CREDIT CARD FRAUD

DETECTING CREDIT CARD FRAUD USING DATA HELD BY FINANCIAL INSTITUTIONS

PRESENTED BY CAROLINE MUGAMBI

INTRODUCTION

- According to globaleconomy.com 6.35% of Kenyans use credit cards. This translates to about 3 million Kenyans .
- As electronic commerce gains rapid growth and significant impact across the country, the Credit Card has become a defacto standard for payment of goods and services.
- Unfortunately, this has led to rapid growth in credit card fraud making it a big problem for consumers, financial institutions and law enforcement agencies.

PROBLEM STATEMENT

- Mastercard reported that the percentage of fraud in all Kenyan commercial banks was approximately 17% credit card holder expenditure (2019).
- Deloitte reported that Banks in East Africa lost Kes 4.05B to fraud in the 18 months ended June 2019.
- Statistics from Central Bank of Kenya's Bank Supervision department show that commercial banks are losing an average of Kes 100m to fraudsters every month with those with the highest number of branches and the most tech-savvy being the worst hit.
- **Objective of this project is to use data that commercial banks already have in their possession, coupled with technology to detect/predict and subsequently prevent credit card fraud.**

DATA SET

- Due to the sensitivity and restrictions on access to data held by commercial banks, I obtained a credit card fraud dataset which closely represented the transactions in most Kenyan banks, from www.Kaggle.com.
- The data contained 101,613 transactions and 11 feature columns; 3 of them being categorical.

Feature Explanation:

Step – unit of time taken in hours

Type – transaction type

nameOrig – transaction originator

Oldbalance – initial balance before transaction

Feature explanation continued….

newbalance – new balance after transaction

nameDest – transaction recipient

Amount – amount of transaction

oldbalanceDest – initial balance of recipient before transaction

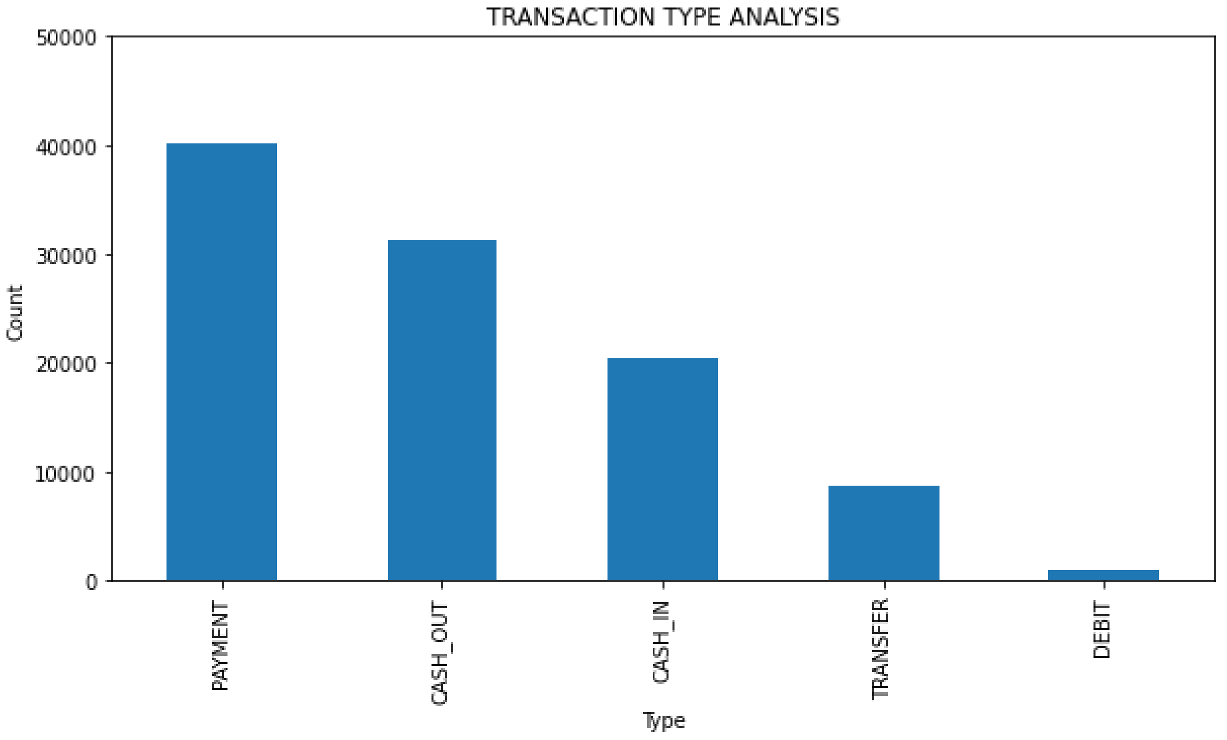
newbalanceDest – new balance of recipient after transaction

isFraud – class 0 transaction is not fraud; class 1 transaction is fraud.

isflaggedFraud - class 0 transaction is suspected not fraud, 1 transaction is suspected fraud.

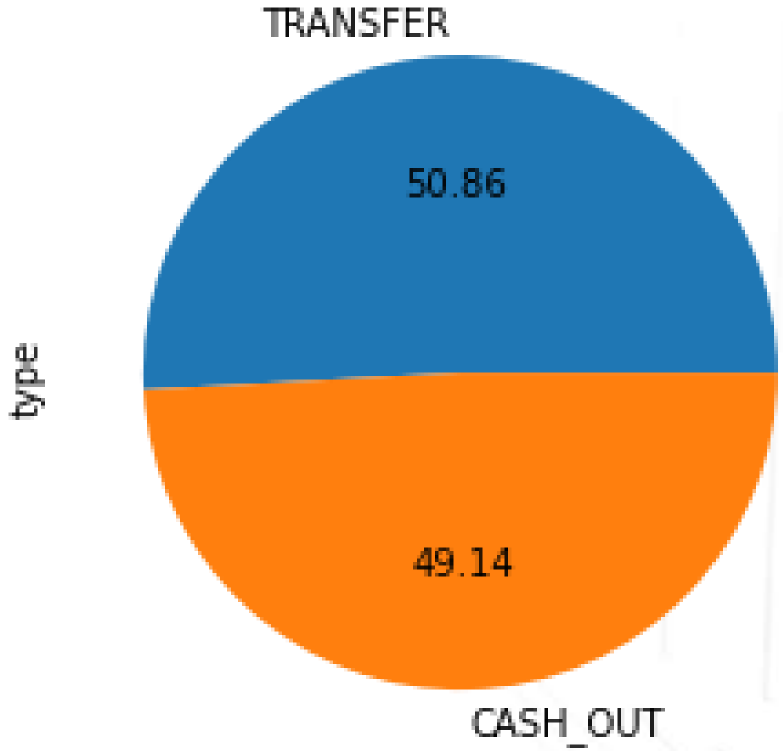
EXPLORATORY DATA ANALYSIS

- The data had no missing values .
- Most credit card transactions related to payment type of transaction as compared to cash, transfer, or auto debit.



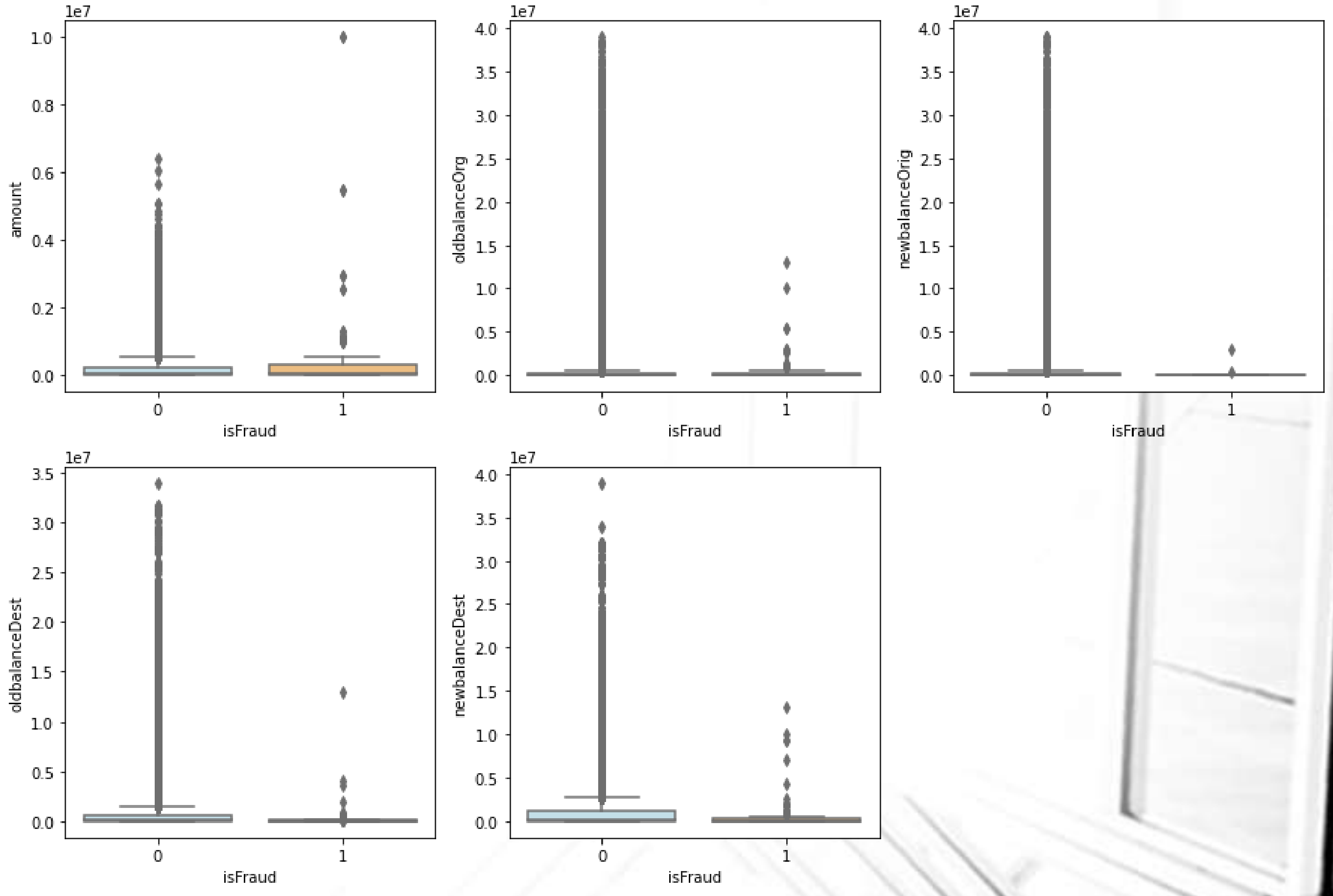
EDA continued.....

- Transfer and Cash Out type of transactions are the most used in fraud cases in this data, with Transfer accounting for 50.86% and Cash Out 49.14% of all the fraud cases.



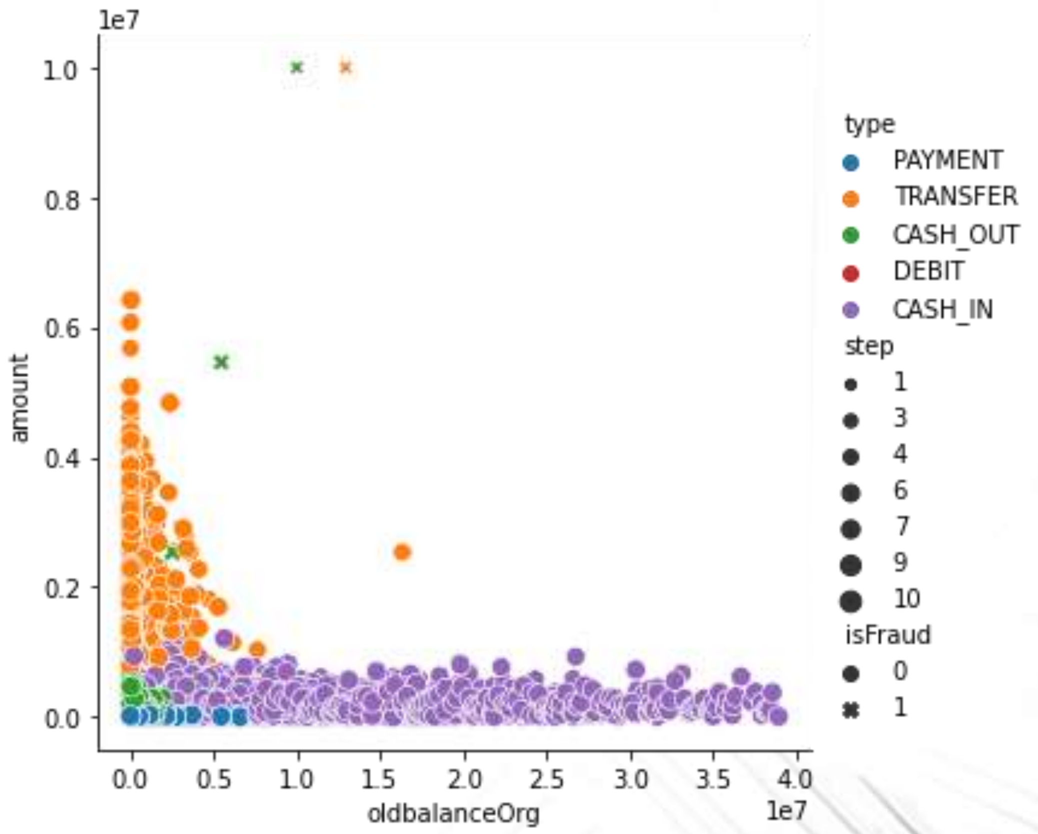
EDA continued…

- The boxplots show how the value of numerical features vary across the target group.
- For example, the balances both old and new and on the originator and destination have distinct difference when target is 0 and when target is 1 suggesting that they are important predictors.
- Amount appears to be less outstanding as the boxplot distribution is similar between target groups.



EDA continued...

- The plot shows that we have very few outliers in our dataset. This shows that the outliers represent natural variation in our population. As such, we leave them as is, and do not expect that they will negatively impact on the accuracy of our conclusions.

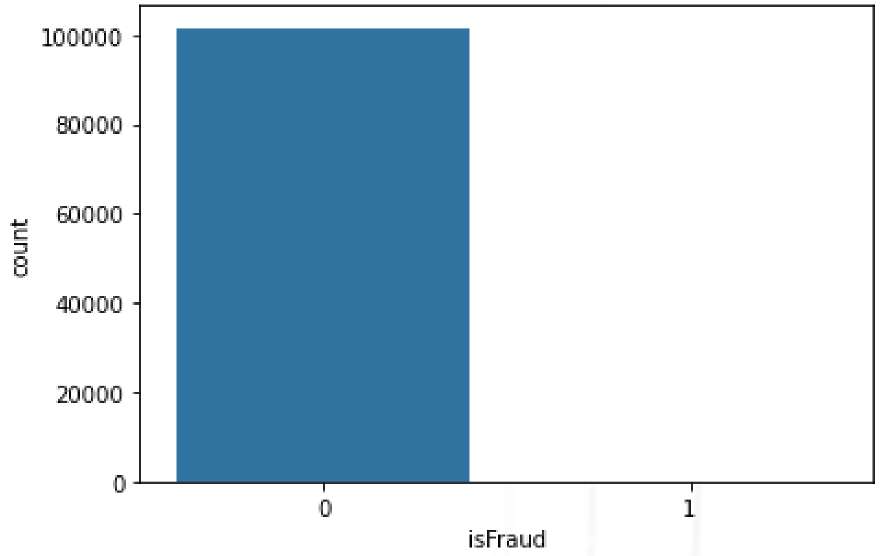


EDA continued…

- Most of the features in our data set have a low correlation with the exception of 2; oldbalanceOrig and newbalanceOrig. This shows that most of our features are independent.

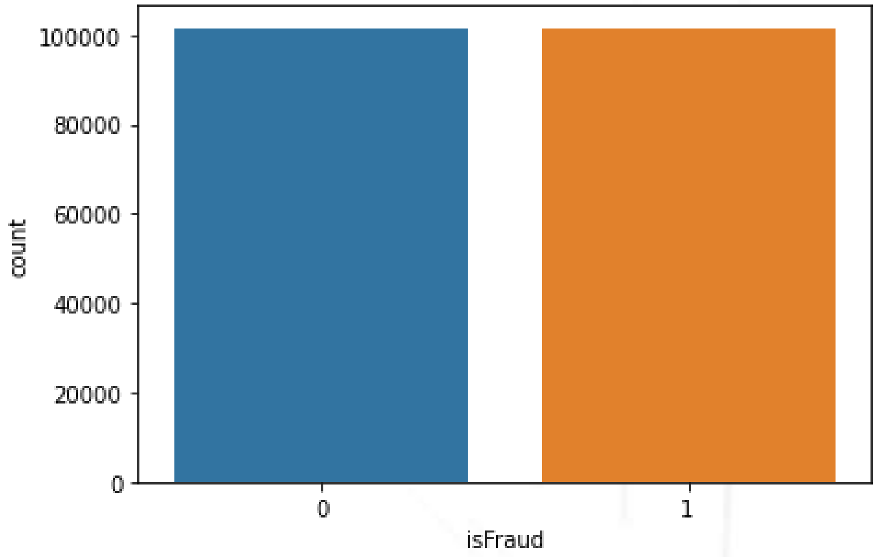
METHODOLGY

- Using label encoder, I assigned a unique integer to our non numerical observations to enable us fit them into the models.
- I then checked if the label for my target variable ('isFraud') is balanced or not.



METHODOLGY continued…

- I then oversampled the minority class to balance the label



METHODOLGY continued…

- I then checked the correlation of the up sampled Data Frame and dropped the ‘isflaggedFraud’ column.
- We train and test split the data
- For each algorithm (Decision Tree, Random Forest and Support Vector Machine) I performed a param-grid to determine the best hyper parameters and then used them to fit the model.

FINDINGS/RESULTS

DECISION TREE

Accuracy Score – 98.27%

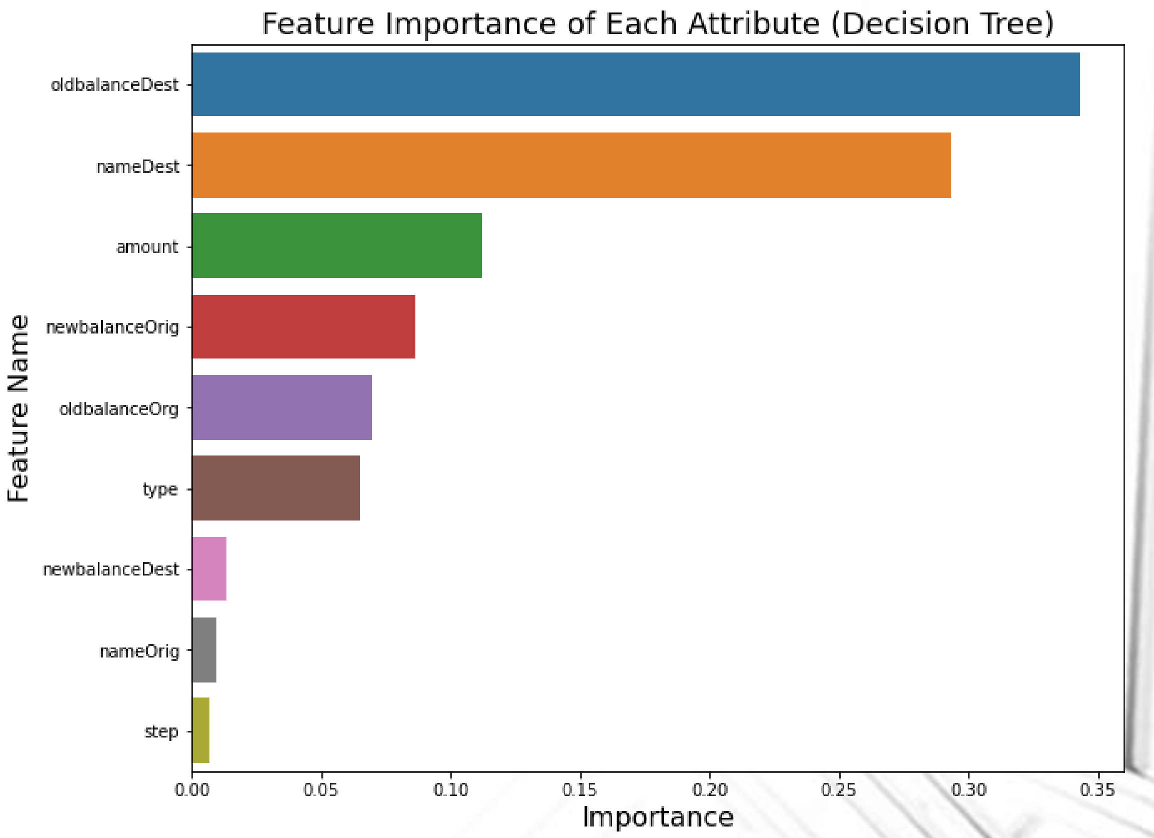
F-1 Score – 0.9827089337

Precision Score – 0.9827089337

Recall Score – 0.9827089337

Jaccard Score – 0.966055665

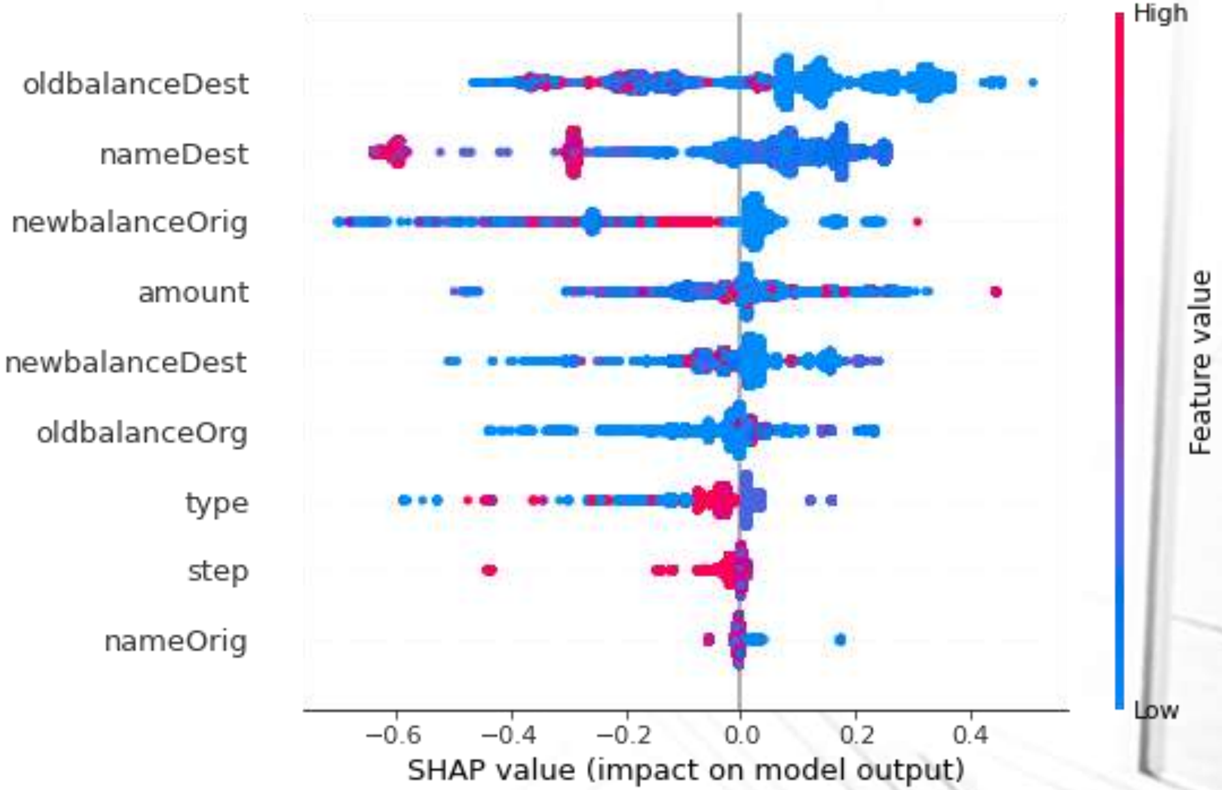
Log Loss – 0.59722



FINDINGS/RESULTS

DECISION TREE

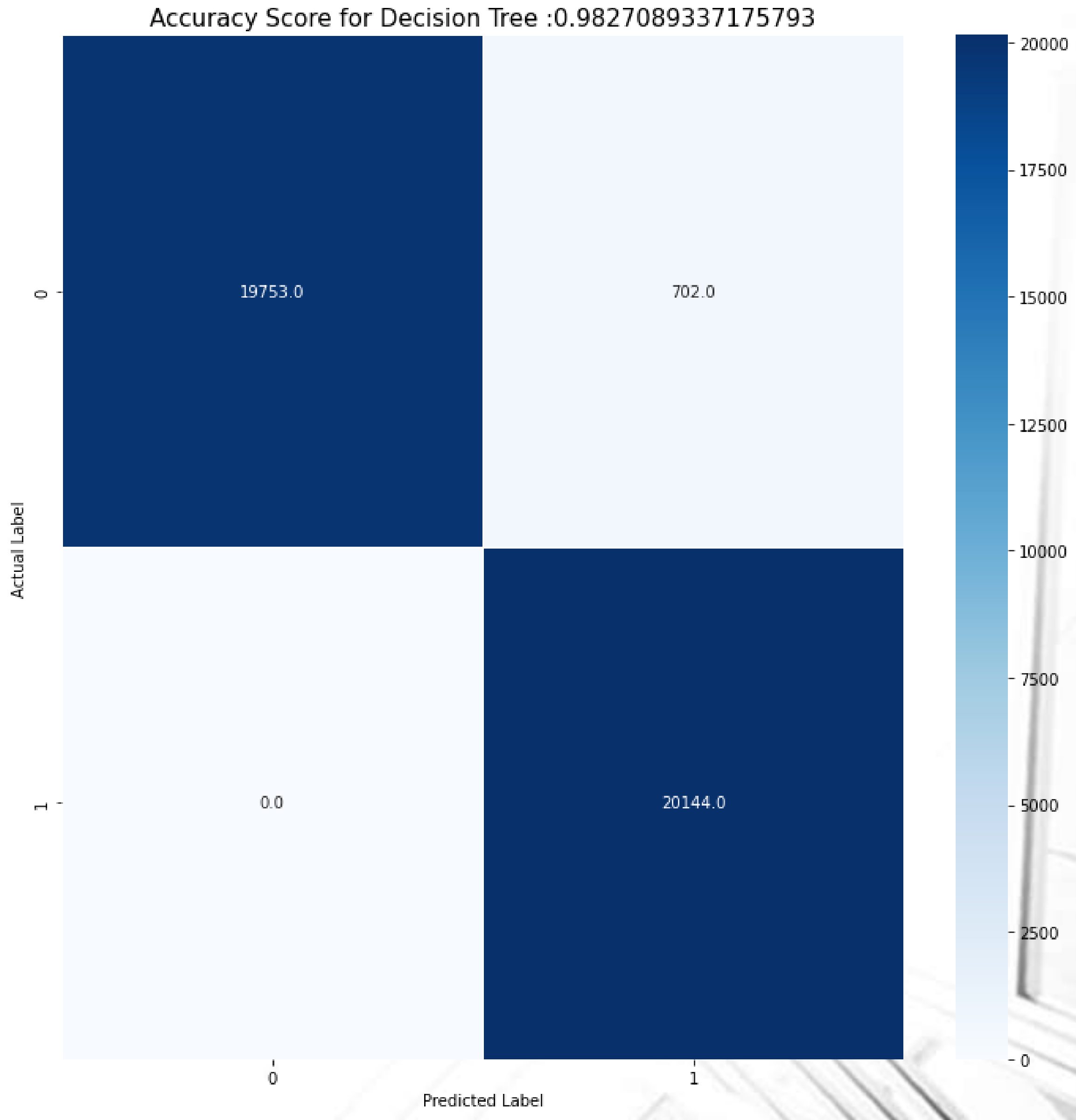
Computed SHAP values (Shapely Additive Explanations) shows the contribution or the importance of each feature on the prediction of the model.



FINDINGS/RESULTS

DECISION TREE

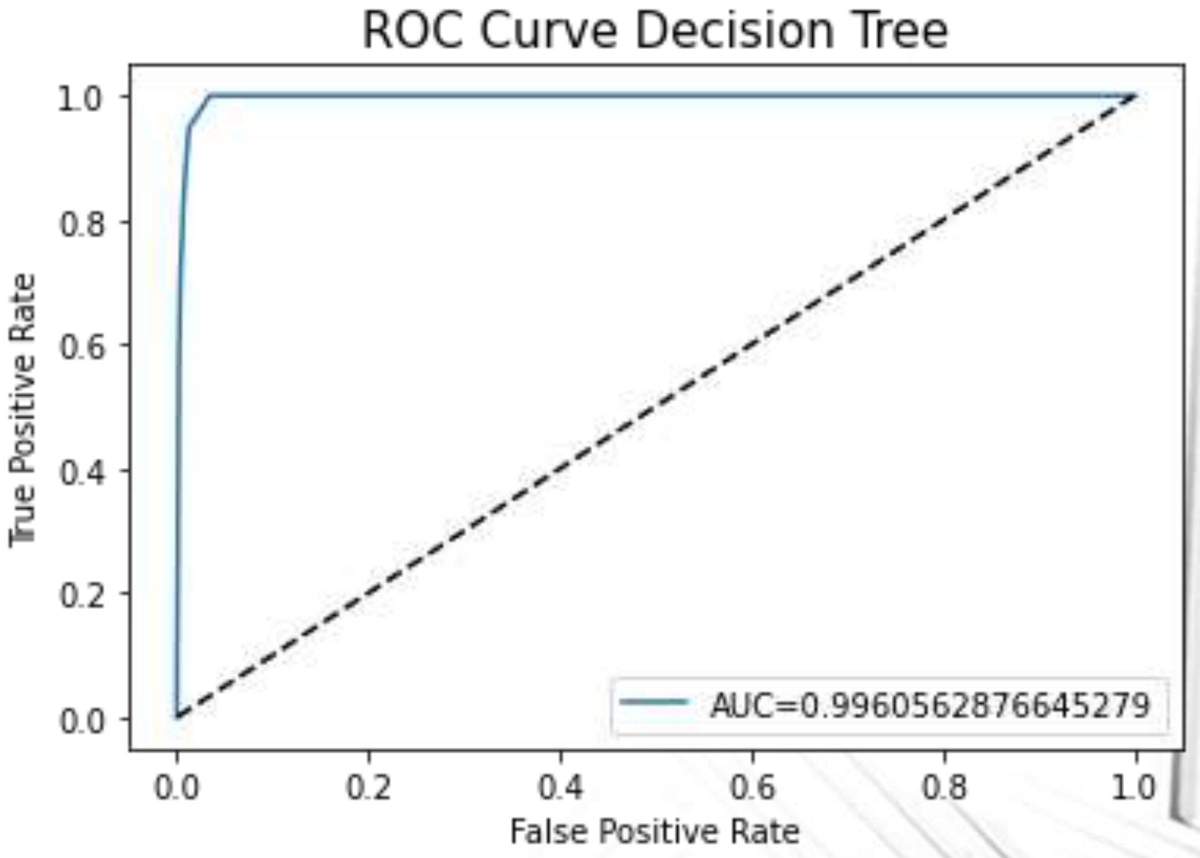
Finally, I generated the Confusion Matrix to measure recall, precision, accuracy and AUC-ROC, and plotted the ROC curve.



FINDINGS/RESULTS

DECISION TREE

Finally, I generated the Confusion Matrix to measure recall, precision, accuracy and AUC-ROC, and plotted the ROC curve.



FINDINGS/RESULTS

RANDOM FOREST

Accuracy Score – 99.81%

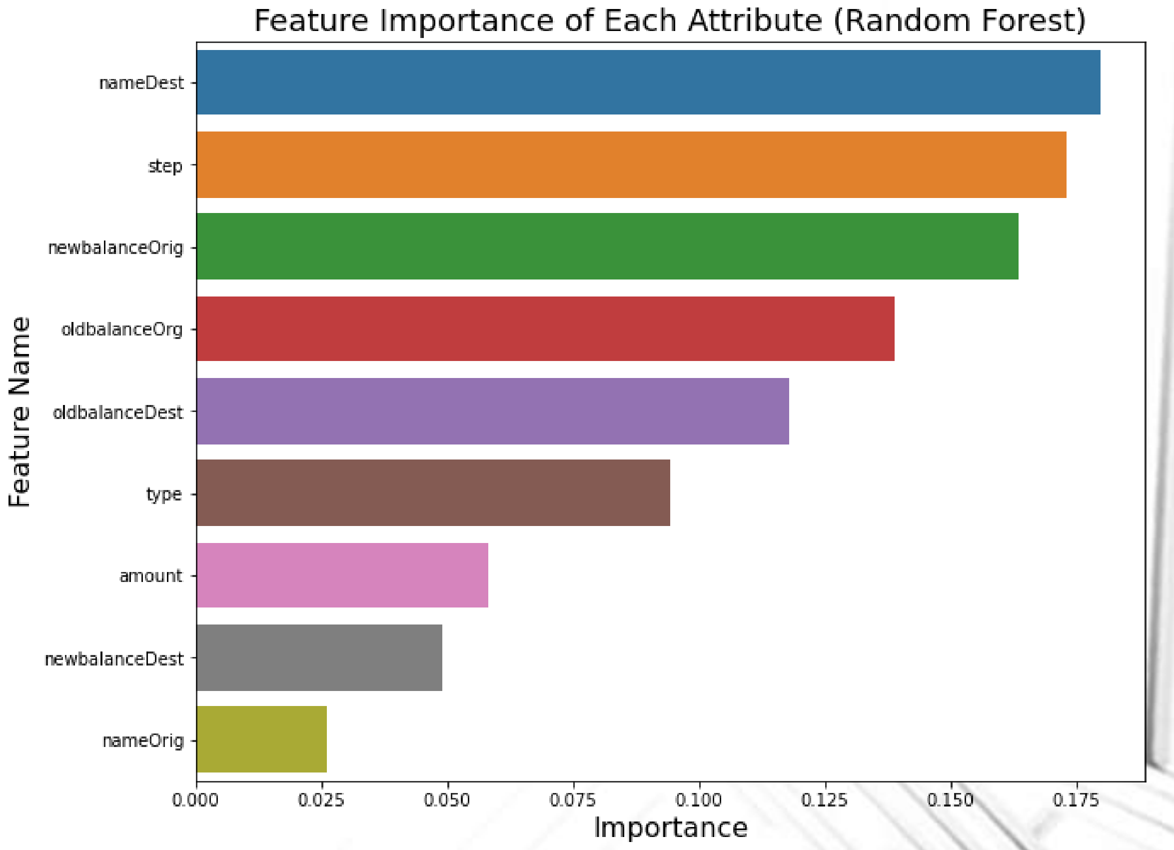
F-1 Score – 0.998054139

Precision Score – 0.998054139

Recall Score – 0.998054139

Jaccard Score – 0.9961158365

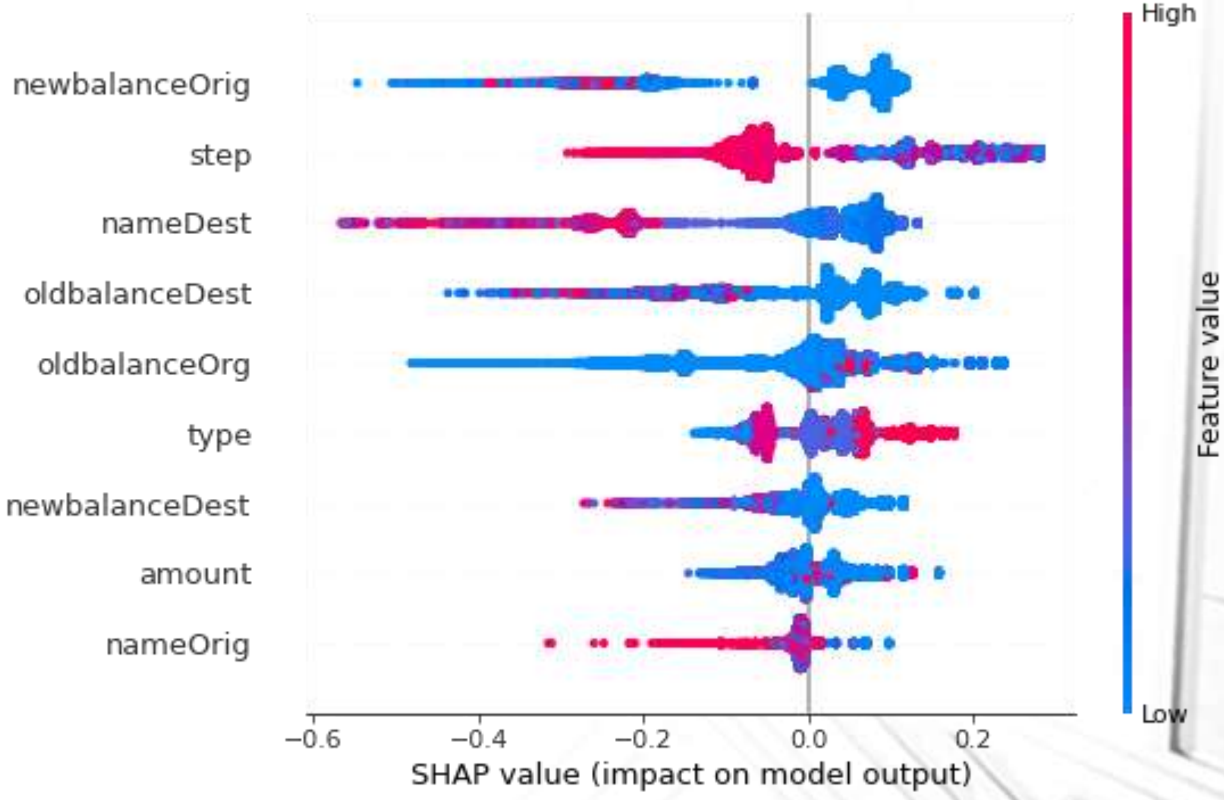
Log Loss – 0.06702920



FINDINGS/RESULTS

RANDOM FOREST

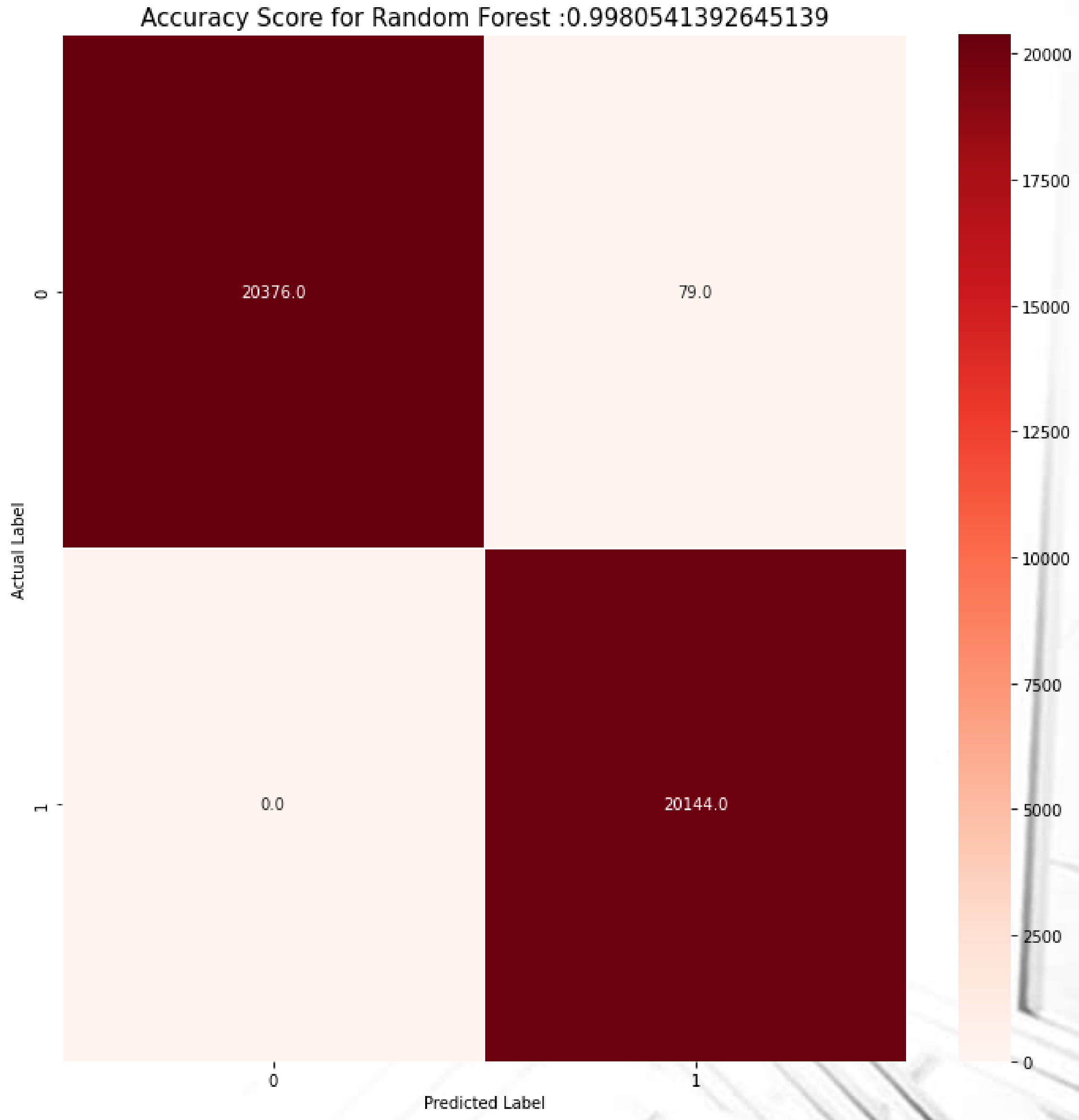
Computed SHAP values (Shapely Additive Explanations) shows the contribution or the importance of each feature on the prediction of the model.



FINDINGS/RESULTS

RANDOM FOREST

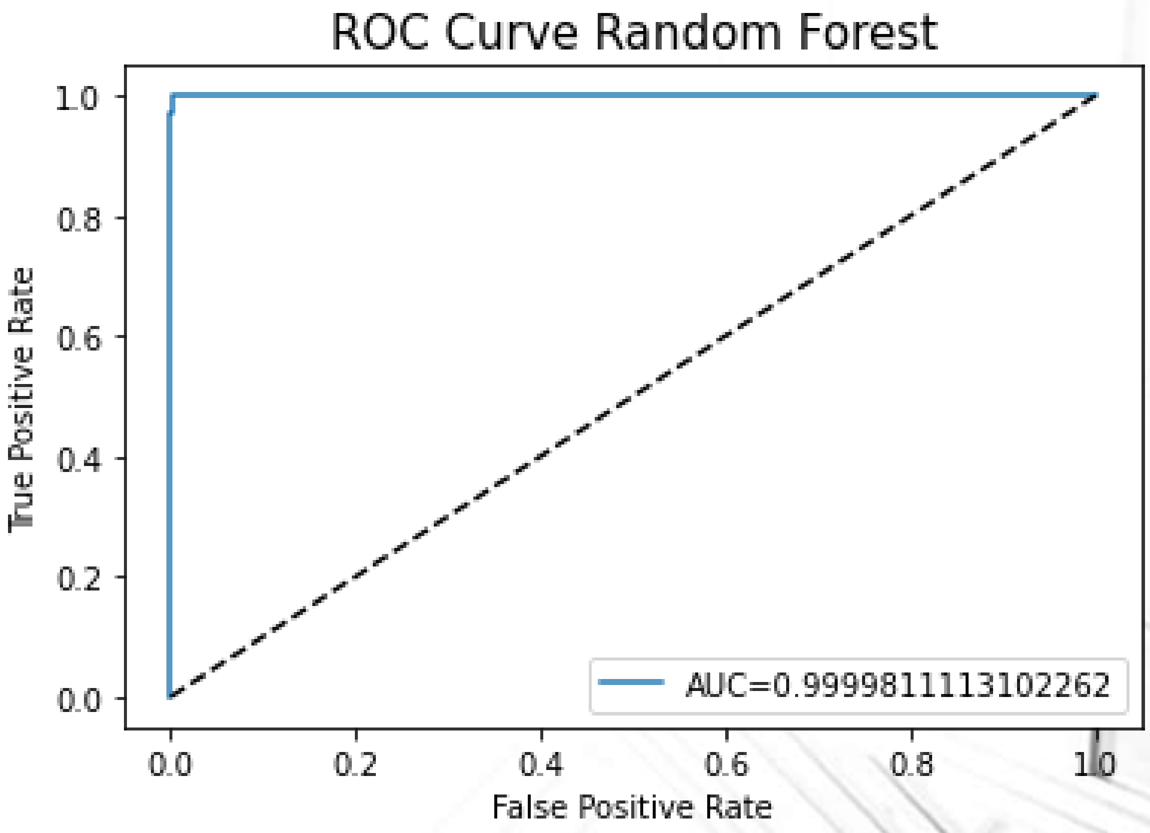
Finally, I generated the Confusion Matrix to measure recall, precision, accuracy and AUC-ROC, and plotted the ROC curve.



FINDINGS/RESULTS

RANDOM FOREST

Finally, I generated the Confusion Matrix to measure recall, precision, accuracy and AUC-ROC, and plotted the ROC curve.



FINDINGS/RESULTS

SUPPORT VECTOR MACHINE

Accuracy Score – 89.39%

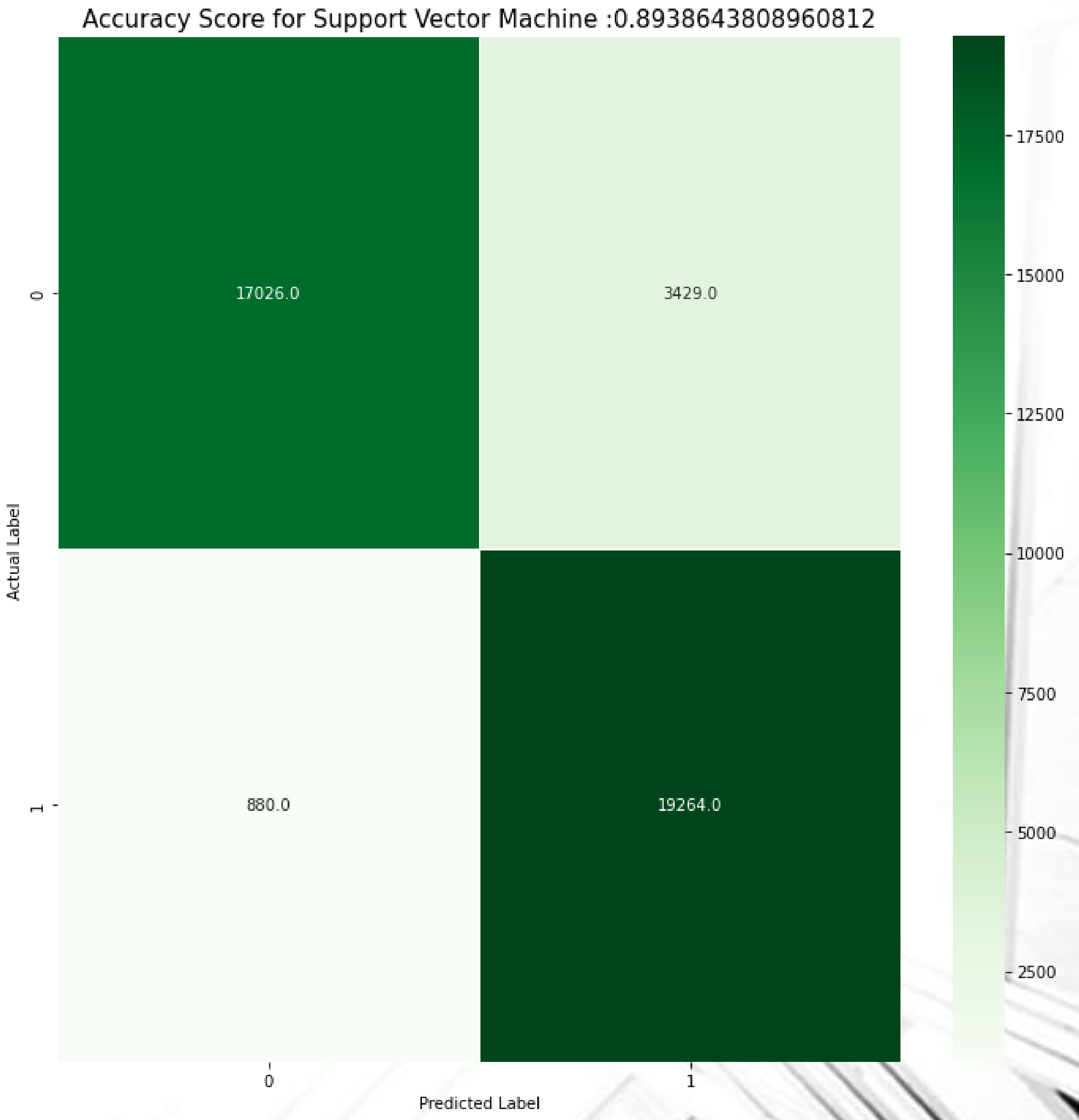
F-1 Score – 0.89386438089

Precision Score – 0.89386438089

Recall Score – 0.89386438089

Jaccard Score – 0.80809655

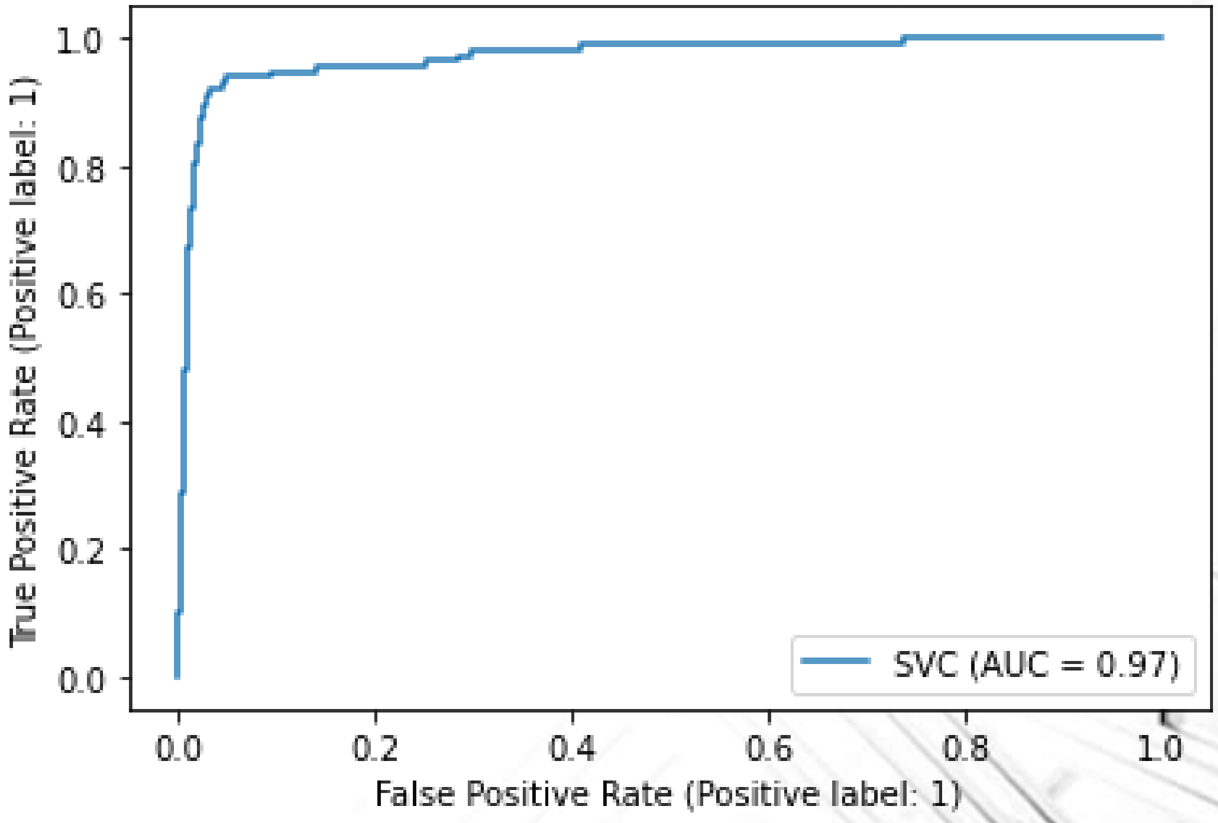
Log Loss – 3.6658619499



FINDINGS/RESULTS

SUPPORT VECTOR MACHINE

Finally, I generated the Confusion Matrix to measure recall, precision, accuracy and AUC-ROC, and plotted the ROC curve.



CONCLUSION

For this dataset, Random Forest algorithm helps us build the most accurate model for detecting/predicting credit card fraud

Limitations;

Our dataset was missing pertinent feature variables such as; gender, age, known income, and employment status of both originator and recipient that would have enabled us to gain in-depth insights on current and emerging trends as far as credit card fraud is concerned.

REFERENCES

1. www.Kaggle.com
2. www.money254.co.ke
3. www.ijcbss.org
4. www.centralbank.go.ke