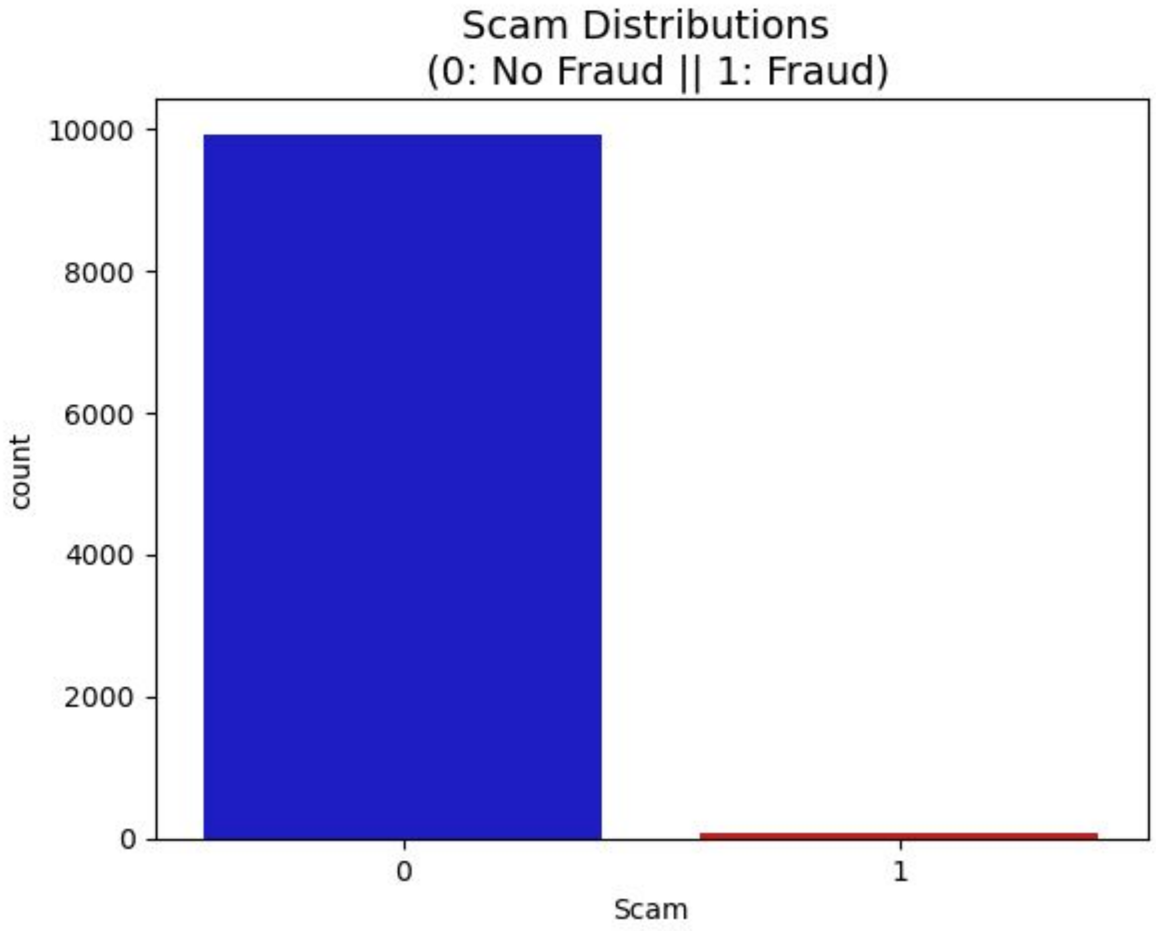
tripadvisor®

Bank Fraud Detection

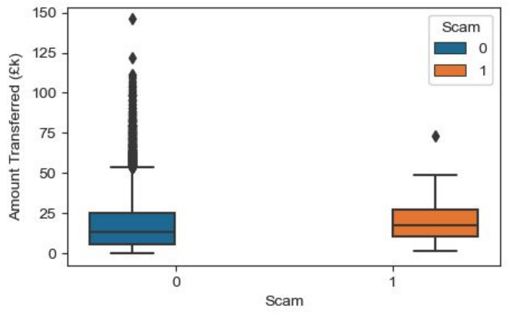
Imbalance Data Set



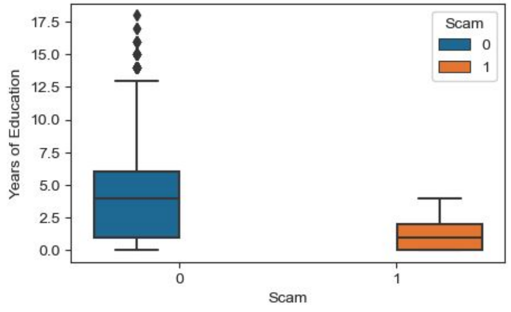
1. 99.19% of dataset is not fraud. Only 0.81% of the dataset is fraud
2. This data set is relatively small (only 10000 rows)

Explanatory Analysis 1

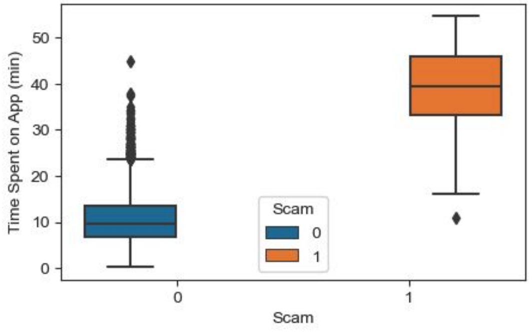
```
non fraud 13.12474397
fraud 17.67900139
```



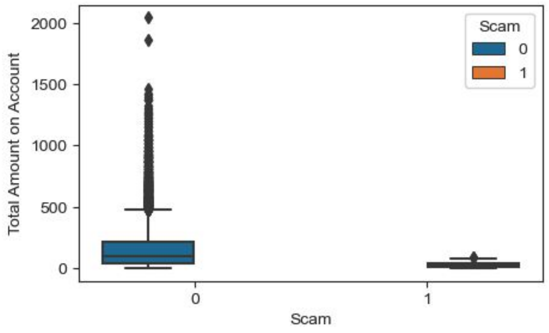
```
non fraud 4.0
fraud 1.0
```



```
non fraud 9.680762736
fraud 39.30743174
```



```
non fraud 103.9257531
fraud 22.90620078
```



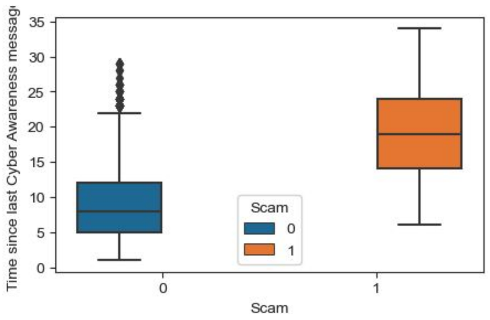
Median	Scam	Non Scam
Amount Transferred	1.76	13.1
Years of Education	1	4
Time Spent on App (mins)	39.3	9.7
Total Amount of Account	23	104

Takeaway

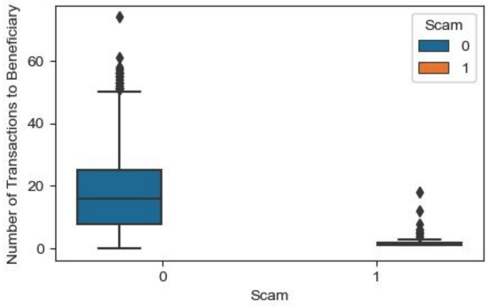
1. Years of Education, Time Spent on App and Total amount on account can be good predictors for fraud since the median between scam and non scam seem different.

Explanatory Analysis 2

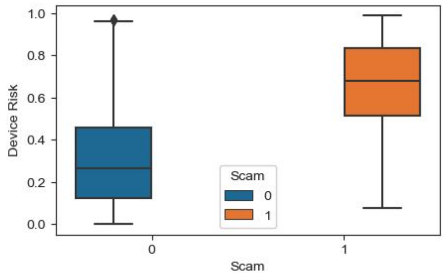
```
non fraud 8.0
fraud 19.0
```



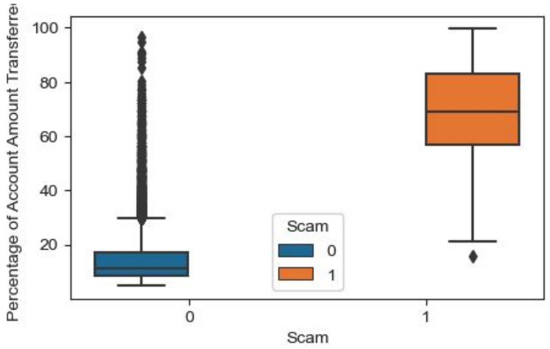
```
non fraud 16.0
fraud 2.0
```



```
non fraud 0.2641968
fraud 0.681043496
```



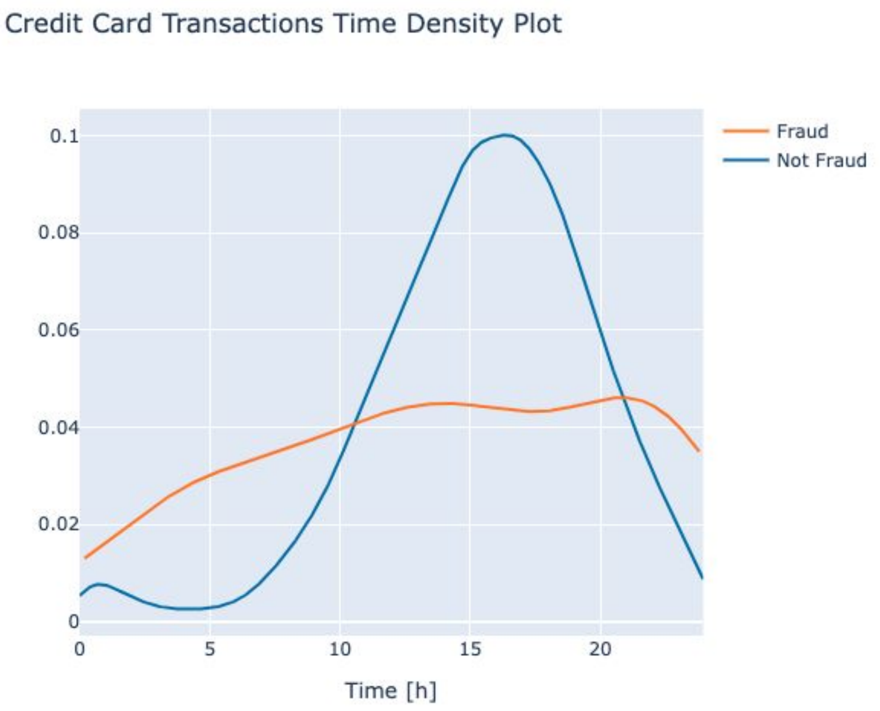
```
non fraud 11.42363902
fraud 68.96023321
```

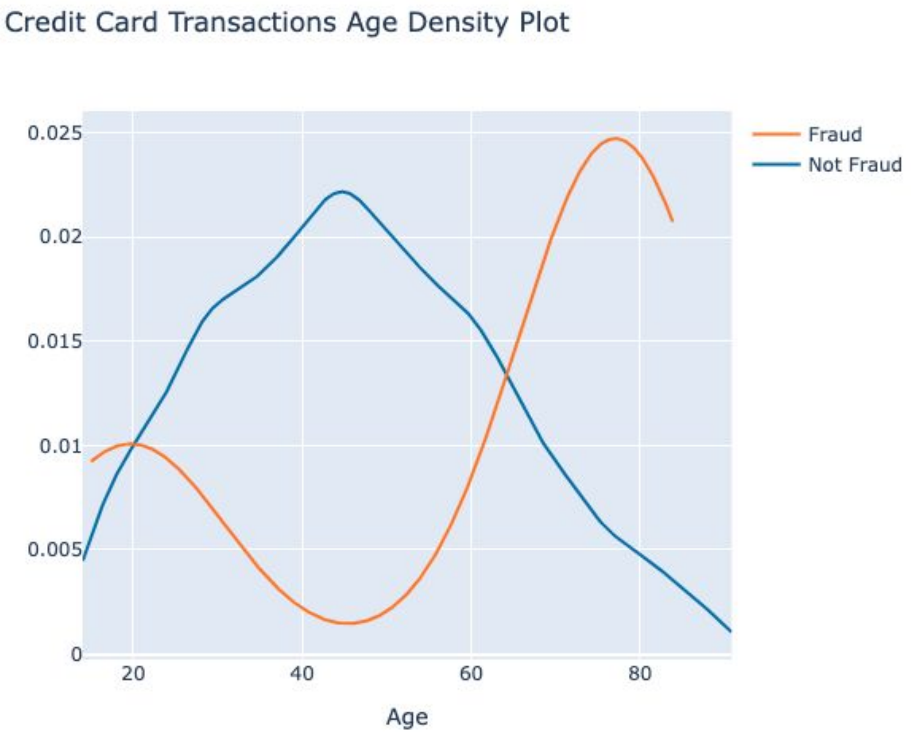


Median	Scam	Non Scam
Time since last Cyber Awareness message	19	8
Number of Transactions to Beneficiary	2	16
Device Risk	0.68	0.26
Percentage of Account Amount Transaction	69.0	11.4

1. Takeaway: Time since last Cyber Awareness message, Number of Transactions to Beneficiary, Device Risk, and Percentage of Account Amount Transaction can be good predictors for fraud since the median between scam and non scam seem different

Explanatory Analysis 3





1. Time: Most of non fraud transactions happen between 10 and 22. However, some fraud transactions happen after 22 and before 10.
2. Age: Fraud Transaction happens to have ages bigger than 65 and less than 20 while non fraud transactions have ages between 20 and 65.

Important Metrics

Problem of Accuracy: We cannot use accuracy to evaluate this model. For example, in this case, if a model predicts all the transactions as non-fraud, it gets 99.19% accuracy since 99.19% of data is non-fraud

Recall: Recall estimates how many of the actual frauds our model captures.

Precision: : Precision estimates how many of the predicted frauds are actual frauds.

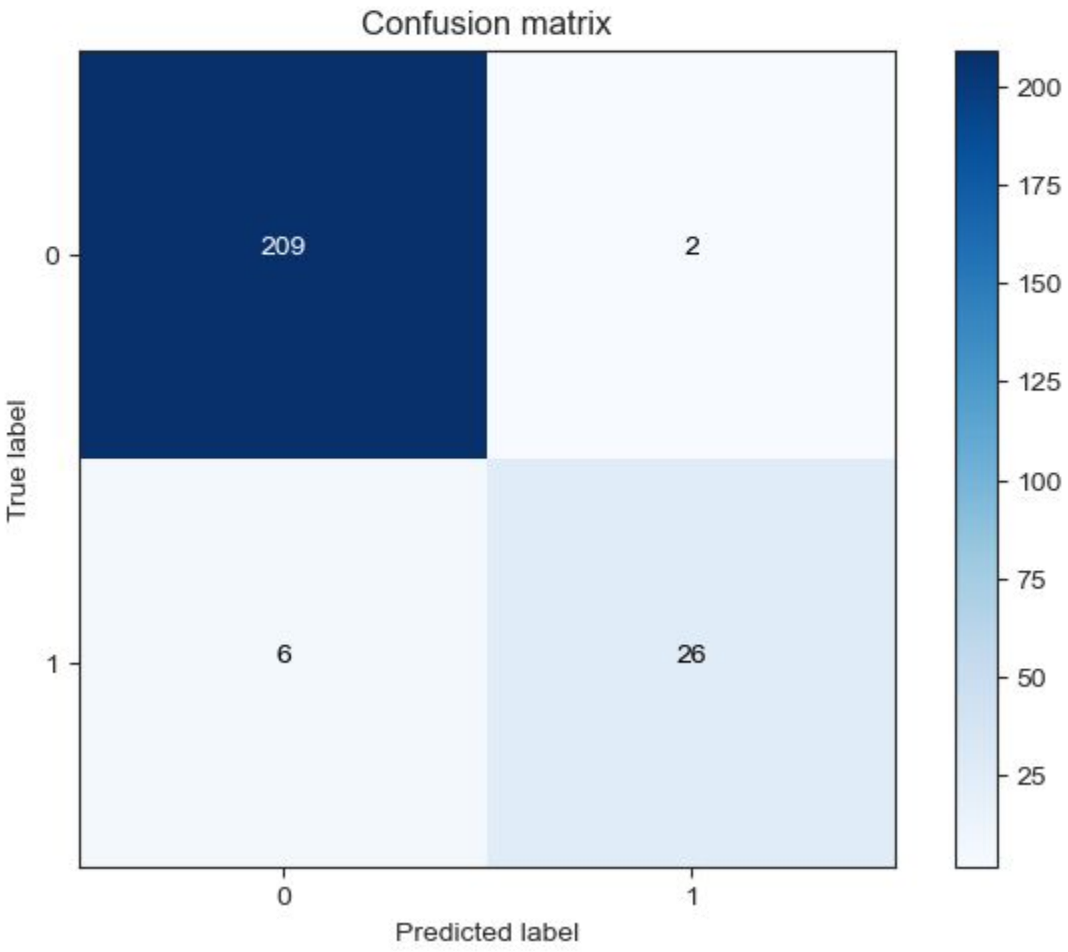
In this case, recall is the most important metrics since we don’t want to miss any actual frauds

Model Selection
(This slide is for technical audience)

Steps to choose a model:

- This data set is very imbalanced and has only 10000 rows.. 99.19% of dataset is not fraud. Only 0.81% of the dataset is fraud
- I undersampled the non fraud dataset so that fraud data set consists 10% of transactions in order to have a more **balanced dataset** and thus avoiding a model to overfitting
- I picked logistic regression with L1 penalty for the following reasons
  a. since time is limited to try more complicated models
  b. there is not much data to fit a more complex model like Xgboost or Random Forest
  c. L1 removes unnecessary variables and makes it easy to interpret
- Used 5-Fold Cross Validation to pick the best parameter

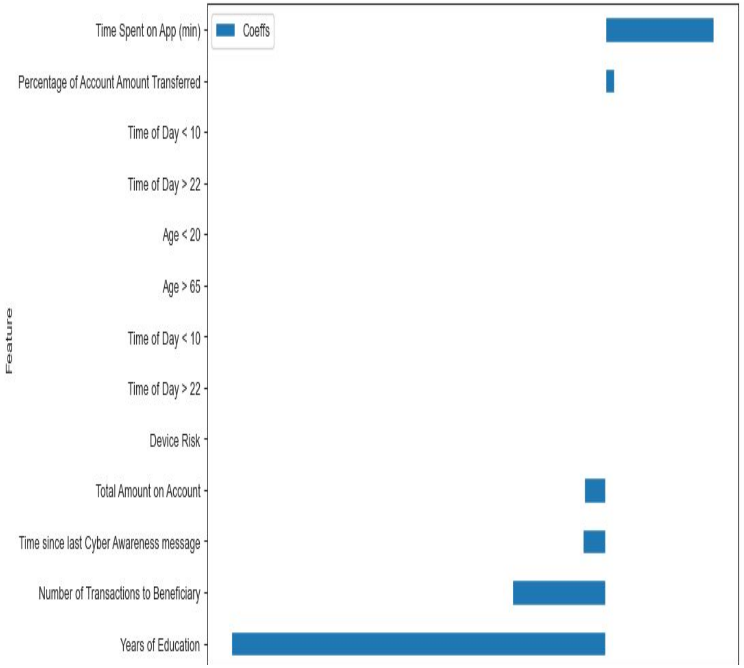
Model Performance on Test Set

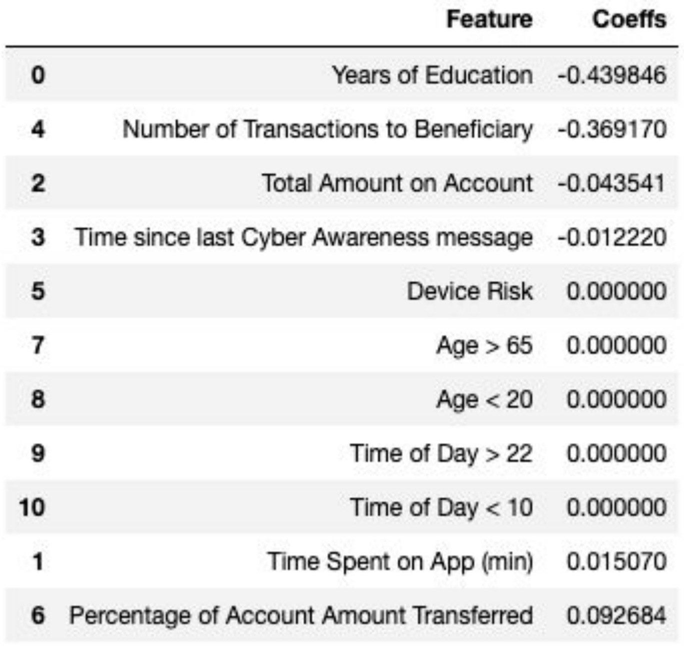


Precision: Out of 32 actual frauds, 26 (81%) were successfully detected by a simple model.

Recall: Out of 28 predicted frauds, 26 were actual fraud (92.8%)

Important Features to Detect Fraud





	Feature	Coeffs
0	Years of Education	-0.439846
4	Number of Transactions to Beneficiary	-0.369170
2	Total Amount on Account	-0.043541
3	Time since last Cyber Awareness message	-0.012220
5	Device Risk	0.000000
7	Age > 65	0.000000
8	Age < 20	0.000000
9	Time of Day > 22	0.000000
10	Time of Day < 10	0.000000
1	Time Spent on App (min)	0.015070
6	Percentage of Account Amount Transferred	0.092684

Important Variables:

1. Years of Education
2. # of Transactions on Beneficiary
3. Time Since last Cyber Awareness Message
4. Total Amount on Account
5. Time spent on app
6. Percentage of account amount transferred

Example Interperation

Time Spent on App: An increase of 1 in Time Spent on App is associated with an increase of 1.5% in the odds of being fraud (*assuming that all the variables remains fixed*).

Years of Education: An increase of 1 in Years of Education is associated with an decrease of 36% in the odds of being fraud (*assuming that all the variables remains fixed*).

Number of Transactions to Beneficiary: An increase of 1 in Number of Transactions to Beneficiary is associated with an decrease of 31% in the odds of being fraud (*assuming that all the variables remains fixed*).

Can we use any other data (internal or external) to help us identify whether the likelihood of scam transactions will increase

Ideas about increasing the accuracy

1. Get more data (10000 rows is not enough)
2. User’s address (far-apart locations in a short time frame)
3. Date of data. Is fraud decreasing or not over time
4. Is fraud decreasing nationwide (external data). Then which area?
5. Device id. Even if a fraudulent user creates a new account, we can flag based on device id
6. Behavioral biometrics: We can analyze user behavior patterns, such as typing speed, swipe gestures, or app usage, to verify the user’s identity and detect any anomalies that may suggest fraud.