Data Science 4th Assignment

Dataset: Credit Card Transactions Fraud

Mahan Madani – 99222092

1. Overview

This dataset is comprised of about 1.3 million records of legitimate and fraudulent transactions from the time period of 1st Jan 2019 - 31st Dec 2020. It contains information in 23 different columns.

The records in the dataset were generated using the Sparkov Data Generation tool. Transactions were generated across various different profiles and then merged together to simulate a more realistic representation of transactions.

2. Dataset Exploration & Analysis

The dataset stores 23 different attributes. Each record in the dataset belongs to one specific transaction.

Below you can find a list of all the attributes:

- **Unnamed: 0:** The index of the record, starting from 0.
- **trans_date_trans_time:** The date and time of the transaction.
- **cc_num:** A unique numeric as the credit card number.
- **merchant:** The name of the seller entity in the transaction.
- **category:** The category of the sold product.
- **amt:** The amount of money transferred in the transaction.
- **first:** The first name of the credit card owner.
- **last:** The last name of the credit card owner.
- **gender:** The gender of the credit card owner.
- **street:** The name of the credit card owner's street of residence.
- **city:** The name of the credit card owner's city of residence**.**
- **state:** The name of the credit card owner's state of residence**.**

- **zip:** The zip code of the credit card owner.
- **lat:** The latitude of the credit card owner's location.
- **long:** The longitude of the credit card owner's location.
- **city_pop:** The population of the city
- **job:** Credit card owner's job.
- **dob:** Credit card owner's date of birth.
- **trans_num:** A unique numerical identifier for a transaction.
- **unix_time:** The time of the transaction (Unix time).
- **merch_lat:** The latitude of the merchant's location.
- **merch_long:** The longitude of the merchant's location.
- **is_fraud:** A Boolean value indicating whether or not a transaction was done with a fraudulent credit card.

3. Data Preprocessing

The dataset is generated by a simulator, so it is mostly clean and processed. There is no need to check for:

- Duplicate records
- Null values
- Outlier values

3.1 Feature Generation:

Based on the available data, the following features can be added to further improve the dataset:

- **Year:** The year based on the transaction's datetime.
- **Month:** The month based on the transaction's datetime.
- **Day:** The day based on the transaction's datetime.
- **Hour:** The hour based on the transaction's datetime.
- **Odd hours:** A Boolean indicating whether or not a transaction occurred during an odd hour (between 22 and 6).
- **Gender_binary:** A Boolean to represent male and female customers.
- **Age:** The age of the credit card owner.

3.2 Feature Transformation:

Training and testing a model require numerical data. Some of the features must be transformed so that they can be used in machine-learning tasks.

Categorical features like 'category' or 'state can be vectorized using **one-hot encoding**, creating a new column for each unique category in them.

Feature Scaling: All of the numerical attributes must be scaled similarly. Otherwise, having features with varying degrees of magnitude and range will cause different step sizes for each feature. I used a **Standard Scalar** to scale all numerical features.

4. Visualization

Figure 4.1: Bar plot of the categories of fraudulent transactions

This plot displays the most frequently repeated categories of products sold in fraudulent transactions.

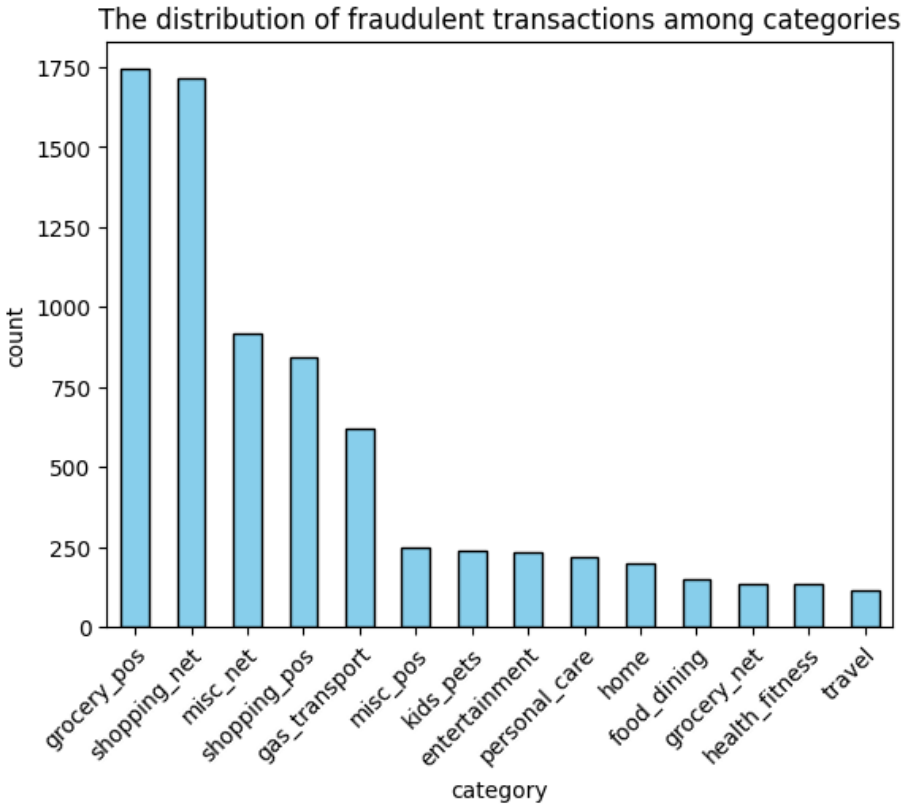


Figure 4.2: Pie chart of the distribution of gender on fraudulent transactions

This plot displays that an almost equal amount of men and women have committed credit card fraud in our dataset, rendering the gender feature less valuable.

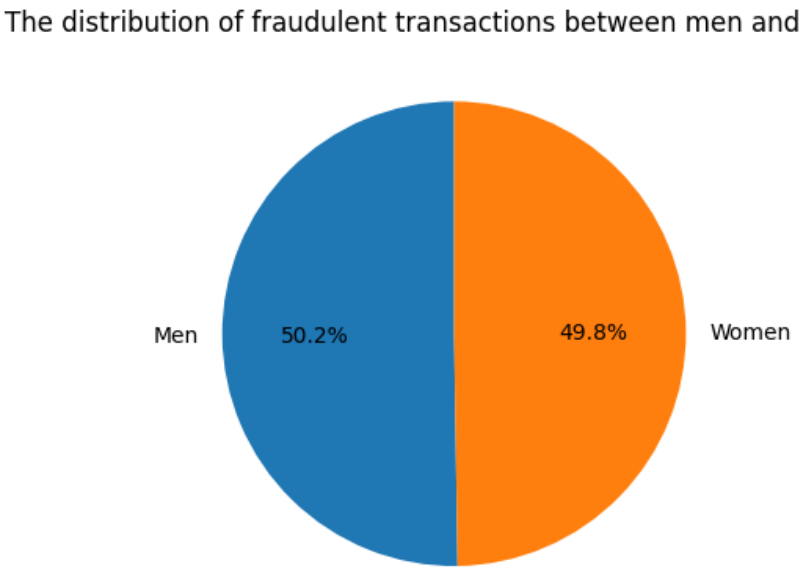


Figure 4.3: The Age Distribution of fraudulent credit card owners

We can see that age can be an important factor when attempting to detect credit card fraud.

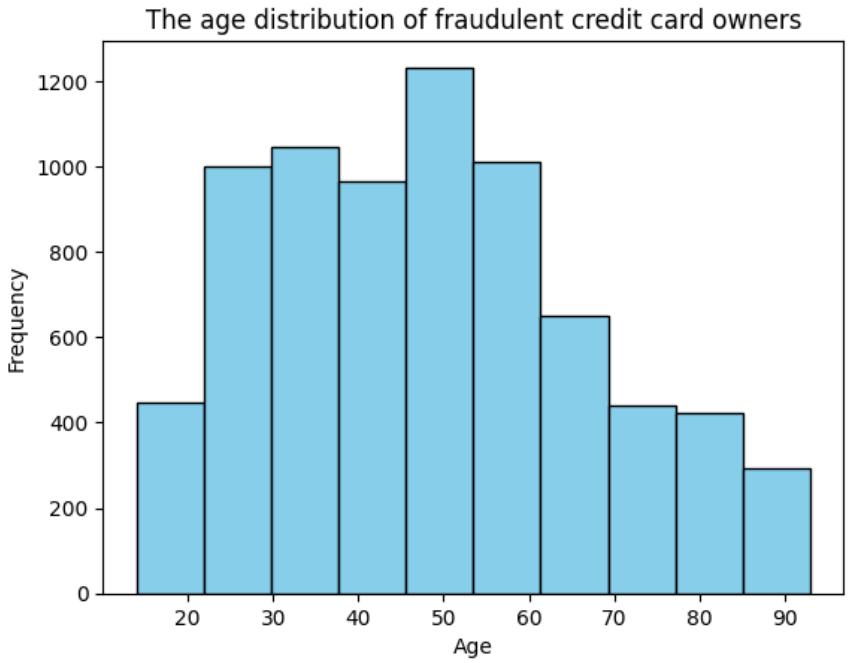
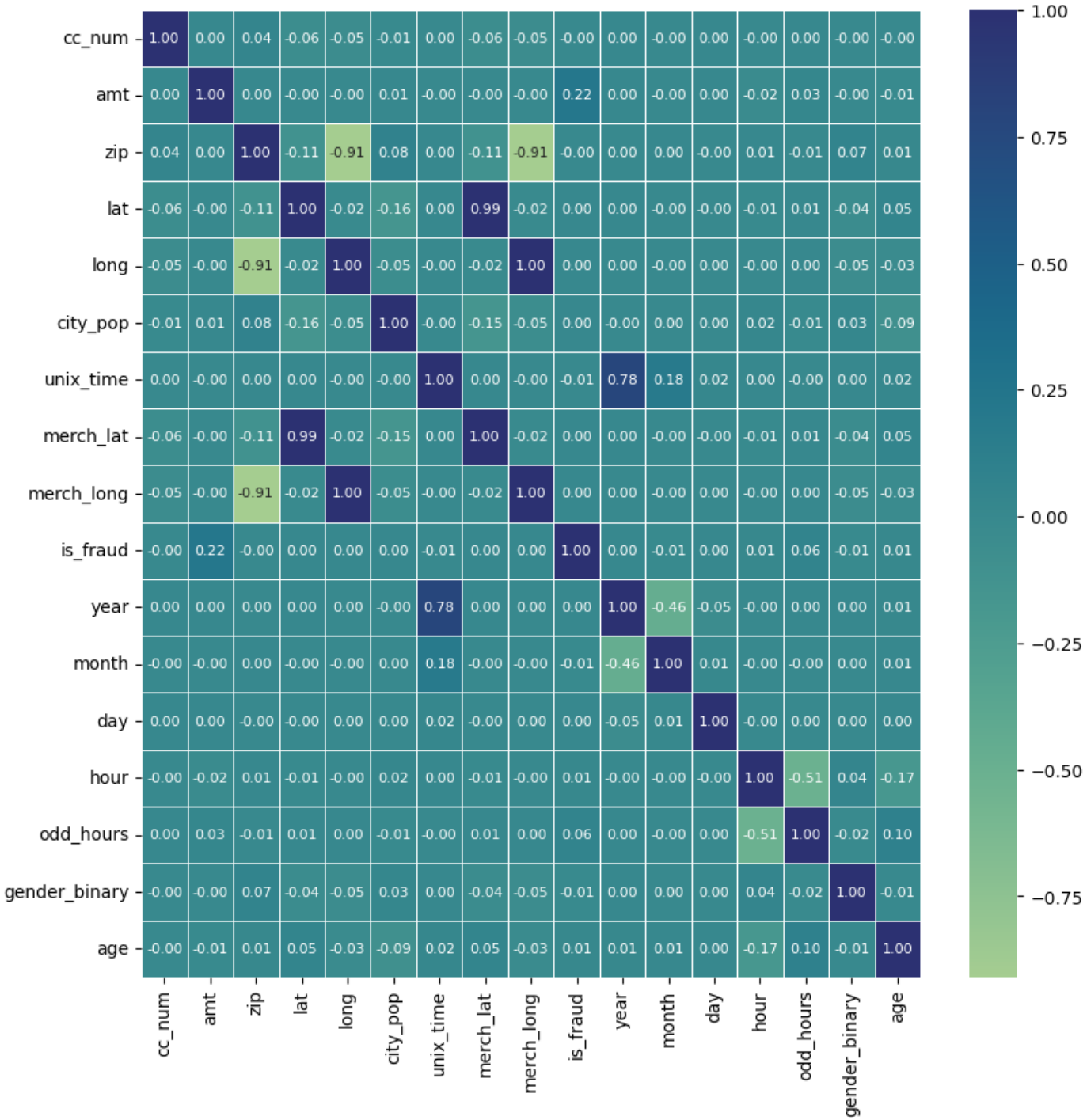


Figure 4.4: Heatmap of the Correlation Matrix

This heatmap visualizes the correlation between the various attributes of the dataset. The important correlations here are the ones between the 'is_fraud' attribute and the rest of the features.

'amt' and 'odd hours' seem to be the most correlated features.



5. Modeling

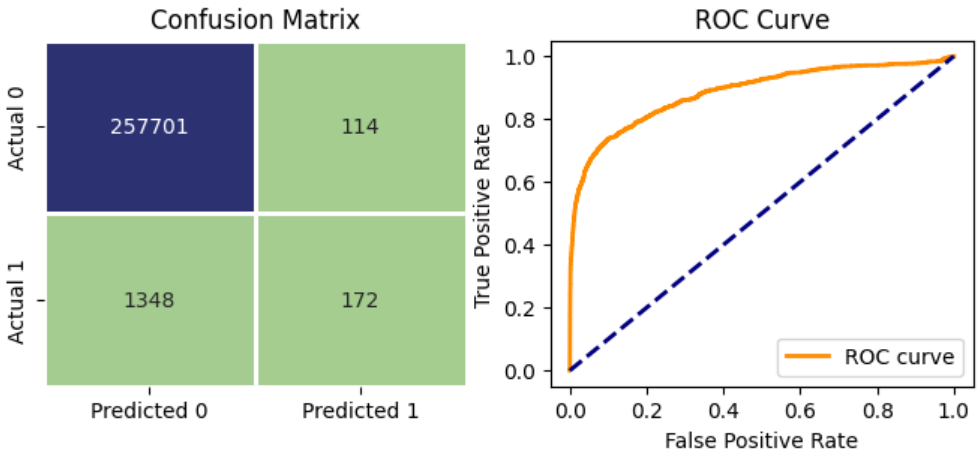
*I found the **Random Forest Classifier** model to be the all-around best choice.

5.1 Logistic Regression Classifier:

One of the simplest models available, but it's less accurate in the long run.

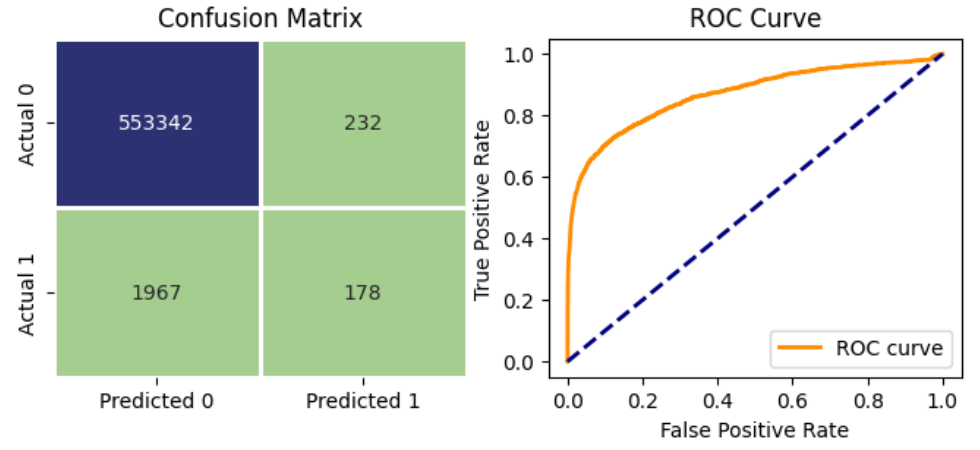
Testing the model on the train data set:

Accuracy: 0.994363
AUC: 0.884959
F1 Score: 0.190476



Testing the model on the test data set:

Accuracy: 0.996043
AUC: 0.868663
F1 Score: 0.139335

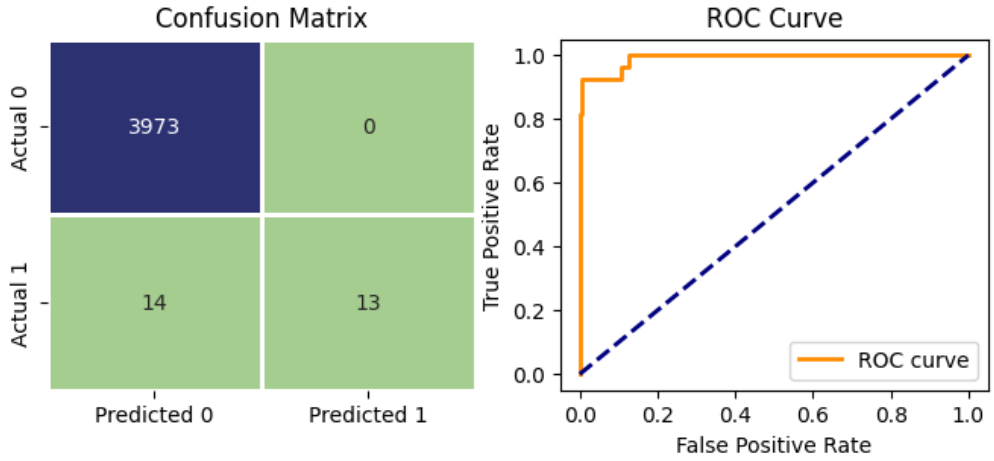


5.2 Support Vector Machine (SVM) Classifier:

Training the SVM model requires high processing power and takes a long time.

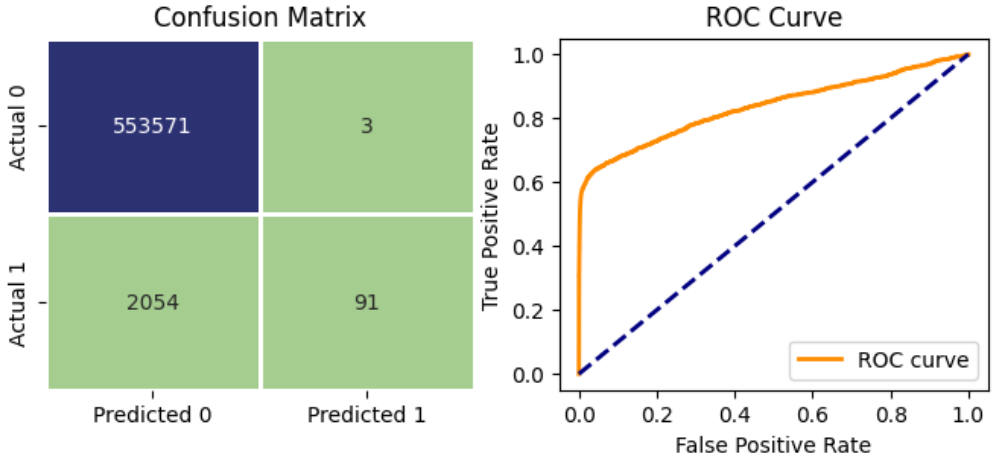
Testing the model on the train data set:

Accuracy: 0.9965
AUC: 0.990538
F1 Score: 0.65



Testing the model on the test data set:

Accuracy: 0.996298
AUC: 0.83602
F1 Score: 0.0812863

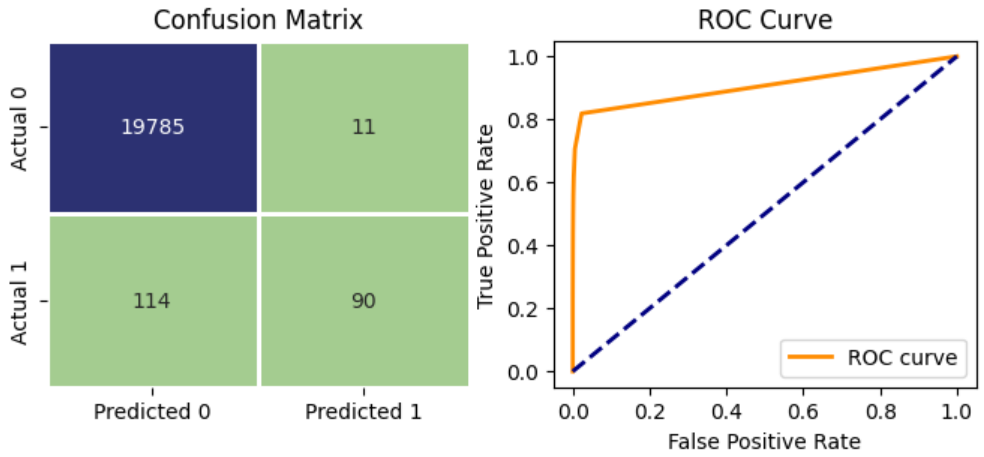


5.3 K-nearest neighbor (KNN) Classifier:

Training the KNN model requires an average amount of processing power.

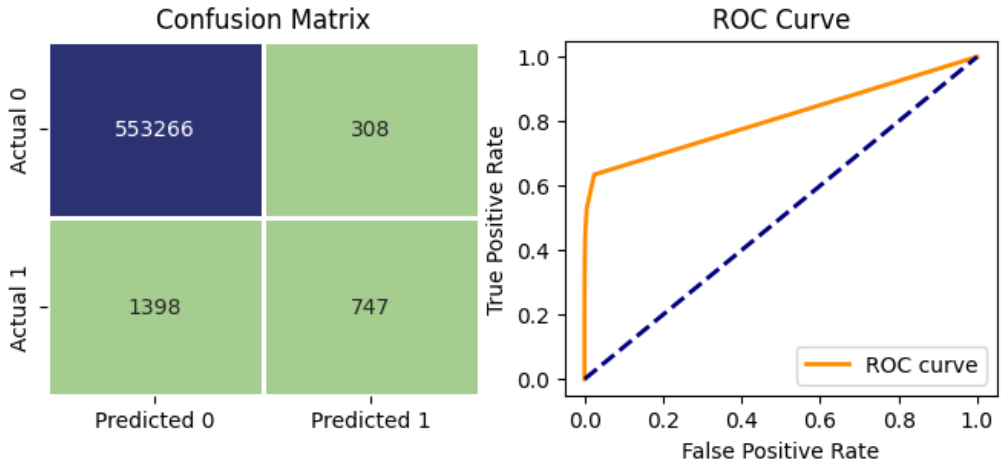
Testing the model on the train data set:

Accuracy: 0.99375
AUC: 0.904991
F1 Score: 0.590164



Testing the model on the test data set:

Accuracy: 0.99693
AUC: 0.810707
F1 Score: 0.466875

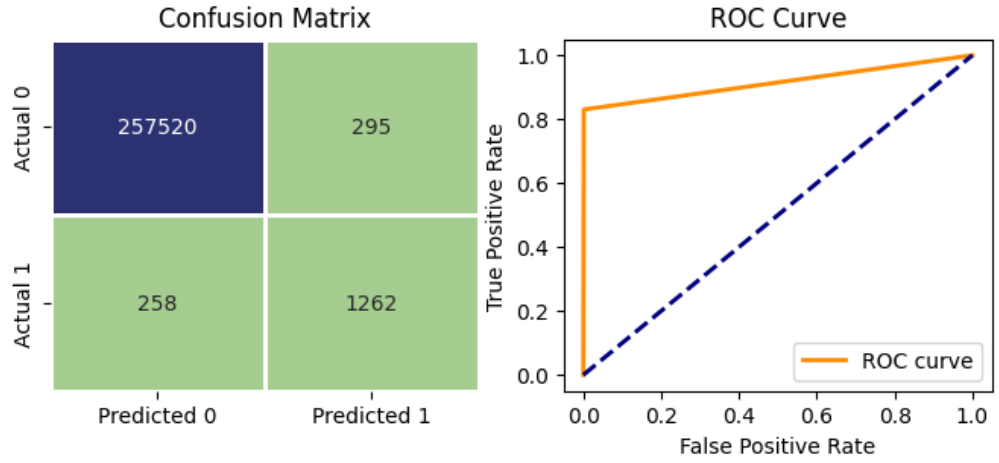


5.4 Decision Tree Classifier:

An extremely lightweight model but more accurate than logistic regression.

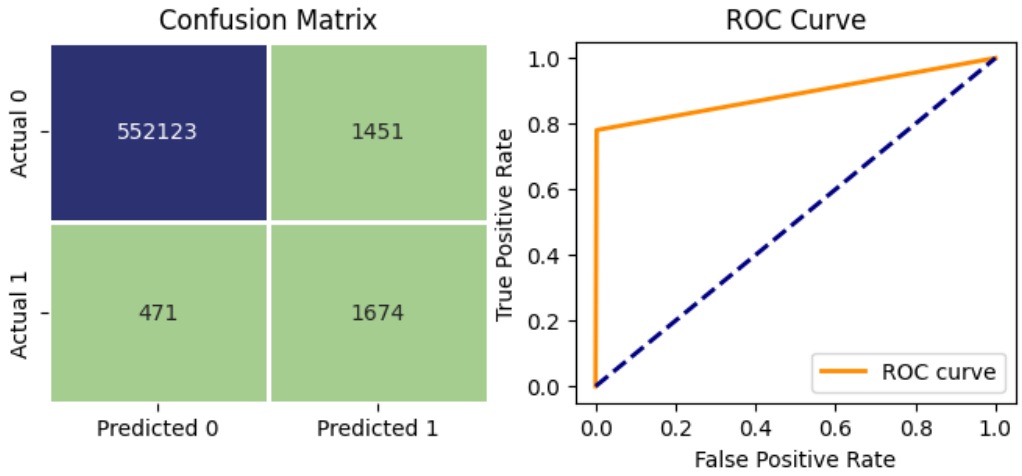
Testing the model on the train data set:

Accuracy: 0.997868
AUC: 0.914559
F1 Score: 0.820279



Testing the model on the test data set:

Accuracy: 0.996541
AUC: 0. 888899
F1 Score: 0.635294

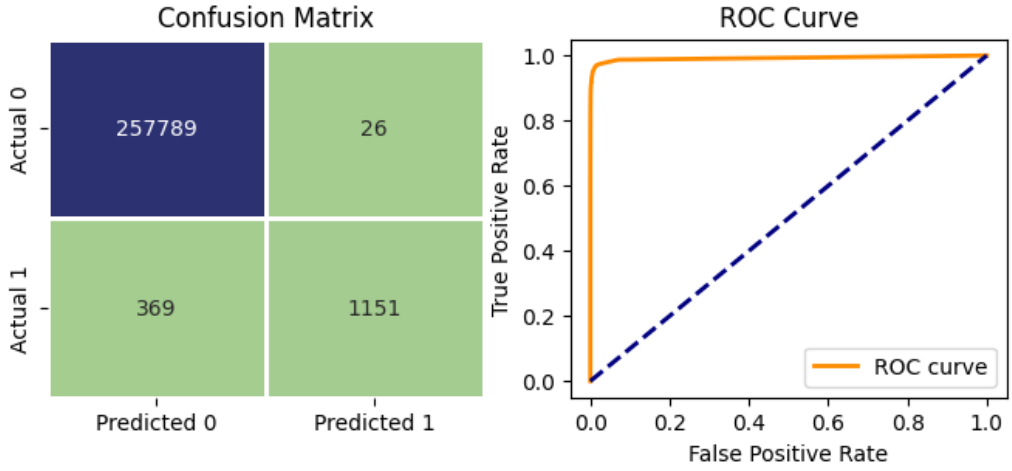


5.5 Random Forest Classifier:

The most accurate models so far, while also being rather lightweight.

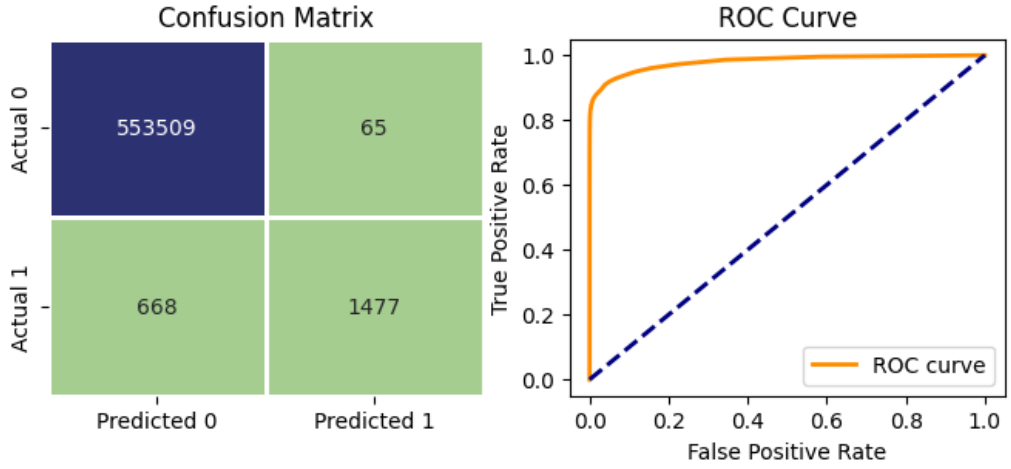
Testing the model on the train data set:

Accuracy: 0.998477
AUC: 0.991849
F1 Score: 0.853541



Testing the model on the test data set:

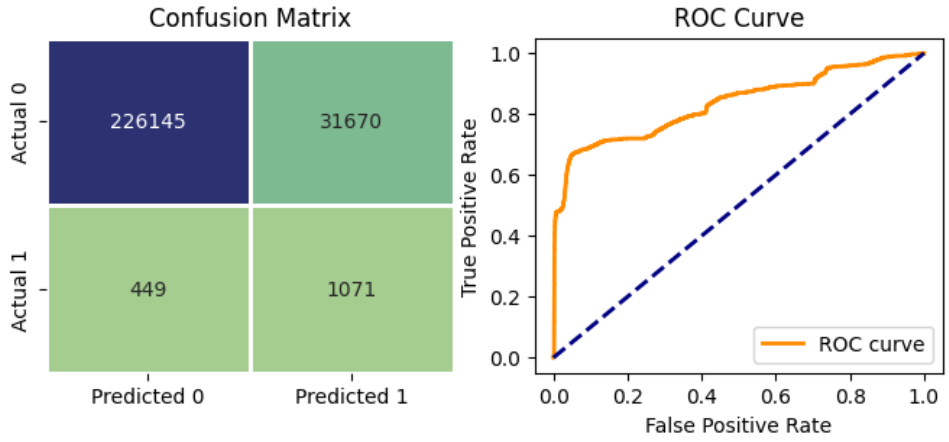
Accuracy: 0.998681
AUC: 0.9801
F1 Score: 0.801193



5.6 Naive Bayes:

Testing the model on the train data set:

Accuracy: 0.876149
AUC: 0.835145
F1 Score: 0.0625201



Testing the model on the test data set:

Accuracy: 0.869274
AUC: 0.811384
F1 Score: 0.0384631