



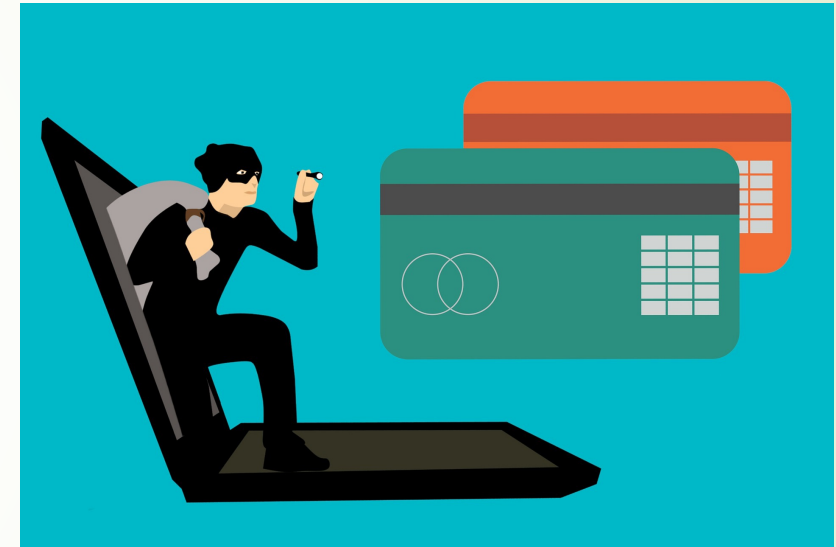
# Credit Card Fraud Detection

**Data Science Capstone Project 3**

**by Zhiling Xie**

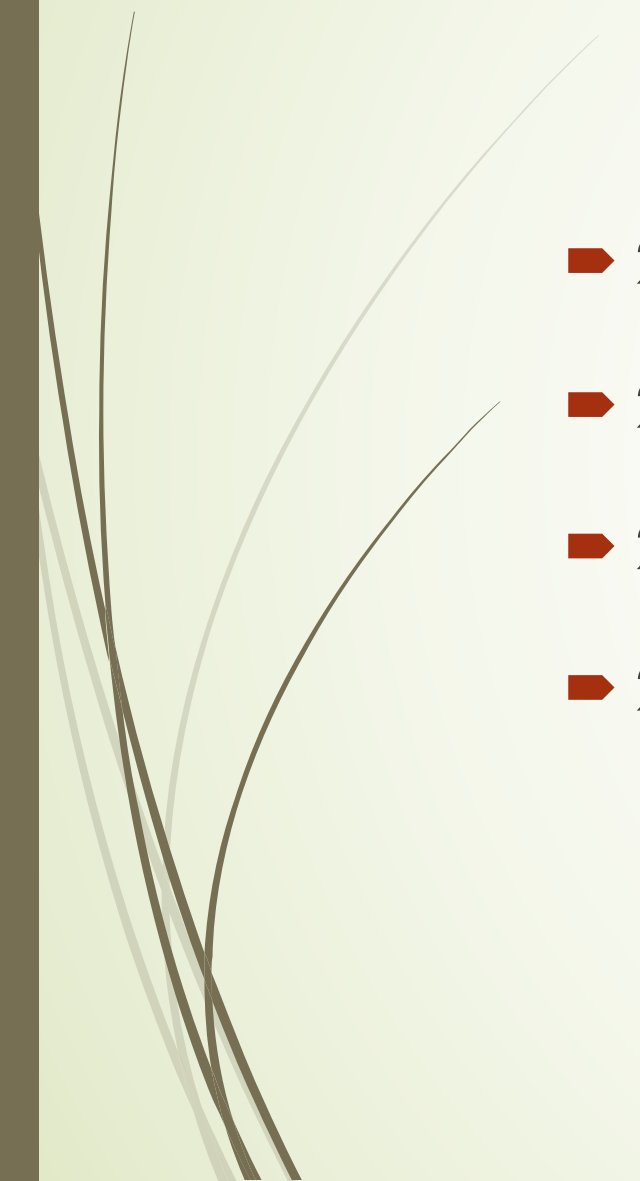
# 1. Problem

- **Client:** European credit card company
- **Objective:** Identify fraudulent transactions
- **How:** More accurate detection algorithm
- **Project Goal:**
  - Develop several classification (supervised) and outlier detection (unsupervised) models.
  - The target is to build models with highest combination of precision and recall.





## 2. Approach

- 2.1 Data Wrangling
  - 2.2 Exploratory Data Analysis (EDA)
  - 2.3 Modeling Part I : Classification (Supervised)
  - 2.4 Modeling Part II: Anomaly Detection (Unsupervised)
- 



## 2.1 Data Wrangling

### ➤ Data information

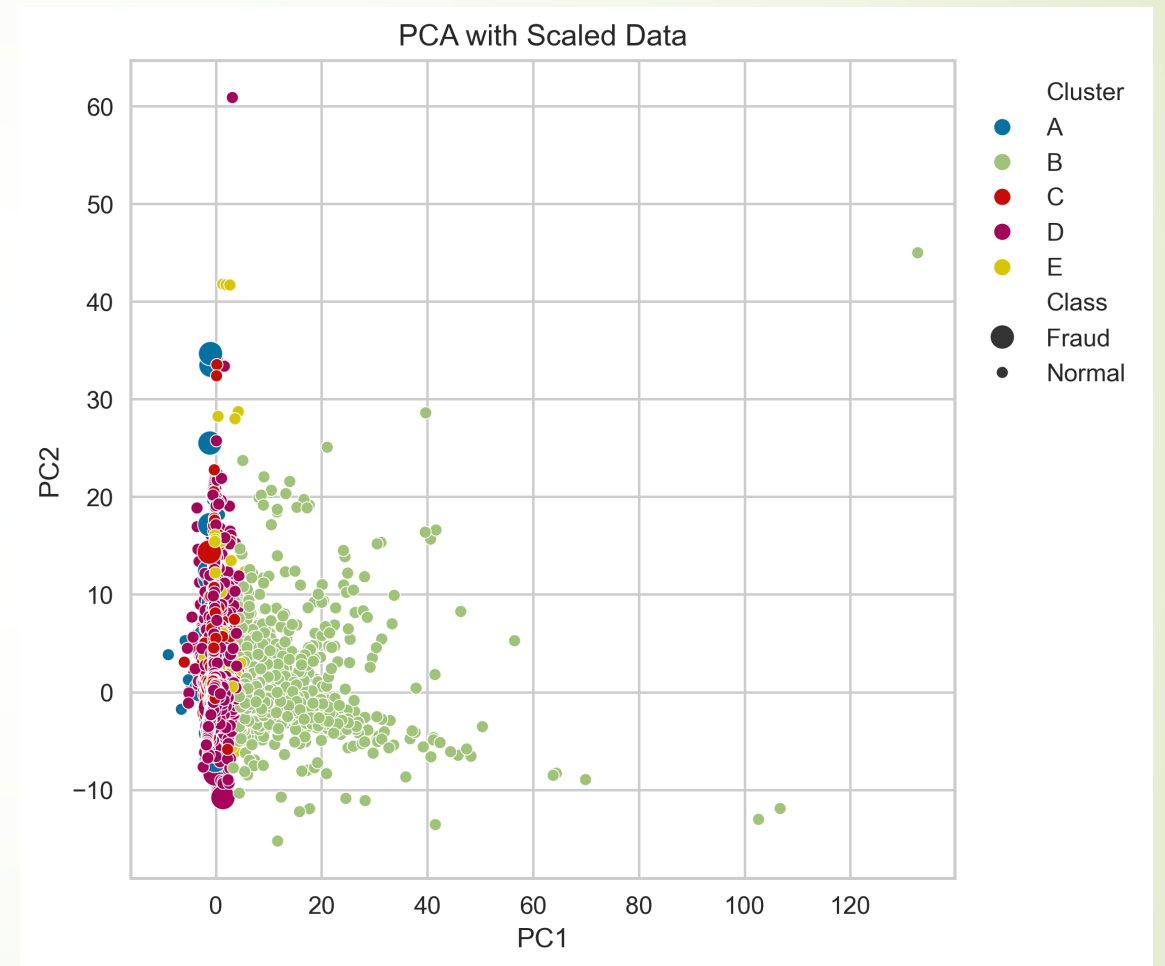
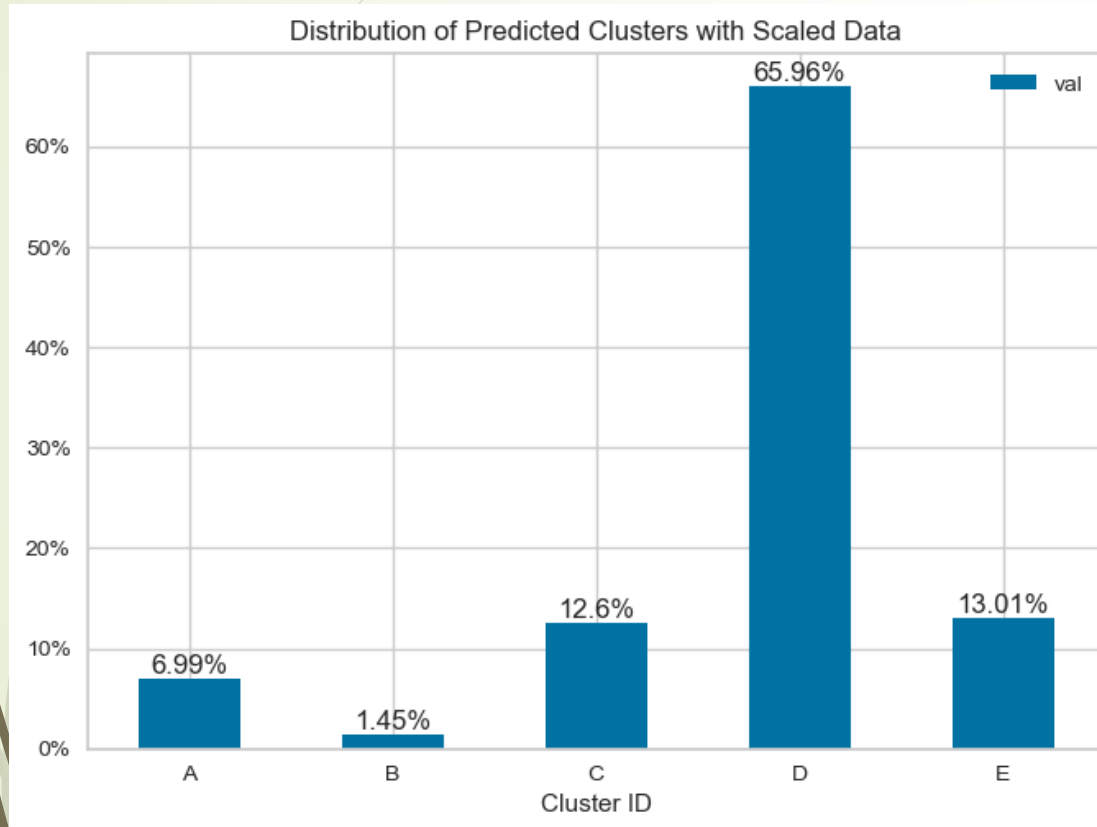
- 1) The dataset contains 492 frauds out of 284,807 transactions. The dataset is highly imbalanced, the positive class (**Fraud**) account for only **0.172%** of all transactions.
- 2) Due to confidentiality issues, features '**V1**', '**V2**',..., '**V28**' are the principal components obtained by applying PCA (rather than the original information).
- 3) '**Time**' contains the seconds elapsed between each transaction and the first transaction in the dataset. '**Amount**' is the transaction amount in dollars. '**Class**' is the response variable (**Target**), and it takes value 1 in case of Fraud and 0 otherwise.

### ➤ Data Cleaning – missing values and duplicated rows

- Dataset do not have missing values. All values are numeric.
- Duplicated rows account for only 0.65% of the total rows - Delete the duplicated rows .




## (2) K-Means Clustering (K=5)





## 2.3 Modeling Part I : Classification (Supervised)

- Logistic Regression: F1 Score = 0.69
  - XGBoost: F1 Score = 0.85
  - Random Forest model: F1 Score = 0.56
- 



## 2.4 Modeling Part II: Anomaly Detection (Unsupervised)

- Isolation Forest: Recall = 0.8
- Local Outlier Factor: Recall=0.84

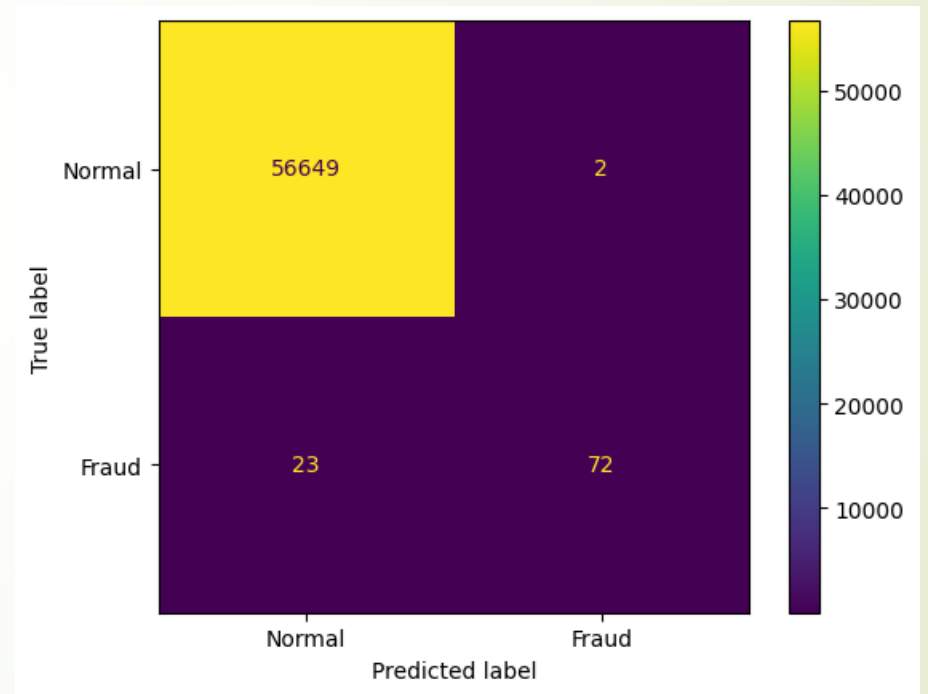


### 3. Findings

#### Modeling Results Comparison

	Precision	Recall	F1 Score
<b>Logistic Regression</b>	0.85	0.58	0.69
<b>LR w/ Resampling</b>	0.06	0.87	0.11
<b>XGBoost</b>	0.97	0.76	0.85
<b>XGB w/ Resampling</b>	0.96	0.77	0.85
<b>LightGBM</b>	0.71	0.46	0.56
<b>LGBM w/ Resampling</b>	0.94	0.76	0.84
<b>Isolation Forest</b>	0.04	0.80	0.07
<b>Local Outlier Factor</b>	0.03	0.84	0.05

Best Model: XGBoost



## 4. Conclusions and Future Work

- Considering the combination of Recall and Precision and computational efficiency, XGBoost is the best model. No resampling is needed, it has the highest F1 Score and Precision, and the computing speed is fast.
- LightGBM with Resampling has similar scores but is considerably slower to compute.
- If only Recall is considered (i.e. focusing on not treating "Fraud" as normal transactions), the Logistic Regression and two unsupervised models (Isolation Forest and LOF) have high Recall.
- **Current Issue and Future Work:**
- Even the best model has Recall smaller than 0.8. The predictive skills should be further improved.