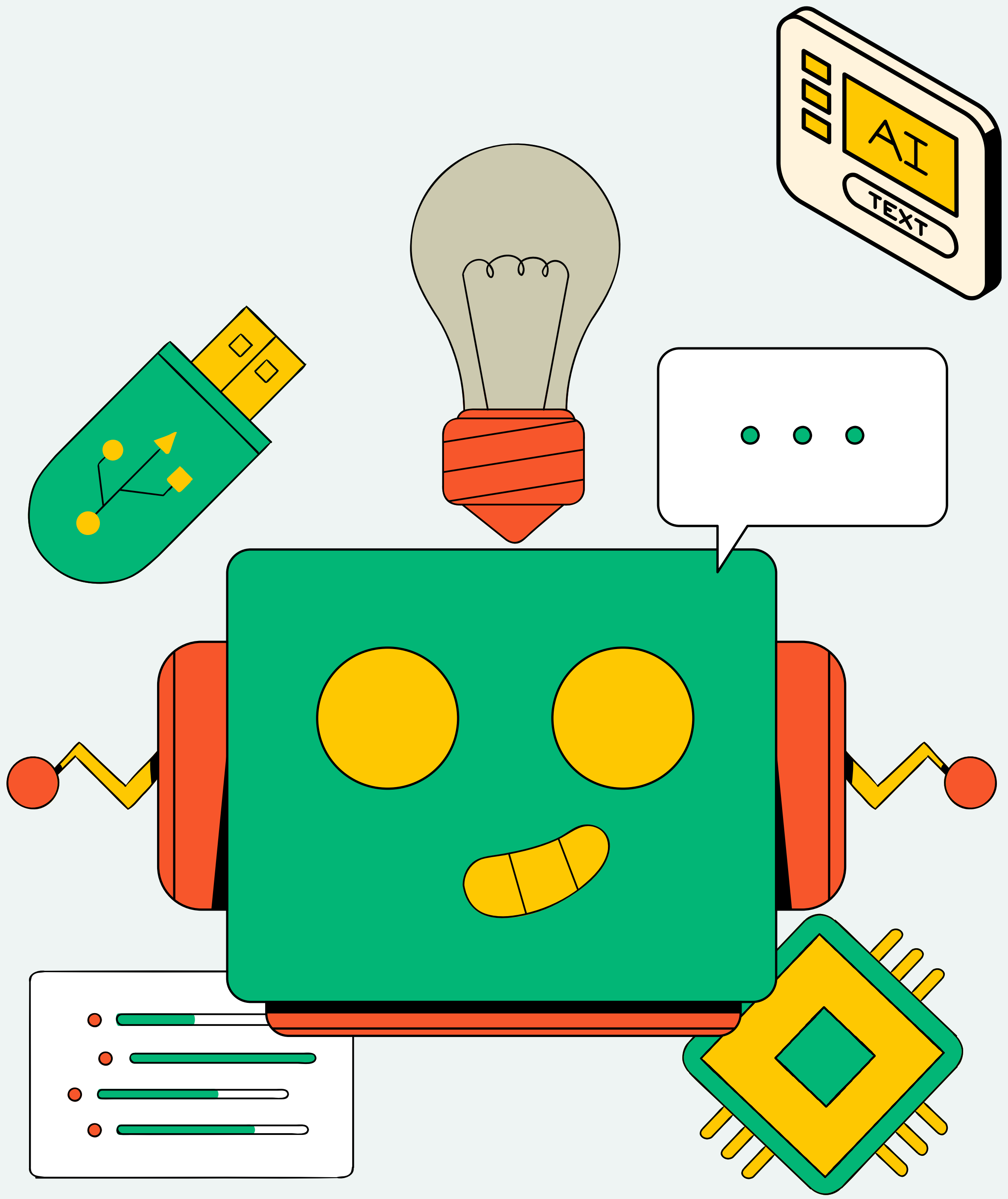
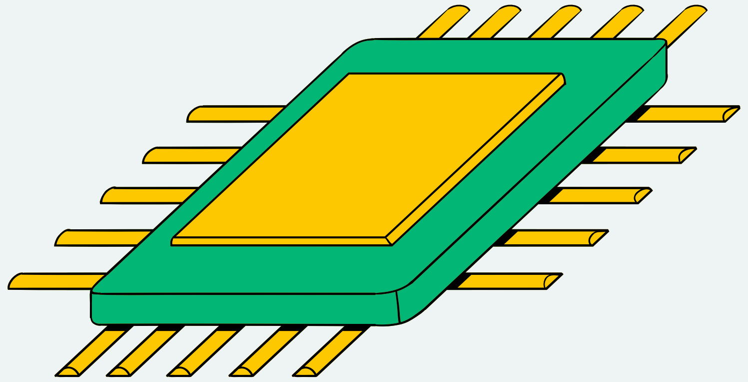
SML CSE 342: STATISTICAL MACHINE LEARNING

FRAUD DETECTION IN FINTECH BANKING TRANSACTION



PRESENTED BY:

VASHU (2022606)
SHOBHIT RAJ (2022482)



INTRODUCTION

The surge in internet and e-commerce drives the prevalence of online payment transactions, with thousands occurring every second. However, this rise also escalates the risk of financial fraud, reducing the integrity of FinTech banking transactions. Fraud detection plays a crucial role in safeguarding customers and merchants from substantial financial losses and preserving trust in online payment systems.

In this project, we propose a fraud detection system for online payments that uses machine learning techniques to identify and prevent fraudulent transactions.

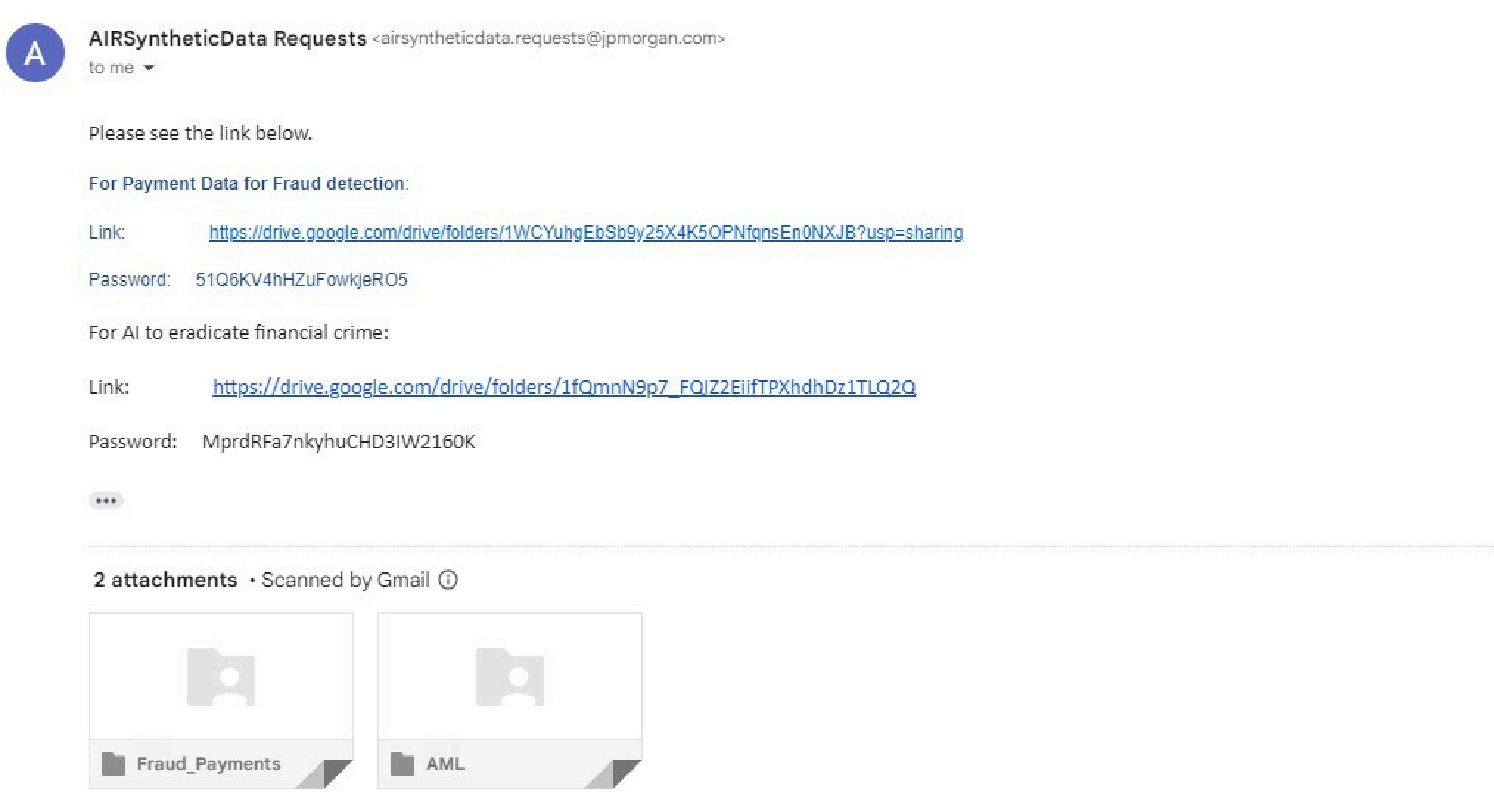
PROBLEM STATEMENT

"Detecting and preventing financial fraud in FinTech banking transactions using Statistical Machine Learning techniques."

DATASET DETAILS

- JP MORGAN DATASET

The synthetic-data used in the project is especially provided by JP Morgan for research purposes which replicate the intricacies of real transactional data.
The Dataset contains 14,98,177 transaction details with 13 parameters for each transaction labelled Fraud and Non-Fraud.


AIRSyntheticData Requests <airsyntheticdata.requests@jpmorgan.com>
to me

Please see the link below.

For Payment Data for Fraud detection:

Link: https://drive.google.com/drive/folders/1WCYuhgEbSb9y25X4K5OPNfqnsEn0NXJB?usp=sharing

Password: 51Q6KV4hHZuFowkjeRO5

For AI to eradicate financial crime:

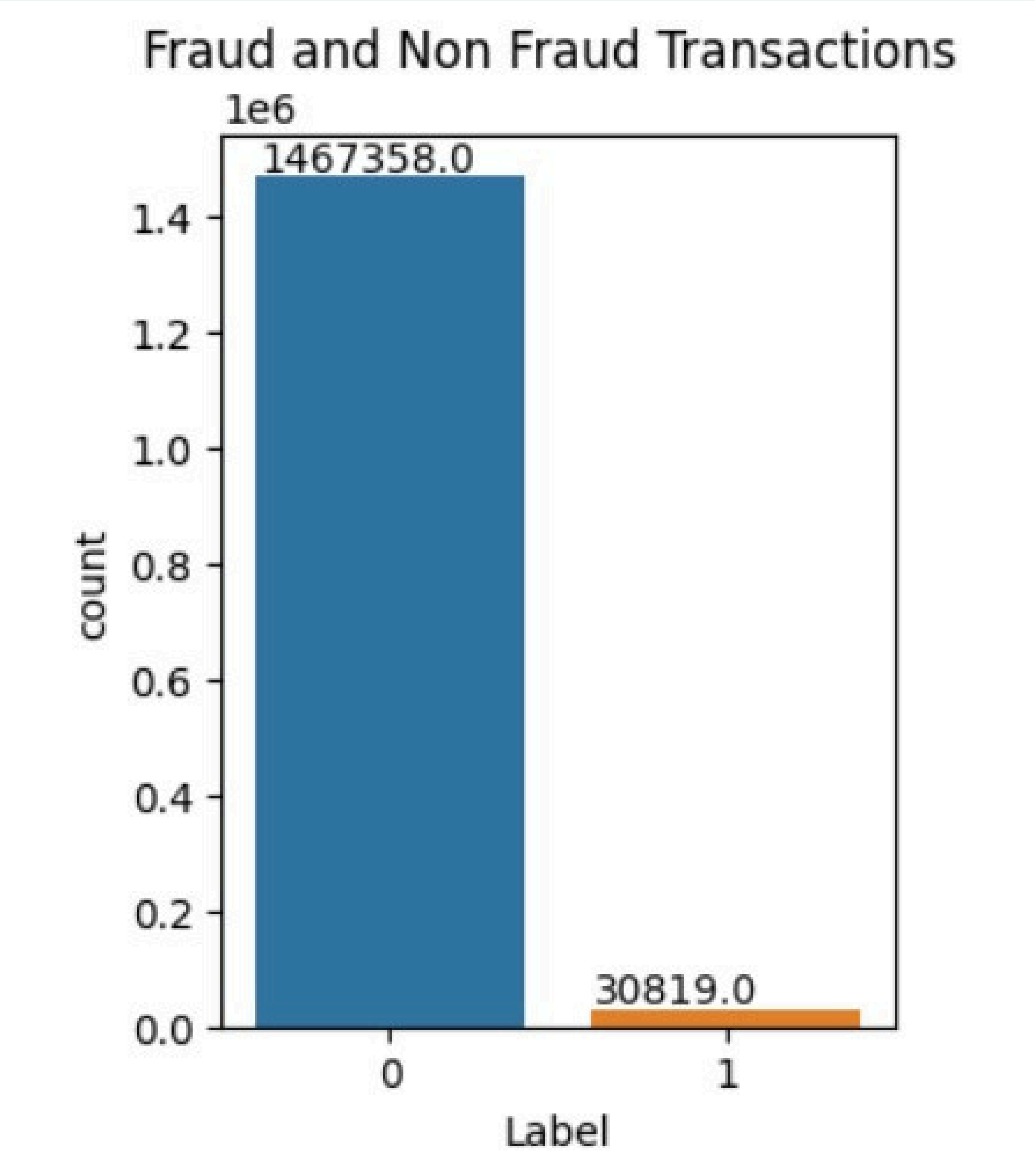
Link: https://drive.google.com/drive/folders/1fQmnN9p7_FQIZ2EiifTPXhdhDz1TLQ2Q

Password: MprdRFa7nkyhuCHD3IW2160K

2 attachments • Scanned by Gmail

Fraud_Payments

AML

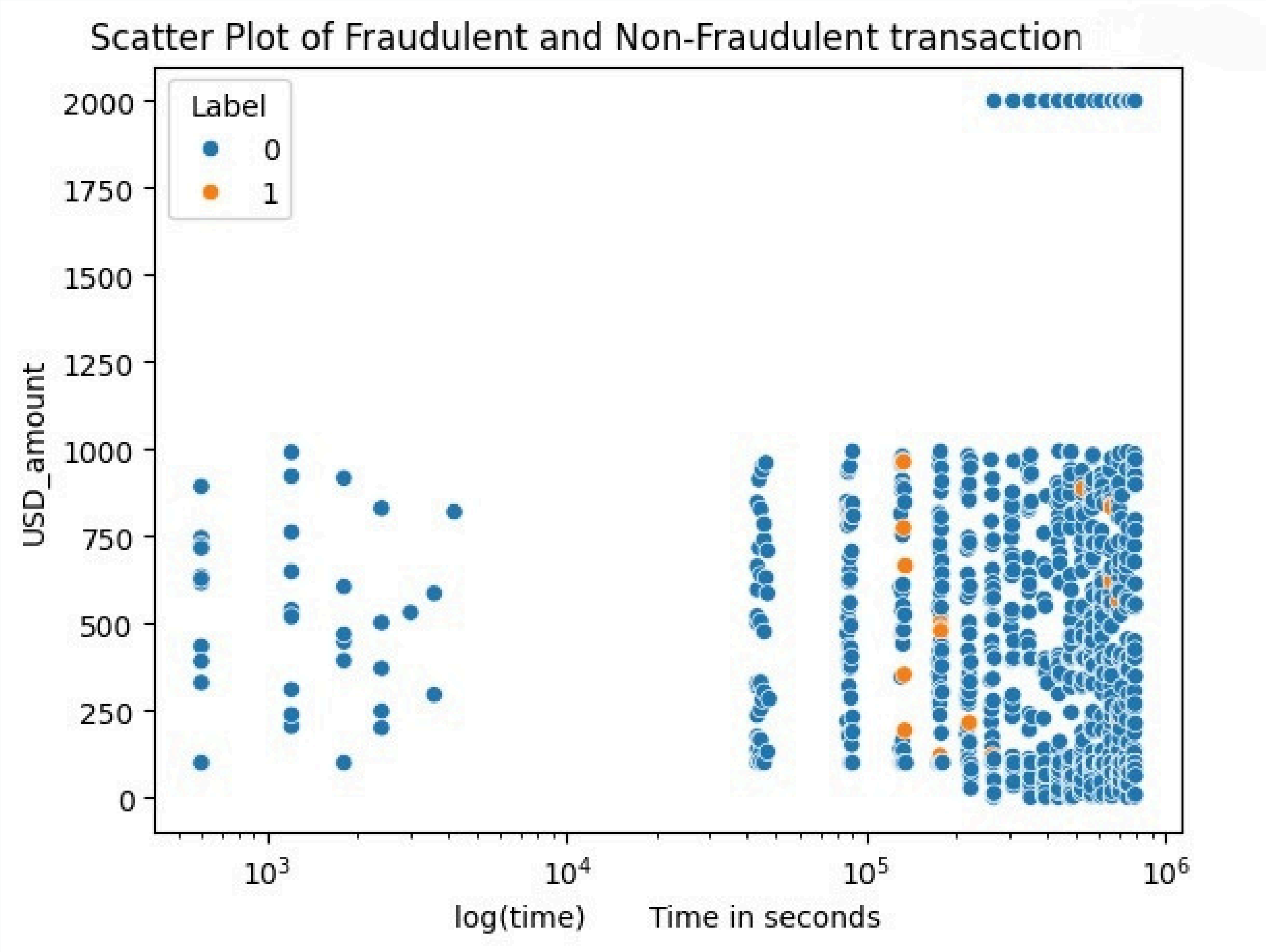


OBJECTIVES WITH THE DATASET

Analyze different patterns in the dataset and discern complex patterns

Exploring the nuances of the dataset with severe class imbalance

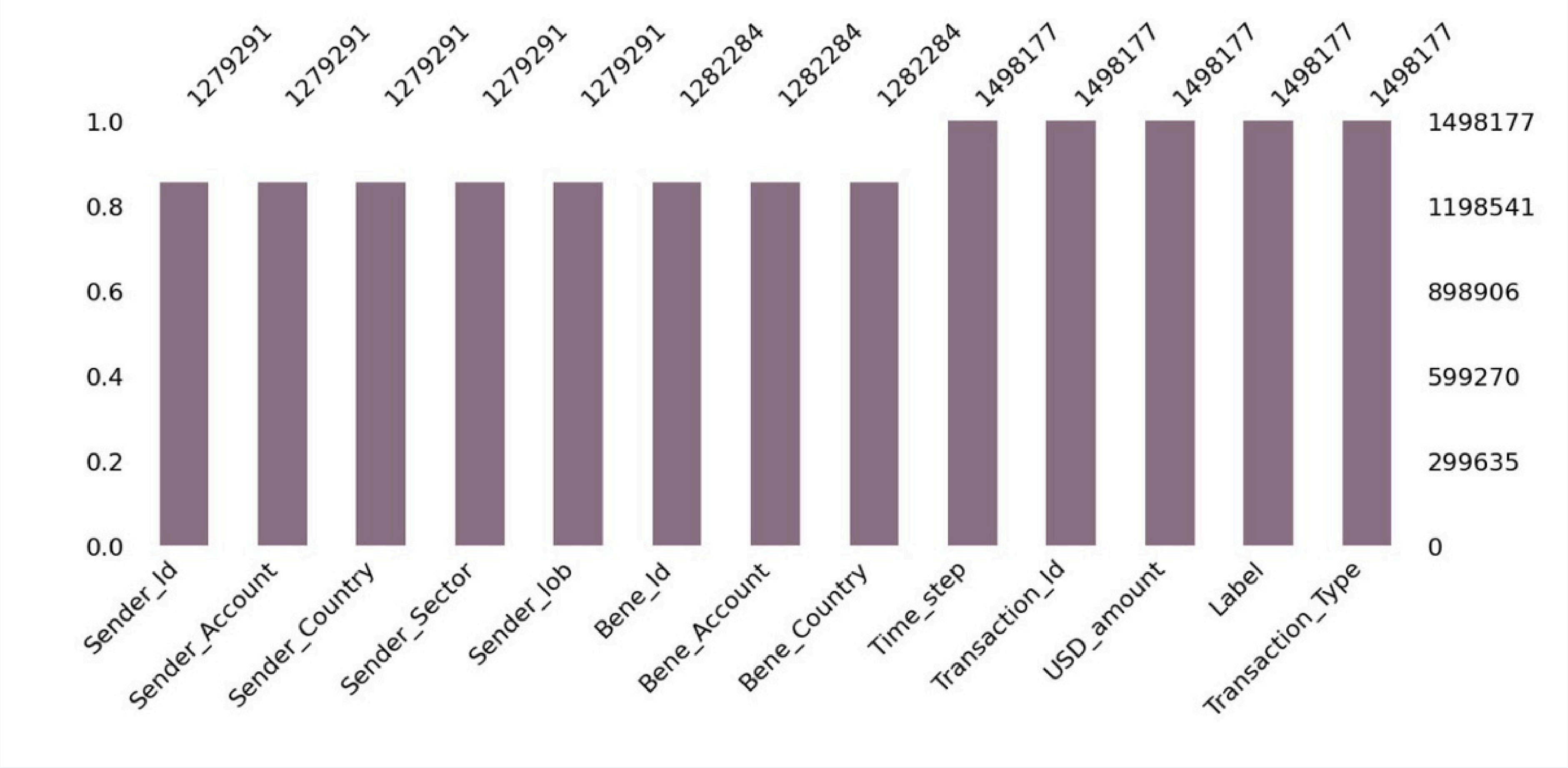
DATASET VISUALIZATIONS - JP MORGAN



WORK DONE - JP MORGAN

The data contains 4,50,000(approx.) transactions with null-fields. We used imputation strategy to induce the values in the null-fields.

Sender_Id	Bene_Id	Sender_Account	Bene_Account	Sender_Country	Sender_Sector	Sender_lob	Bene_Country
JPMC-CLIENT-10098	NaN	ACCOUNT-10108	NaN	USA	35537.0	CCB	NaN
JPMC-CLIENT-10098	CLIENT-10100	ACCOUNT-10109	ACCOUNT-10106	USA	15287.0	CCB	CANADA
NaN	JPMC-CLIENT-9812	NaN	ACCOUNT-9826	NaN	NaN	NaN	USA
JPMC-CLIENT-9812	JPMC-CLIENT-9814	ACCOUNT-9825	ACCOUNT-9824	USA	38145.0	CCB	USA
NaN	JPMC-CLIENT-9789	NaN	ACCOUNT-9800	NaN	NaN	NaN	USA



statistics of Null-Values

```
Sender_Id           4121
Bene_Id             5136
Sender_Country      4121
Sender_Sector       4121
Bene_Country        5136
USD_amount             0
Label                  0
Transaction_Type       0
dtype: int64
```

WORK DONE – JP MORGAN

The Null-fields in the column “Sender-Sector” were assigned avg. of fraud and Non-Fraud Transactions of “sender-sector” field.

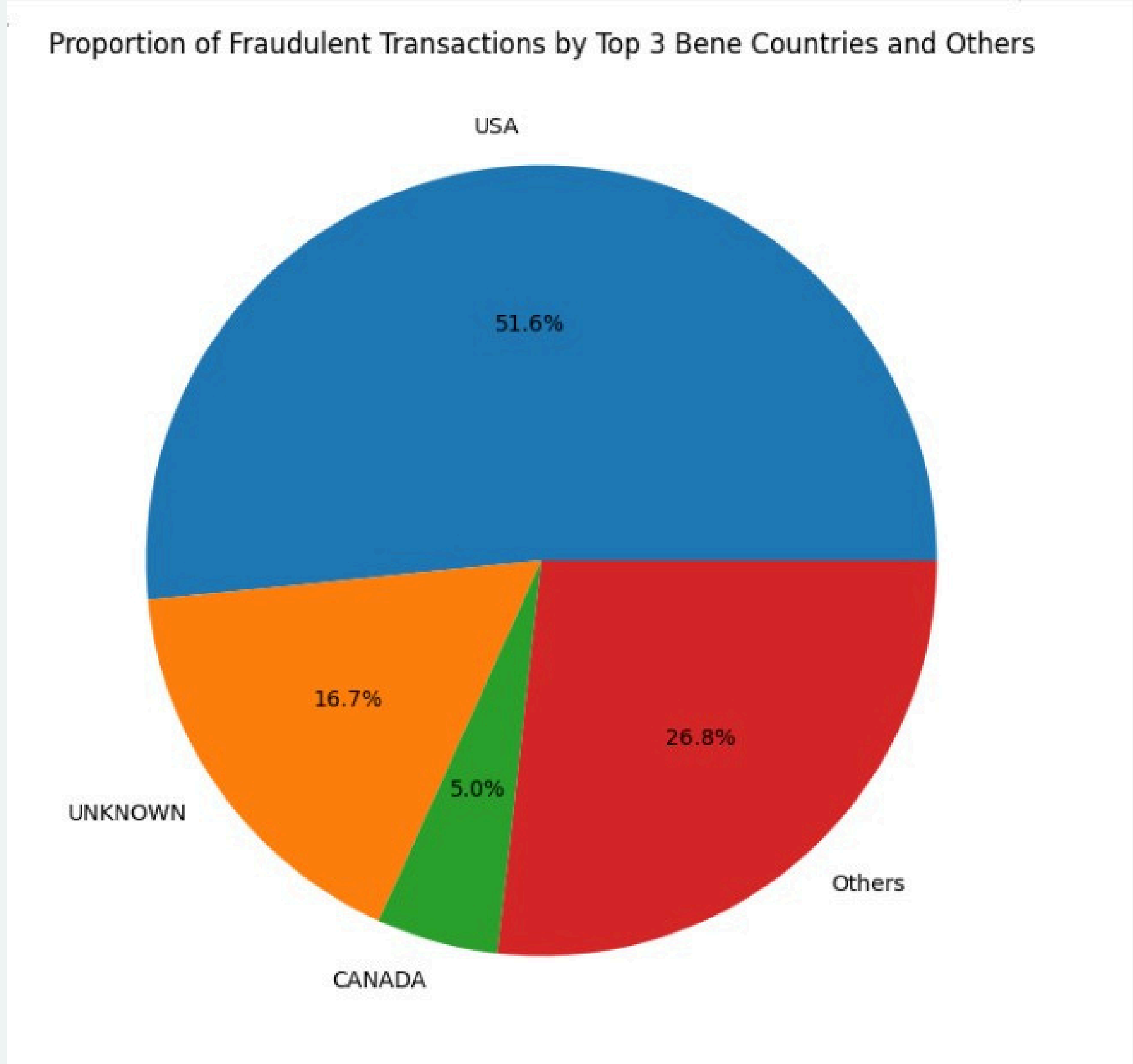
The Null-fields in the column “Bene-Country” and “Sender-Country” were assigned “UNKNOWN” TAG.

Sender_Id	Bene_Id	Sender_Country	Sender_Sector	Bene_Country	USD_amount	Label	Transaction_Type
JPMC-CLIENT-10098	NaN	UNKNOWN	35537.000000	UNKNOWN	558.43	0	WITHDRAWAL
JPMC-CLIENT-10098	CLIENT-10100	CANADA	15287.000000	CANADA	622.78	0	QUICK-PAYMENT
NaN	JPMC-CLIENT-9812	USA	25020.183491	USA	802.54	0	DEPOSIT-CASH
JPMC-CLIENT-9812	JPMC-CLIENT-9814	USA	38145.000000	USA	989.09	0	PAY-CHECK
NaN	JPMC-CLIENT-9789	USA	25020.183491	USA	786.78	0	DEPOSIT-CHECK

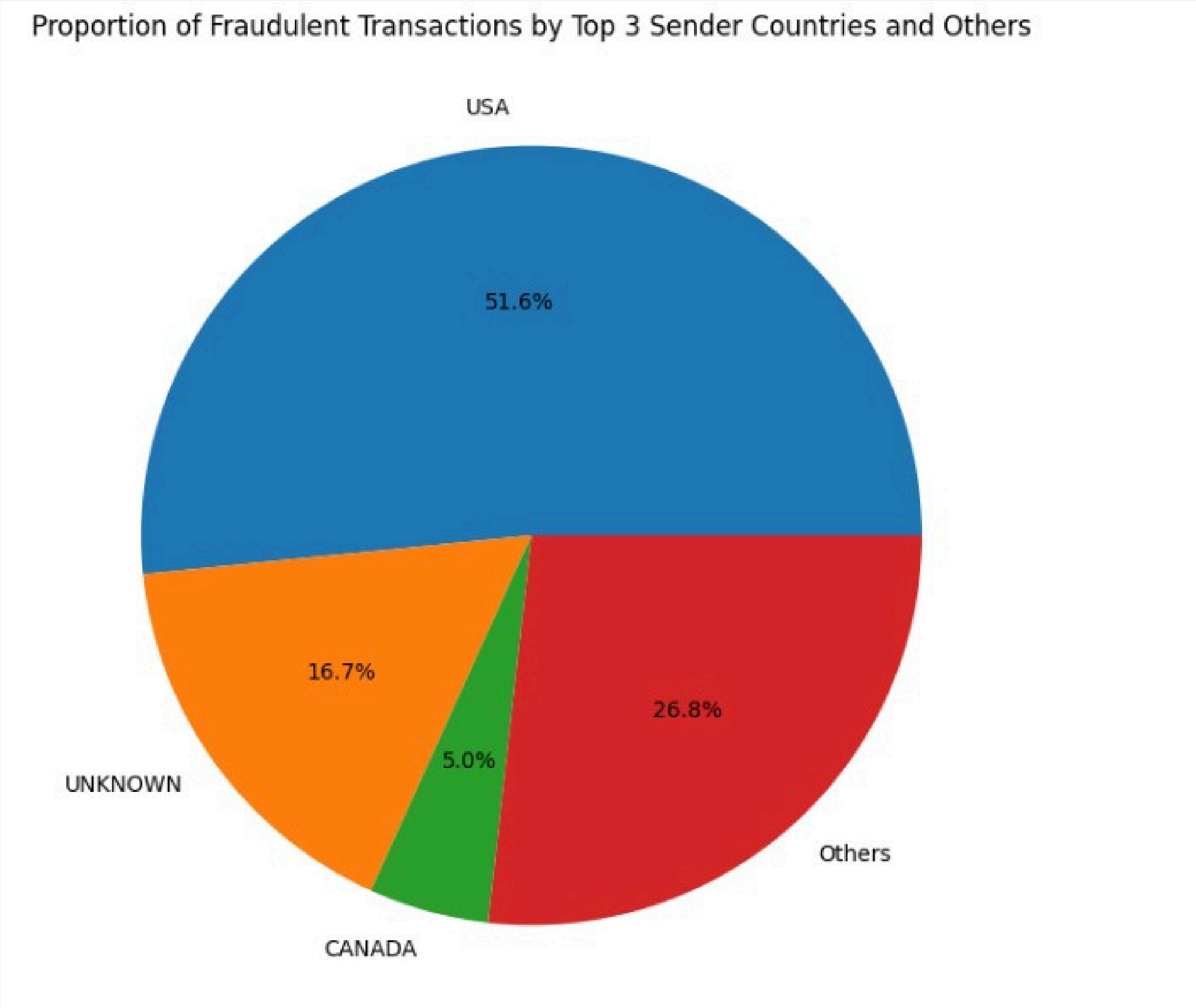
WORK DONE - JP MORGAN

Fraud Transactions

Bene country



Sender country



WORK DONE - JP MORGAN

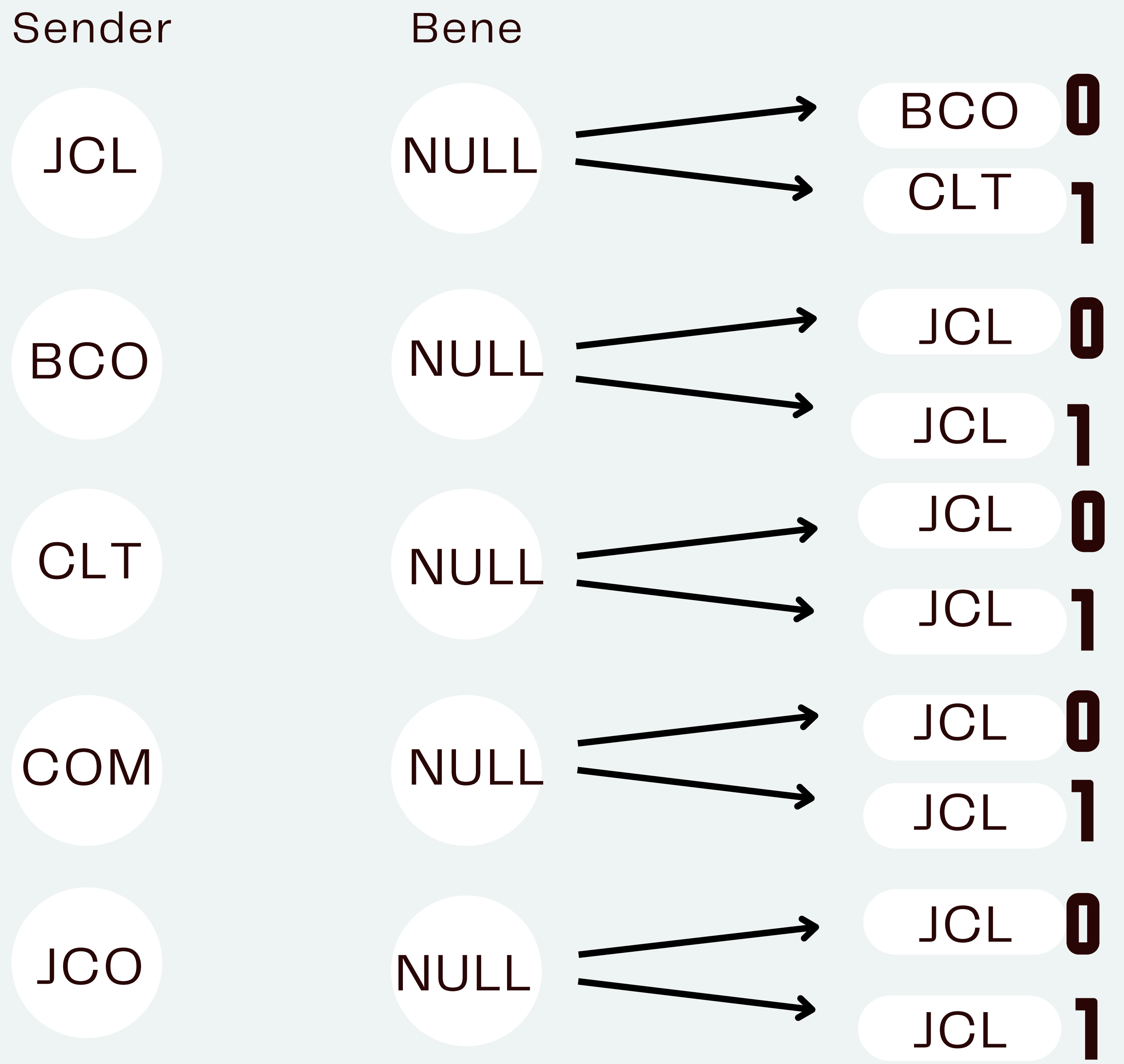
Assigning Values to the Null-values in the Sender_Id and and Bene_Id was a more complex process.It improved the model performance.

There were majorly five parties involved in the transactions

Bill Company(BCO), Company(CO), Client(CLT), JPMC-Client(JCLT), JPMC-Company(JCO)

Data Obtained after feature engineering not considering rows with null values

```
Number of fraud transactions according to type are below:
 send_rec
JCLCLT    8883
JCLBCO    3805
CLTJCL    3676
JCLJCL    1734
JCLCOM    1731
JCLJCO    723
BCOJCL    619
COMJCL    270
JCOJCL    121
Name: count, dtype: int64

Number of valid transactions according to type are below:
 send_rec
JCLBCO    358360
JCLCLT    298159
CLTJCL    106932
JCLJCL    86698
JCLCOM    61753
COMJCL    58934
JCLJCO    26322
JCOJCL    24892
BCOJCL 19786
```

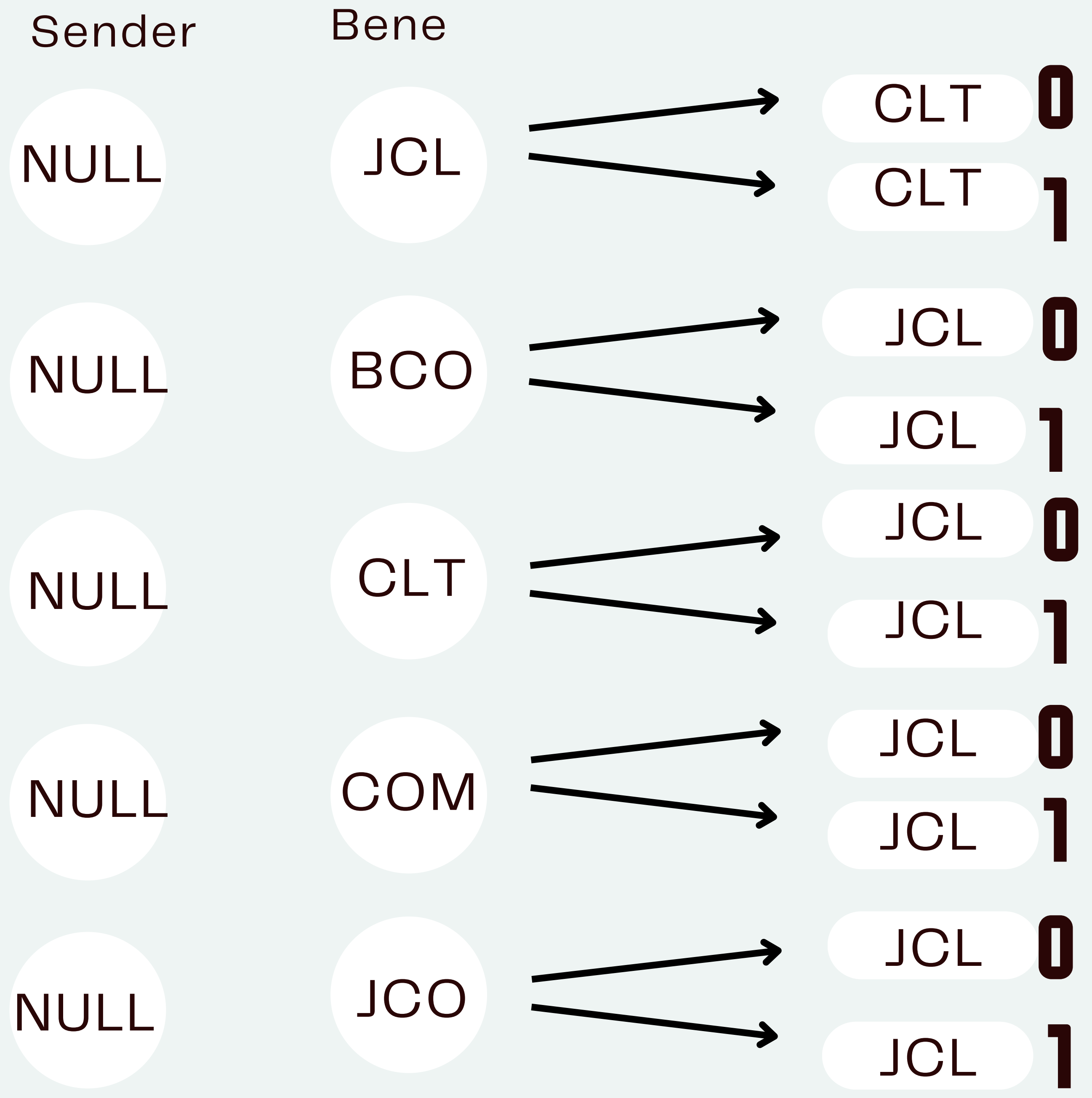


WORK DONE - JP MORGAN

Assigning Values to the Null-values in the Sender_Id and and Bene_Id was a more complex process.It improved the model performance.

Data Obtained not considering rows with null values

```
Number of fraud transactions according to type are below:
 send_rec
JCLCLT    8883
JCLBCO    3805
CLTJCL    3676
JCLJCL    1734
JCLCOM    1731
JCLJCO     723
BCOJCL     619
COMJCL     270
JCOJCL     121
Name: count, dtype: int64

Number of valid transactions according to type are below:
 send_rec
JCLBCO    358360
JCLCLT    298159
CLTJCL    106932
JCLJCL     86698
JCLCOM     61753
COMJCL     58934
JCLJCO     26322
JCOJCL     24892
BCOJCL 19786
```



WORK DONE - JP MORGAN

Feature Engineering

Five parties involved in the transactions traced from “Sender_Id” and “Bene_Id” were -

1. Bill-Company(BCO)
2. Company(CO)
3. Client(CLT)
4. JPMC Client(JCLT)
5. JPMC Company(JCO)

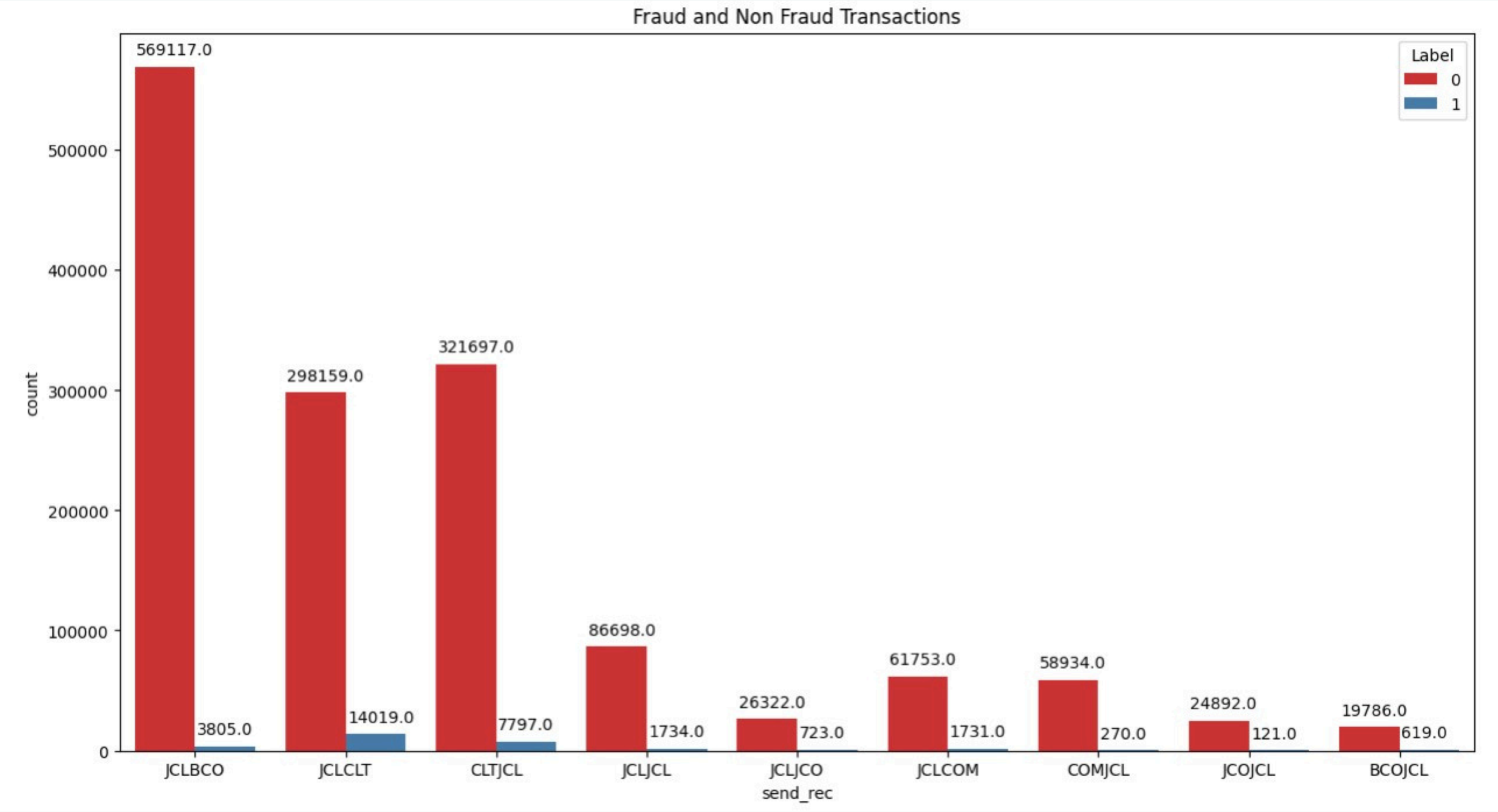
we decided to merge the features and make a single feature “send_rec” which stores the transaction carried between two parties.

Sender_Id	Bene_Id
JPMC-CLIENT-10098	NaN
JPMC-CLIENT-10098	CLIENT-10100
NaN	JPMC-CLIENT-9812
JPMC-CLIENT-9812	JPMC-CLIENT-9814
NaN	JPMC-CLIENT-9789

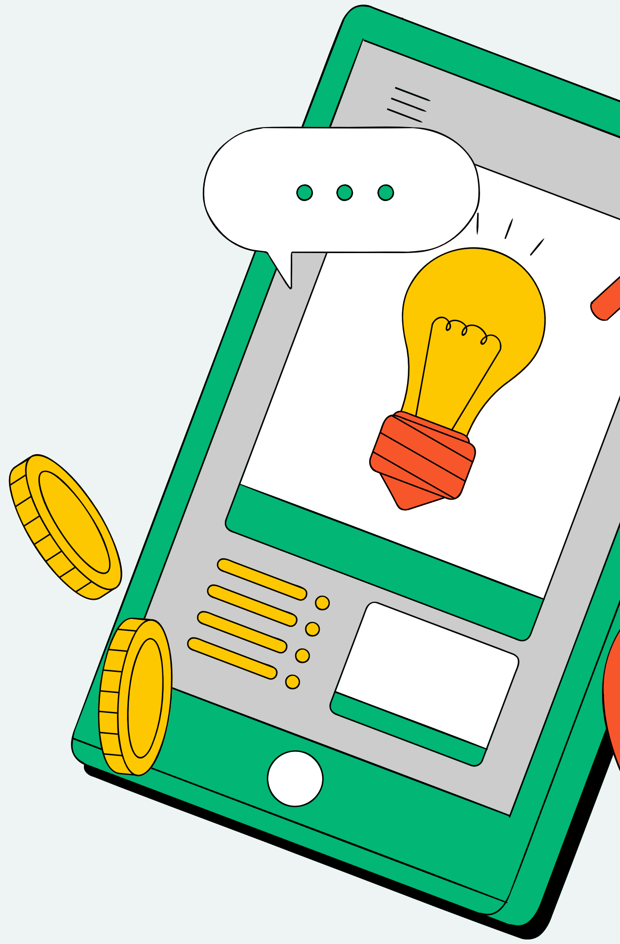
send_rec
JCLBCO
JCLCLT
CLTJCL
JCLJCL
CLTJCL

WORK DONE - JP MORGAN

Fraud transaction analysis



WORK DONE - JP MORGAN

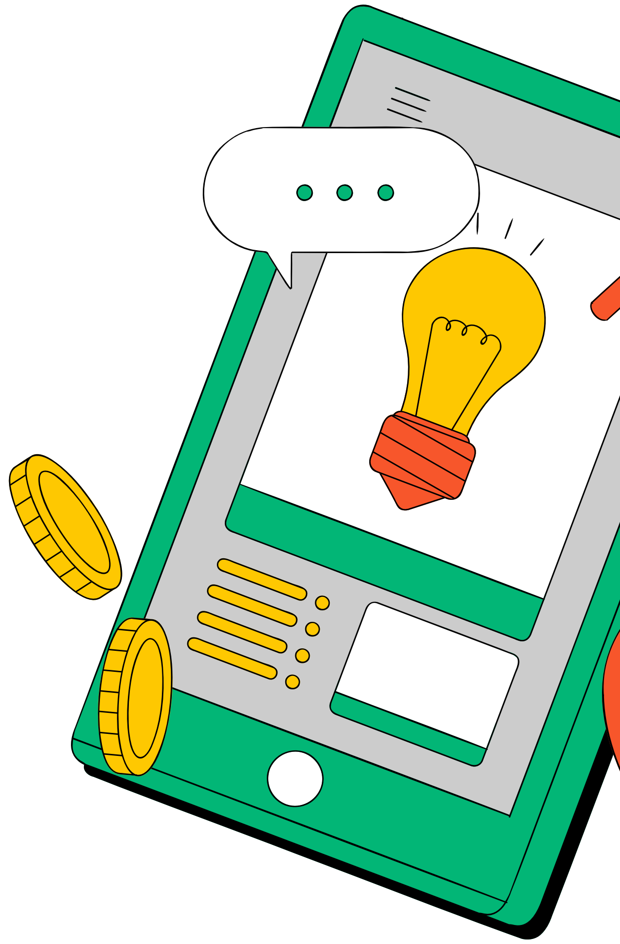


New Feature – “surge indicator” – Creates a new column which has 1 if the transaction amount is greater than the threshold else it will be 0.

Threshold* – 75th percentile

```
surge
0    1497969
1        208
```

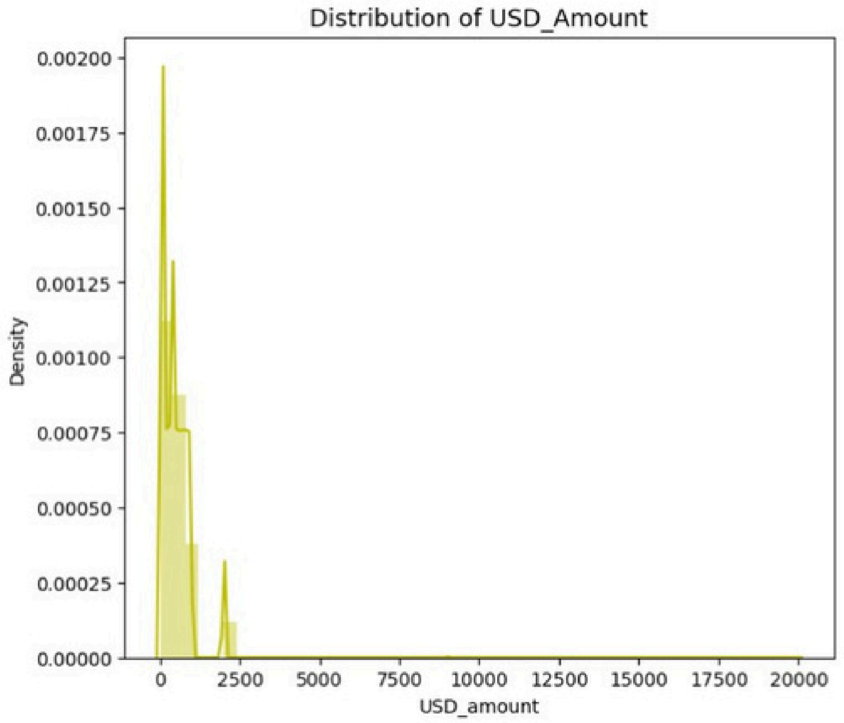
WORK DONE - JP MORGAN

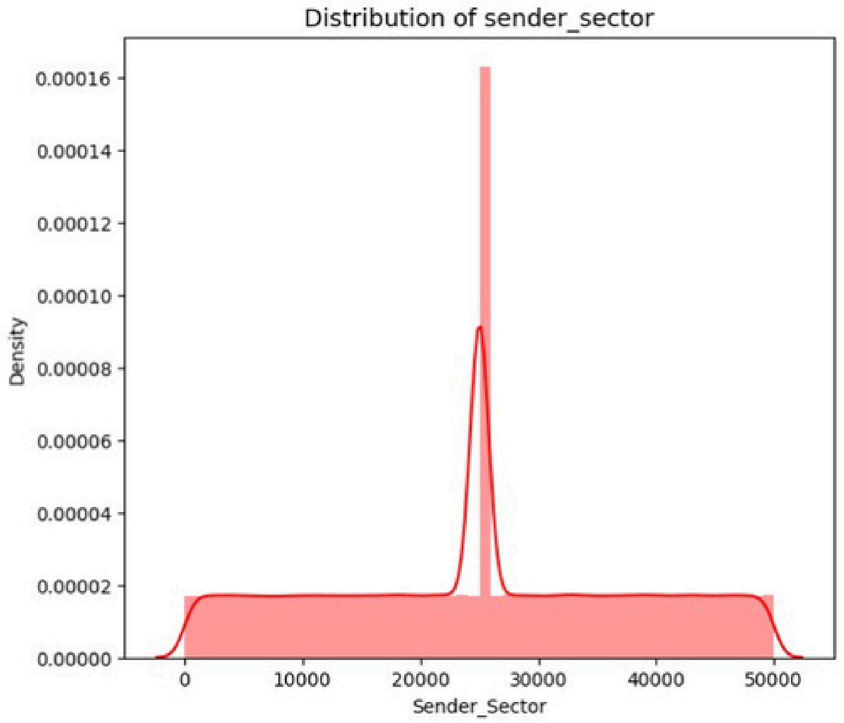


Data Pre-processing

The Data was categorical, so we need to employed label encoding. Followed by Normalization. The Data was separated into train and test by 30% ratio.

send_rec	Sender_Country	Bene_Country	Transaction_Type	Sender_Sector	USD_amount	Label	surge
3	238	238	7	35537.000000	558.43	0	0
4	40	40	6	15287.000000	622.78	0	0
1	240	240	0	25020.183491	802.54	0	0
6	240	240	5	38145.000000	989.09	0	0
1	240	240	1	25020.183491	786.78	0	0





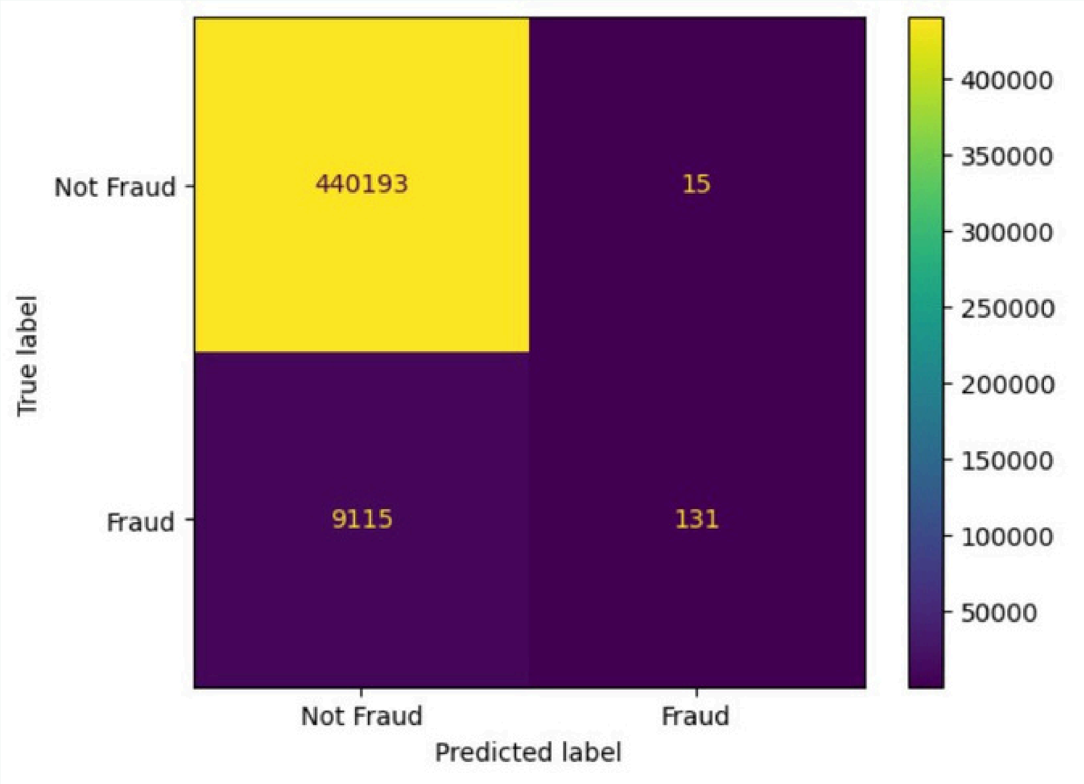
MODEL TRAINING - JP MORGAN

Model Training

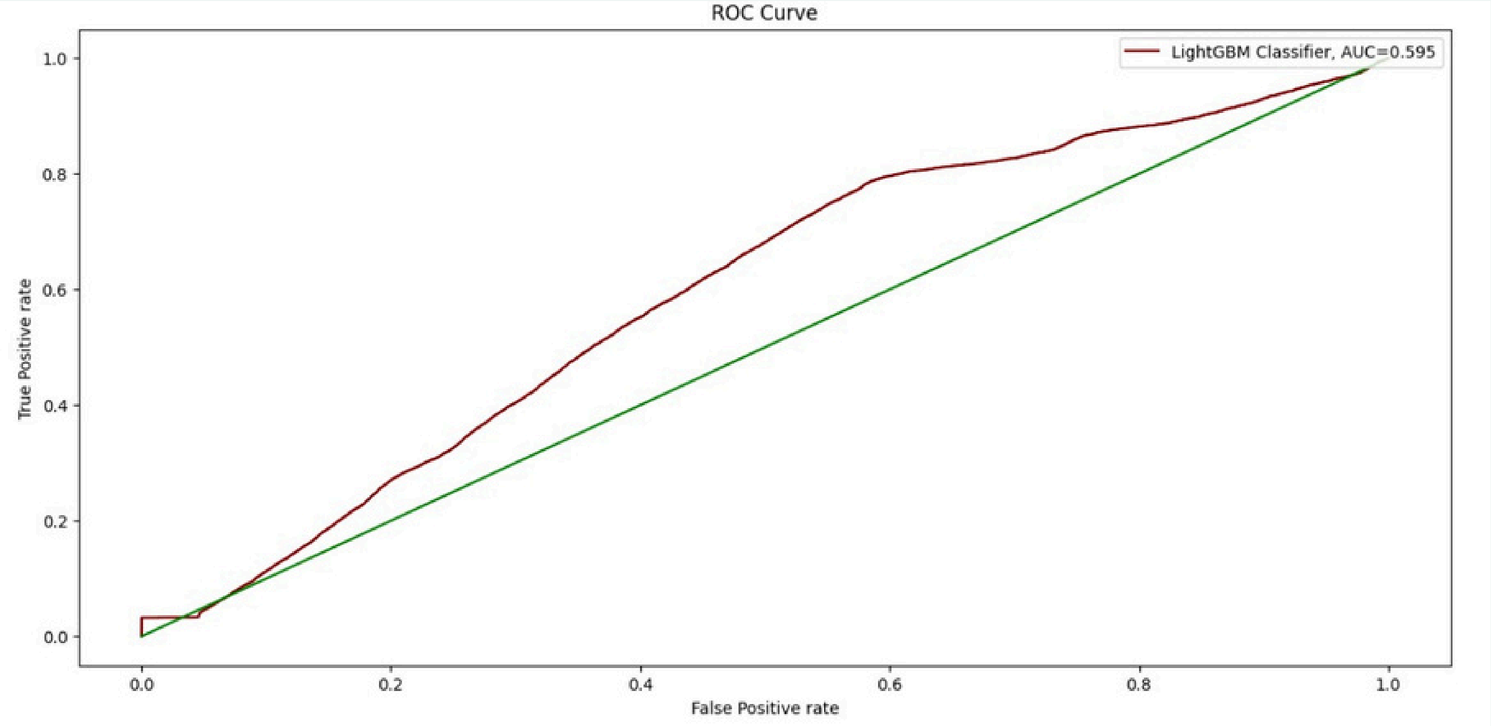
1. Logistic Regression

Why Logistic Regression?
It is a standard technique used in binary-classification problems.we try to draw a separating “S-shaped” curve which separates two classes.

Why Logistic Regression Failed?
The Model was not able to separate two classes based on the given features, which shows the complexity in the transactions data.



	precision	recall	f1-score	support
Not Fraud	0.98	1.00	0.99	440208
Fraud	0.87	0.02	0.03	9246
accuracy			0.98	449454
macro avg	0.93	0.51	0.51	449454
weighted avg	0.98	0.98	0.97	449454

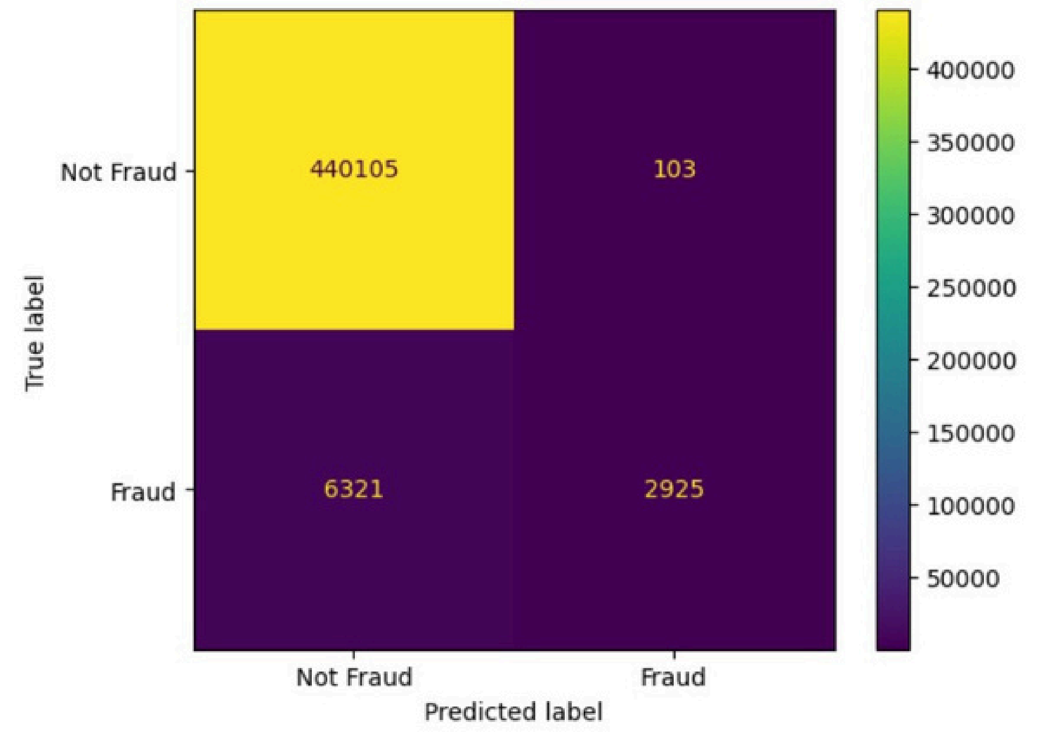


MODEL TRAINING - JP MORGAN

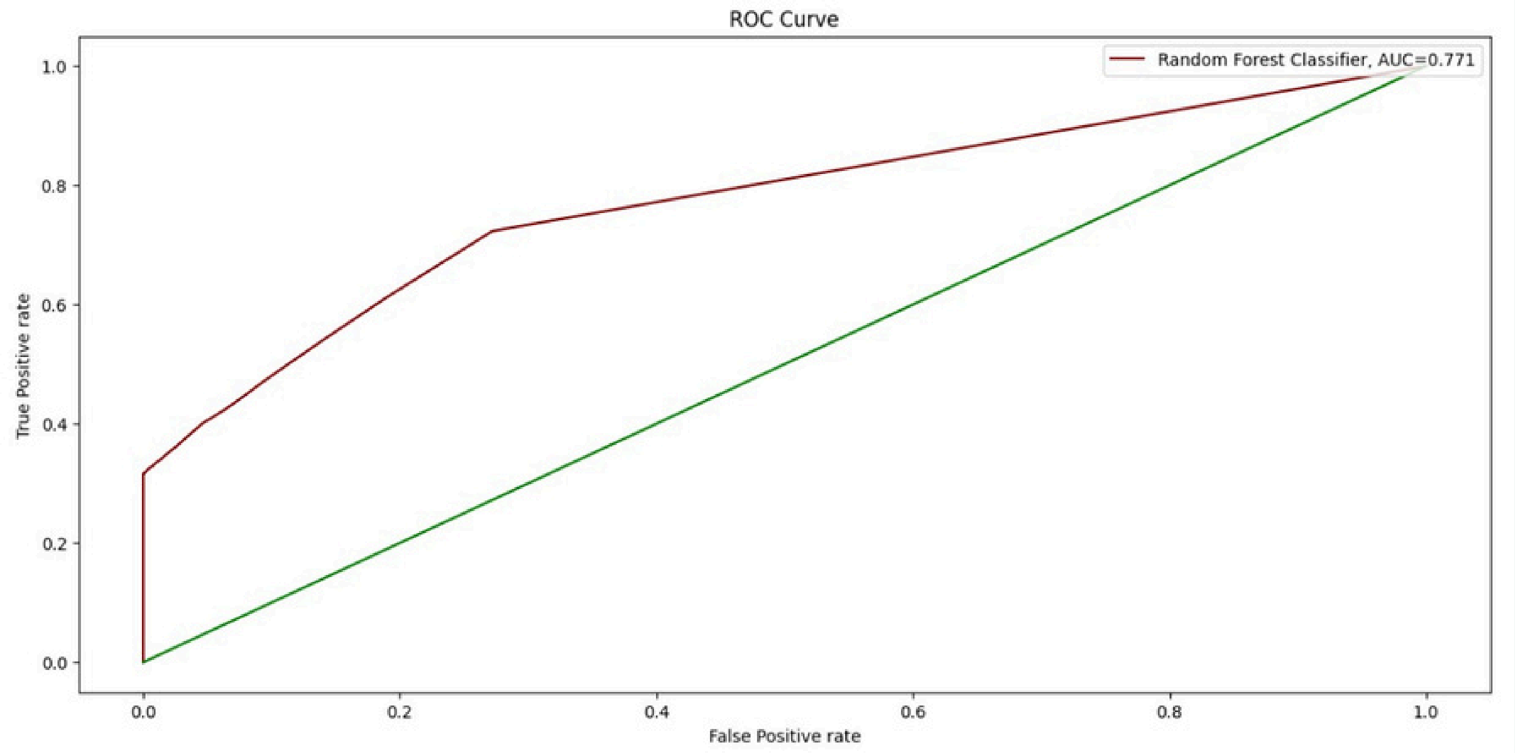
2. Random Forest Classifier

Why Random Forest Classifier?
Growing trees over different features can potentially be a good method for classification task[ref. Bhattacharyya et al.]. The model grew 100 trees to the default depth and the results remarkably improved.

Can random Forest predict better?
The classifier works by building multiple decision trees which are then combined to create a single unified model. Each decision tree takes into account different parameters and is able to make predictions based on these parameters. Combining all the predictions gives more accurate and reliable results.



	precision	recall	f1-score	support
Not Fraud	0.99	1.00	0.99	440208
Fraud	0.97	0.32	0.48	9246
accuracy			0.99	449454
macro avg	0.98	0.66	0.73	449454
weighted avg	0.99	0.99	0.98	449454

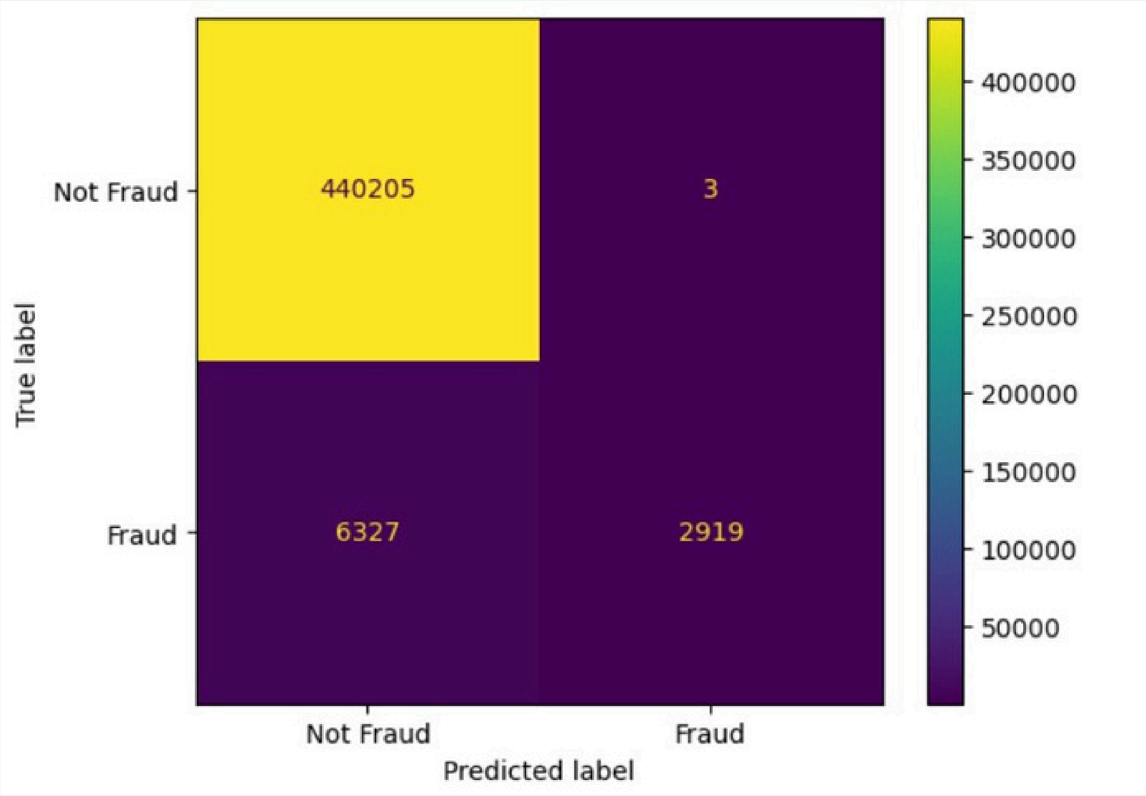


MODEL TRAINING - JP MORGAN

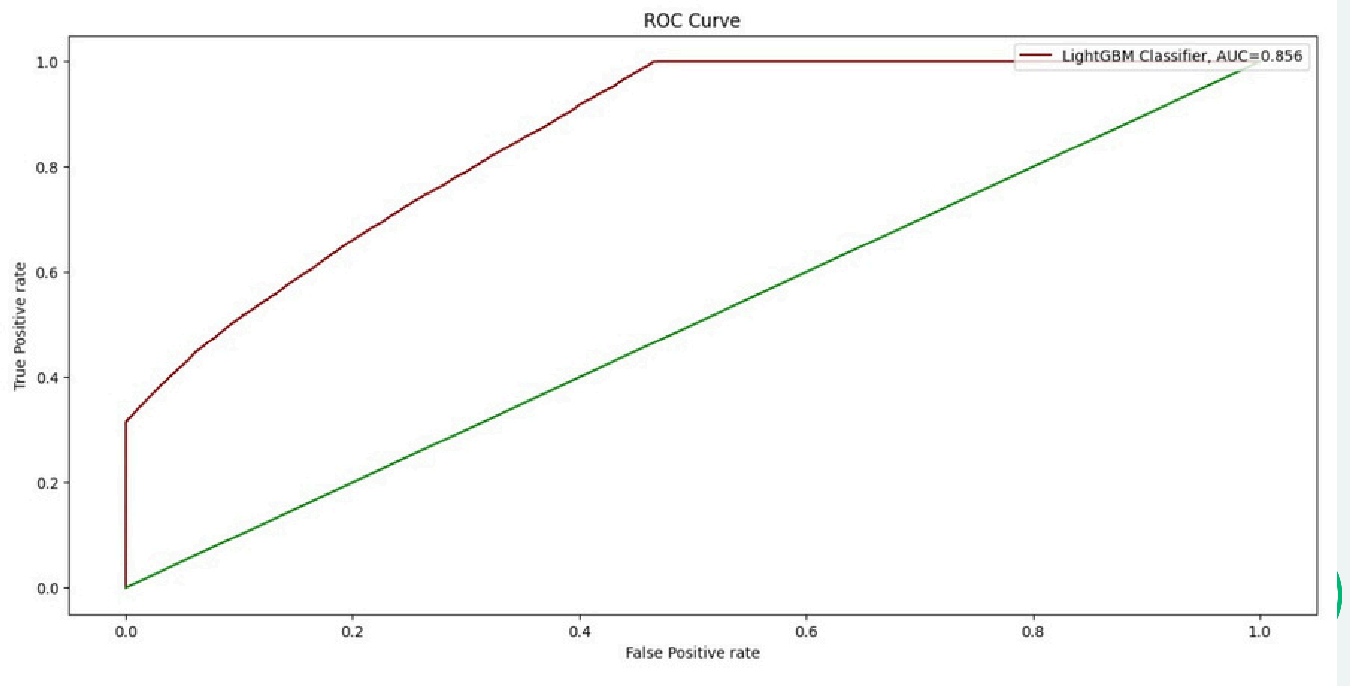
3. Light GBM Classifier

Why Light GBM Classifier?
Light GBM Classifier have three advantages from traditional boosting strategies.
a. Bin-wise splitting
b. Exclusive feature building
c. GOSS(Gradient based one side saampling)
These advantages helps LGBM to ran faster compared to random forest classifier and boosting algorithm.

How Light GBM performs better?
LGBM recall and precision rate is slightly higher than tree based methods. LGBM is faster than tree based approaches and is not computational expensive.



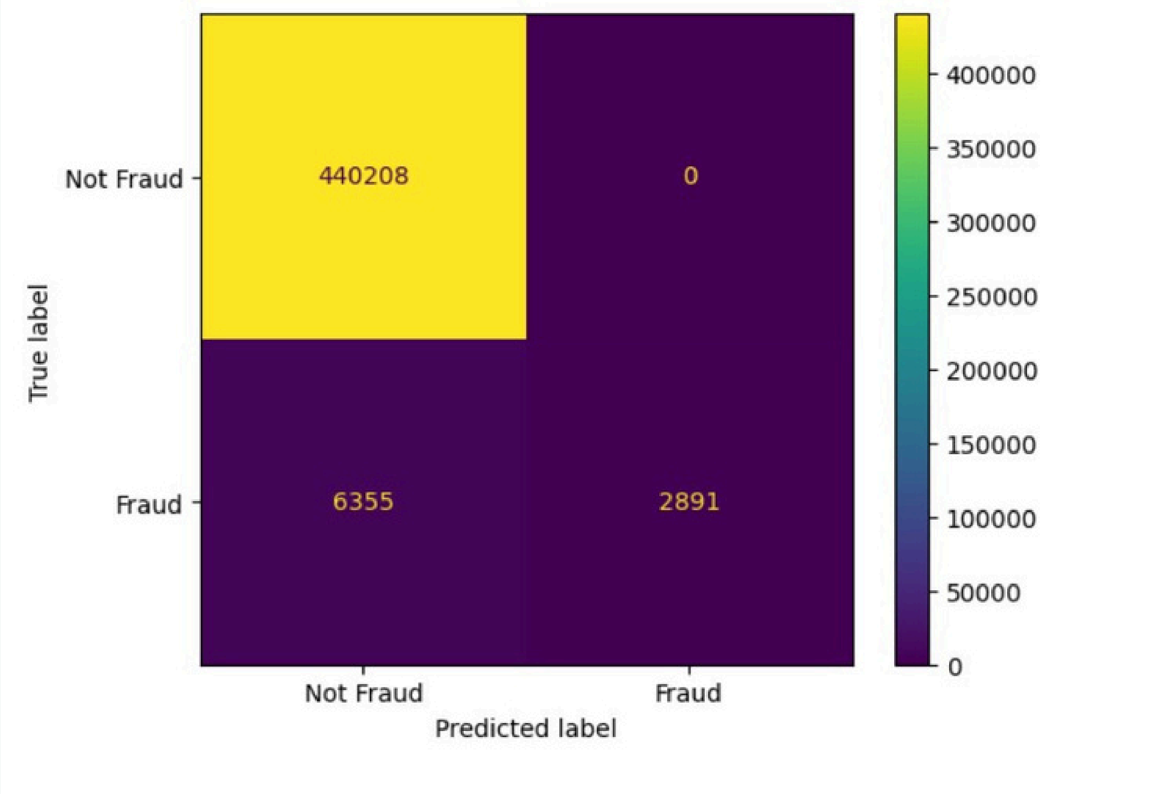

```
              precision    recall  f1-score   support

   Not Fraud       0.99      1.00      0.99    440208
       Fraud       1.00      0.32      0.48      9246

    accuracy                           0.99    449454
   macro avg       0.99      0.66      0.74    449454
weighted avg       0.99      0.99      0.98    449454

col_0       0     1
Label
0      440208     0
1        6287  2959
```



MODEL TRAINING - JP MORGAN

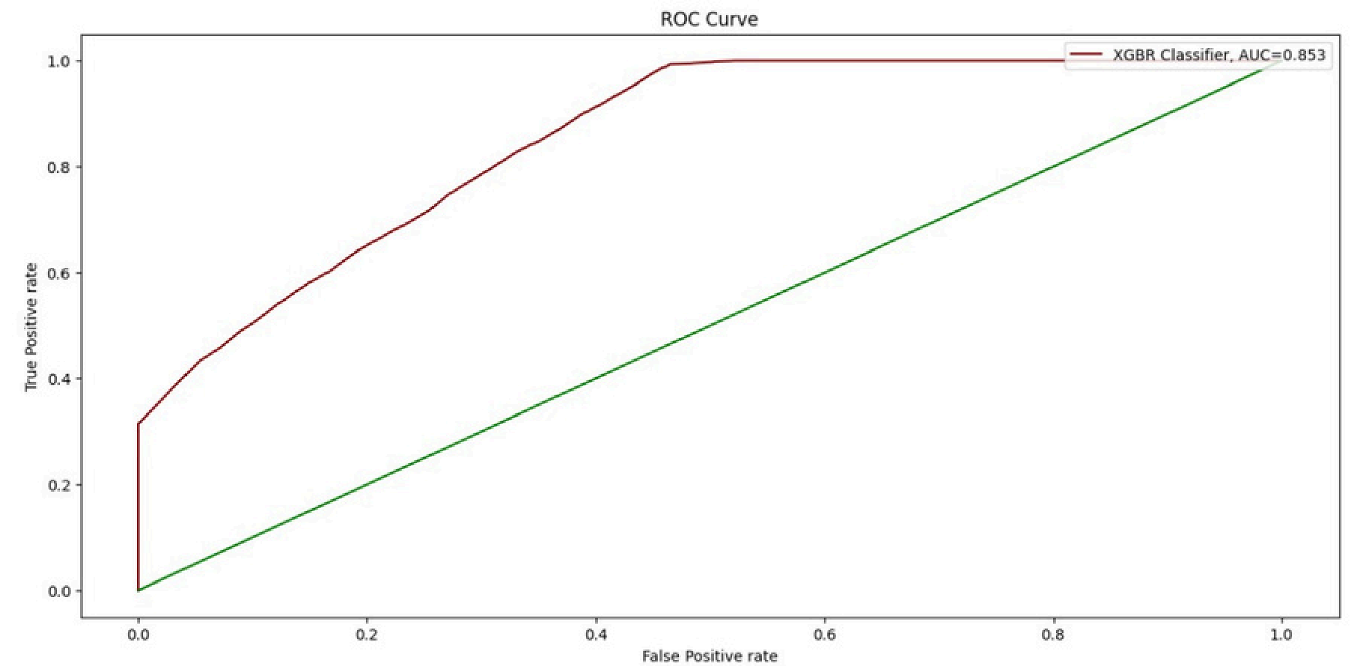
4. XG Boost Classifier

Why XG Boost Classifier?
It works by sequentially adding simple models to correct the errors made by previous models.XGBoost has an in-built routine to handle missing data.



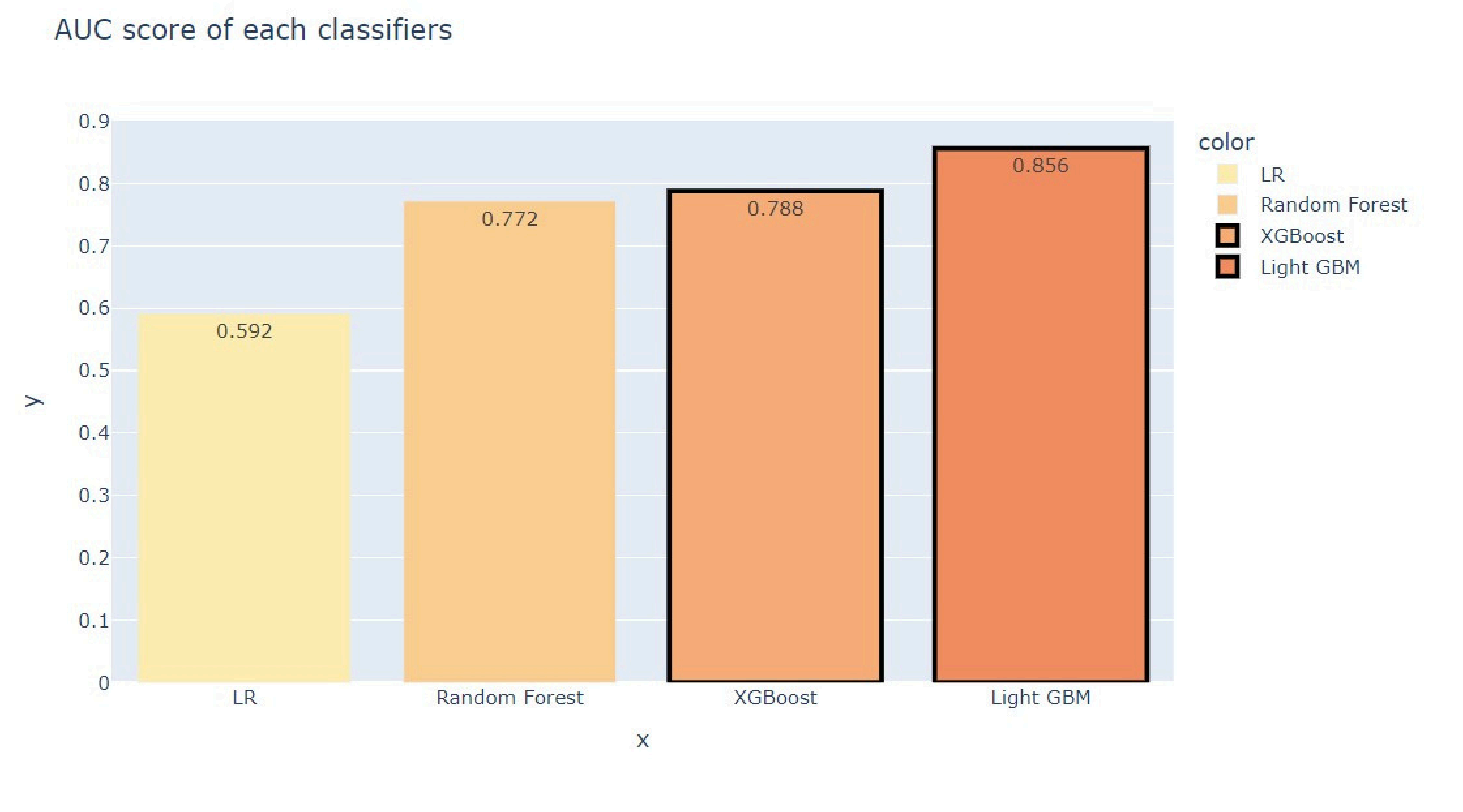

```
              precision    recall  f1-score   support

   Not Fraud       0.98      1.00      0.99    440208
       Fraud       0.98      0.19      0.31      9246

    accuracy                           0.98    449454
   macro avg       0.98      0.59      0.65    449454
weighted avg       0.98      0.98      0.98    449454

col_0       0     1
Label
0      440176    32
1        7532  1714
```

How XGBoost performs better?
XG-Boost performs equivalent to LightGBM and Random Forest Classifier in precision but it's recall/sensitivity is less compared to LightGBM and Random forest classifier.If we exclude "surge" feature the recall increases.



MODEL TRAINING - JP MORGAN

Model Training Comparisons

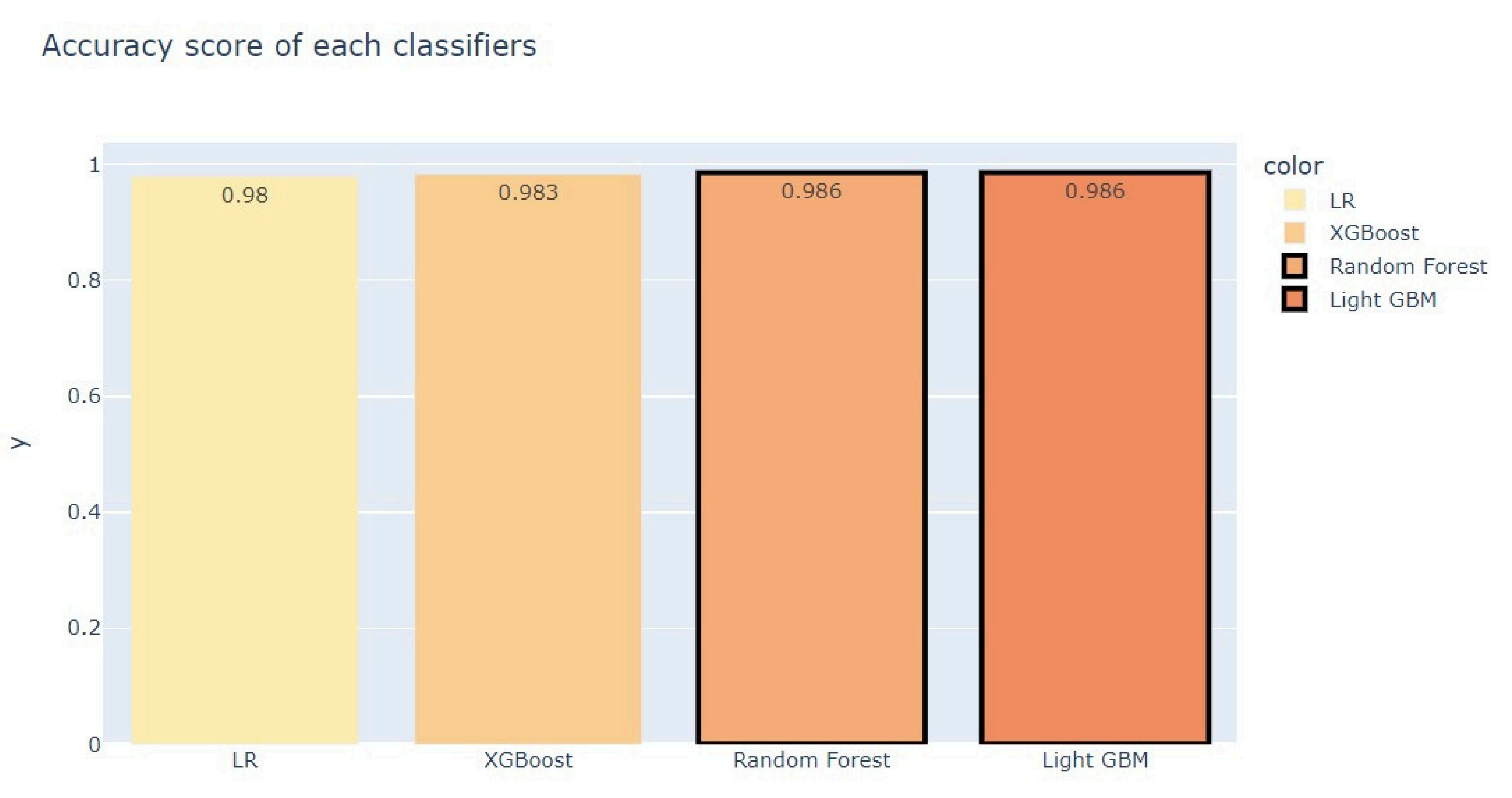


Light GBM AND XGBoost gave the best AUC SCORE of 85.6 and 78.6 among all the classifiers.
Following Machine Learning community standards -
An AUC-Score between 80-90 is considered very good which has been successfully achieved by the Light-GBM model.

MODEL TRAINING - JP MORGAN

Model Training

Comparisons



“Accuracy” of all the classifiers was almost at the saturation of 98-99%.
Why?
Because there are very large no of transactions that are not fraud, implies “Accuracy” is not a good metric for this classification problem.Thus, Other Metrics like AUC-ROC score, Precision, Recall/sensitivity, F1-score are more reliable metrics.

MODEL TRAINING - JP MORGAN

Conclusions

“Metrics like AUC-ROC score, Precision, Recall/sensitivity, F1-score are more reliable metrics than accuracy” for our classification problem.

LightGBM was the most suitable model for the given classification problem.

The Dataset actually mimics the complexity of the real-time data. After extensive feature engineering, the Model has reached this performance. Future works can be done on feature engineering by analysing customer behavior pattern using time parameter efficiently as a periodic variable.

Problems Faced-

1. The transactions that are fraud does not show any remarkable distinctive feature when compared to non-fraud transaction. so, it required us to do feature engineering to make more distinctive features for the model to train on. Without feature engineering the F1-score was 10%. After feature engineering it boosted upto 28%
2. The Dataset contained a significant amount of NULL-Values. The Null-value were imputed using different strategies for each column. After imputation, we saw a significant improvement in the model performance F1-score got boosted from 28% to 48-49%.

THANK YOU!