Statement of Work

Credit Card Fraud Detection

Name: Maviya Shaikh
Student number: 100766785
Date: November 1st,2020

Table of Contents

1. 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.

2. Problem Statement

The credit card companies should be able to recognize fraudulent credit card transactions so that users are not charged for items that they did not purchase. For this, "**The analysis will be conducted to identify whether transaction occurred is fraudulent or not**".

3. Analytics Rationale Statement

*"***The model will be developed to predict whether the transaction carried out is normal payment or a fraud***"*. Early detection of this cases can help prevent fraud and save the money of the customer.

4. 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.

5. 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.

6. Data Analysis Approach

The analysis approach that will be used to identify the fraudulent cases are as follows:

1. Exploratory Data Analysis (EDA) will be conducted to understand all the features and characteristics of dataset
2. Data Preparation will be done after conducting EDA to deal with outliers or unbalanced dataset
3. Dataset will be divided into training, validation and test data to build the model
4. Various base models and optimized models will be developed using different techniques to achieve the goal of the project
5. The models are evaluated based on the test process to select the final model

In terms of software tools to run the models, Jupyter Notebook (Anaconda 3) will be used and for coding purpose, python programming will be used to build the models for this project.

7. 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)
- The necessary software and hardware tools require to complete the project are available

8. 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

9. 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.

10. Test Process

To evaluate the performance of the model, the below analytical score card will be used to measure its performance. Various metrics such as accuracy, precision, recall, f1-score, confusion matrix and AUC-ROC curve will be used to analyse the model. In order to use the model, the targets are set for the metrices that should be achieved.

Table 1: Analytical Score Card

Sr. No.	Metrics	Explanation	Its use in the project	Target
1	Accuracy	It represents the percentage of predictions that model got right. This metric will give the correct result only if there is equal distribution of the class.	This metric will be used to compare the result of different models. Thus, it will help in selecting the best model.	It should be greater than 90% (to select the model)
2	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)
3	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)
4	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)
5	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.
6	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.)

Table 1: Analytical Score Card

11. Project Plan

The project plan includes all the deliverables with the date. This includes all the task that will be carried out during this project which is as described below:

Phase	Milestone	Submit Date
1	Business Understanding and Problem Discovery	November 1,2020
2	Data Acquisition and Understanding	November 18,2020
3	ML Modelling and Evaluation	November 23,2020
4	Deployment	December 18,2020

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

Appendix

Principal Component Analysis (PCA): - It is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set. Here, the data is masked with this technique to protect the privacy of the user at the same time variables are scaled that will make these variables useful in our analysis for predicting fraud and no fraud cases.