Ayesha Ishrath
ayesha.ishrath95@gmail.com

Credit Card Fraud Detection

15th December 2020

Context

It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.

Content

The datasets contain 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. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

It contains only numeric input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, … V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

THE DATA

The data is already available to us in the form of a Kaggle:
https://www.kaggle.com/mlg-ulb/creditcardfraud

Therefore, there isn't a need for additional data mining or web scraping. The data will however have to be cleaned and wrangled before any analysis is performed on it.

The data provided by is in the form of CSV files : credicard.csv

Class va;ue of ‘0’ means normal transaction and ‘1’ means a fraudulent transaction.

APPROACH AND MILESTONES

The approach to solving this problem is subject to change as I progress with the Career Track and learn new concepts and approaches. The tentative broad overview of solving the problem is explained below.

Data Wrangling

The first step would be to load the given datasets into Pandas dataframes and clean them. This would be followed by various wrangling methods to arrive at data on which analysis and prediction can be performed.

Data Visualisation

The second step is to create visualisations for the cleaned data and try to come up with a few preliminary hypotheses. The graphs, charts and other visualisations will also be a very important part in creating our story for the project.

Statistical Analysis

Based on the hypothesis formed in the previous step, the next step would be to perform various statistical analysis on it to test if the hypothesis is indeed correct.

Machine Learning

The next step would be to build a predictive model using elementary ML methods and, if necessary, deep learning. The model will be incrementally improved.

Data Story, Results and Conclusion

The final step is to report the results of the analysis performed and the accuracy of the predictive model. This step will involve creating a story around the initial problem, the problems it aims at solving and the insights gained from the data. This will be followed by explaining the intuitions involved in building the ML model, the incremental improvements involved, the accuracy and future prospects of improvement.

DELIVERABLES

The following should be considered as deliverables as part of the project:

1. Jupyter Notebook: Contains all the code involved as part of wrangling, analysis and building predictive models.

2. Project Report: A Document highlighting the entire process of the project.

3. Presentation: A Slide Deck to be presented to the clients as the final product of the analysis performed and the model built.