CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING

Project Overview

- This project focuses on the creation of a machine learning model for credit card fraud detection for all stakeholders involved in the credit card company looking to improve its fraud detection system.

Business Understanding

- Credit card companies must recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.

Problem Statement

- High rate of false positives in the current system, which leads to unnecessary inconvenience for customers and additional workload for the fraud investigation team.

Objectives

- General Objective

Detection of credit card fraudulent transactions

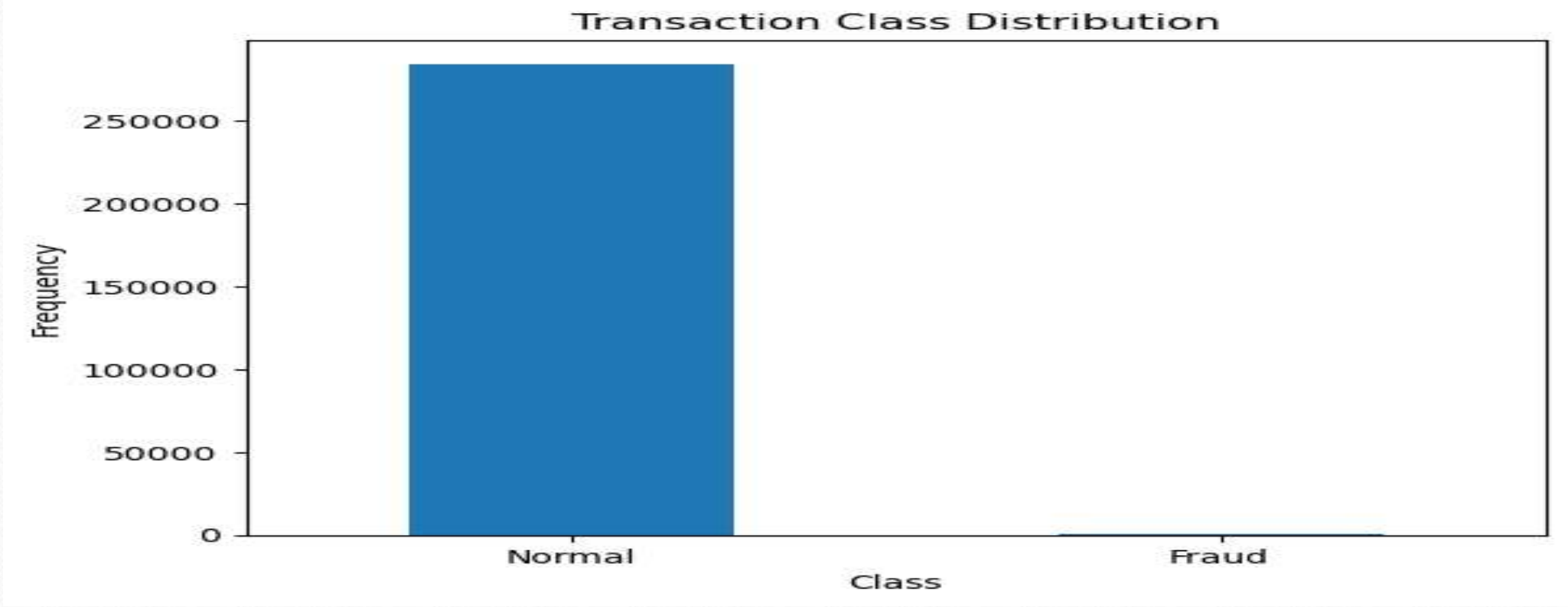
- Specific Objective

Build a fraud detection model on credit cards. Use the transaction and their labels as fraud or non-fraud to detect if new transactions made by the customer are fraud or not.

Data Understanding

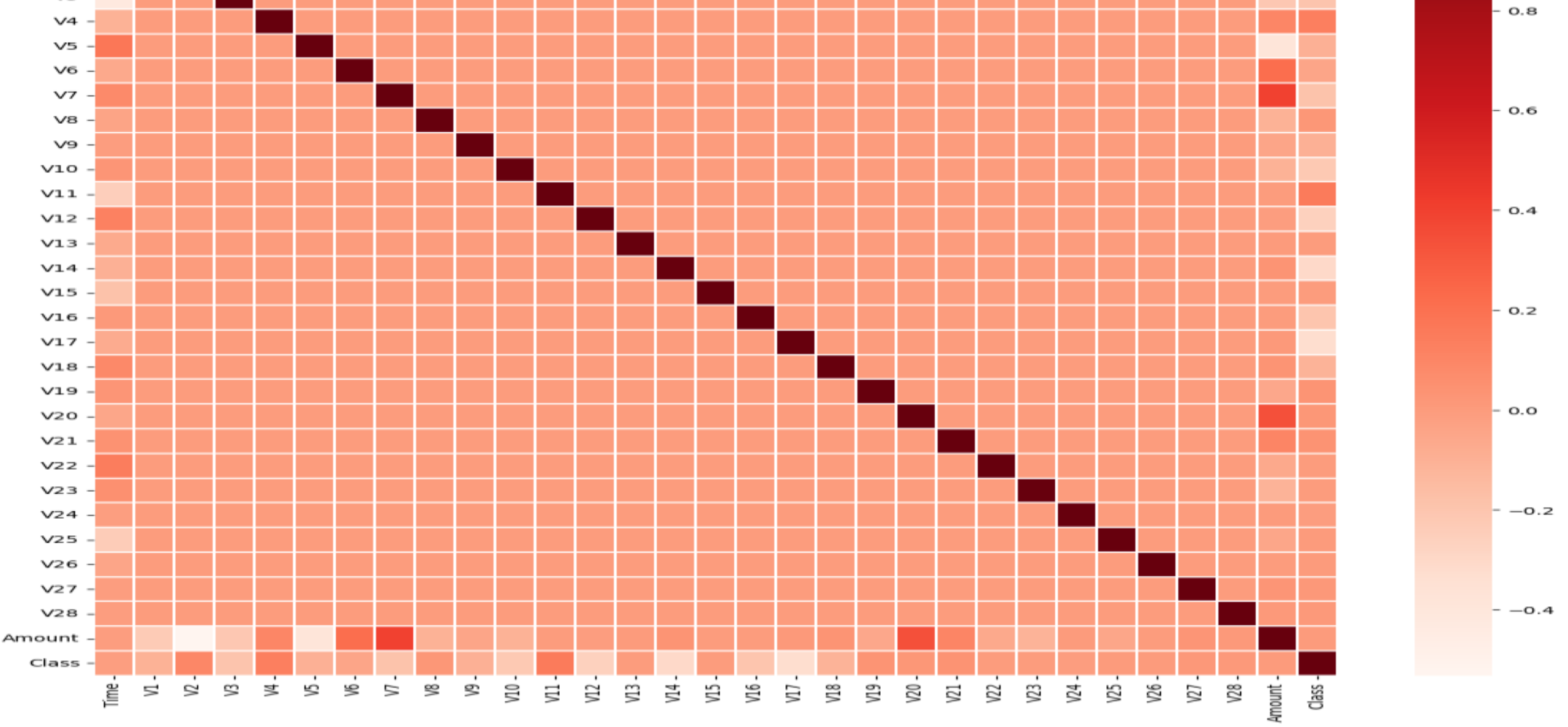
- The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.
- The dataset which consist of transactions made by European cardholders during month of July 2015 through various credit cards. The dataset consists of transactions occurred in span of two days, where we have 492 fraudulent transactions out of 284,807 total transactions
- It contains only numerical input variables which are the result of a PCA transformation. Features V1, V2, … V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'.

Visualizations

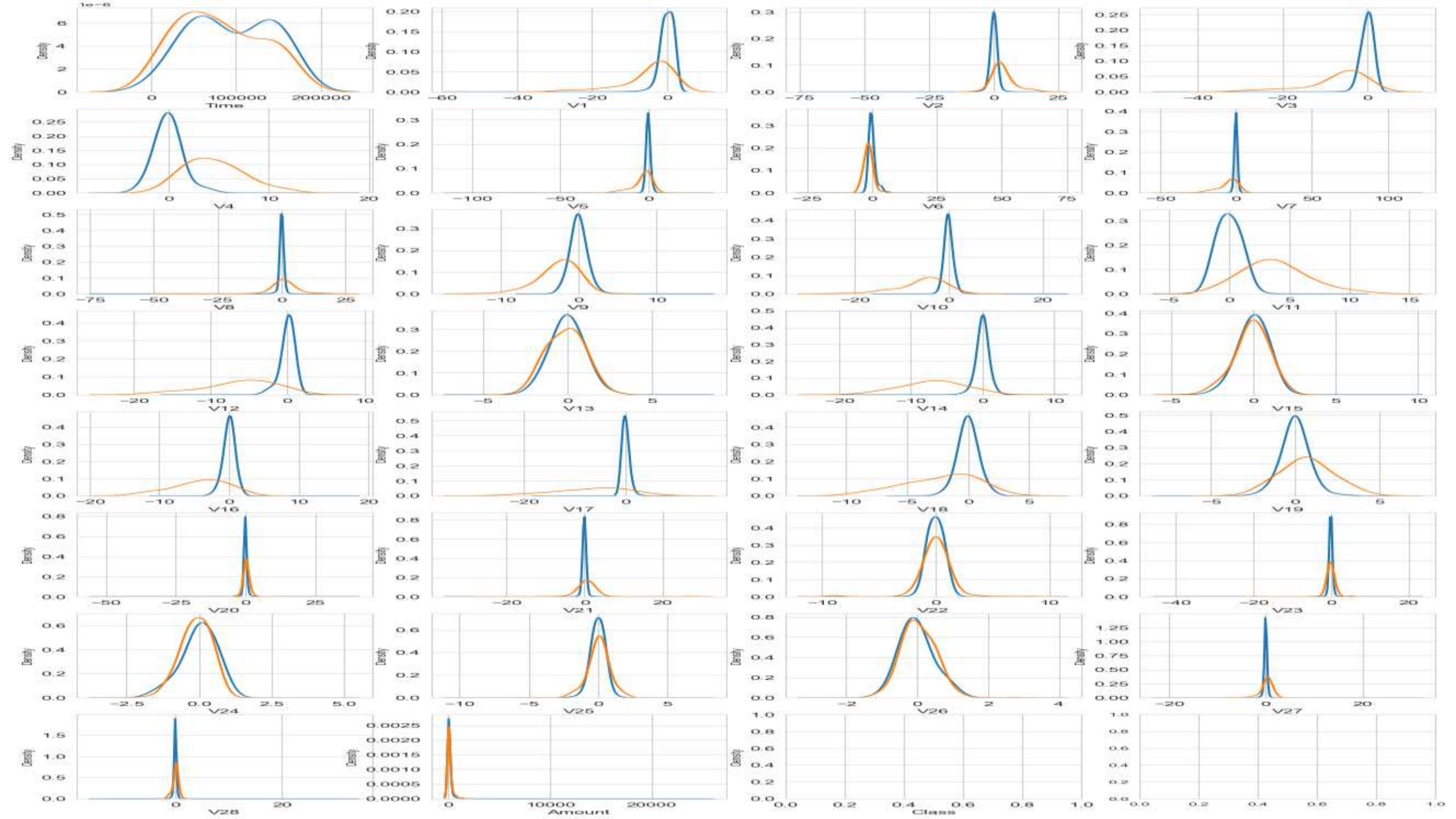


- Only 492 of transaction are fraudulent.
- The data is highly unbalanced with respect with target variable Class.

- By Describing both classes 0 and 1 we can clearly see that the real transactions has a larger mean value, larger Q1, smaller Q3 and Q4 and larger outliers; fraudulent transactions have a smaller Q1 and mean, larger Q4 and smaller outliers.



- As expected, there is no notable correlation between features V1-V28. As there are certain correlations between some of the features and Time (inverse correlation with V3) and Amount (direct correlation with V7 and V20, inverse correlation with V1 and V5).


Time
V1
V2
V3
V4
V5
V6
V7
V8
V9
V10
V11
V12
V13
V14
V15
V16
V17
V18
V19
V20
V21
V22
V23
V24
V25
V26
V27
V28
Amount
Class
Density

- Some of the features we can observe a good selectivity in terms of distribution for the two values of Class: V4, V11 have clearly separated distributions for Class values 0 and 1. V12, V14, V18 are partially separated.
- V1, V2, V3, V10 have a quite distinct profile whilst V25, V26, V28 have similar profiles for the two values of Class.
- In general, with just few exceptions (Time and Amount), the features distribution for legitimate transactions (values of Class = 0) is centered around 0, sometime with a long queue at one of the extremities. In the same time, the fraudulent transactions (values of Class = 1) have a skewed (asymmetric) distribution

Modelling

- We begin our modelling with a baseline logistic regression model followed by an intermediate Random forest model and cap it off with a hyper parameter tuned complex model. The models will be evaluated based on their precision, accuracy, recall and F1 score. The model with the best metrics will be chosen for use and deployment.

Model Evaluation

- Hyper parameter tuning was only performed on the Random Forest model and which had a higher F1 score compared to the logistics regression and it was the one chosen for the modelling.
- Indicating an improvement in the model's overall performance, particularly in terms of balancing precision and recall.
- This suggests that the tuning process and increased model complexity have resulted in a model that is better at distinguishing between legitimate and fraudulent transactions.

Challenges

- Imbalanced Data: Fraudulent transactions are typically rare compared to legitimate ones, leading to imbalanced datasets.
- Evolution of Fraud Patterns: Fraudulent activities are dynamic and can evolve over time. Fraudsters continually adapt their techniques to avoid detection.
- Feature Engineering: Identifying relevant features for fraud detection can be challenging.
- High-Dimensional Data:
- Credit card transaction datasets can be high-dimensional, with numerous features. Managing and processing large amounts of data efficiently can be computationally demanding.
- Cost of False Positives: False positives (incorrectly flagging a legitimate transaction as fraudulent) can inconvenience customers and result in financial losses for the business.

Conclusion

- In conclusion, developing a robust credit card fraud detection model involves navigating through various challenges inherent to the dynamic and complex nature of financial transactions. The imbalanced distribution of fraudulent and legitimate transactions, the evolving tactics of fraudsters, and the need for interpretability and compliance present significant hurdles.