Coding/Machine Learning

A company's ability to recognize fraudulent transactions is very important so that customers are not charged unduly. The accompanying dataset contains transactions made by credit cards in two days. It contains only numeric input variables which are the result of a Principal Component Analysis (PCA) transformation, due to confidentiality issues. The only features that are not transformed are 'Time' and 'Dollar_amount'.

- Features P1, P2, … P28 are the principal components obtained with PCA
- 'Time' is the seconds elapsed between each transaction and the first transaction in the dataset.
- 'Dollar_amount' is the transaction amount
- 'Outcome' is the response variable, 1 in case of fraud and 0 otherwise.

Through this exercise, we want you to perform the following:

a. Data processing, if needed
b. Show if there is any need to perform feature selection
c. Build a model which can predict whether a transaction is fraudulent or not
d. Show us how you settled on any hyper-parameters (if relevant)
e. Show us how you evaluated the model you built

Expectations:

- You can use any ML libraries (for e.g. scikit-learn) for your analysis; however AutoML libraries (e.g. Auto-Sklearn) is not encouraged
- We do not expect you to spend more than 3-4 hours on this exercise
- You will NOT be evaluated on obtaining the highest performance metric but on the steps and choice of approach and methodology
- You will be evaluated on code readability and clear explanation of the steps taken in your work
- You will have ~10 mins to explain your work in this section at the live technical interview. Please aim to present only key actions/insights in a clear and concise manner. You can use an IDE of your choice.
- Python is the preferred programming language (as this is what we use internally) but feel free to use other languages

*The dataset has been collected and analysed during a research collaboration of Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) of ULB (Université Libre de Bruxelles) on big data mining and fraud detection.