Credit Card Fraud Detection

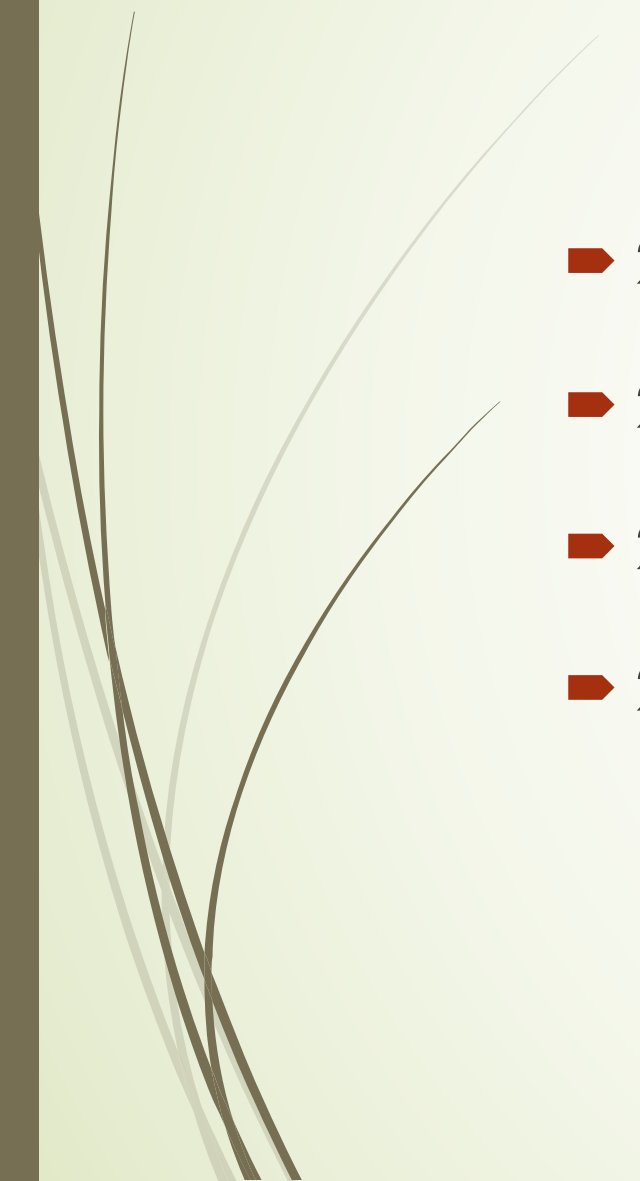Data Science Capstone Project 3

by Zhiling Xie

1. Problem

- **Client:** European credit card company

- **Objective:** Identify fraudulent transactions

- **How:** Use 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**

o The dataset contains 492 frauds out of 284,807 transactions.

o **Time** contains the seconds between each transaction and the first transaction.

o Features **V1, V2, ... , V28** are the principal components obtained by applying PCA.

o **Time** contains the seconds between each transaction and the first transaction.

o **Amount** is the transaction amount.

o **Class** is the response variable (**Target**), with a value 1 as Fraud and 0 otherwise. Highly imbalanced: Class 1 accounts for only **0.172%**.

▶ **Data Cleaning – Missing values and duplicated rows**

o This dataset does not have missing values. All variables are numeric.

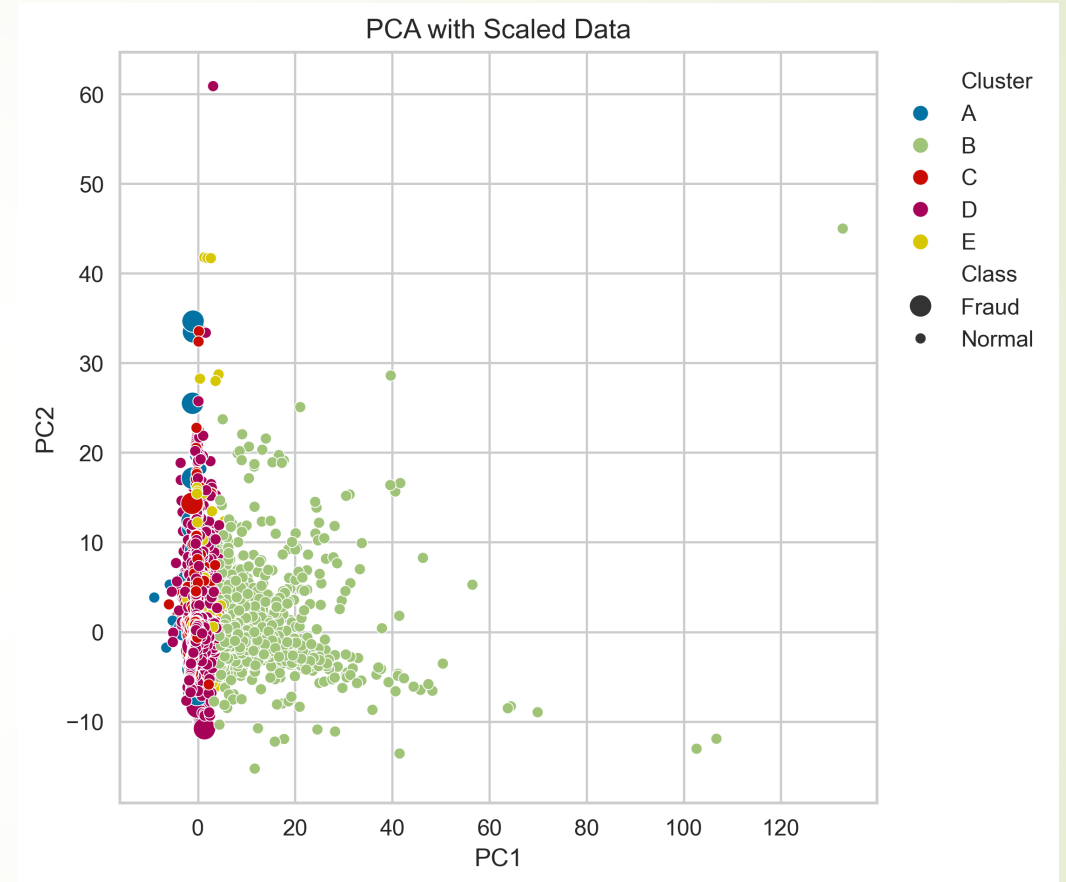o Duplicated rows account for only 0.65% of the total rows – just deleted.

2.2 EDA

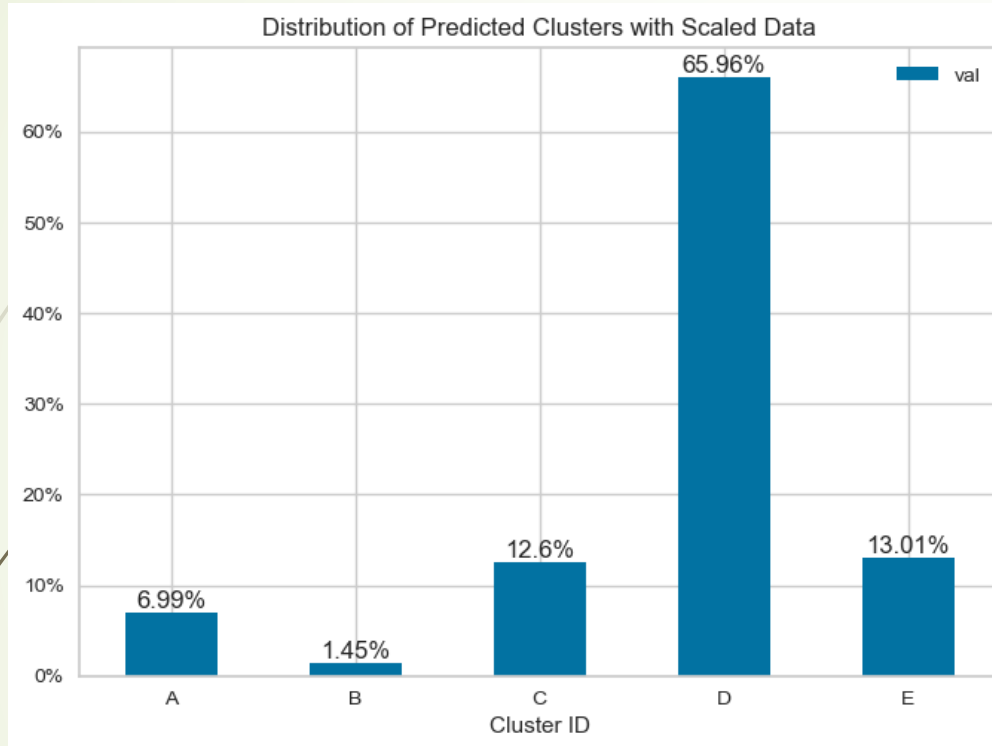➡️ **Feature correlation heatmap**

- **V1** to **V28** are not correlated with each other because they are PCs.

- **Class/Amount** has some positive or negative correlations with several **V** features.

- **Class** and **Amount** are not correlated.
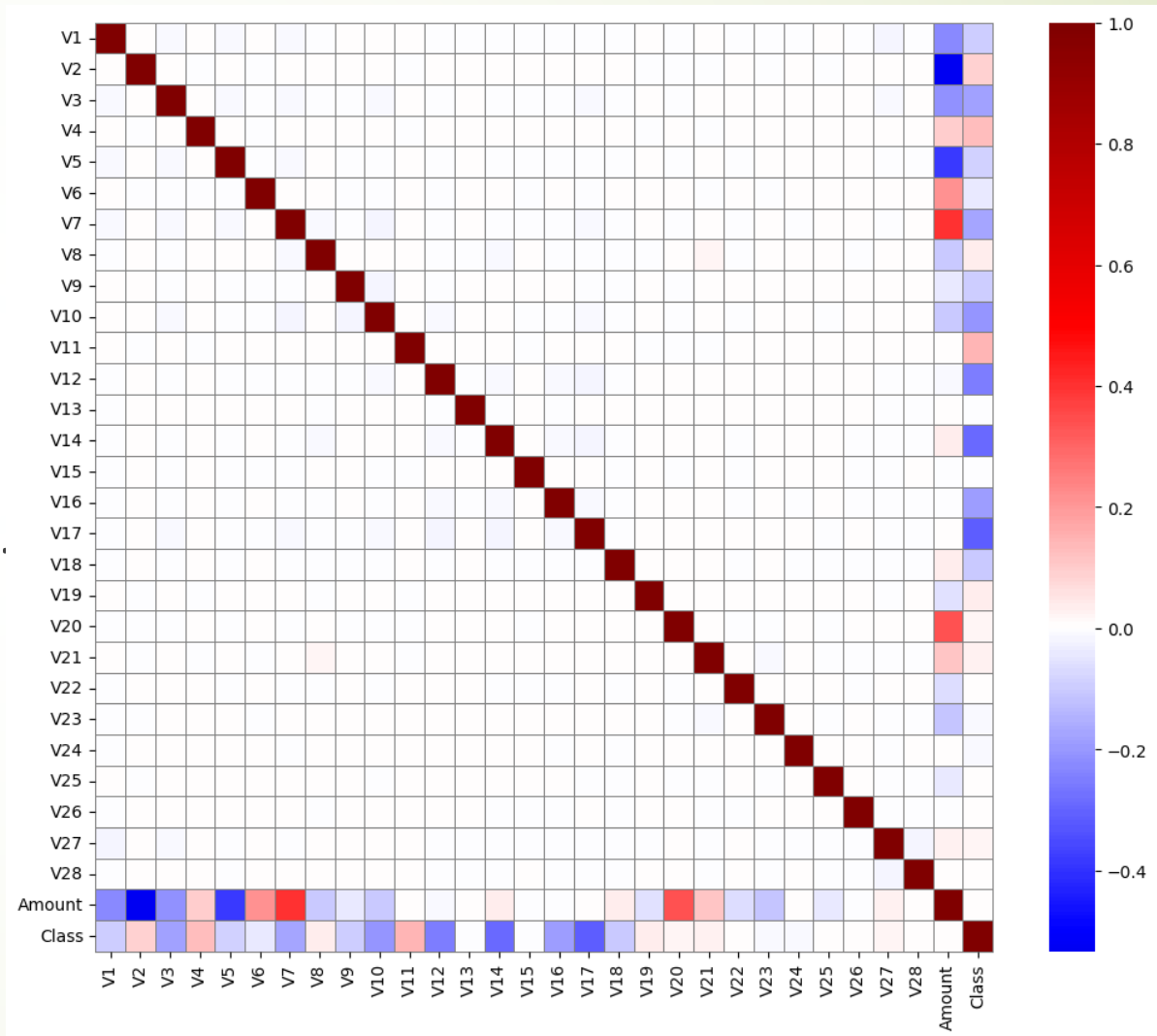
K-Means Clustering (K=5)

- Cluster D accounts for more than 65% of the data, following by clusters E and C.
- In the PCA visualization figure, the large dot denotes the **Fraud** Class. Some large dots can be seen in clusters A, C and D.

2.3 Modeling Part I : Classification (Supervised)

➡ The classification models are built using two sets of labeled data (one is the original data, and the other is the resampled data). Random oversampling is used to obtain resampled data.

o **Logistic Regression (LR)** : The F1 score is 0.69 without resampling and 0.11 with resampling.

o **XGBoost**: The F1 score is 0.85 regardless of resampling.

o **LightGBM**: The F1 score is 0.56 without resampling and 0.84 with resampling.

2.4 Modeling Part II: Anomaly Detection (Unsupervised)

➡ The anomaly detection models are built with unlabeled data (without the **Class** column for training). The following two models have very low F1 scores but high Recall.

o **Isolation Forest:** The F1 score is only 0.07 but the Recall is 0.80.

o **Local Outlier Factor (LOF):** The F1 score is only 0.05 but the Recall is 0.84.

3. Findings

Modeling Results Comparison: Prediction Scores

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

o **Logistic Regression**: High Precision and fair Recall. After resampling, F1 score and Precision become much worse, but Recall becomes very high.

o **XGBoost**: Highest Precision (0.97), highest F1 score (0.85), and very good Recall (0.76).

o **LightGBM**: Resampling improves the result, with higher Precision, Recall and F1 score, almost the same as XGBoost.

o **Isolation Forest and LOF**: Both have very high Recall but very low Precision. Isolation Forest is very computationally efficient, while computation of LOF is much slower.

Confusion Matrix
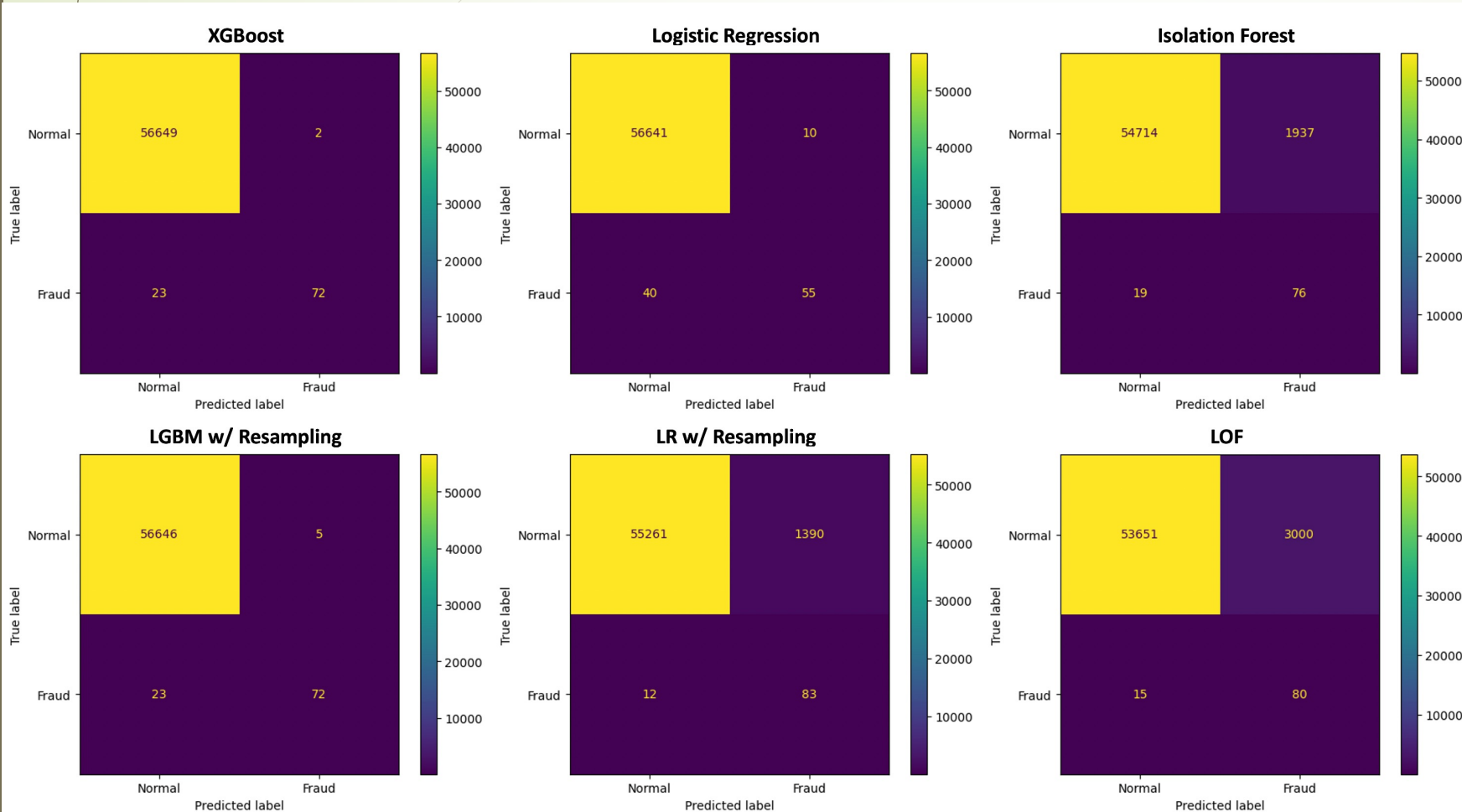
- **XGBoost** & **LightGBM** **with Resampling** (left two):
- Very few FALSE Positives, and relatively small number of FALSE Negatives.

- **Logistic Regression** (middle):
- Resampling significantly reduces FALSE Negatives but remarkably increases FALSE Positives.

- **Isolation Forest** and **LOF** (right):
- A very large number of FALSE Positives and a small number of FALSE Negatives.

4. Conclusions and Discussion

➡ **Best model for fraud detection:**

o **XGBoost** is the **best model,** considering both Recall and Precision and computational efficiency.

o If we only consider **Recall**, Logistic Regression with Resampling and two anomaly detection models (Isolation Forest and LOF) have very high scores.

➡ **Practical implications:**

o In Fraud Detection, identifying a fraudulent transaction as a normal one can result in significant financial losses, which means the cost of **False Negatives** is very high. In this case, **Recall** is a better metric. Therefore, those models with high Recall may also be good choices.

o On the other hand, a low **Precision** score indicates lots of **False Positives**, meaning that many normal transactions are identified as fraudulent. This situation can increase complaints from credit card users and cause credit card companies to lose customers.