Credit Card Fraud Analysis

Digital payments are evolving, but so are cyber criminals.

According to the Data Breach Index, more than 5 million records are being stolen on a daily basis, a concerning statistic that shows - fraud is still very common both for Card-Present and Card-not Present type of payments.

In today's digital world where trillions of Card transaction happens per day, detection of fraud is challenging.

Following analysis has been done using this dataset: https://www.kaggle.com/datasets/dhanushnarayananr/credit-card-fraud

Problem Statement:

1. Fraud Detection Based on Transaction Location:

Analyze how the distance from home and distance from the last transaction affect the likelihood of fraud. Are transactions that occur farther from home or the last transaction more likely to be fraudulent?

2. Impact of Transaction Value on Fraud:

Investigate the relationship between the ratio of the transaction's purchase price to the median purchase price and the likelihood of fraud. Do unusually high or low purchase values correlate with higher fraud risk?

3. Retailer and Fraud Correlation:

Study the effect of repeat purchases from the same retailer on fraud occurrence. Is fraud more common in repeat transactions from the same retailer or in transactions from new retailers?

4. Effect of Chip and PIN Usage on Fraud:

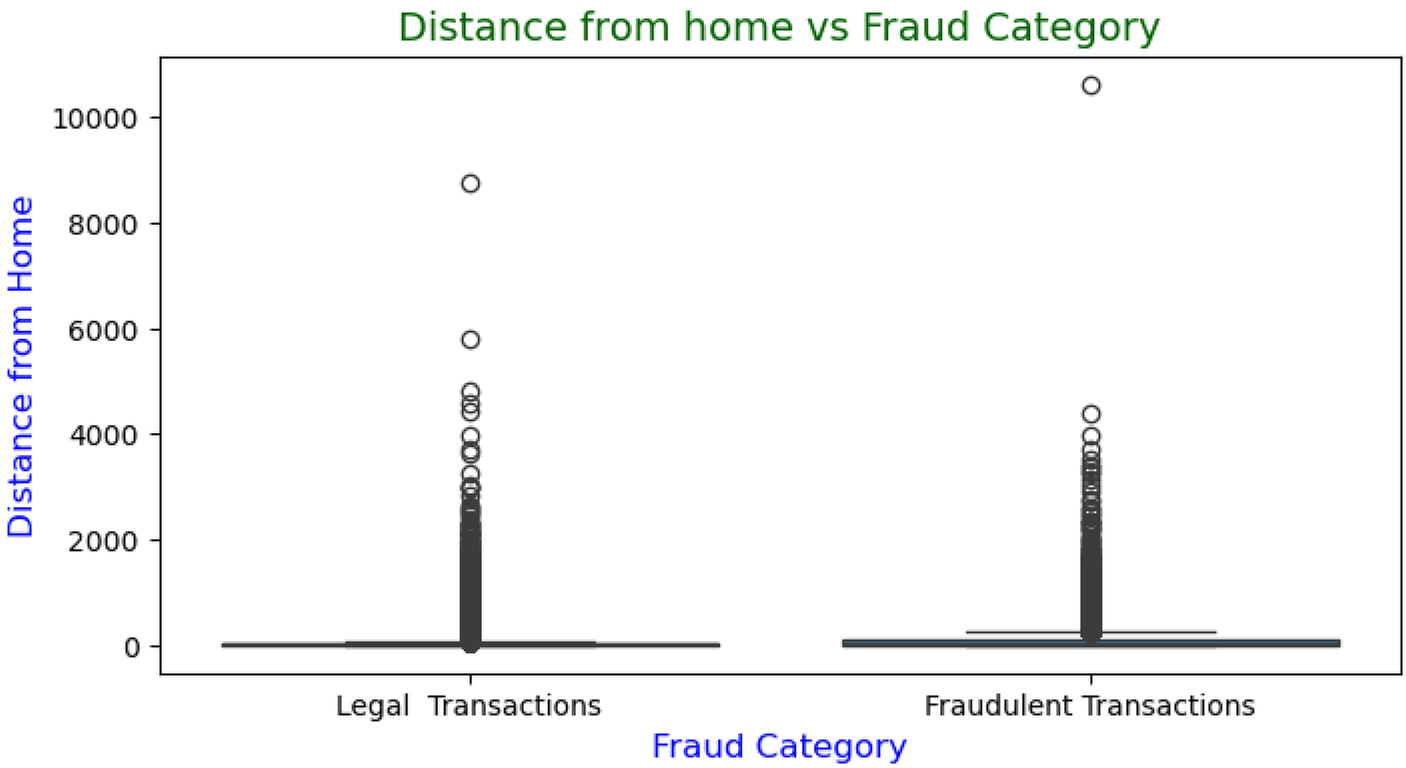
Examine whether transactions using a chip or PIN are less likely to be fraudulent compared to those without Chip or PIN usage. Can Chip and PIN usage be reliable indicators of legitimate transactions? And comapred to PIN, how safe is using Chip (Credit Card)?

5. Online vs. Offline Fraud Patterns:

Analyze whether online orders are more susceptible to fraud than in-store transactions. Are there distinct patterns in fraud rates between online and offline orders?

Exploratory Data Analysis (EDA)

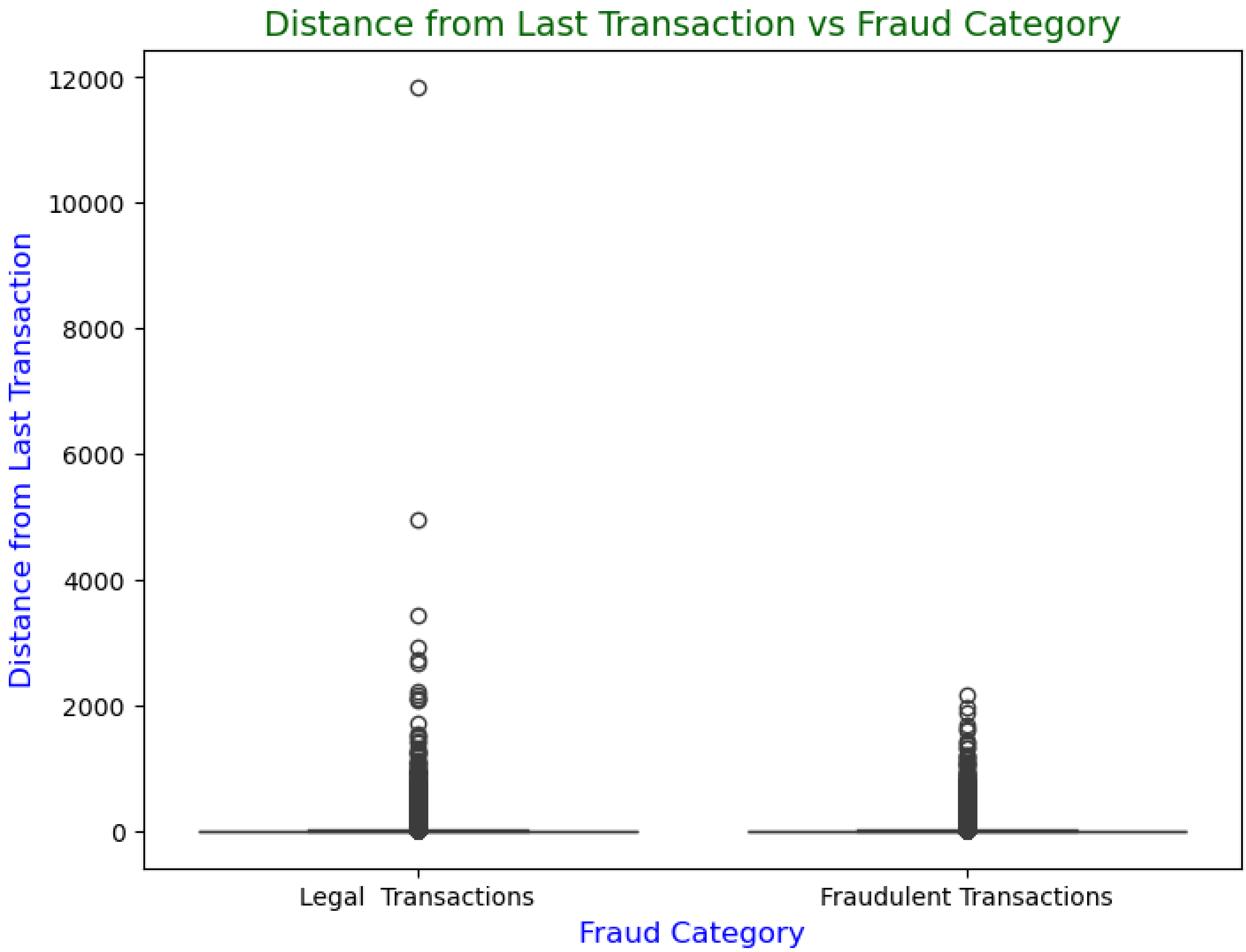
Fraud Detection Based on Transaction Location:



No clear pattern in the values of the column "distance_from_home" to indicate possibilty of fraud.

As we can see fraud can take place at all possible distance values. There is no way we can say that fraud takes place only closer to home or farther away. Thus this cannot be used as a parameter to understand fraudulent transactions.

Fraud Detection Based on Transaction Location:

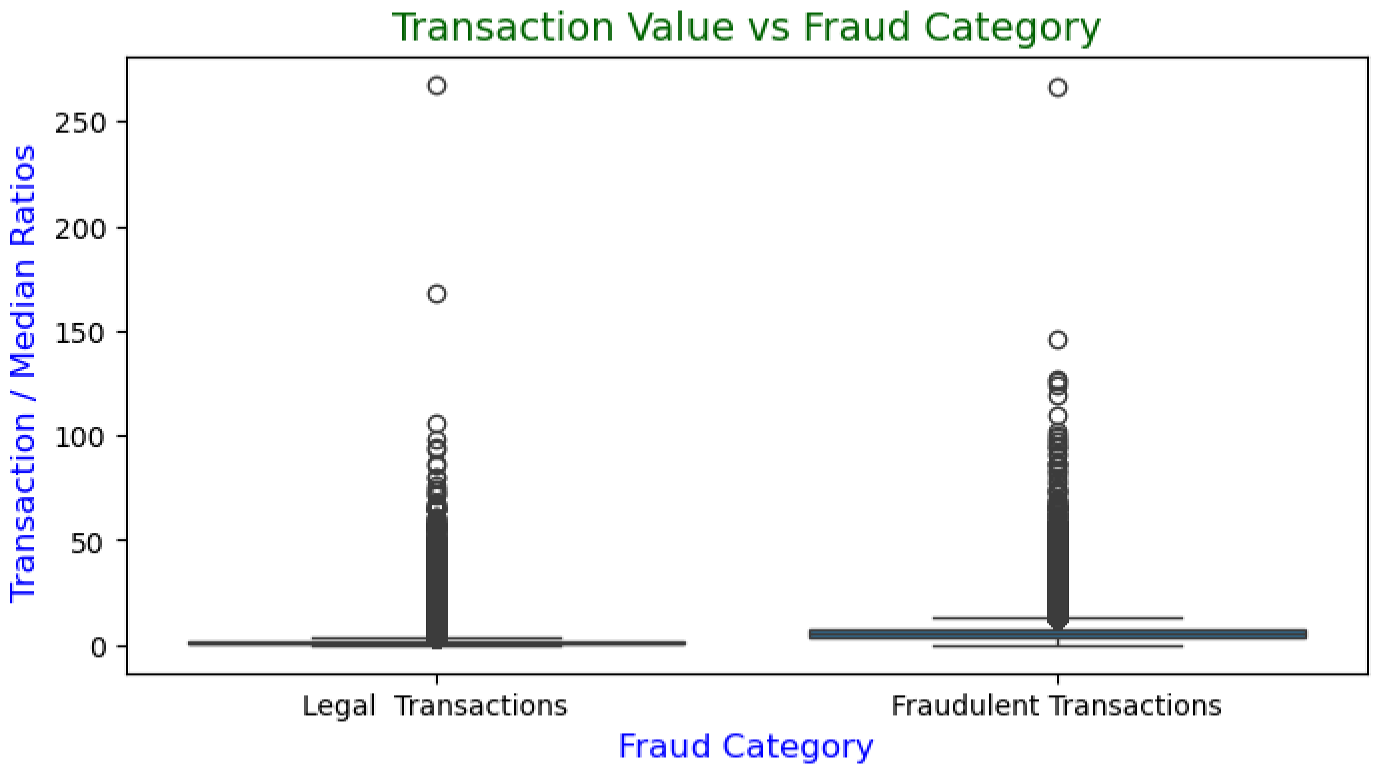


No clear pattern in the values of the column "distance_from_last_transaction" to indicate possibilty of fraud.

Fraudelent transactions are more concentrated towards lower distances as compared to legal transactions. But this is only nominal as there are very few values at larger distances which are legal transactions.

Considering concentration of larger sample size, it is not possible to draw any conclusion about fraudulent transactions from the column values "distance_from_last_transaction". So we cannot use this as a parameter to understand fraudulent transactions.

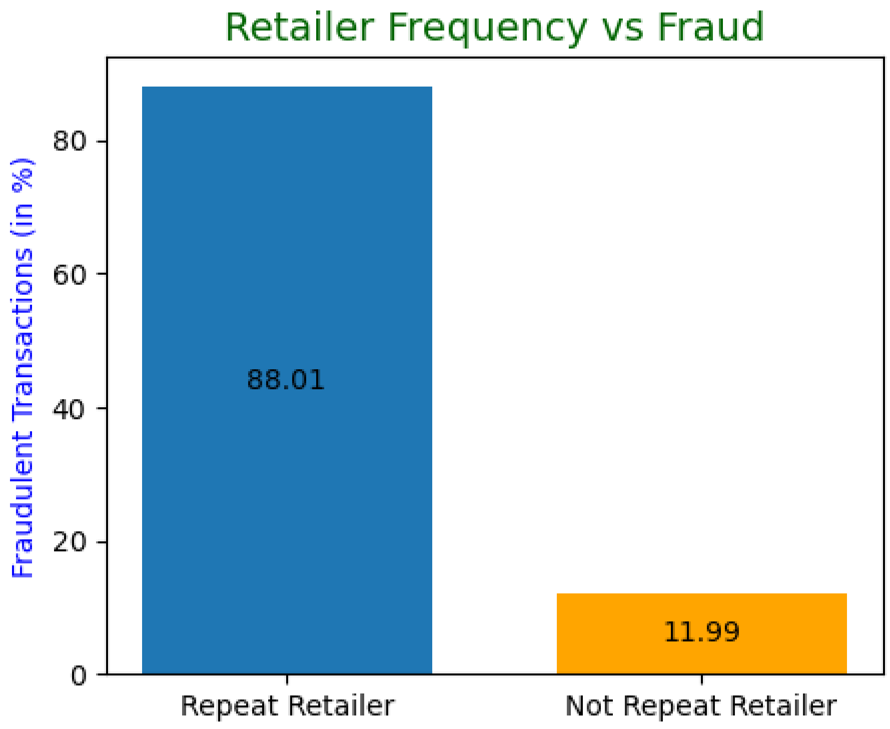
Impact of Transaction Value on Fraud:



Ratio of purchase price to median cannot be used as a paramter to understand fraudulent transactions.

The boxplot does not show any visible pattern to highlight transactions that are fraudulent. All values are highly concentrated in similar groups for both fraudelent and legal transactions.

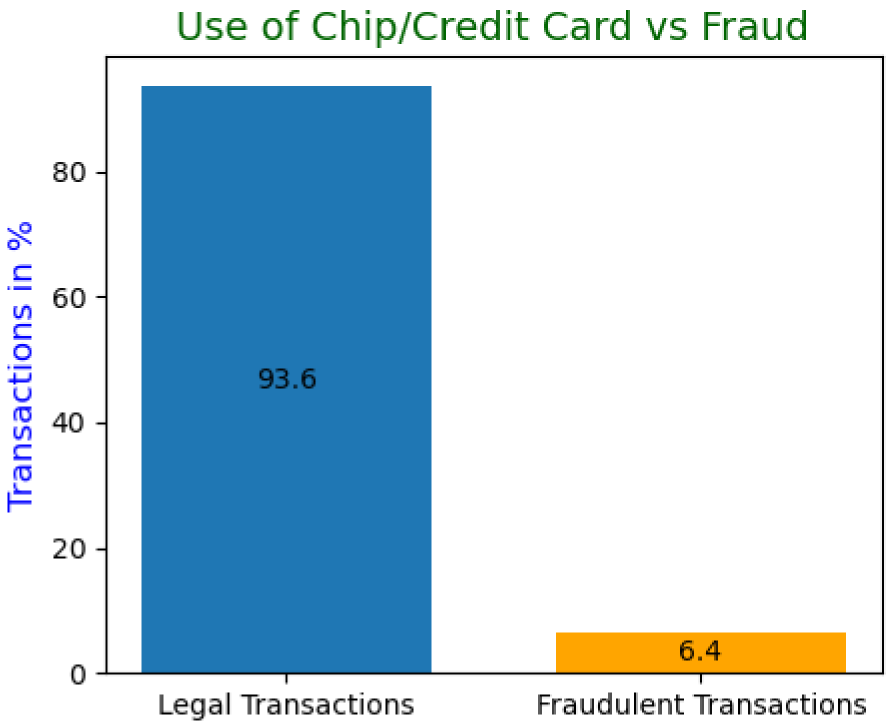
Retailer and Fraud Correlation:



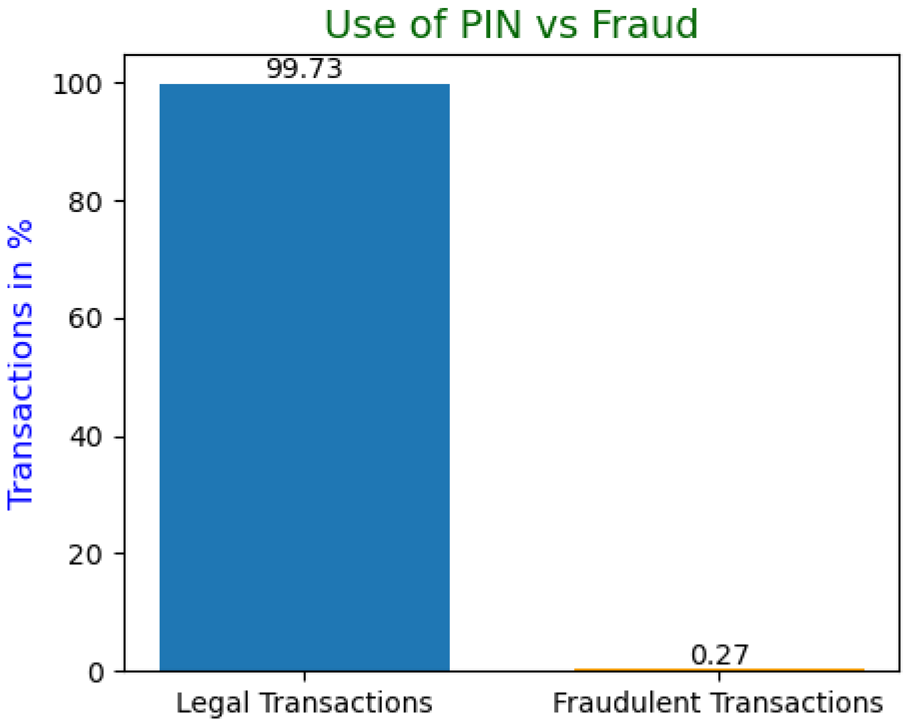
From the above bar plot we can see that about 88% of all fraudulent transactions have taken place when the retailers were repeated that is when the transaction happened from the same retailer. Remaining 12% of fraud happened when transactions were made with new retailers.

So this explains a relationship between retailer frequency and occurence of fraudulent transactions. Chances of transactions becoming fraudulent increases when the transactions take place with same retailer. Repeated Retailers therefore can be used as a possible parameter for identifying chances of fraudulent transactions.

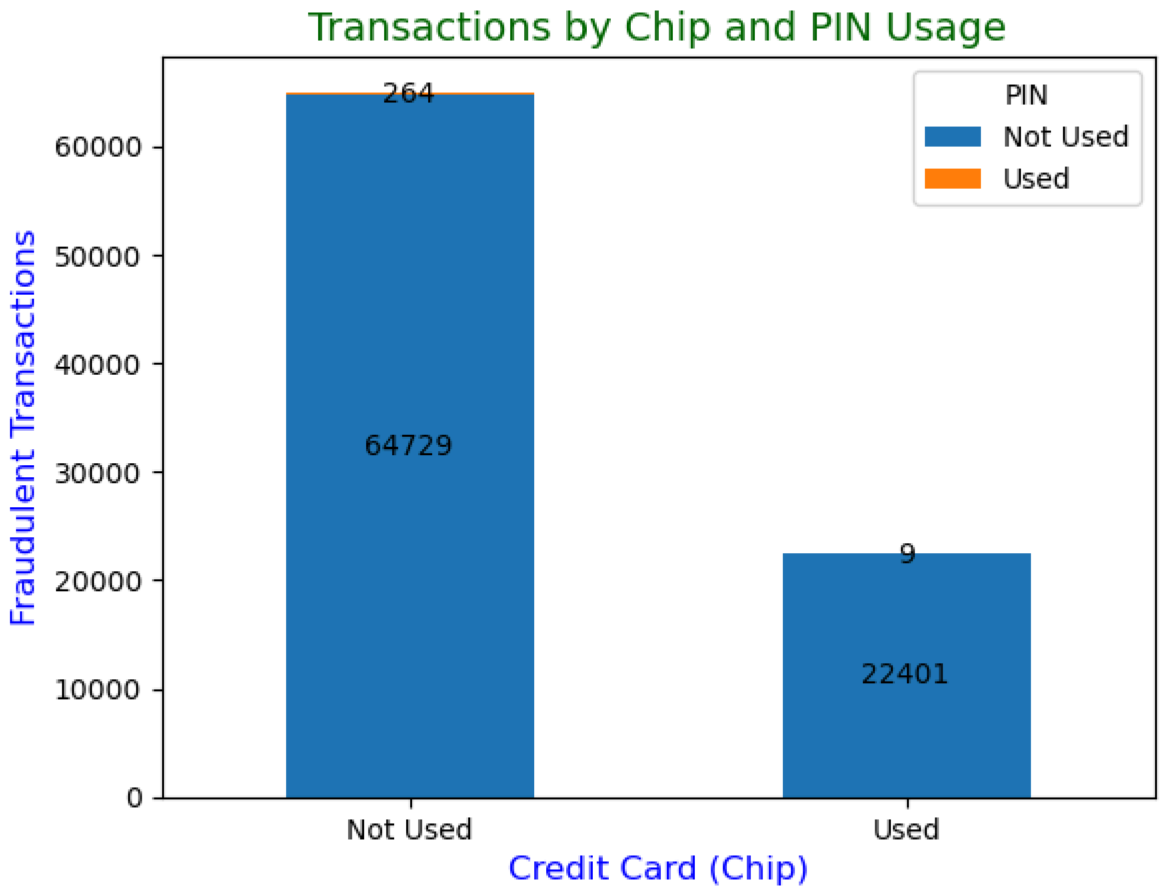
Effect of Chip and PIN Usage on Fraud:



When **Chip/Credit Card is used** the fraudulent transactions are just **6.4%** as opposed to **93.6%** fraudulent transactions when **Chip/Credit Card is not used**.



On use of **PIN**, fraudulent transactions are reduced to a meagre **0.27%**



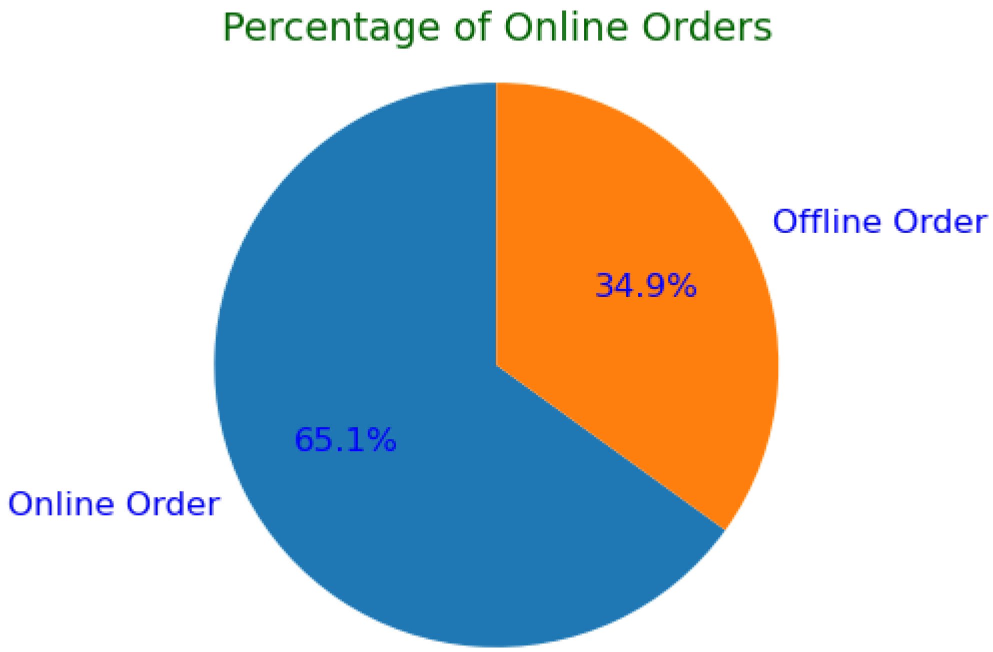
Data obtained from Graph:

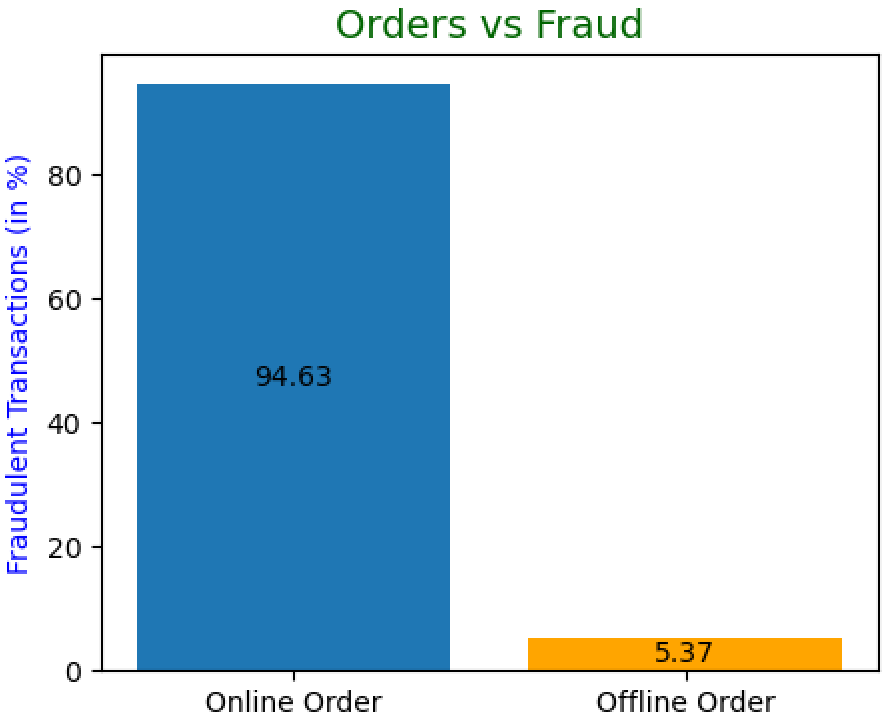
Around 74% of fraudulent transactions involves payments not using chip/credit cards or PIN.
Approximately 25.6% of fraudulent transactions use credit card payment without using PIN as additional authentication method. Among remaning transactions that are fraudulent use of credit card with PIN amounts to only 0.009%.

The above figures highlight that use of chip with PIN is the most secure way to reduce fraudulent transactions. And non-chip transactions that does not involve PIN authentication accounts for about 3/4th of total fraudulent cases in this dataset.

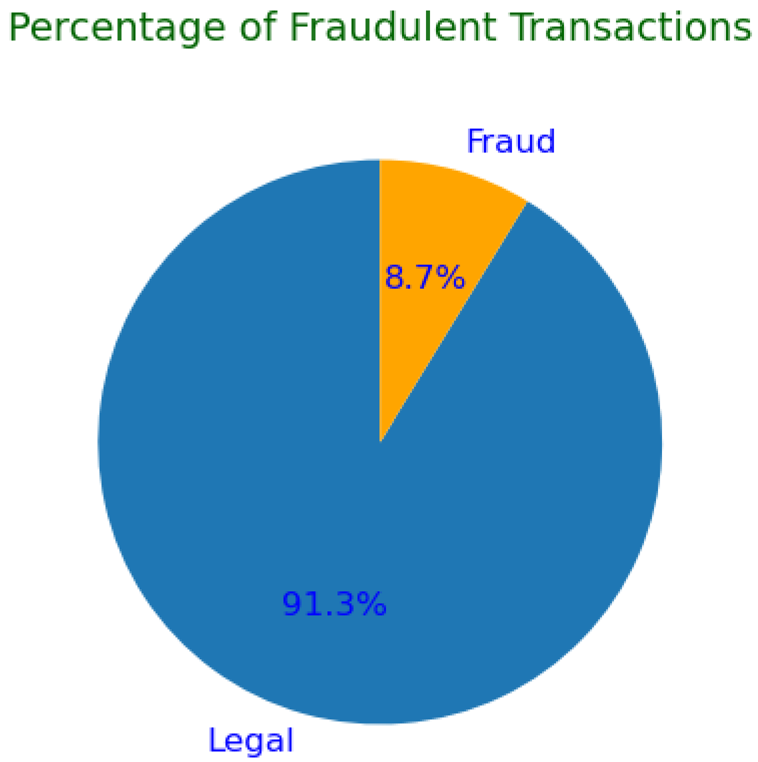
Another highlight is use of PIN with chip payments reduces fraudulent transactions from 1/4th to meagre 0.009% making it least number of cases among fraudulent transactions.

Online vs. Offline Fraud Patterns:

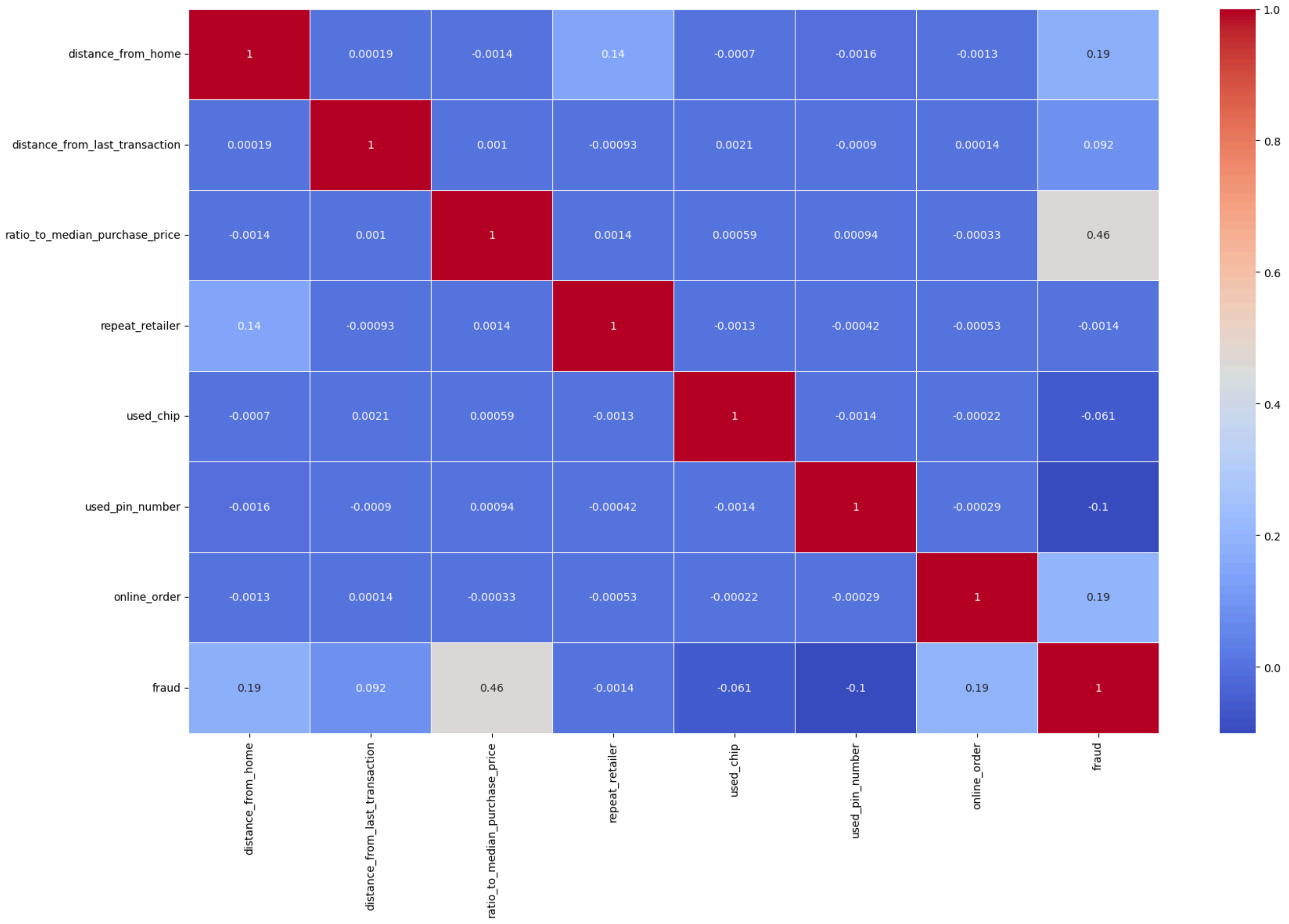




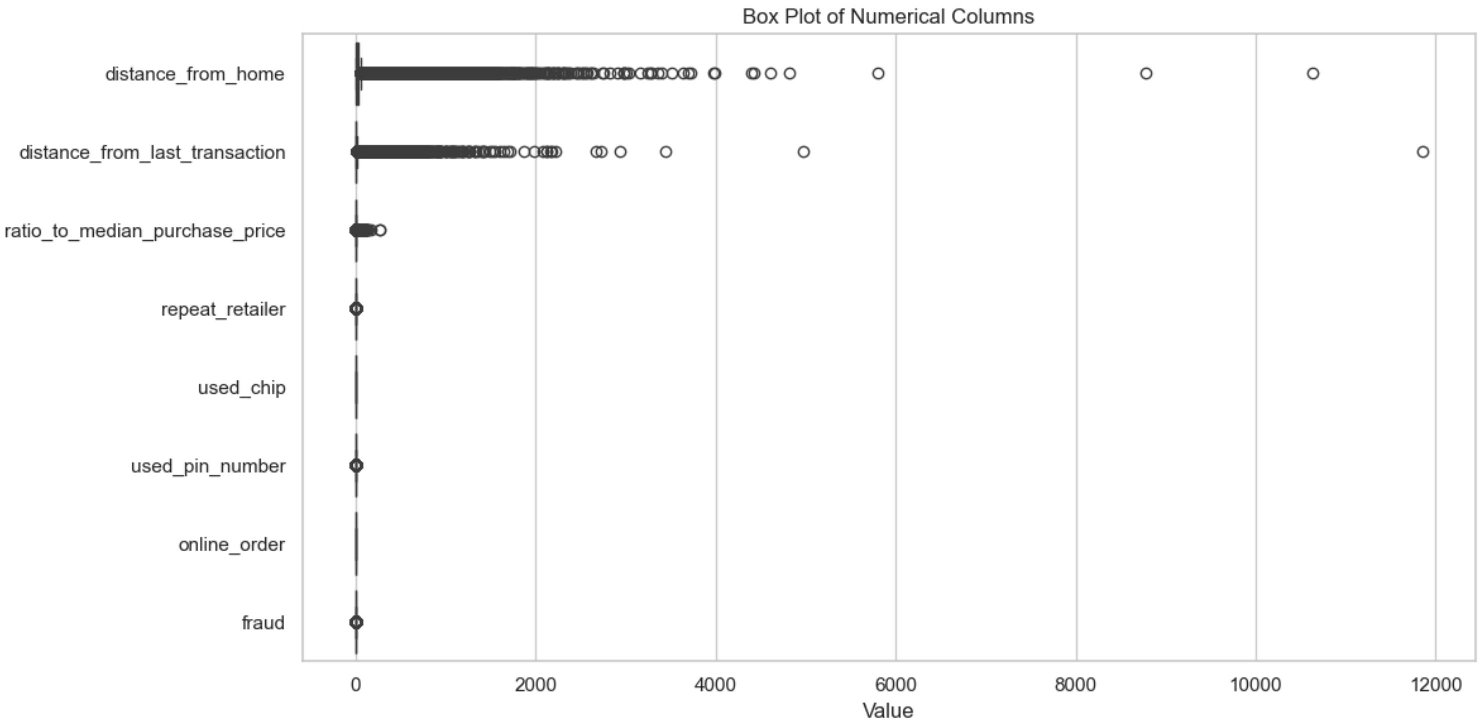
About 95% of fraudulent transactions have taken place through online orders which clearly shows that online orders do have a part to play in increasing chances of a transaction becoming fraudulent transaction.



Correlation HeatMap



Preprocessing (Checking and removing Outliers)

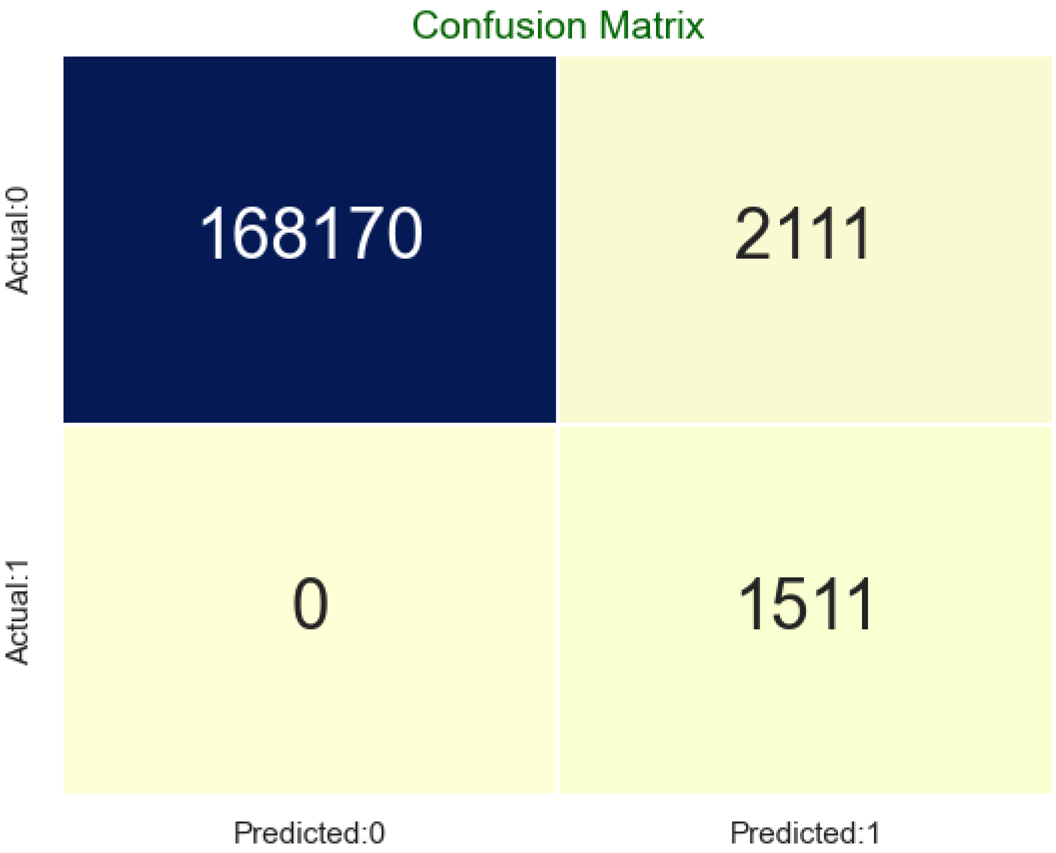


```
Total number of deleted outliers: 141044

Total number after deleted outliers: 858956
```

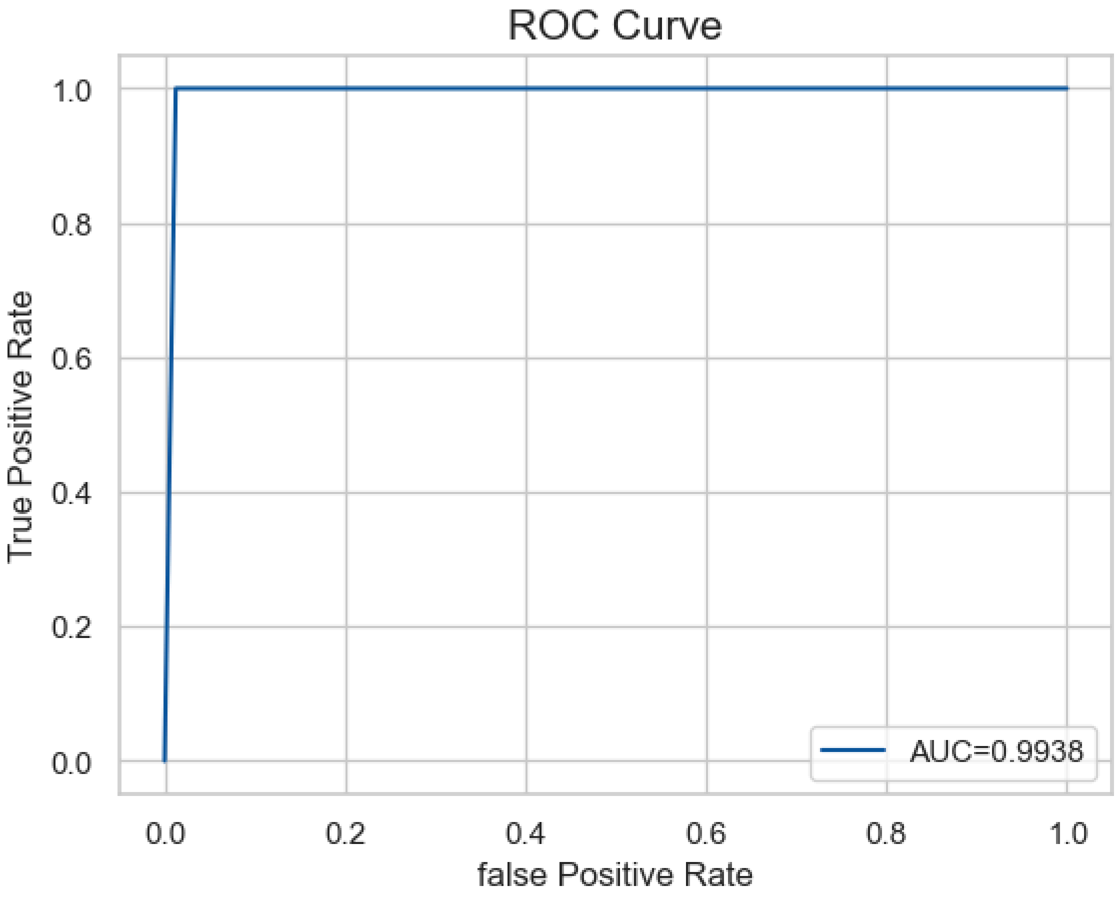
ML Algorithms

Logistic Regression

```
Accuracy: 0.9877118841389587
              precision    recall  f1-score   support

         0.0     1.0000    0.9876    0.9938    170281
         1.0     0.4172    1.0000    0.5887      1511

    accuracy                         0.9877    171792
   macro avg     0.7086    0.9938    0.7913    171792
weighted avg     0.9949    0.9877    0.9902    171792
```

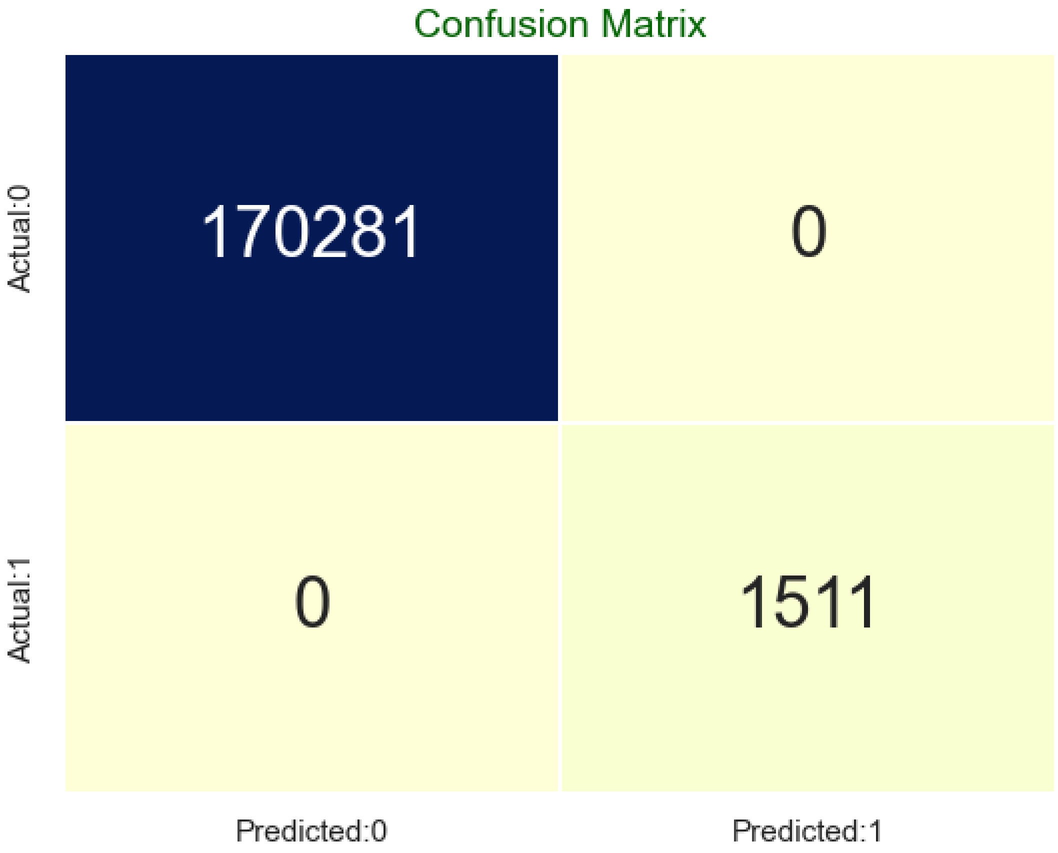


```
<matplotlib.legend.Legend at 0x1f320003aa0>
```

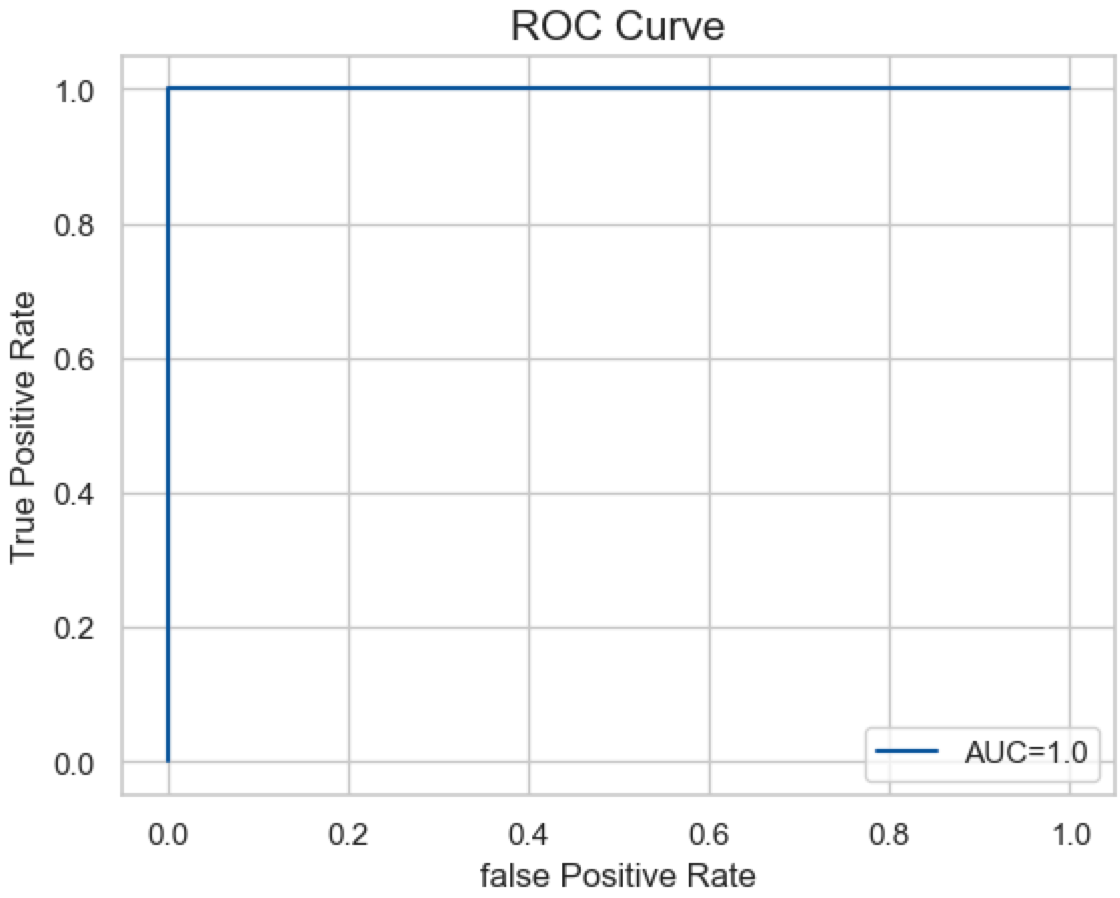


Decision Tree

```
Accuracy: 1.0
              precision    recall  f1-score   support

         0.0     1.0000    1.0000    1.0000    170281
         1.0     1.0000    1.0000    1.0000      1511

    accuracy                         1.0000    171792
   macro avg     1.0000    1.0000    1.0000    171792
weighted avg     1.0000    1.0000    1.0000    171792
```

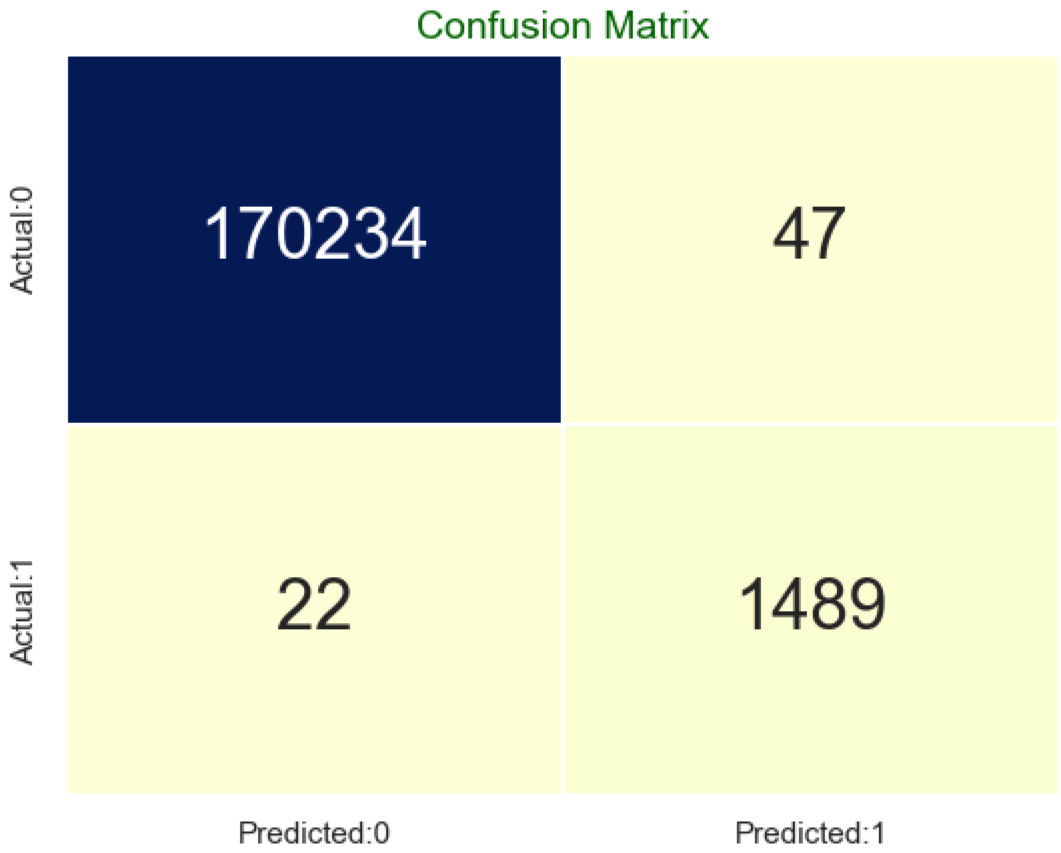


```
<matplotlib.legend.Legend at 0x1f302edcb60>
```

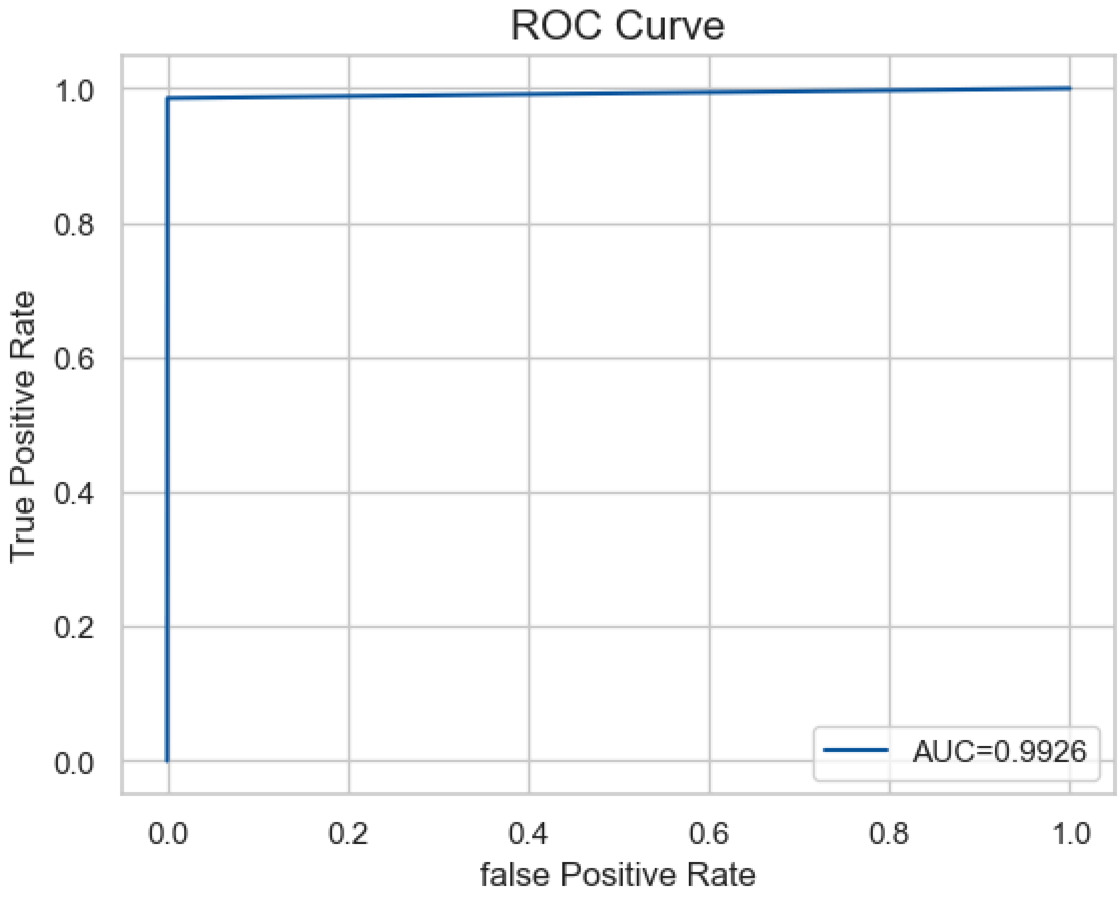


XGBoost

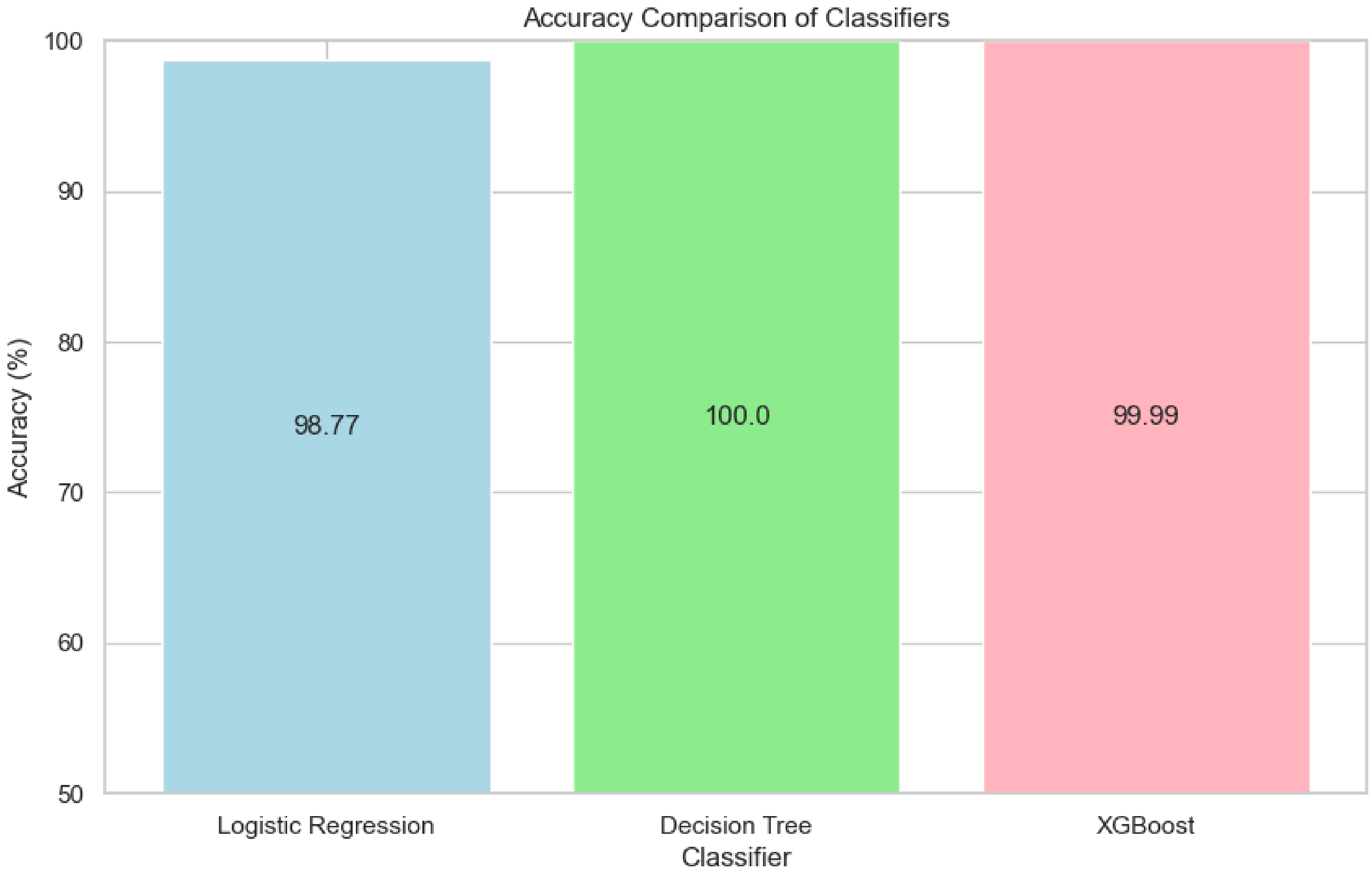
```
Accuracy: 0.9995983514948 31
              precision    recall  f1-score   support

         0.0     0.9999    0.9997    0.9998    170281
         1.0     0.9694    0.9854    0.9774      1511

    accuracy                         0.9996    171792
   macro avg     0.9846    0.9926    0.9886    171792
weighted avg     0.9996    0.9996    0.9996    171792
```


Confusion Matrix
Actual:0
Actual:1
170234
47
22
1489
Predicted:0
Predicted:1

```
<matplotlib.legend.Legend at 0x1f302efd130>
```


ROC Curve
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
false Positive Rate
AUC=0.9926

Conclusion:



In this fraud analysis project, we implemented three machine learning classifiers—**Logistic Regression, Decision Tree, and XGBoost**—to detect fraudulent activities in a dataset with a significant class imbalance (91.3% non-fraudulent/legal and 8.7% fraudulent data). The classifiers achieved impressively **high accuracy scores of 98.77%, 100.00%, and 99.99%, respectively**.

Given the high imbalance, scaling and SMOTE have been used to improve the detection of non-fraudulent transactions. Along with it additional metrics like Precision, Recall, F1 score and ROC-AUC have been assessed that are more sensitive to the minority class (fraudulent transactions) in this imbalanced dataset.