Barclays Data Science Exercise

Fraud detection is a topic applicable to many sectors, including financial services and insurance.

In this exercise, you will be asked to build a predictive model for predicting fraud and a simple application to simulate the model running on unseen data.

The target variable 'isFraud' should be used and the 'isFlaggedFraud' variable must be ignored for the purpose of this analysis.

Data

You will be using the PaySim dataset (E.A. Lopez-Rojas , A. Elmir, and S. Axelsson - 2016)

Feature	Definition
step	maps a unit of time in the real world. In this case 1 step is 1 hour of time. Total steps 744 (30 days simulation).
type	CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER.
amount	amount of the transaction in local currency.
nameOrig	customer who started the transaction
oldbalanceOrg	initial balance before the transaction
newbalanceOrig	new balance after the transaction
nameDest	customer who is the recipient of the transaction
oldbalanceDest	initial balance recipient before the transaction. Note that there is not information for customers that start with M (Merchants).
newbalanceDest	new balance recipient after the transaction. Note that there is not information for customers that start with M (Merchants).
isFraud	This is the transactions made by the fraudulent agents inside the simulation. In this specific dataset the fraudulent behavior of the agents aims to profit by taking control or customers accounts and try to empty the funds by transferring to another account and then cashing out of the system. **This is the target variable**
isFlaggedFraud	**This variable should be deleted from the dataset**

Instructions

1. Download the PaySim fraud dataset from here: https://www.kaggle.com/ntnu-testimon/paysim1/data
2. **Exploratory Data Analysis**: With the overall objective in mind, conduct exploratory data analysis on the dataset. It may be presented in any form of your choosing (e.g. Jupyter notebooks, Latex, PowerPoint).
3. **Feature Engineering**: Informed by the above analysis, create any features you think will be informative in predicting the target variable ('isFraud')
4. **Modelling**: Using the dataset and the above analysis, build a model to predict transactions which are fraudulent
    a. You will be assessed primarily on the model choice and the features selected for this. There is no need to perform extensive parameter / hyperparameter optimization.
5. **Scoring**: Build a simple streaming application which takes transactions **one by one** from a given dataset and then classifies them as fraudulent/not-fraudulent using the model created in step 4.
    a. You may assume that the data takes the same format as the training dataset (and the application should be tested with this).
    b. Incoming transactions should have interarrival times following an exponential distribution with mean 1
    c. Only the code needs to be submitted for this part - Scores do not need to be submitted
    d. This part should be built with **fewer than 20 lines of code** (excluding comments)

Further Comments

- You should use Python, R or Scala (appropriate analytics and visualisation packages / libraries may be used)
- Code and other outputs should be shared through email. No solutions should not be posted in the public domain (e.g. git)
- If your home computer is not able to process the 187mb file, feel free to work on a subset of the file

You will primarily be assessed on:

- The end to end data science workflow
- The quality of code written (efficiency, conciseness, readability and documentation/commentary)

The Solutions to this exercise will be discussed in the initial telephone interview for the role.