Meteorite Group Report

z

Inter-Uni Datathon 2024

Member: Tong Zhu (Judy), Kevin Shen, Kevin Xu, Shubhanker Rawat

z

Introduction & Aim

- Skyline Financial Services (SFS) faced an unexpected dip in monthly revenue.
- Internal investigations revealed fraud slipping past current detection systems
- Our task is to analyze transaction data and detect potential fraudulent activities

Data Understanding & Exploration

Data Overview:

train.csv: contains transactions with the IsFraud label.

test.csv: contains similar transactions but without IsFraud label for predictions.

sample_submission.csv: provides the format for submitting fraud predictions

z

Preprocessing Method

<u>Methods</u>

For preprocessing, we utilized multiple different data preprocessing methods, which includes:

- Text Processing(extracting domains from email and removing $ symbols)
- Currency Exchange(from AED or GBP to AUD)
- Data Clipping(transform outliers into -1 for age)
- Grouping(gender, transaction location and device type)
- Label encoding(occupation, education level, marital status …)
- …

Preprocessing Method Data Example

Clipping age outliers

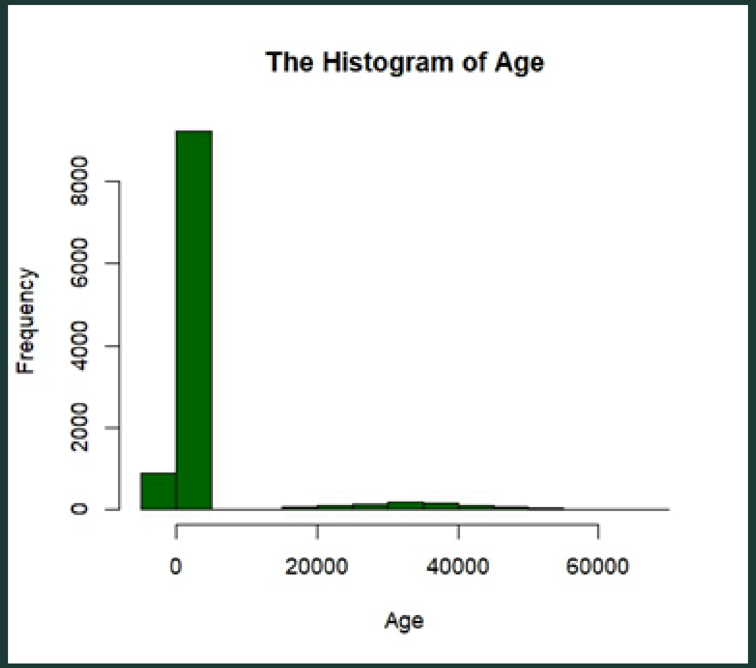
 - Improve accuracy(consider ages above 100 or below 0 as misinput)

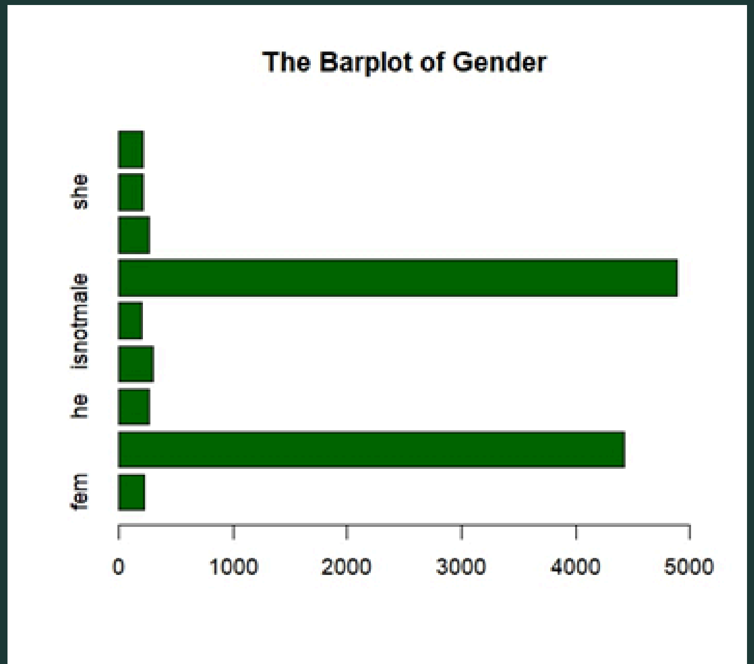
Transforming different monetary units into AUD(income, expenditure, gift)

-1 AED = 0.4 AUD

-1 GBP = 2.0 AUD

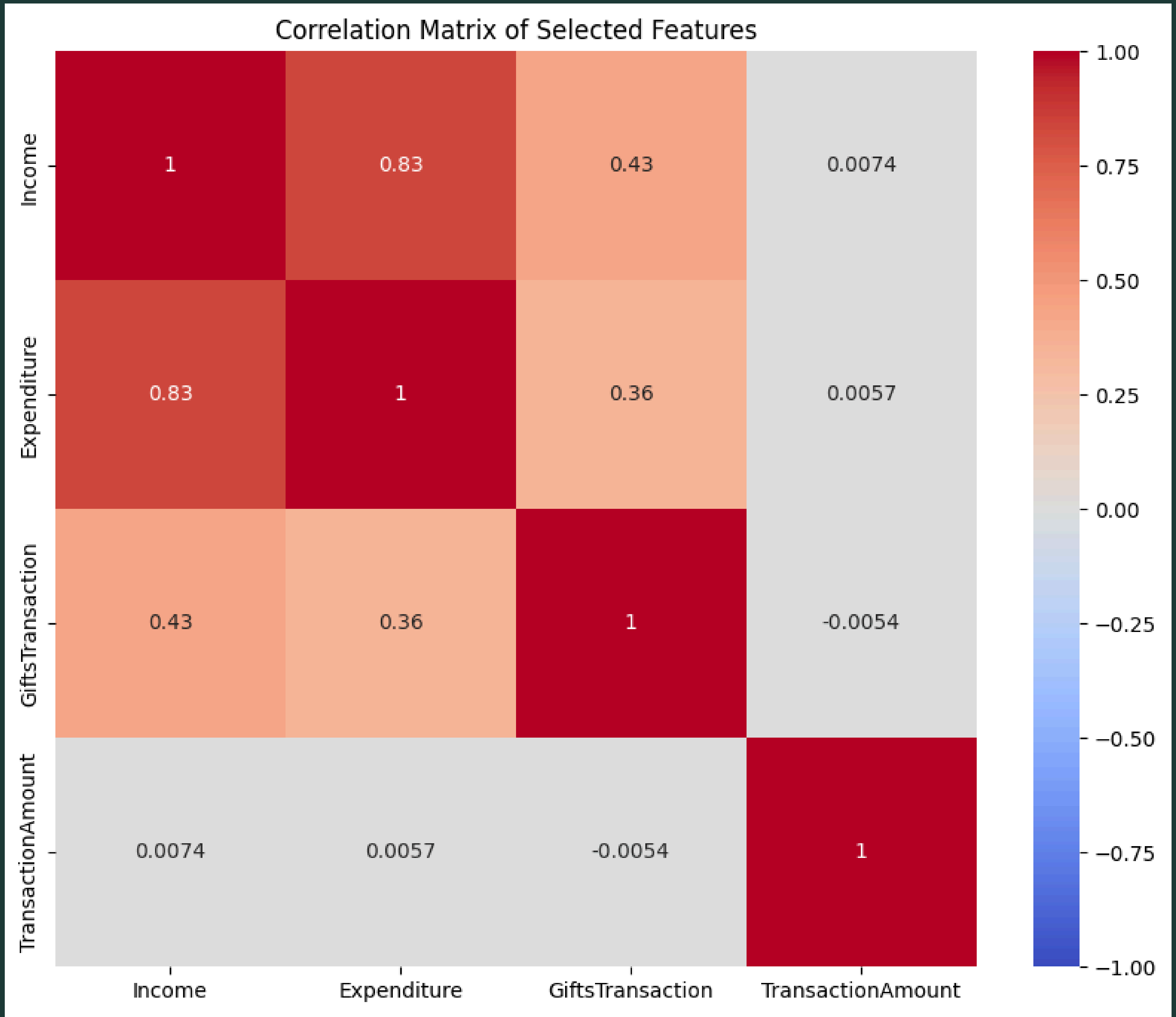
•Grouping iphone 15 as Mobile for device type





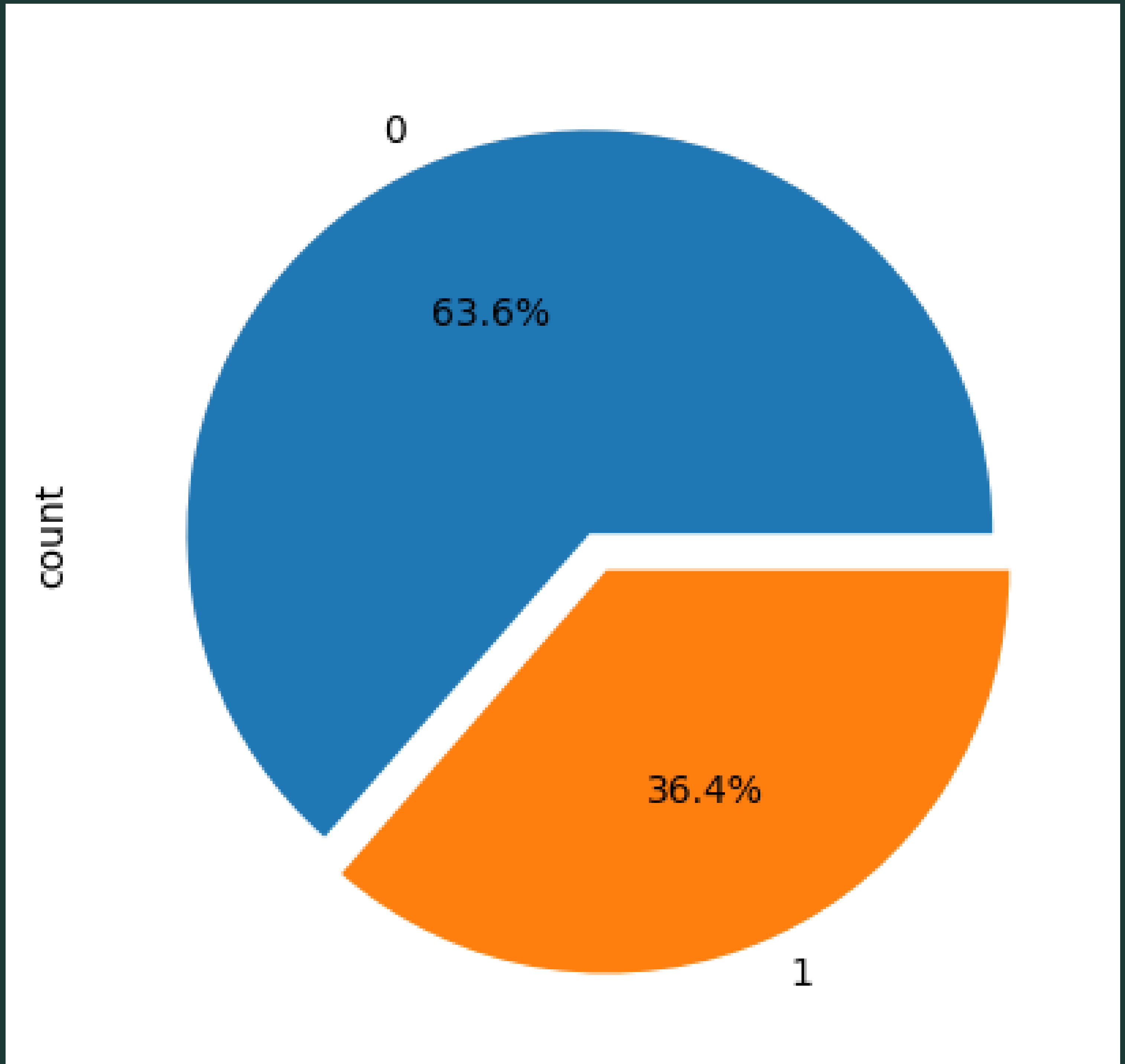
Data Analysis - Correlations

The correlations among all numeric variables are relatively low.



Data Analysis - Label Distribution

There are 36.4% Fraud and 63.6% not Fraud in the train datatset.



Data Analysis - Label Encoding

Before Label Encoding ->

Gender	Occupation	EducationLevel	MaritalStatus	NumDependents	Income	Expenditure	...	MerchantID	TransactionType	TransactionLocation	DeviceType
Female	Professional	Bachelor	Widowed	3	28884.43 AUD	14610.61 AUD	...	M006	Withdrawal	Adelaide	Mobile
Male	Student	High School	Married	4	AU$ 54919.07	39169.49 AUD	...	M002	Withdrawal	Canberra	Mobile
Male	Unemployed	Master	Married	2	AU$ 74728.57	55873.76 AUD	...	M008	Purchase	Brisbane	Mobile
Male	Professional	High School	Married	3	AU$ 55712.62	AED 89649.04	...	M001	Purchase	Darwin	iphone 15
Male	Professional	High School	Single	4	53004.7 AUD	AED 43601.02	...	M001	Withdrawal	MLB	Tablet
...	...	...	...	...	...	...	...	...	...	...	...
Male	Unemployed	High School	Single	3	64488.68 AUD	AU$ 21813.53	...	M007	Purchase	Canberra	Mobile
Female	Professional	High School	Married	2	80403.31 AUD	AU$ 63429.08	...	M003	Purchase	Hobart	iphone 15

After Label Encoding ->

Gender	Occupation	EducationLevel	MaritalStatus	NumDependents	Income	Expenditure	GiftsTransaction	TransactionAmount	TransactionType	TransactionLocation
0	2	1	3	3	28884.43	14610.610	2100.02	258.140	3	0
1	0	0	1	4	54919.07	39169.490	9939.42	34.940	3	2
1	1	2	1	2	74728.57	55873.760	2299.70	323.820	1	1
1	2	0	1	3	55712.62	35859.616	4335.70	12.996	1	3
1	2	0	0	4	53004.70	17440.408	4763.48	456.300	3	5
...	...	...	...	...	...	...	...	...	...	...
1	1	0	0	3	64488.68	21813.530	5489.06	182.510	1	2
0	2	0	1	2	80403.31	63429.080	382.42	137.500	1	4

We use label encoding method to transfer categorical to numerical type.

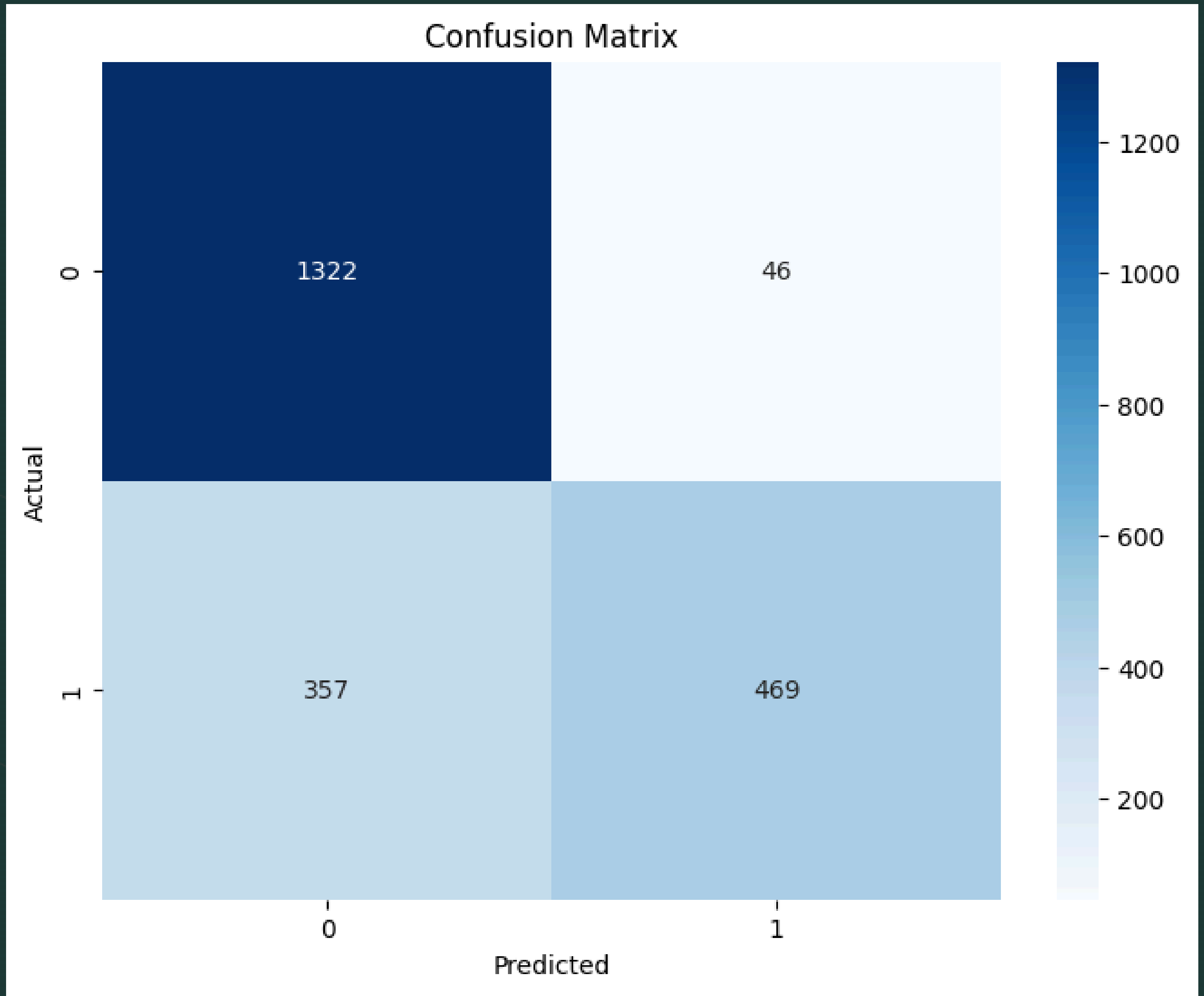
Model Evaluation & Testing

For Machine Learning, we utilized multiple different machine learning methods, which includes:

- **Random Forest Classifier**
- **Linear Regression**
- **Logistic Regression**
- **XGBoost**

Random Forest Classifier

- **Classification**
  - Dividing dataset into a single train set (80%) and a test set (20%)
- **Accuracy :** Around 82%.
- **AUC-ROC**: Around 87%.
- **Confusion Matrix**



XGBoost

- **Classification**
- Dividing dataset into a single train set (70%) and a test set (30%)
- Accuracy : Around 85%.
- AUC-ROC: Around 83%.

AUC-ROC Score: 0.8346641490258302

	precision	recall	f1-score	support
0	0.79	0.94	0.86	2091
1	0.85	0.58	0.69	1199
accuracy			0.81	3290
macro avg	0.82	0.76	0.77	3290
weighted avg	0.81	0.81	0.80	3290

z

Linear Regression & Logistic Regression

- Linear Regression
  - Low R squared score

```
Mean Squared Error: 0.1863773827213508
R^2 Score: 0.20603691515753342
```

- Logistic Regression
  - Low accuracy

Classification Report:

	precision	recall	f1-score	support
0	0.63	0.97	0.77	1368
1	0.57	0.07	0.12	826
accuracy			0.63	2194
macro avg	0.60	0.52	0.44	2194
weighted avg	0.61	0.63	0.52	2194

Regression is not suitable.

Results & Analysis

- Although the model's accuracy is high, there is room for improvement in handling fraudulent transactions. The confusion matrix reveals a noticeable number of false negatives (357 fraud cases classified as non-fraud).
- The AUC-ROC score shows strong overall classification capability, but further optimization is needed, especially for fraud detection, where recall and false negatives are critical.

Thank You!