Credit Card Fraud Detection

Review and explore the credit card fraud detection

Introduction:

- Credit card fraud normally happens when consumers use their credit card number to unfamiliar individuals, when cards are lost or stolen, when mail is diverted from the intended recipient and taken by criminals, or when employees of a business copy the cards or card numbers of a cardholder. In this notebook I will develop a few ML models using anonymized credit card transaction data.

Problem Statement:

- The Credit Card Fraud Detection Problem includes modeling past credit card transactions with the knowledge of the ones that turned out to be fraud. This model is then used to identify whether a new transaction is fraudulent or not. my aim here is to detect 100% of the fraudulent transactions while minimizing the incorrect fraud classifications.

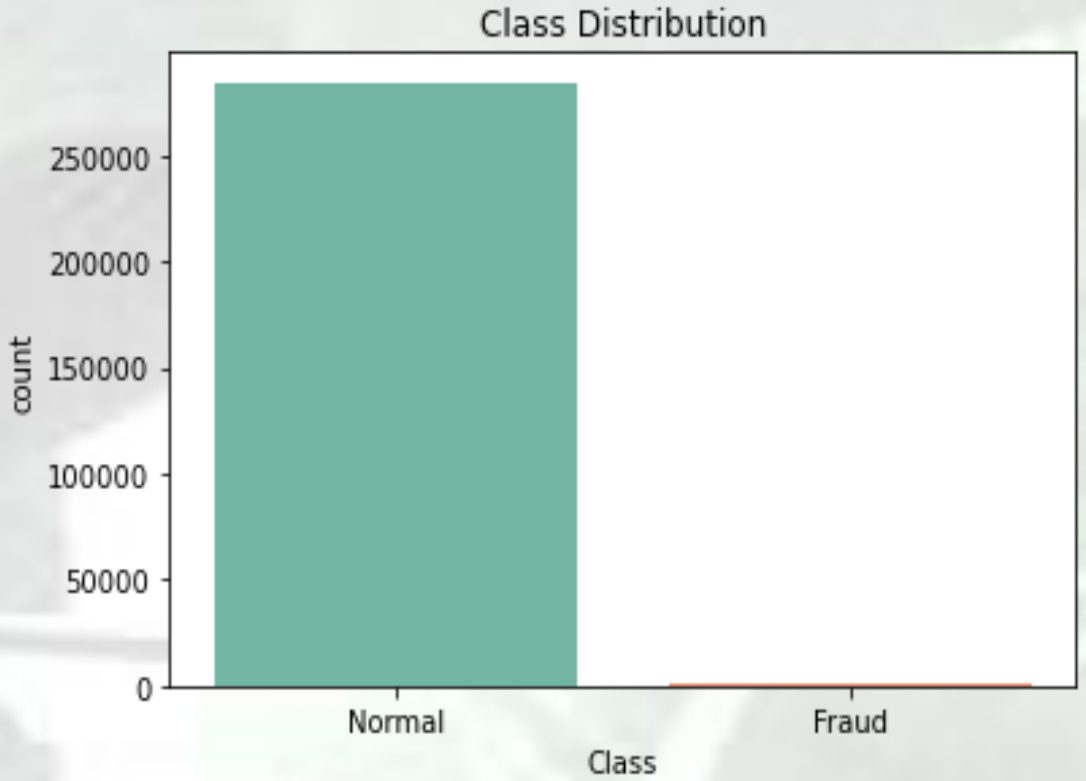
Methodology :

I’ve used the OSEMiN data science workflow, which involves:

- Obtain (import the data)
- Scrub (clean the data, deal with missing values and data types)
- Explore (answer descriptive questions using EDA)
- Model (model selection)
- Interpret (comment on our model and findings)

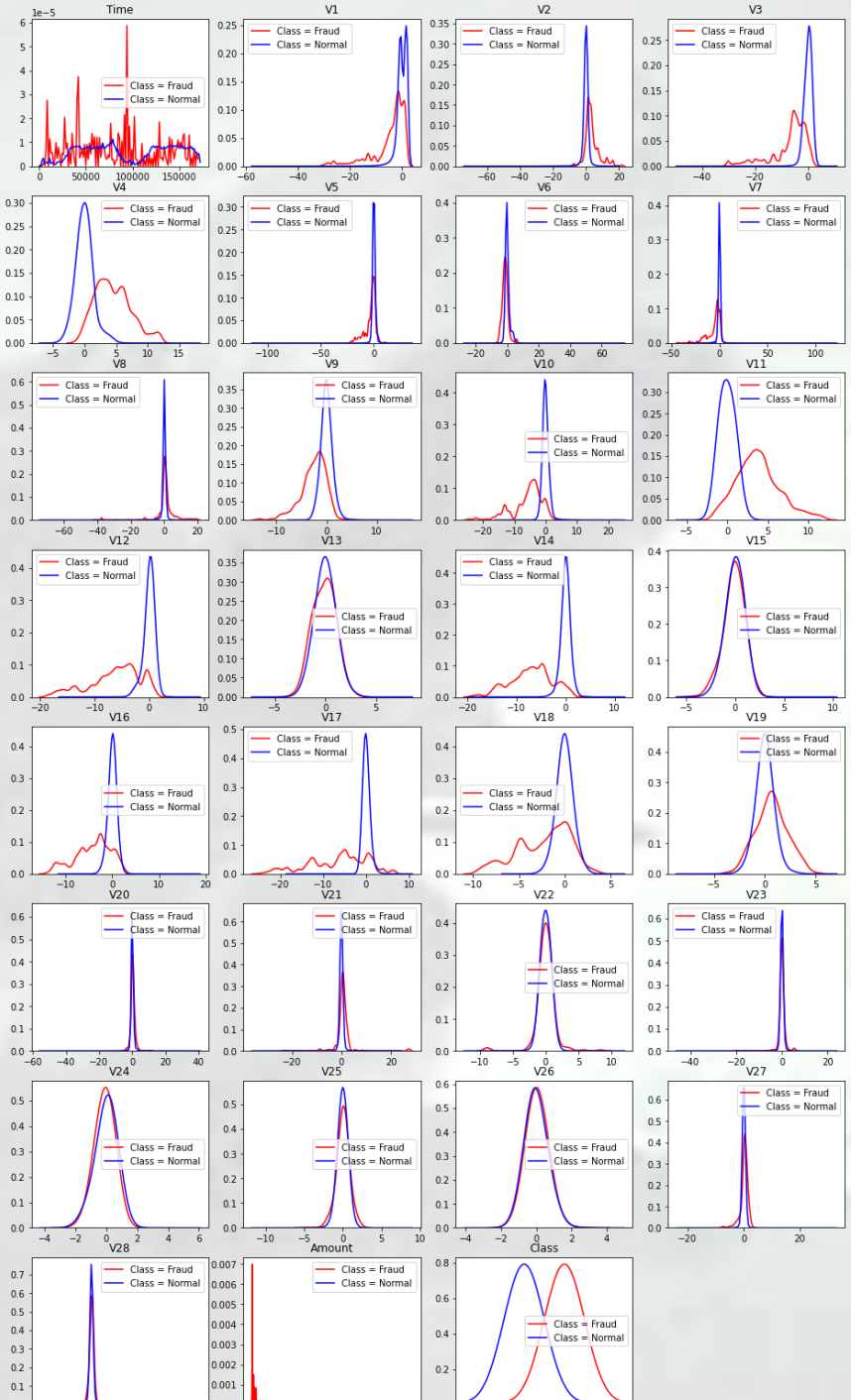
Dataset:

The datasets contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions.

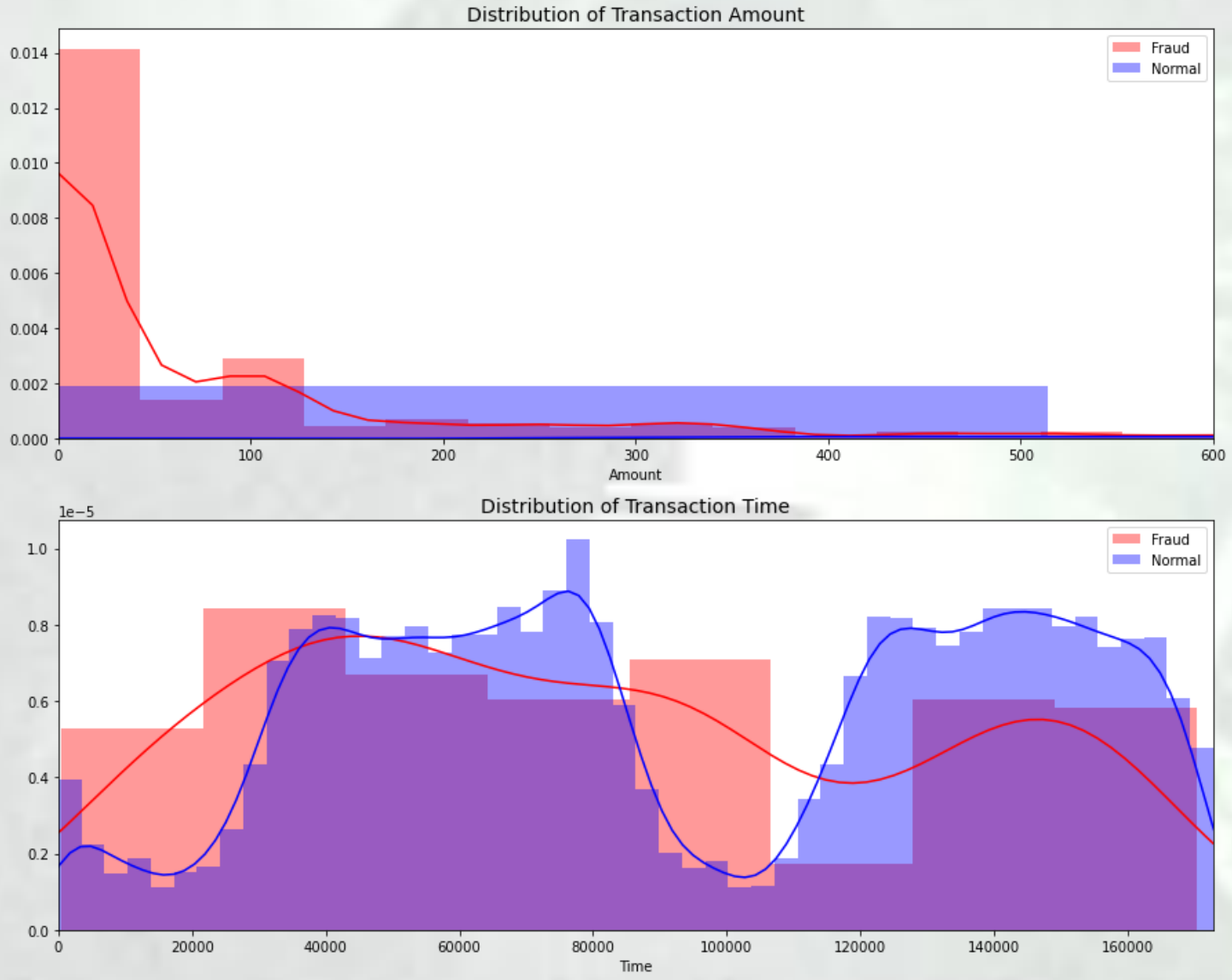


Transaction	Count	percent
Normal	284,315	99.83%
Fraud	492	0.17%

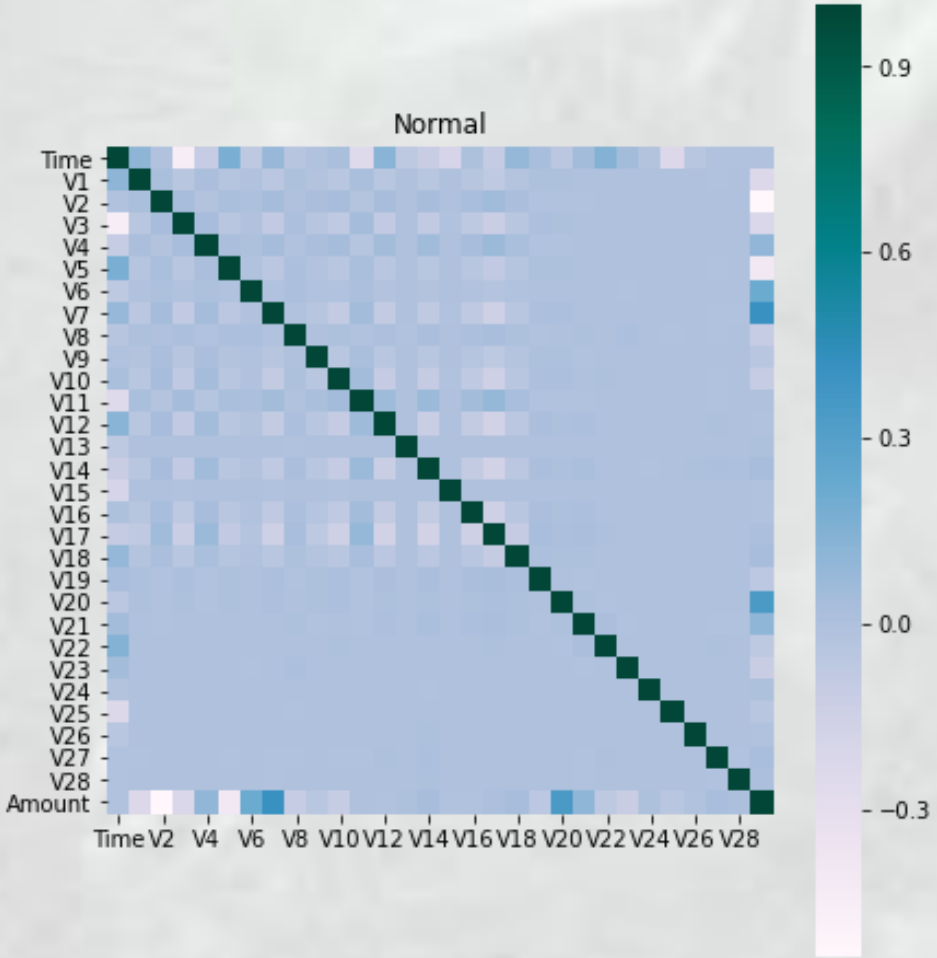
Dataset:

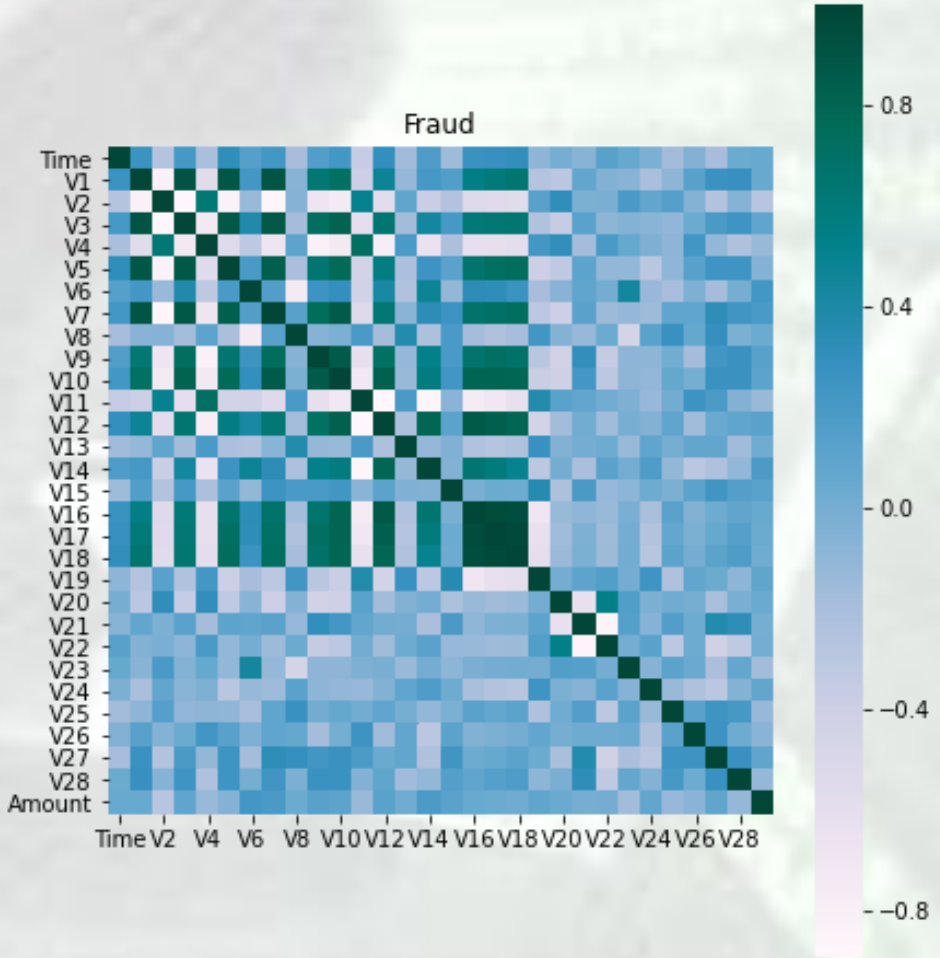

Time
V1
V2
V3
V4
V5
V6
V7
V8
V9
V10
V11
V12
V13
V14
V15
V16
V17
V18
V19
V20
V21
V22
V23
V24
V25
V26
V27
V28
Amount
Class
Class = Fraud
Class = Normal

Distributions of Transaction Time and Transaction Amount:



Correlation Between variables:

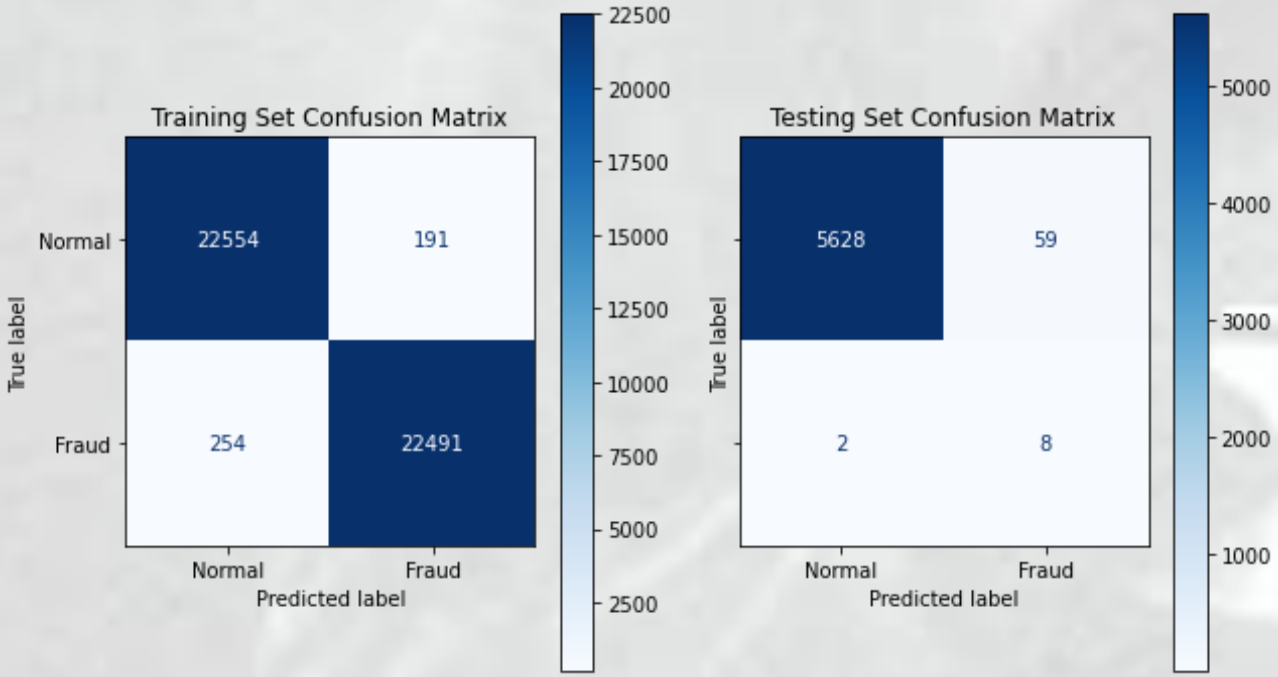


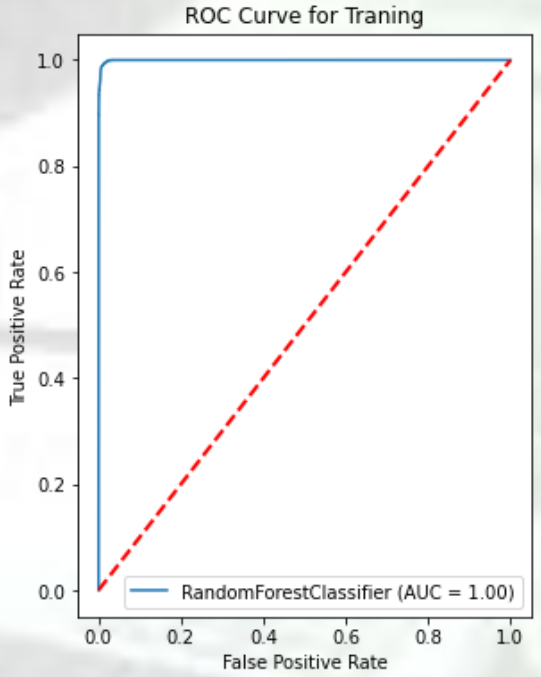


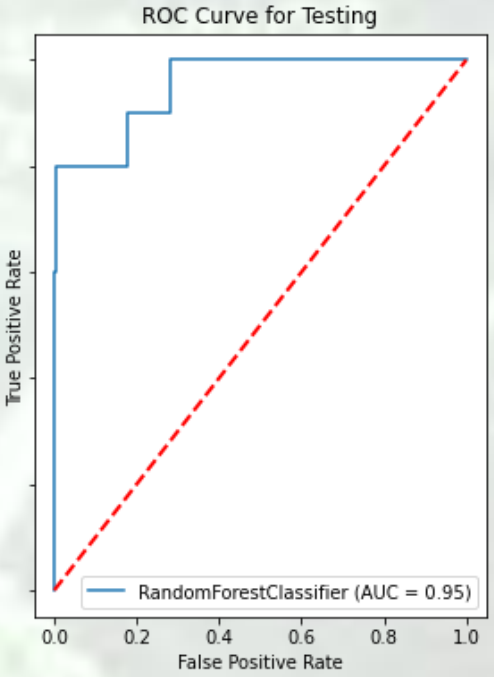
Train and Evaluate Models:

Models	Accuracy	Precision	Recall	F-1	AUC
Random Forest	99.90%	100.00%	70.00%	82.30%	95.00%
Logistic Regression	0.17%	0.17%	100.00%	35.00%	50.00%
Supper Vector Machine	0.17%	0.17%	100.00%	35.00%	50.00%
XGBoosting	99.50%	24.20%	80.00%	37.20%	94.00%
GBM	99.80%	54.00%	60.00%	57.00%	94.00%

Random Forest Classification:







Future Work:

- Improve models by using bagging algorithm.
- Fit LightGBM and CatBoosting to the data set.

Thank You!