PROJECT

FRAUD DETETION

MARCELA CAETANO

CONTENT

PROJECT GOAL

The goal of the project is building a model using Machine learning that detects financial fraud by identifying suspicious activities.

The objective is analyze fraud patterns

CHALLENGES VS SOLUNTIONS

Multicollinearity between features

Solution: Addressed multicollinearity by combining features and eliminating unnecessary ones.

Feature Selection

Solution: Kept features like ZIP, latitude, and longitude in the model,enhancing the feature engineering process for better fraud detection.

DATASET OVERVIEW

01 Shape: The dataset used is a syntecthic data.

02 File Format: The dataset is provided in CSV format.

03 Dataset Source: The dataset is stored oon kaggle.

METHODOLOGY

1.ETL

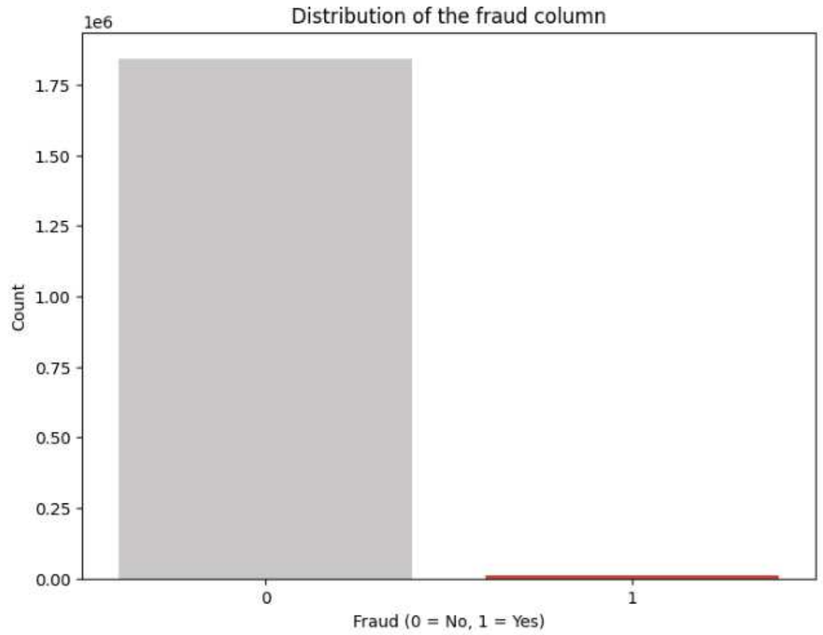
Data Extraction & Cleaning

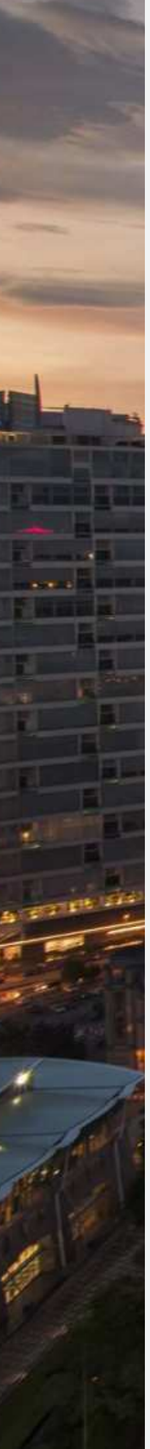
- Data Extraction of the Fraud Datset.
- Handled missing values and duplicates.
- Converted Dates into datatime format

EDA

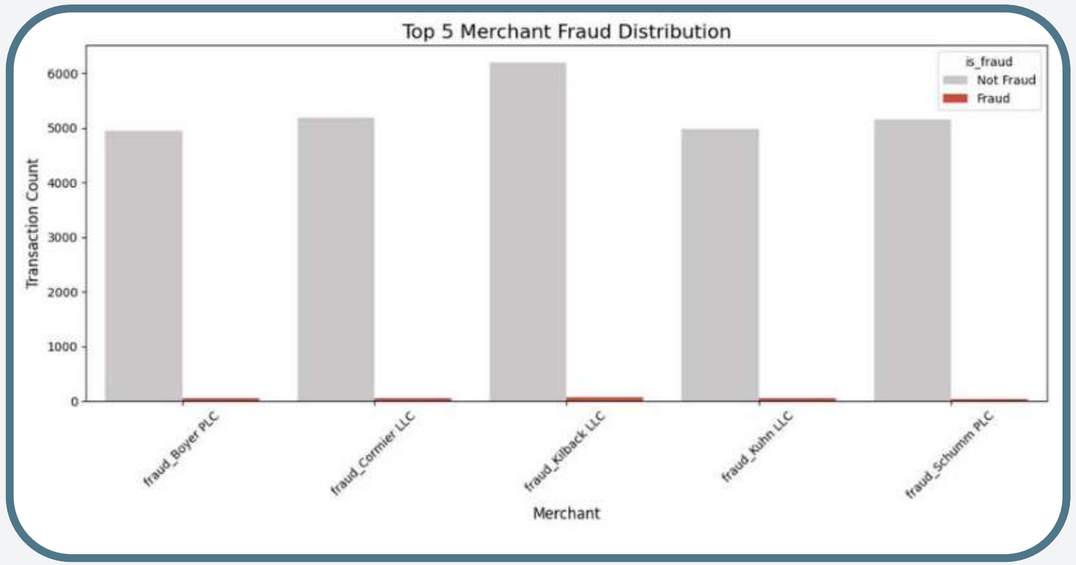
Descriptive Statitics and Distributions

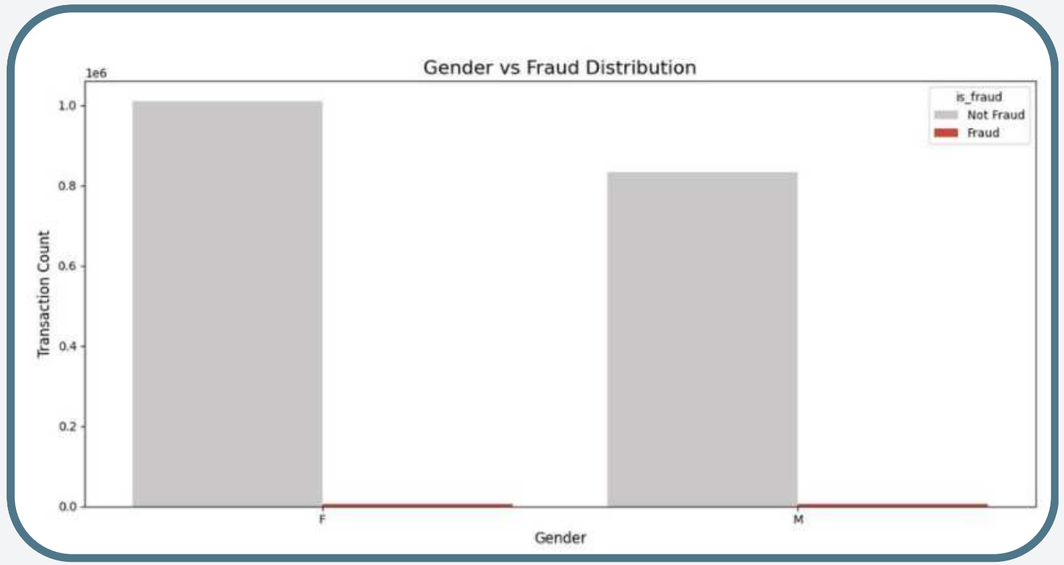
- Data Exploration of the data (Checking the numerical columns distributions))
- Descriptive statitics (statistical summary for numerical columns, including the interquartile range for outlier detection.
- Data Visualizations of categorical columns and the target.





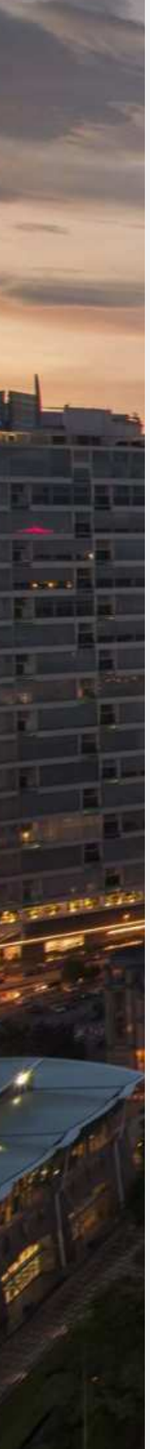
DATA VISUALIZATION





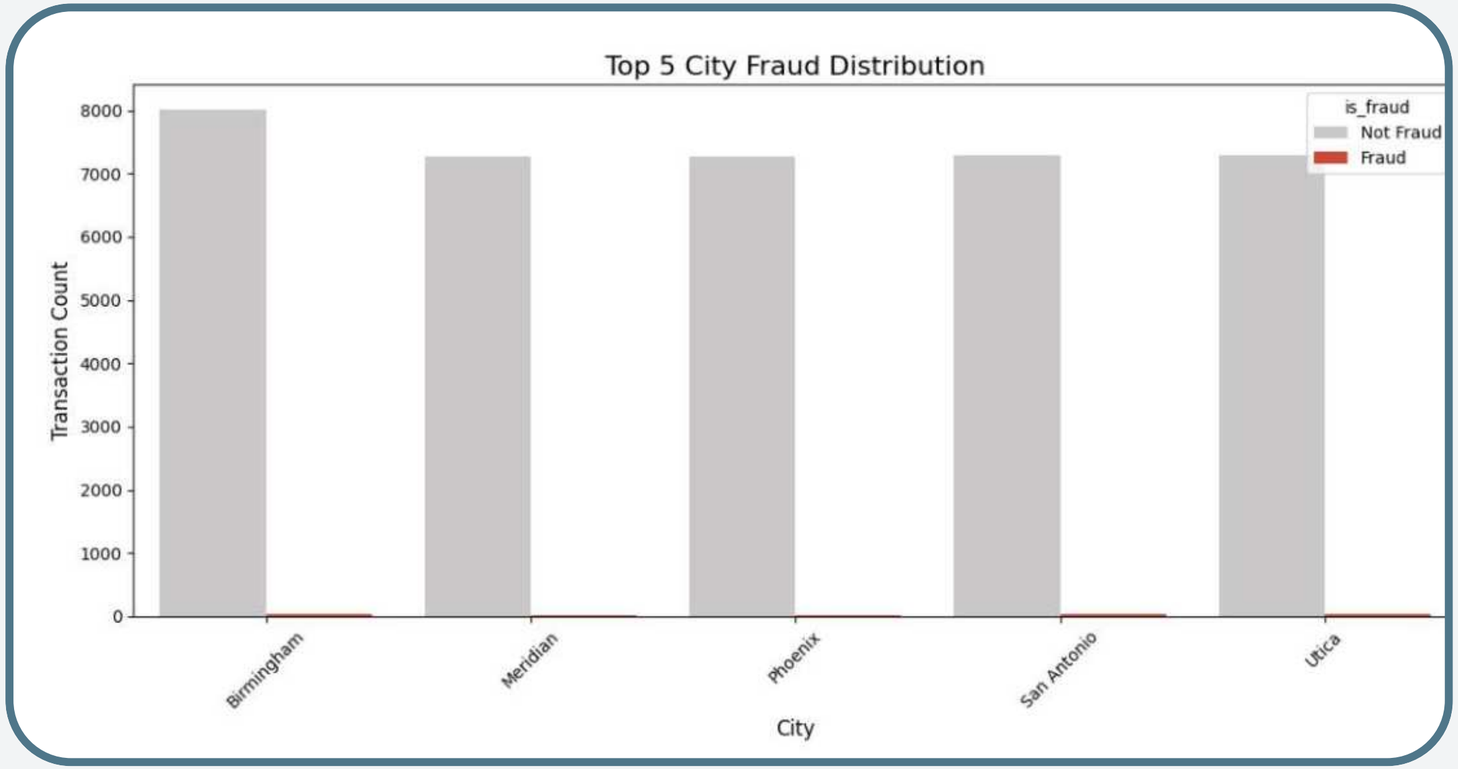
- The first feature we analyzed was the merchant name in relation to the fraud distribution (target variable).
- Kilback LLC, stands out with higher proportion of number transactions compared to others.
- The remaining four merchants appear to have a relatively similar distribution of transactions.

- The second feature we analyzed was gender in relation to the target variable.
- The results showed that the dataset contains a higher proportion of female customers compared to male customers, highlighting gender imbalance.



DATA VISUALIZATION

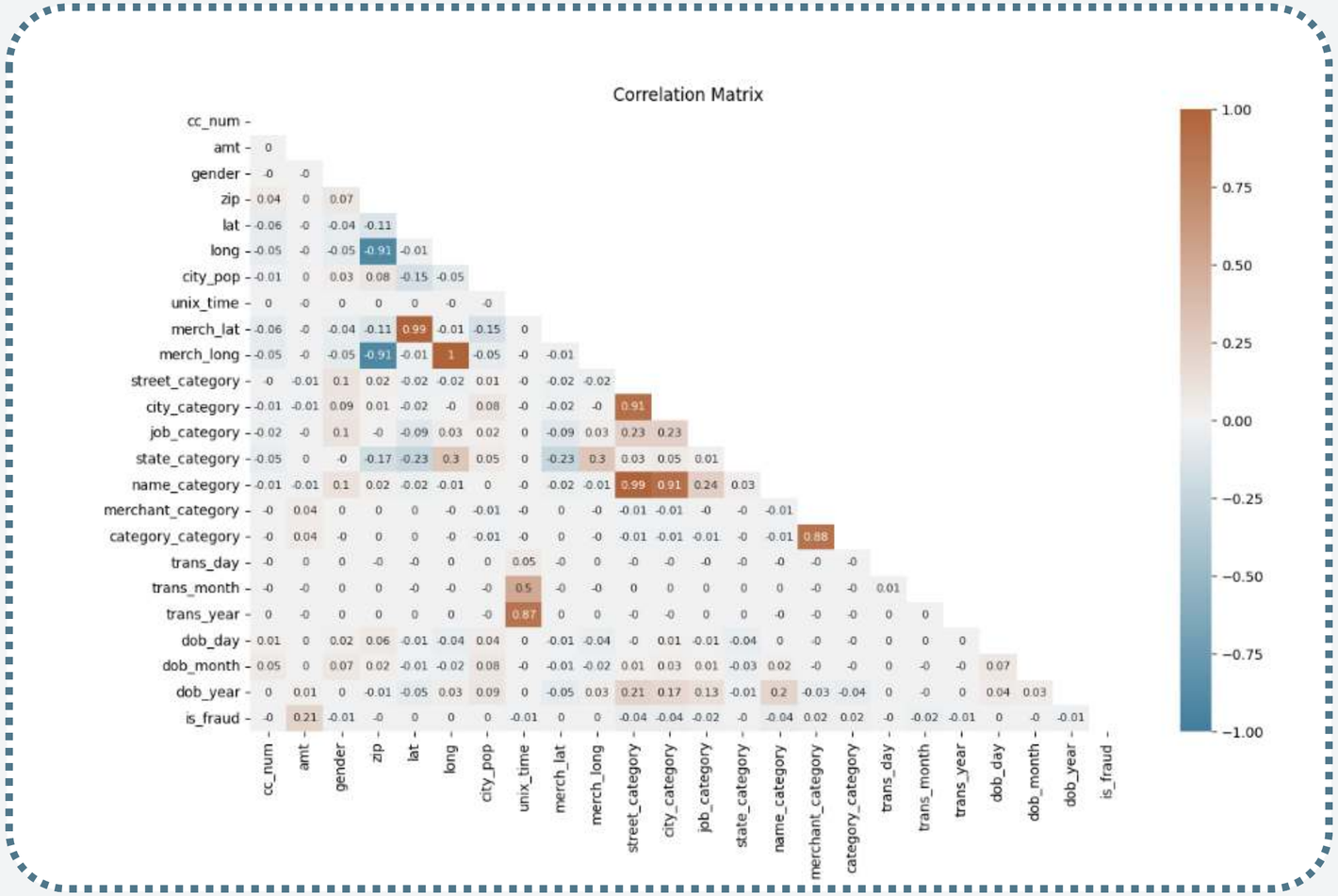
- The third feature we analyzed was city in relation to the target variable.
- Birmingham stands out as having the highest number of non-fraulent transactions.
- The remaining cities have similar transaction volumes, showing less variation in comparison

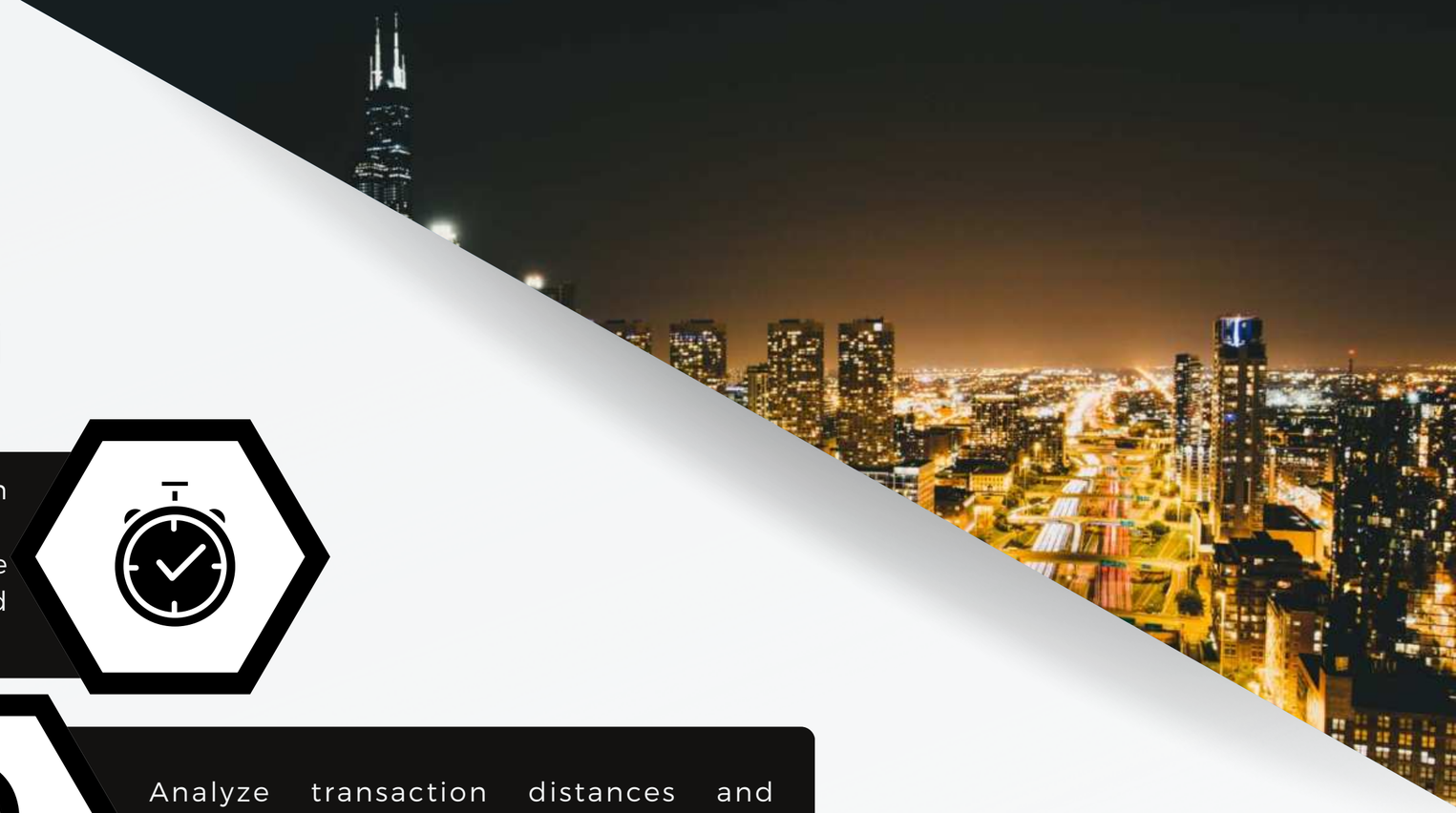


DATA VISUALIZATION

CORRELATION MATRIX

Multicollinearity was detected, with some features showing strong interdependence, while most features exhibit low correlation with the target.





FEATURE ENGINEERING

- Performed one hot enconding on categorical variables.
- Created 8 new features (from the features with lower correlation and with multicolinearity)

Analyze transaction distances and average amounts per card to spot unusual activity.

- Highlight fraud trends by identifying transactions in small versus large cities.
- Track time gaps between transactions and use category details to uncover suspicious patterns.

MODELS USED

Random Forest Classification Report:

	precision	recall	f1-score	support
0.0	1.00	0.99	1.00	552870
1.0	0.38	0.79	0.51	2849
accuracy			0.99	555719
macro avg	0.69	0.89	0.75	555719
weighted avg	1.00	0.99	0.99	555719

Logistic Regression Report:

	precision	recall	f1-score	support
0.0	1.00	0.69	0.81	552870
1.0	0.01	0.55	0.02	2849
accuracy			0.68	555719
macro avg	0.50	0.62	0.41	555719
weighted avg	0.99	0.68	0.81	555719

XGBoost Classification Report:

	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	552870
1.0	0.72	0.80	0.76	2849
accuracy			1.00	555719
macro avg	0.86	0.90	0.88	555719
weighted avg	1.00	1.00	1.00	555719

XGBOOST

MODEL SELECTED

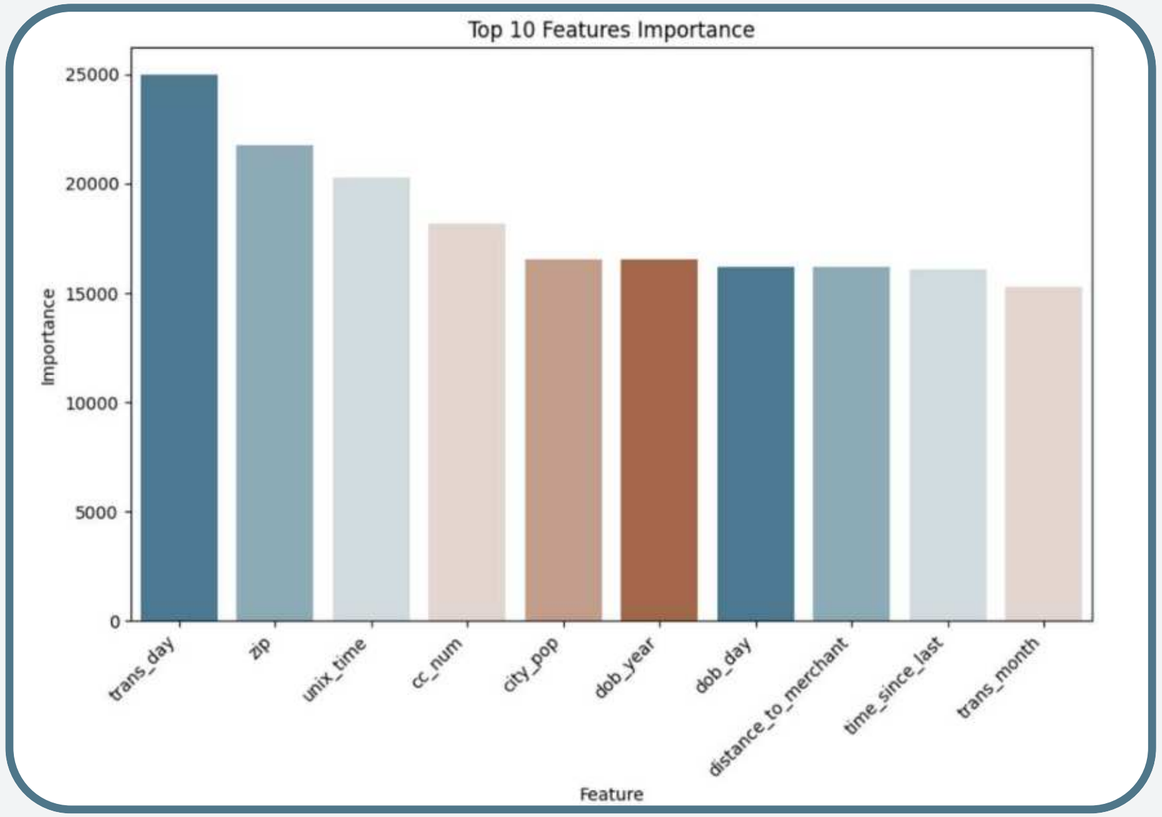
- The model performs well in detecting non fraudulent transactions, with 100% precision and recall.
- However, it only detects 81% of fraudulent transactions, missing 19% and having a 14% false positive rate.
- The high accuracy alone is not a reliable metric for this model due to the imbalanced data between the classes, where non-fraudulent transactions dominate.

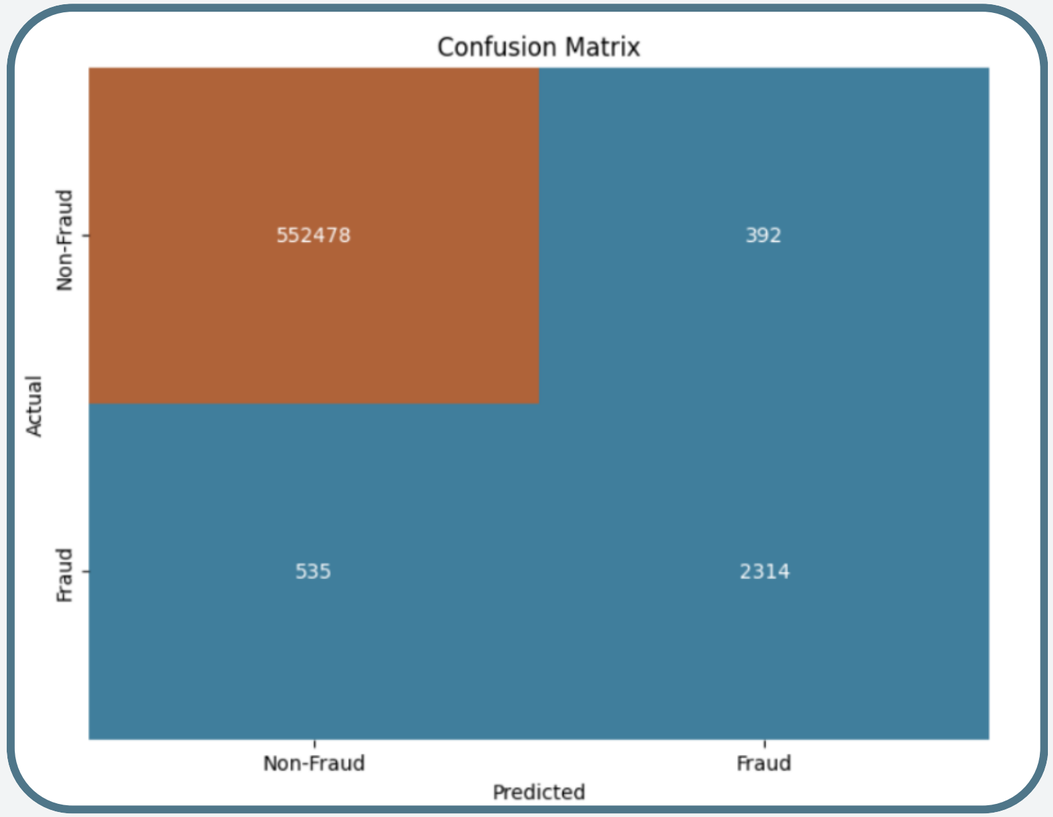
XGBoost Classification Report:

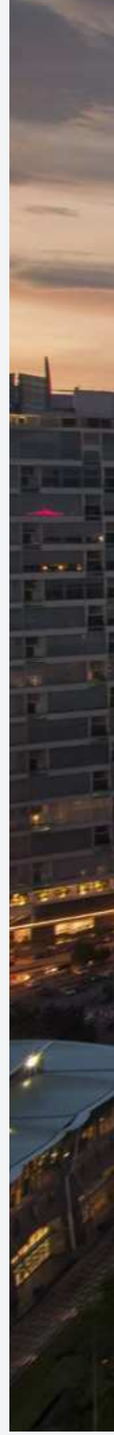
	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	552870
1.0	0.86	0.81	0.83	2849
accuracy			1.00	555719
macro avg	0.93	0.91	0.92	555719
weighted avg	1.00	1.00	1.00	555719

DATA VISUALIZATION & EVALUATION

FEATURE IMPORTANCE &CONFUSION MATRIX



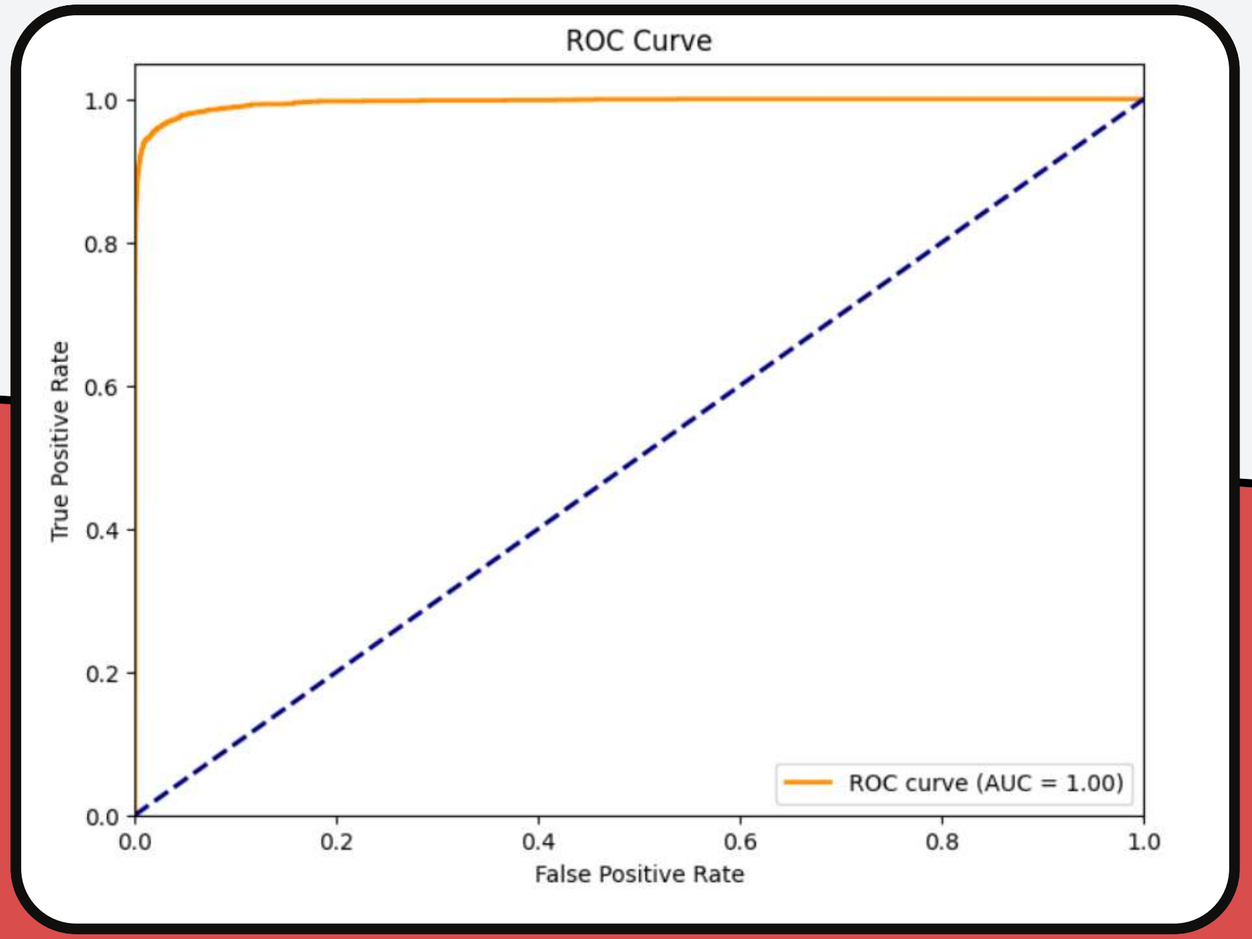




EVALUATION

ROC CURVE

- The model is good at distinguishing between fraudulent and non-fraudulent transactions (performing with very few mistakes)
- The model can correctly classify +90% of the cases based on its predictions.



MODEL OPTIMIZATION

NEXT STEPS

Review the specific false positives and false negatives to understand the model's mistakes and improve data preprocessing.

Performed more different Models and hyperparameter tuning while considering review the Class Imbalance


THANK
YOU