Springboard Data Science Career Track Program
Capstone Project 3 Proposal:

Credit Card Fraud Detection

By: Zhiling Xie
December 2023

1. Problem Statement
In this project, I study the problem of modeling past credit card transactions in order to determine which ones are fraudulent.

2. Context
Credit card fraud is a form of identity theft that involves an unauthorized taking of another's credit card information for the purpose of charging purchases to the account or removing funds from it.

Credit card fraud happens both online and in stores. It is important that credit card companies can recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.

3. Criteria for success
1) Several outlier detection (unsupervised) and classification models (supervised) will be developed.
2) Given the class imbalance ratio, the model performance will be evaluated using the Area Under the Classification Report, Confusion Matrix, and F1 Score. The target is to build models with highest combination of precision and recall.

4. Scope of solution space: Interpretability analyses will be conducted to study the impact of features on the target.

5. Constrains: None identified at this point.

6. Stakeholders: European credit card companies and cardholders.

7.Data sources
1) The dataset contains transactions made by credit cards in September 2013 by European cardholders. The transactions occurred in two days, including 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (**frauds**) account for **0.172%** of all transactions.
2) Due to confidentiality issues, the original features and more background information about the data are not provided. The dataset contains only numeric input variables which are the result of a PCA transformation. **Features V1, V2, … V28** are the principal components obtained applying PCA.
3) 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. Feature **'Class'** is the response variable (**Target**), and it takes value 1 in case of fraud and 0 otherwise.

8. Methods

- **Supervised Classification** (for pre-labeled data): Logistic Regression, XGBoost, and LightGBM classifiers.
- **Unsupervised Outlier Detection** (algorithms in scikit-learn): Isolation Forest, Local Outlier Factor (LOF).

9. Deliverables: A GitHub repo containing the Jupyter notebooks developed for this project, a written final report, and a presentation slide deck.