tells how often the model is correct when it makes the prediction. | This metric will use to analyse the specific class in the dataset. | It should be greater than 90% (to analyse the specific class of the dataset) |
| 2 | Recall | It is the ratio of correctly classified positive or negative instances from total number of positive or negative instances respectively. | This metric will identify all the relevant instances from retrieved instances. Thus, it will use to analyse the specific class. | It should be greater than 90% (to analyse the specific class of the dataset) |
| 3 | F1-score | It is the harmonic mean of precision and recall. If there is an uneven distribution of the class then this metric is considered for evaluation of the algorithm. | This is the main metric that will be used to evaluate the model as our dataset is unbalanced. | It should be greater than 90% (to analyse the overall model) |
| 4 | Confusion matrix | It is a table used to describe the performance of the classification model on the test data. | This will give the number of frauds and no frauds which are correctly or incorrectly identified by a model | Most of the fraud cases should be predicted i.e. at least 90% of fraud cases should be predicted correctly. |
| 5 | AUC-ROC curve (Area Under the Curve – Receiver Operating Characteristics) | It is a performance measurement tool for classification problem to know the performance of the model in differentiating classes of the dataset. | This is also an important metric that will help to know the capacity of the model on how well it distinguishes the classes No fraud and fraud. | AUC should be at least 0.90 (For instance: if AUC=0.90 there is 90% chance that model will be able to distinguish between no fraud and fraud.) |

# Assumptions

* The dataset is valid and come from reliable source
* The dataset provided is legible and comprehensible
* All independent variables in the dataset are useful to predict the outcome variable (fraud or no fraud). However, less correlated features can be removed for better prediction

# Constraints

* The additional data cannot be added in the dataset
* Due to confidentiality issues, original features are transformed with Principal Component Analysis (PCA), therefore there is no background information about the features to better understand it
* The final model to detect fraud and no fraud cases must be built to meet the deadline which is December 18, 2020

# Limitation

The model developed might predict the normal transaction as a fraud, this might be the problem to the customer because their card might be blocked, or they are contacted depending on their banking institute. This will waste the customer’s time and let them go through all the procedures of changing their password or provide their authorization to secure the card.

# 

# References

1. Machine Learning Group -ULB. (2018, March 23). Credit Card Fraud Detection. Retrieved October 30, 2020, from <https://www.kaggle.com/mlg-ulb/creditcardfraud>
2. Narkhede, S. (2019, May 26). Understanding AUC - ROC Curve. Retrieved October 30, 2020, from <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>
3. Jaadi, Z. (n.d.). A Step by Step Explanation of Principal Component Analysis. Retrieved October 31, 2020, from <https://builtin.com/data-science/step-step-explanation-principal-component-analysis>