Credit Card Fraud Detection

Springboard Capstone Project I

Serik Omarov – August 2021

Email: serik.o.omarov@gmail.com

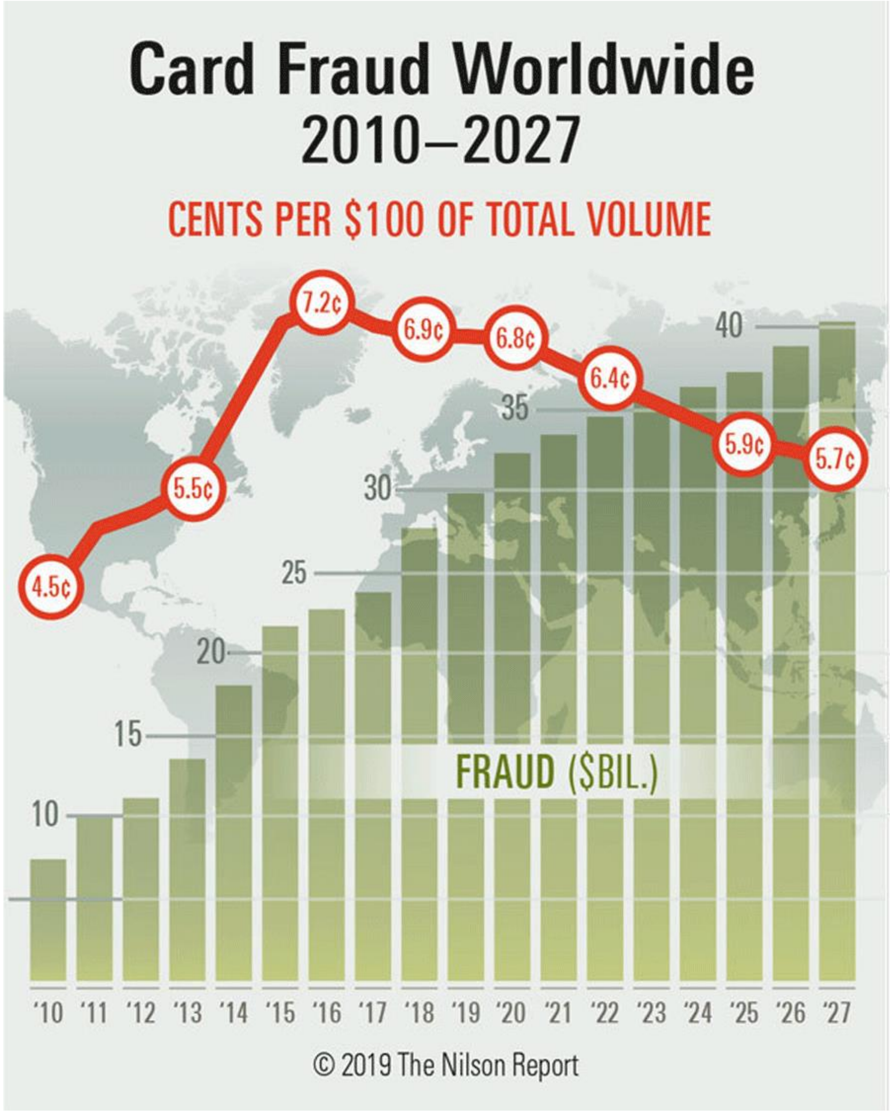
Agenda

- Problem Statement and Introduction
- Exploratory Data Analysis
- Model Selection
- Model Evaluation
- Conclusion

The Problem

"While fraud as a percentage of all card dollar volume declined for the second year in a row, criminals saw double-digit growth in the money they were able to steal from the system. Card fraud netted criminals $3.88 billion more in 2018 than in 2017,"

Publisher of The Nilson Report.



Who does this affect?

Banks

Lost Revenue & Work Hours

Businesses and Consumers

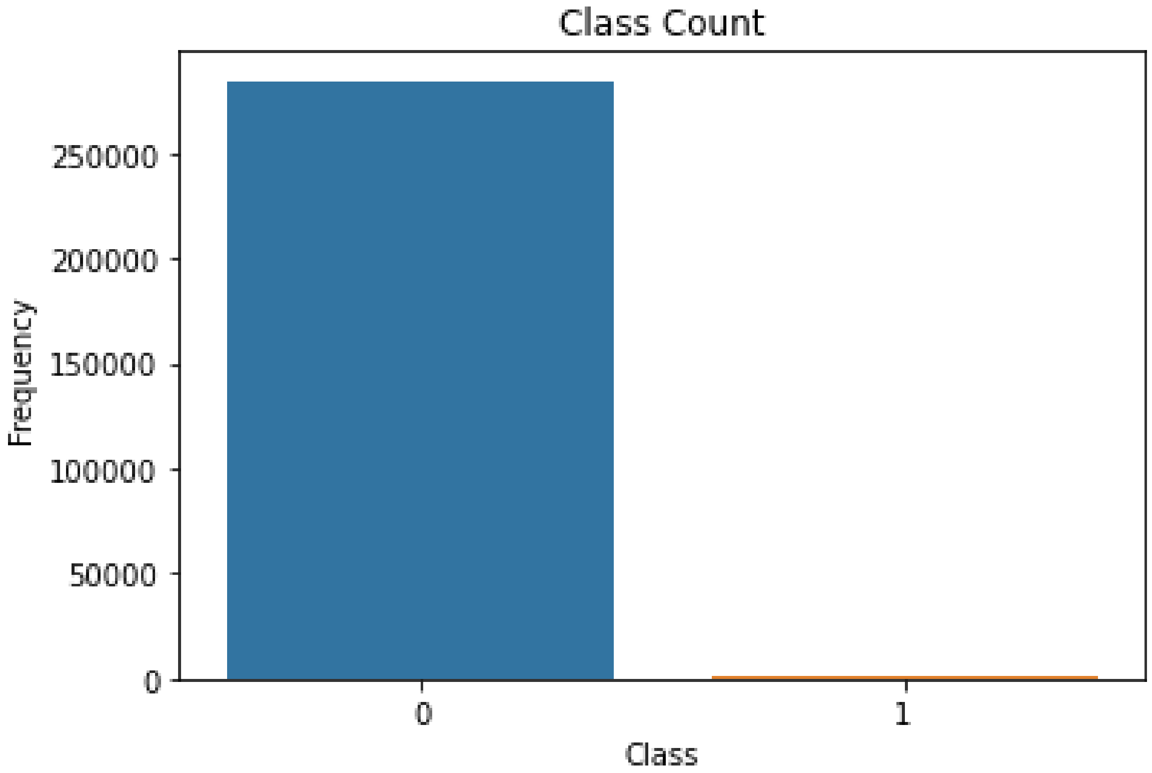
Frustration & Lost Time

The Data

- Dataset from Kaggle
- Credit Card transactions made in September 2013 by European Cardholders
- Highly imbalanced

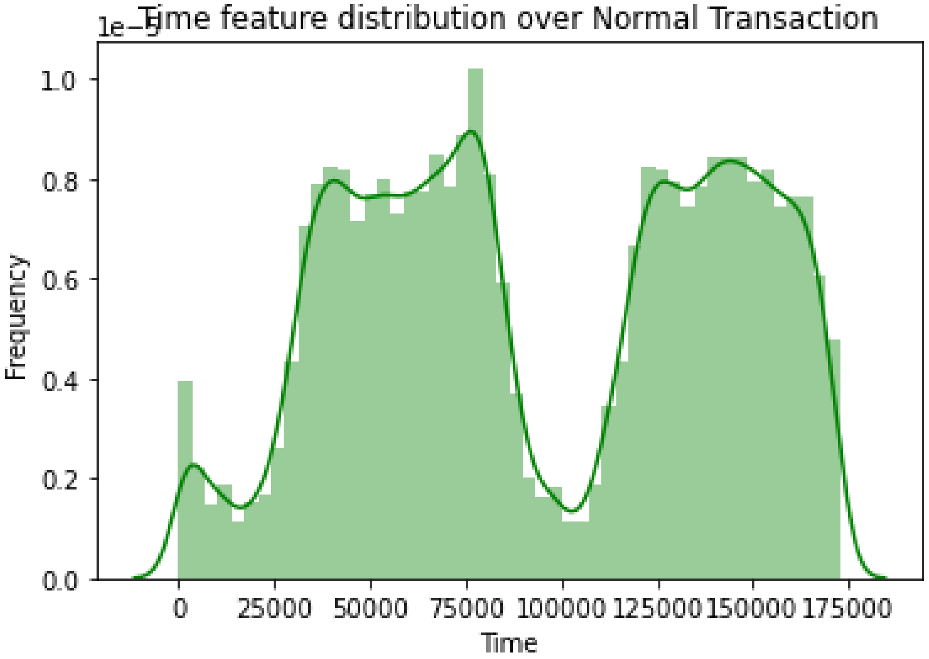
Exploratory Data Analysis

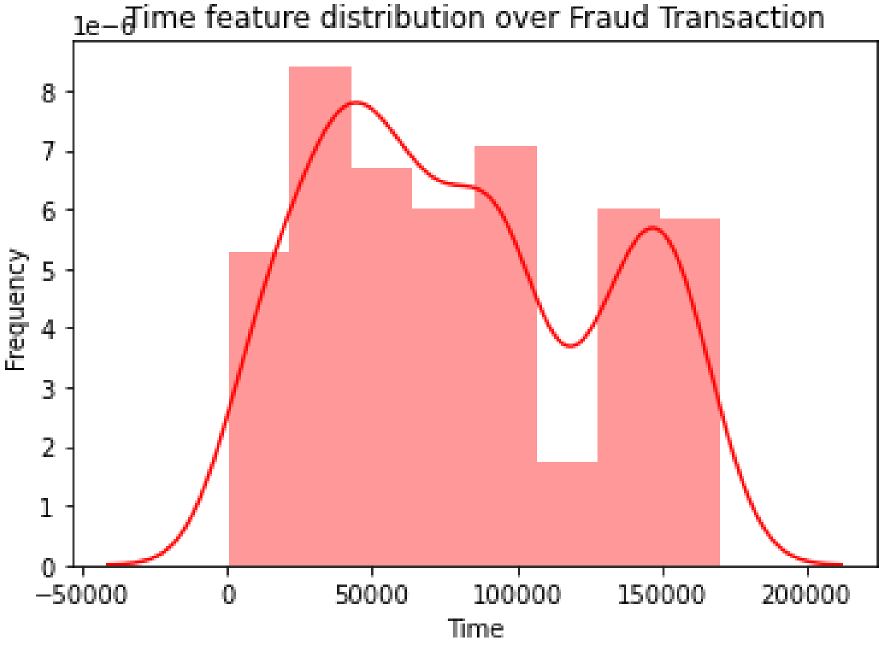
Class Type



- Normal 99.83 % of the dataset
- Frauds 0.17 % of the dataset

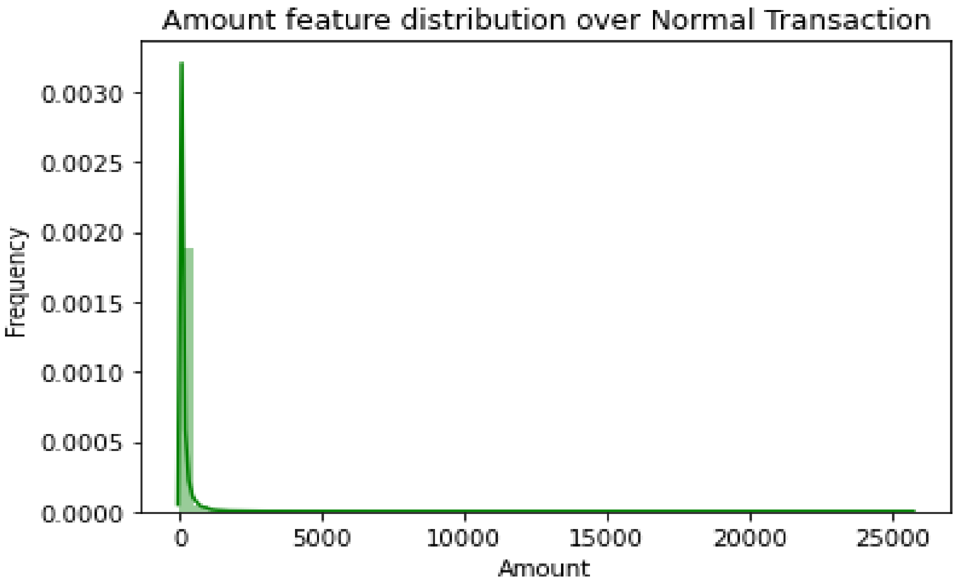
Exploratory Data Analysis

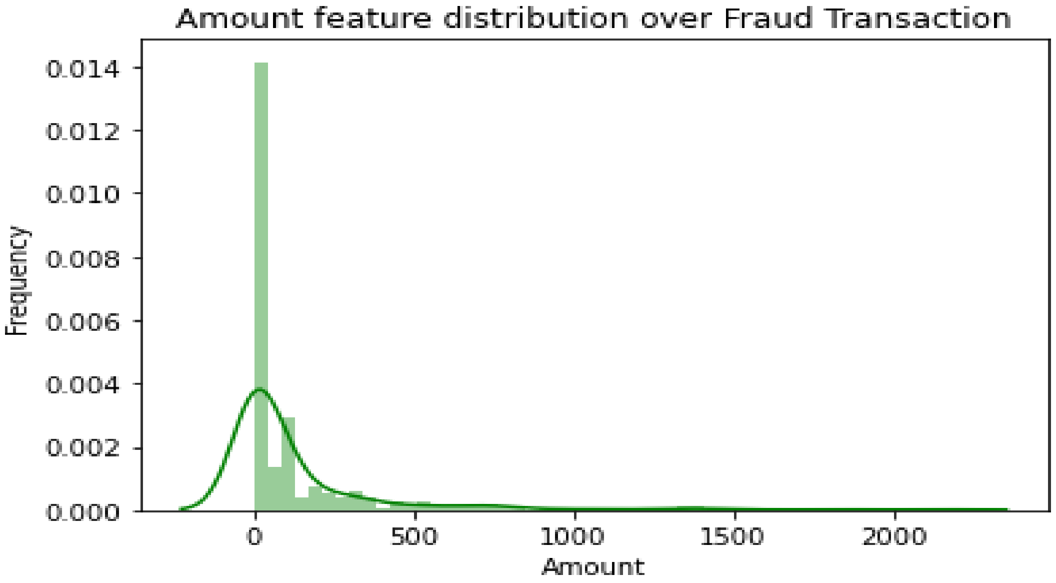




- Distribution of Time over Normal Transactions, we can interpret that they are 2 peaks in the distribution and nothing unusual.
- Distribution of Time over Fraud Transactions, we can interpret that it is a normal distribution and nothing unusual.

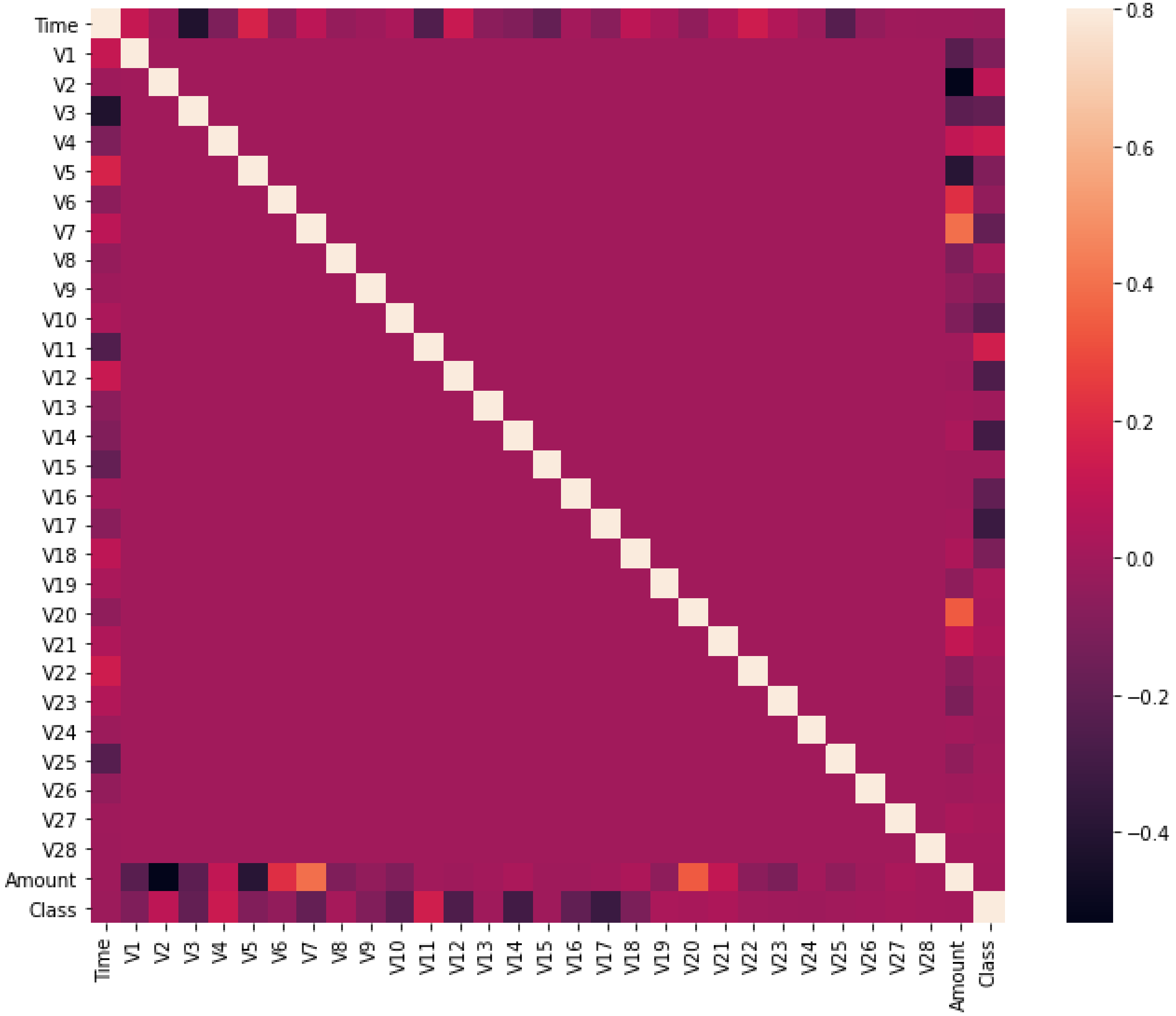
Exploratory Data Analysis





- Distribution of Amount over Normal Transactions, we can interpret there is peak at the beginning, but it becomes flat after the peak
- Distribution of Amount over Fraud Transactions, we can interpret there is huge peak at the beginning, but it becomes flat after 900

Correlation Matrix



Exploratory Data Analysis - Summary

	MEAN	STD	MAX
Normal	88.29	250.1	25691.16
Fraud	122.21	256.68	2125

Exploratory Data Analysis - Class

0.173%

492 fraud cases out of 284,807 transactions

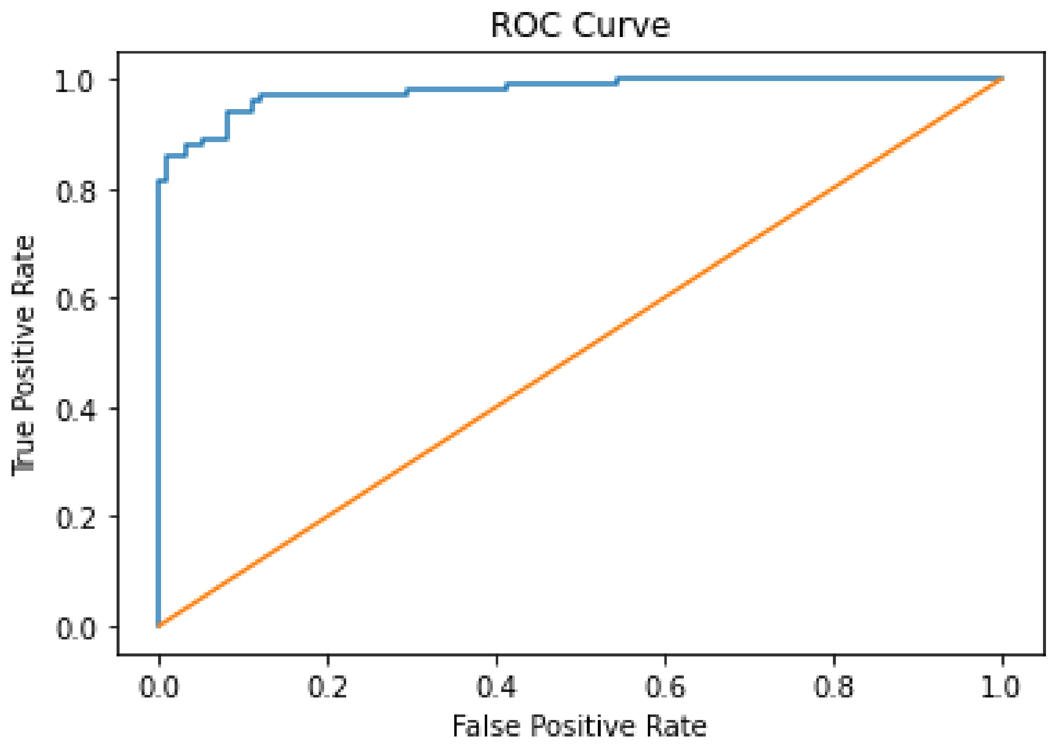
Model Selection

- Supervised Learning
  - Classification model - Binary (0 for non-fraud; 1 for fraud)
  - Highly imbalanced
  - Tools: scikit-learn

Model Accuracy Results

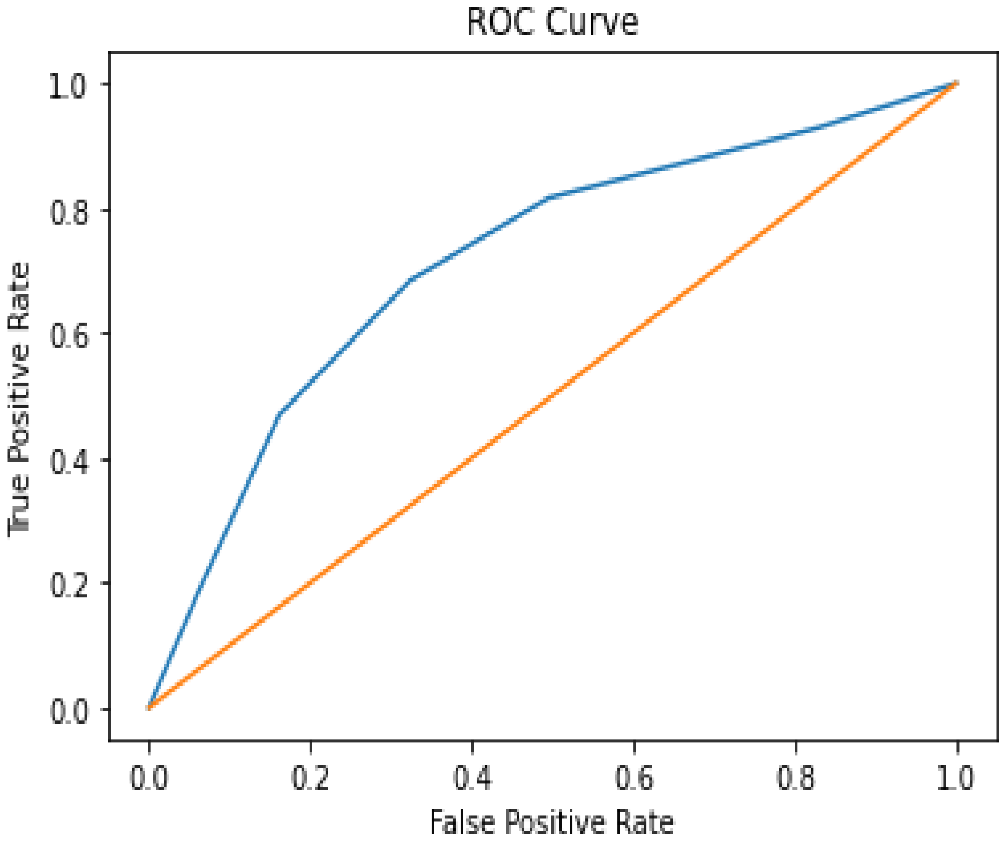
Logistic Regression	94.9%
Support Vector Machine	90.9%
Naive Bayes	86.3%
K-nearest Neighbors	68.5%

Model Evaluation: Logistic Regression



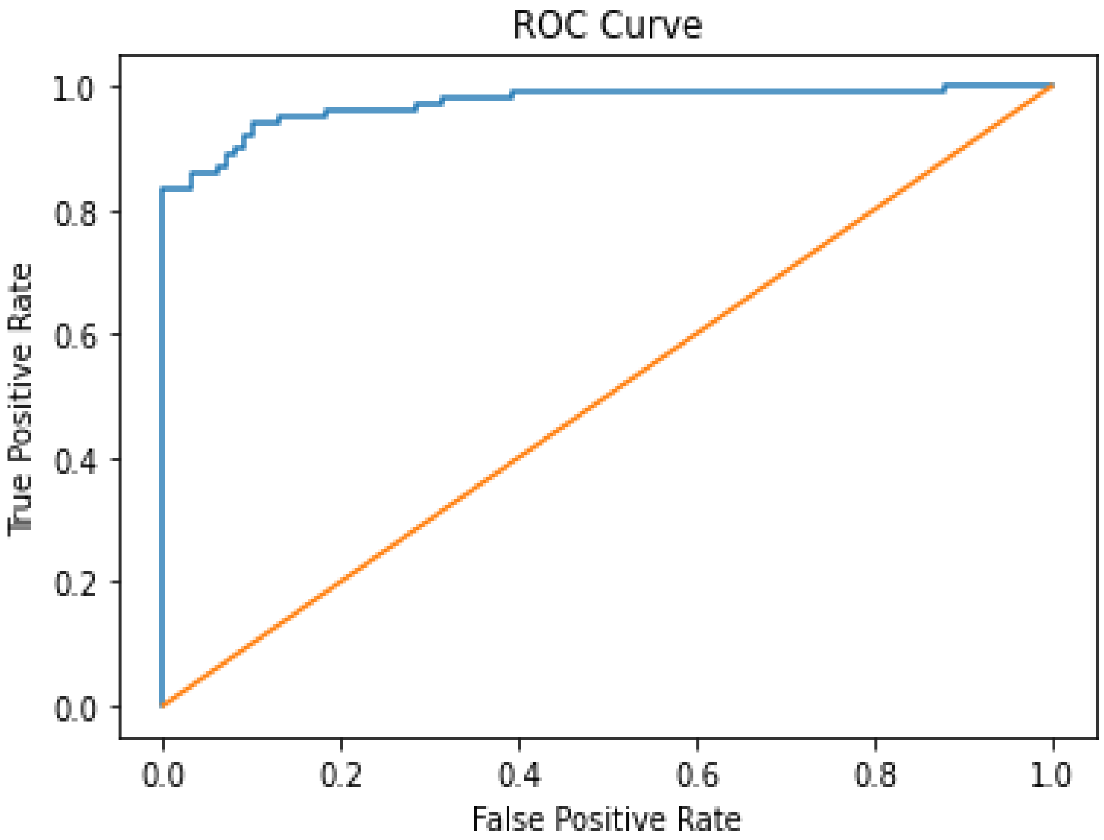
- As we can see that our Logistic Regression model performed very well in classifying the Credit Card Transactions, with an Accuracy score, Precision and Recall of approx. 93%. And the roc curve and confusion matrix are great as well which means that our model is able to classify the labels accurately, with fewer chances of error.

Model Evaluation: K Nearest Neighbors



As we can see that our KNN Classifier model not performed well in classifying the Credit Card Transactions, with an Accuracy score, Precision and Recall of approx 63%. And the roc curve and confusion matrix are not good as well which means that our model is not able to classify the labels accurately

Model Evaluation: Naive Bayes



As we can see that our Naive Bayes model performed well in classifying the credit card transactions, with an Accuracy score, Precision and Recall of approx. 86%. And the roc curve and confusion matrix are good as well which means that Naive Bayes model is able to classify the labels accurately, with fewer chances of error.

Conclusion

The results are highly clear that except KNN all the models are good in detecting the fraud transactions.

Logistic Regression is the most accurate method because its ability to handle binary data.

Also, it performs well even with the presence of principal components of features and is relatively unaffected by them.