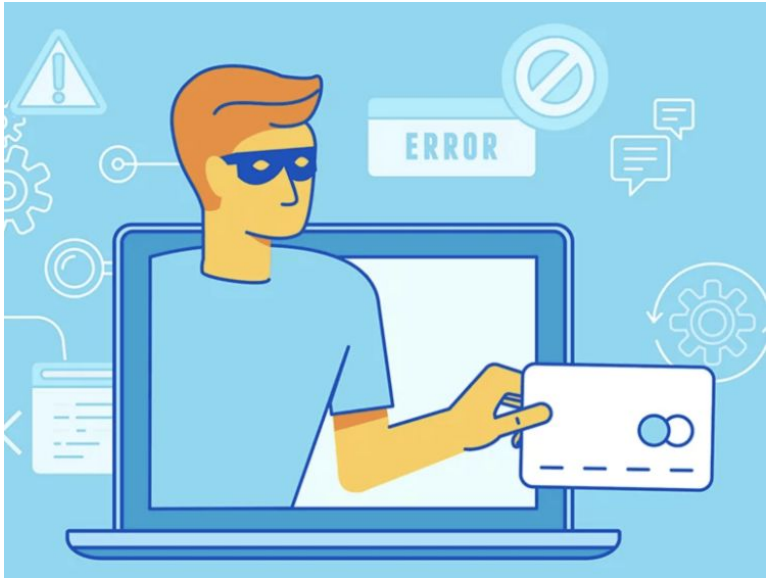
Using ML models to predict fraudulent transaction

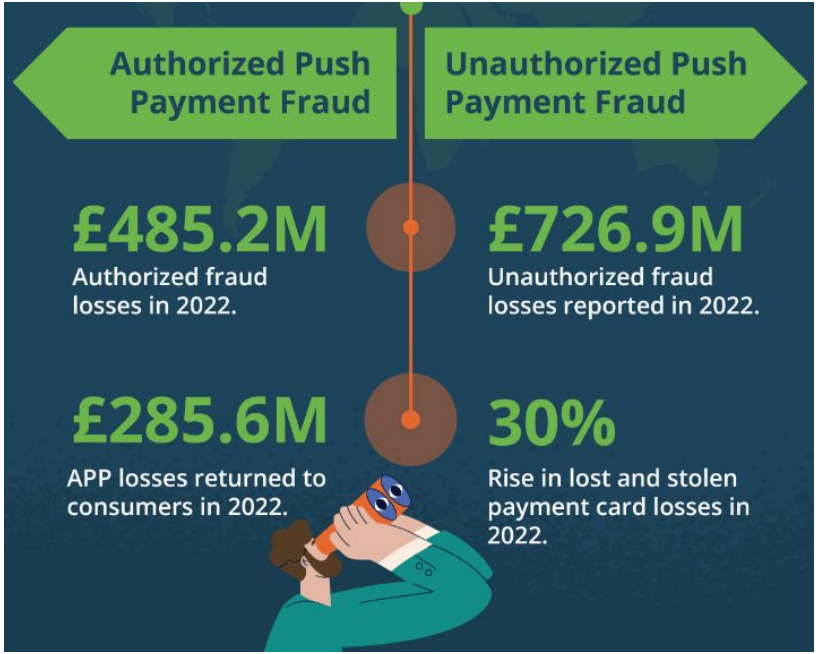
Content

1. What is a fraudulent transaction?
2. Why important?
3. How fraud is detected?
4. The dataset
5. Preprocessing
6. 3 datasets
7. Evaluation of the model
8. Conclusion
9. Appendix



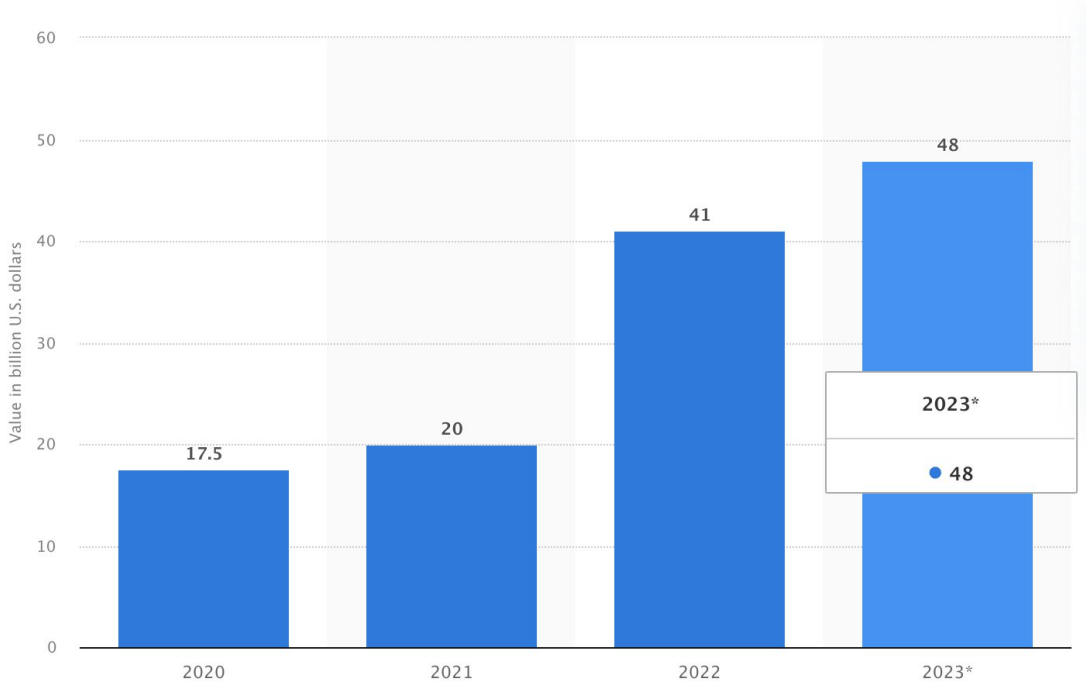
What is a fraudulent transaction?

- any type of purchase which was not authorized by a legitimate user



Why important?

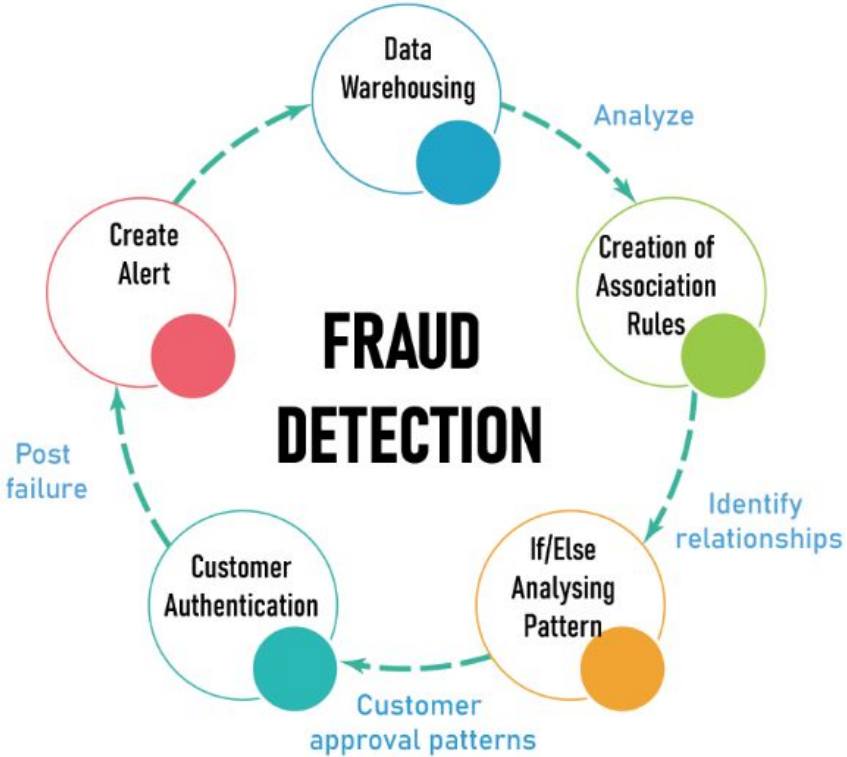
- business losing money
- Reputation at stake
- Personal information stolen



E-commerce payment fraud losses worldwide 2020-2023

From Statistica.com

How fraud is detected?

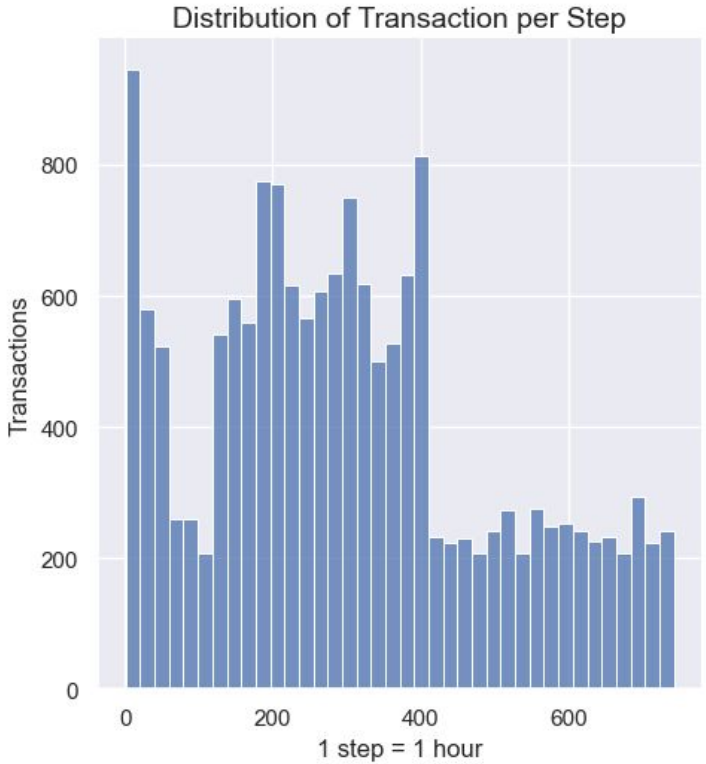


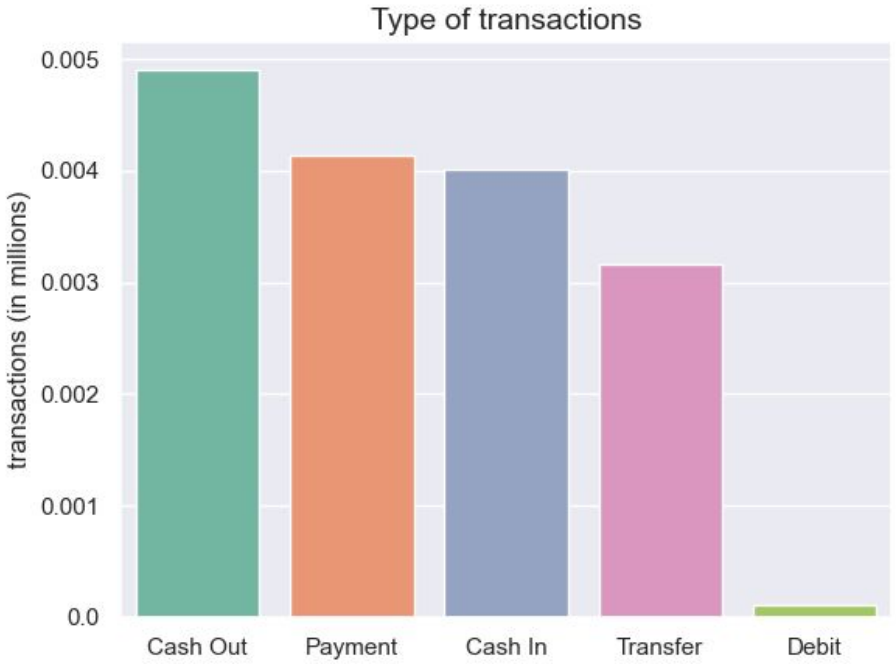
The dataset

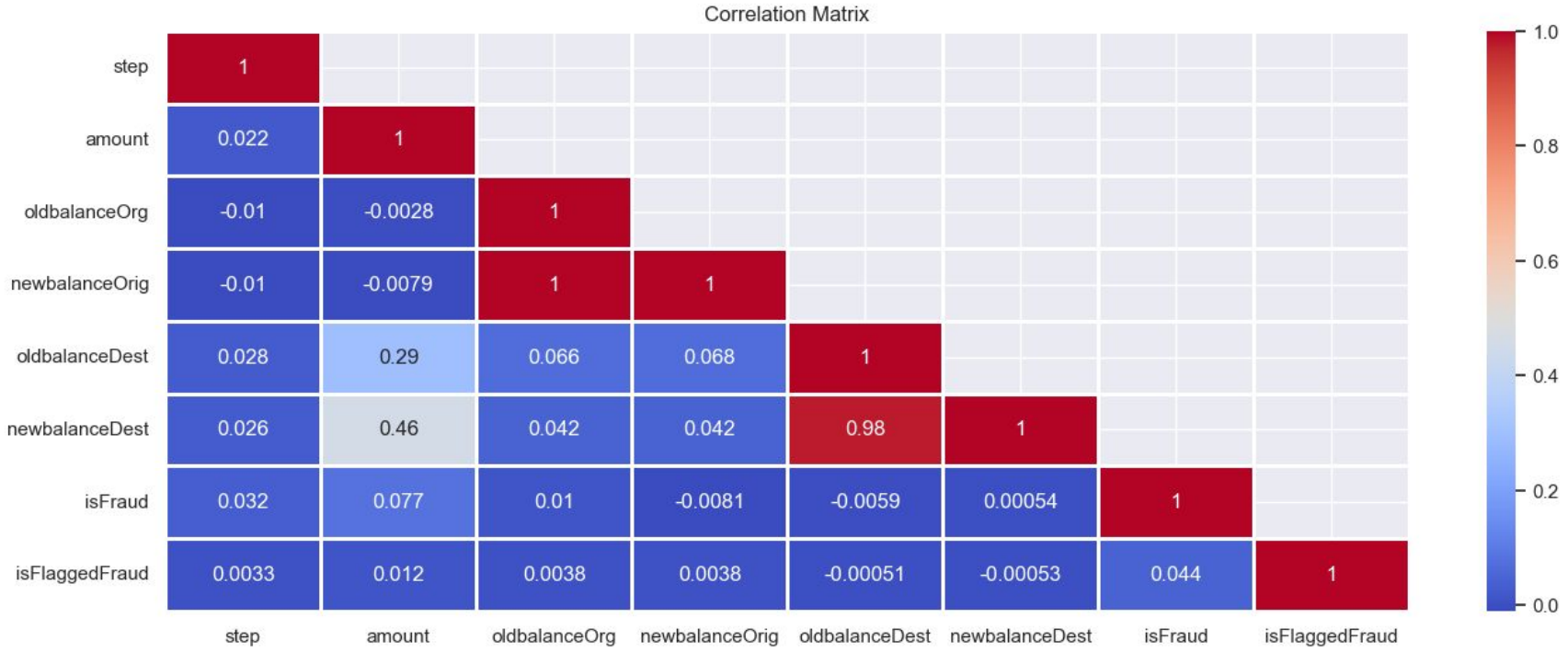
- 6,362,620 records
- 11 columns (10 features)
- 3 categorical
- 8 numerical
- Target variable: isFraud

- No missing values

```
RangeIndex: 6362620 entries, 0 to 6362619
Data columns (total 11 columns):
 #   Column          Dtype
---  ------          -----
 0   step            int64
 1   type            object
 2   amount          float64
 3   nameOrig        object
 4   oldbalanceOrg   float64
 5   newbalanceOrig  float64
 6   nameDest        object
 7   oldbalanceDest  float64
 8   newbalanceDest  float64
 9   isFraud         int64
 10  isFlaggedFraud  int64
dtypes: float64(5), int64(3), object(3)
```


Distribution of Transaction per Step
Transactions
800
600
400
200
0
0
200
400
600
1 step = 1 hour


Type of transactions
transactions (in millions)
0.005
0.004
0.003
0.002
0.001
0.0
Cash Out
Payment
Cash In
Transfer
Debit


Correlation Matrix
step
amount
oldbalanceOrg
newbalanceOrig
oldbalanceDest
newbalanceDest
isFraud
isFlaggedFraud
1
0.022 1
-0.01 -0.0028 1
-0.01 -0.0079 1 1
0.028 0.29 0.066 0.068 1
0.026 0.46 0.042 0.042 0.98 1
0.032 0.077 0.01 -0.0081 -0.0059 0.00054 1
0.0033 0.012 0.0038 0.0038 -0.00051 -0.00053 0.044 1
step
amount
oldbalanceOrg
newbalanceOrig
oldbalanceDest
newbalanceDest
isFraud
isFlaggedFraud
1.0
0.8
0.6
0.4
0.2
0.0

1- Issue with the dataset

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
0	1	PAYMENT	9839.64	C1231006815	170136.0	160296.36	M1979787155	0.0	0.0	0	0

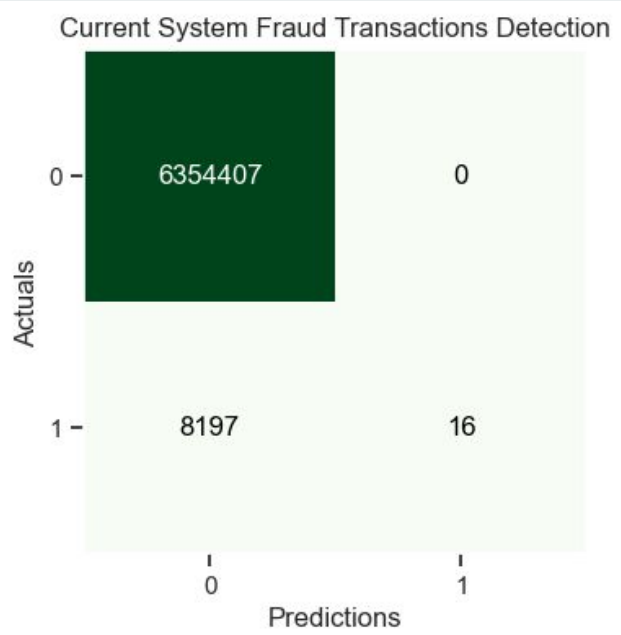
- Transaction amount is not reconciled for the recipient nameDest and its before&after balance

- Removed the mismatched records:

  38% left

	Type	Original dataset	New dataset	Percentage of new dataset
5	TOTAL	6362620	2423175	38.1
0	Cash Out	2237500	239407	10.7
1	Payment	2151495	945843	44.0
4	Cash In	1399284	1186107	84.8
3	Transfer	532909	23281	4.4
2	Debit	41432	28537	68.9

Imbalance dataset: undersample

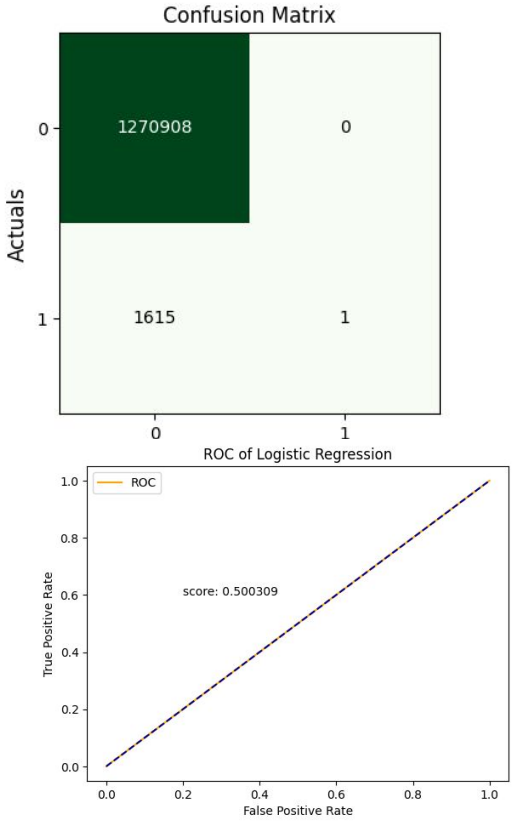


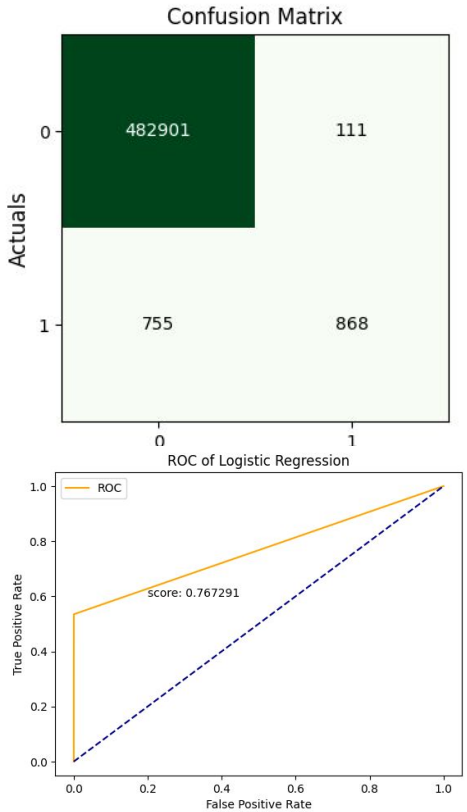
	Type	Original dataset	New dataset	Percentage of new dataset
0	TOTAL	6362620.00	2423175.00	38.1
1	isFraud	8197.00	8168.00	99.6
2	% isFraud	0.13	0.34	261.5

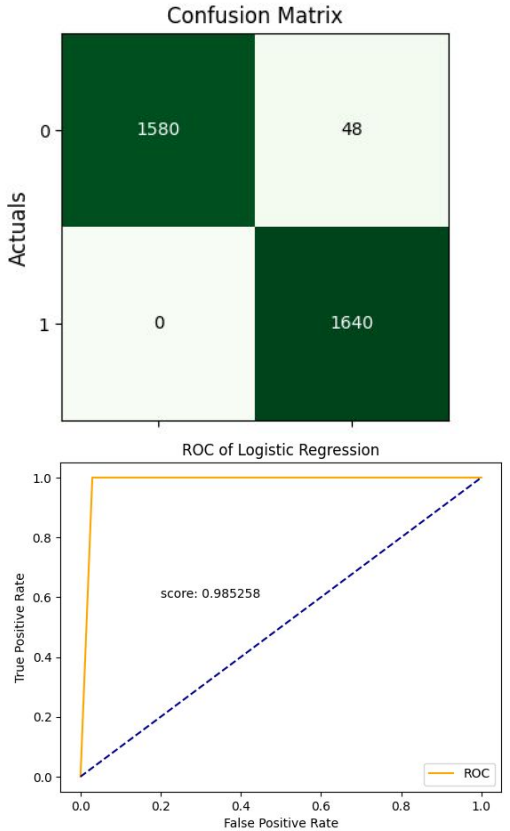
2- Pre-processing

- Dropped nameOrign, nameD
- Replace outliers with median
- One-hot encoding to type
- Scaled with RobustScaler()

Logistic model: original data, data2, data3



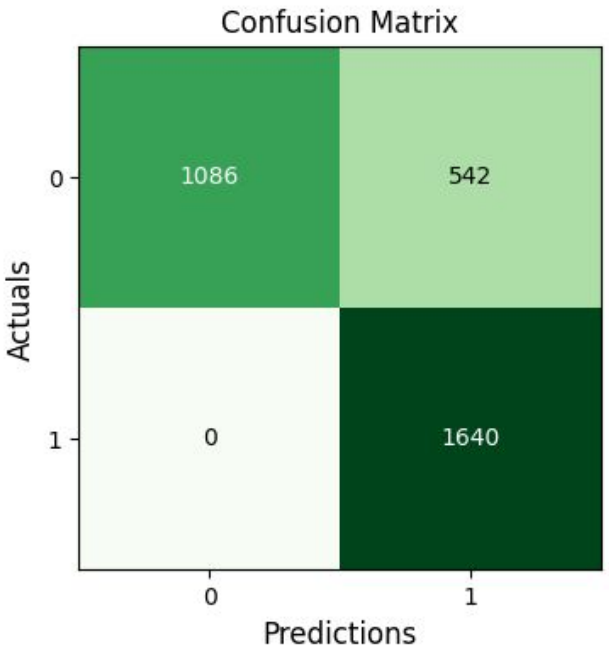


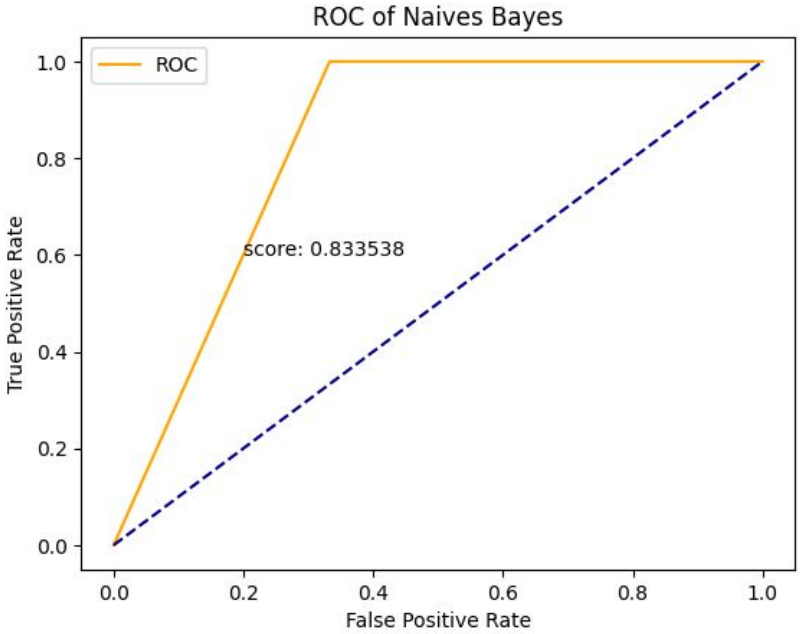


Data 3

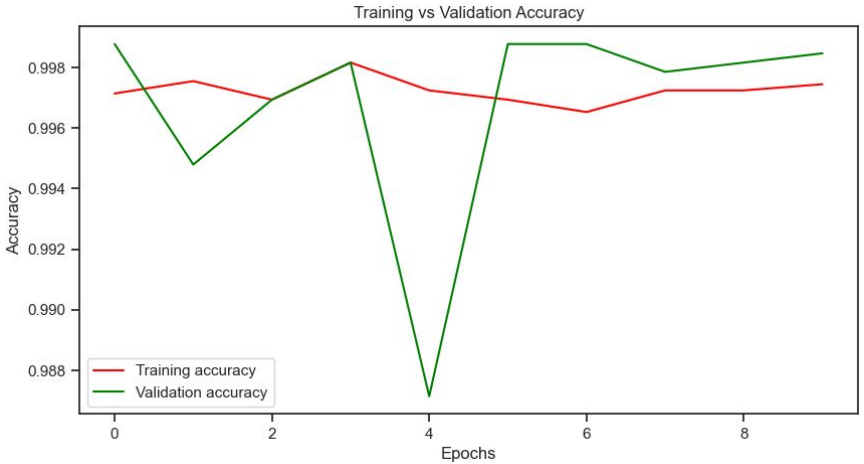

Number of fraud transactions
transactions (in millions)
8000
7000
6000
5000
4000
3000
2000
1000
0
50%
50%
Not Fraud
Fraud

Data 3: Naive Bayes



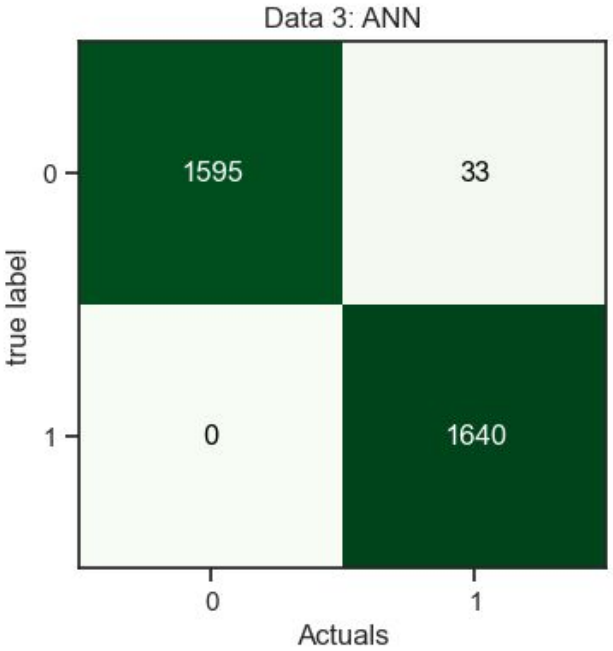


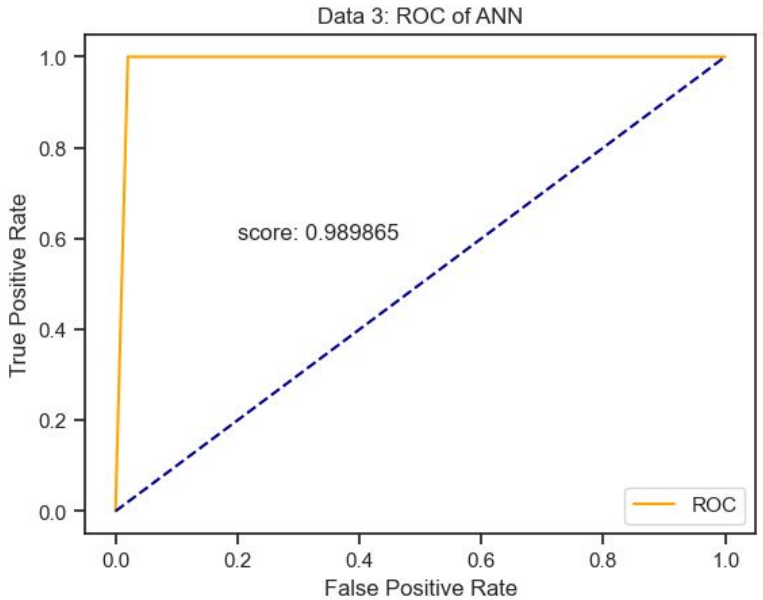
Data 3: ANN




```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_3 (Dense)             (None, 16)                192

 dense_4 (Dense)             (None, 8)                 136

 dropout_1 (Dropout)         (None, 8)                 0

 dense_5 (Dense)             (None, 1)                 9

=================================================================
Total params: 337
Trainable params: 337
Non-trainable params: 0
```

ANN Model 1: loss: 0.0090 - accuracy: 0.9974 - val_loss: 0.0141 - val_accuracy: 0.9985





Conclusion

- What I can explore if I had more time:
  - hyperparatune more the ANN
  - try more models on the other datasets
  - ask an expert for the mismatched transactions
  - try oversampling to maintain big dataset

- Overall, I think the models such as ANN or logistic regression are good, depending on the company's budget
- What we would need to implement the solution: data acquisition, hardware

Appendix

Github:

Date: 23/06/2013

Statistica barplot:
https://www.statista.com/statistics/1273177/ecommerce-payment-fraud-losses-globally/