FASTAG TRANSACTION FRAUD DETECTION - Mentorness

PROBLEM STATEMENT

This internship project focuses on leveraging machine learning classification techniques to develop an effective fraud detection system for Fastag transactions. The dataset comprises key features such as transaction details, vehicle information, geographical location, and transaction amounts. The goal is to create a robust model that can accurately identify instances of fraudulent activity, ensuring the integrity and security of Fastag transactions.

Exploratory Data Analysis (EDA):

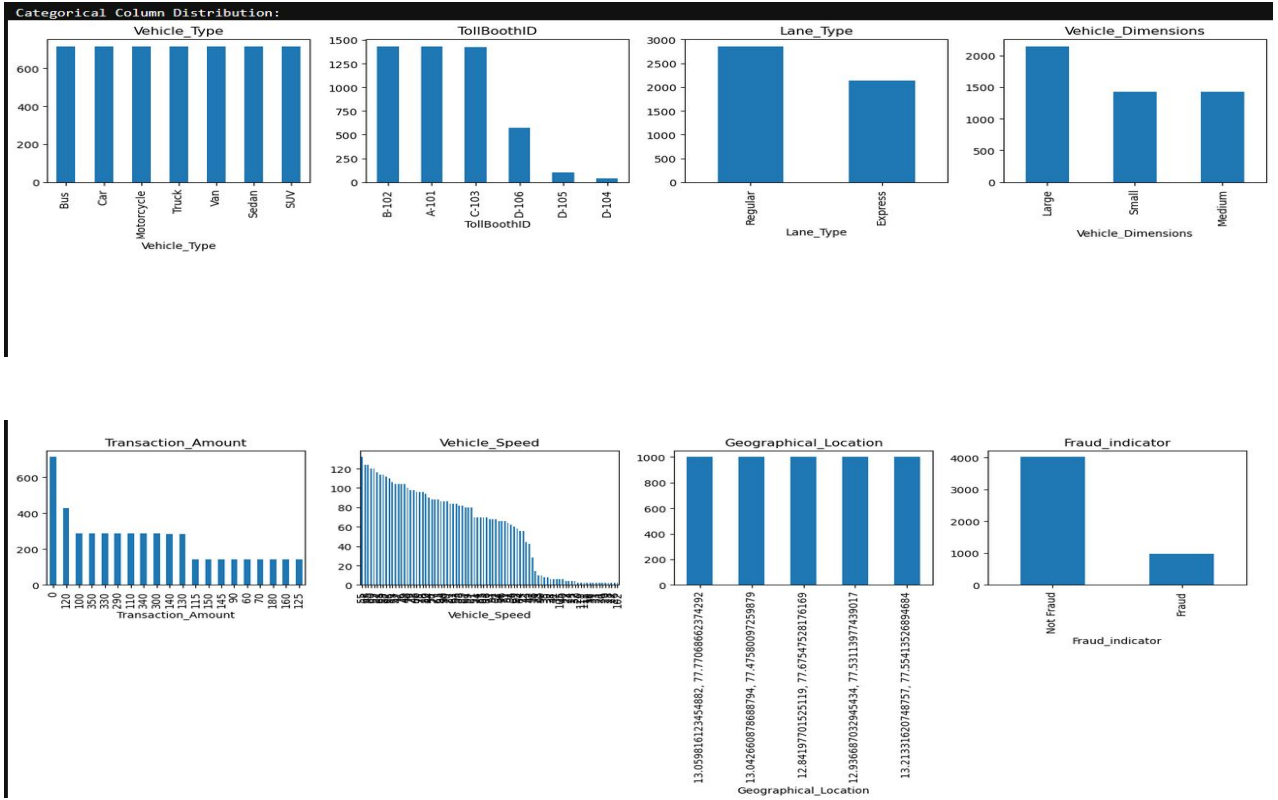
```
Duplicates: 0

Transaction_ID              0
Timestamp                   0
Vehicle_Type                0
FastagID                  549
TollBoothID                 0
Lane_Type                   0
Vehicle_Dimensions          0
Transaction_Amount          0
Amount_paid                 0
Geographical_Location       0
Vehicle_Speed               0
Vehicle_Plate_Number        0
Fraud_indicator             0
dtype: int64
```

The data set comprised **5000 rows** and **13 columns**.
NO duplicates.
FastagID column was dropped.

Features Distribution:

The diagram shows the distribution of features in the dataset.

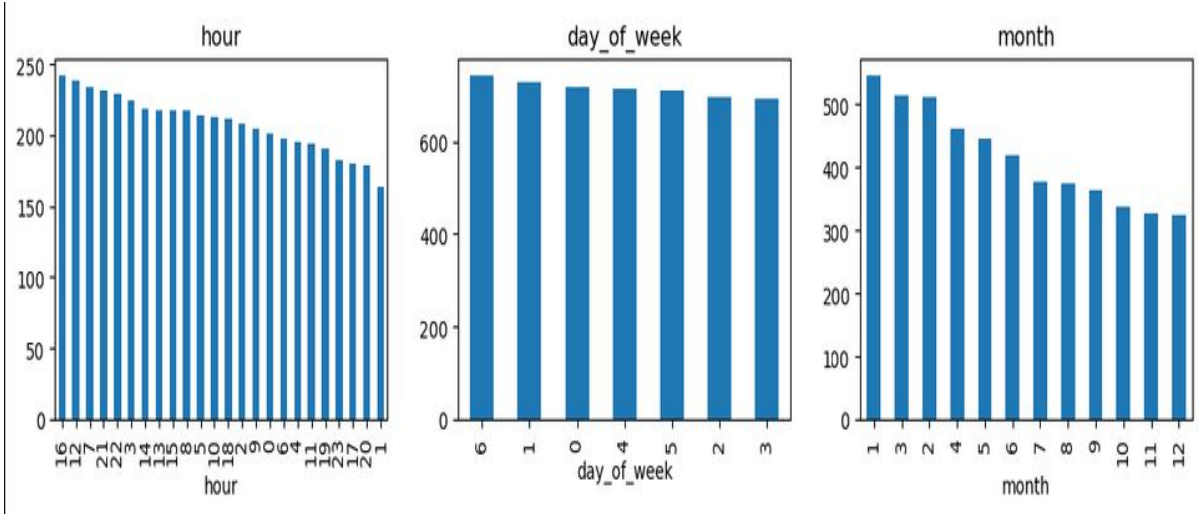
New features such as **“day”, “hour”, “day_of_week”, and“month”** were also created from the **“Date”** column.



Features Distribution:

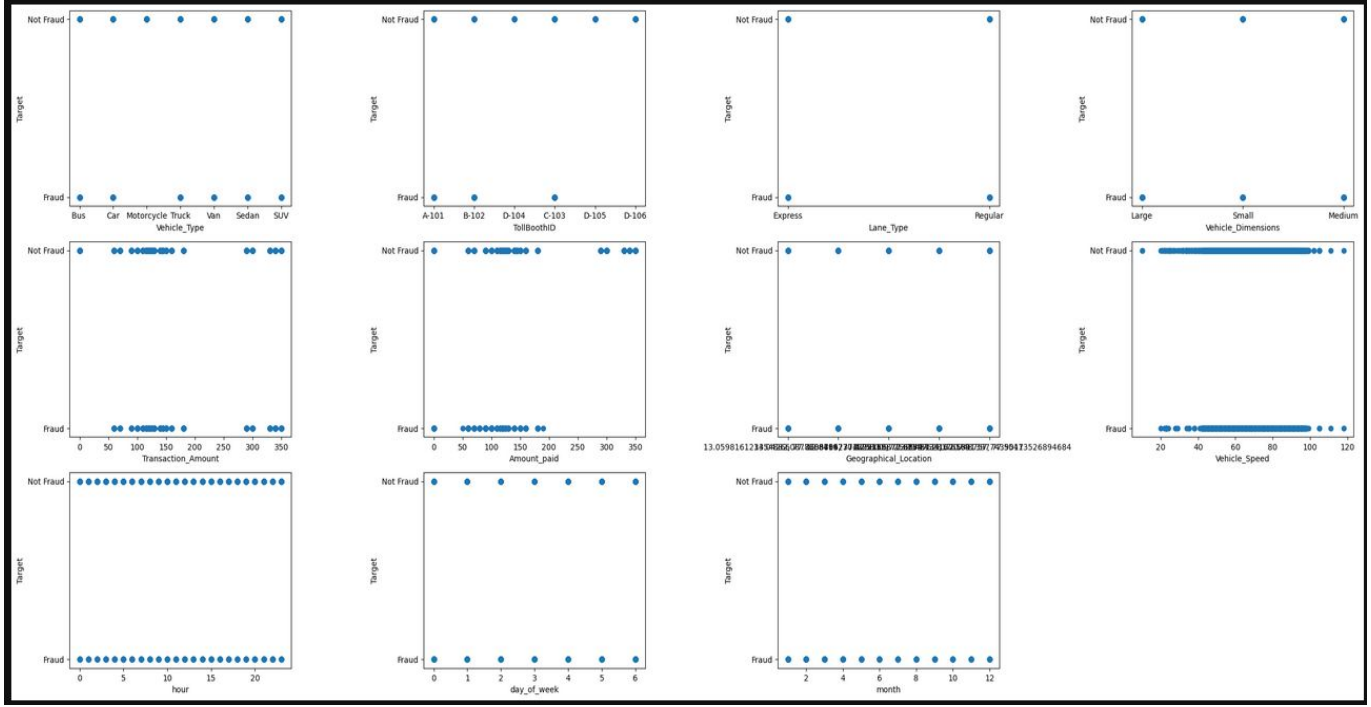
The diagram shows the distribution of features in the dataset.

New features such as **“day”, “hour”, “day_of_week”, and“month”** were also created from the **“Date”** column.

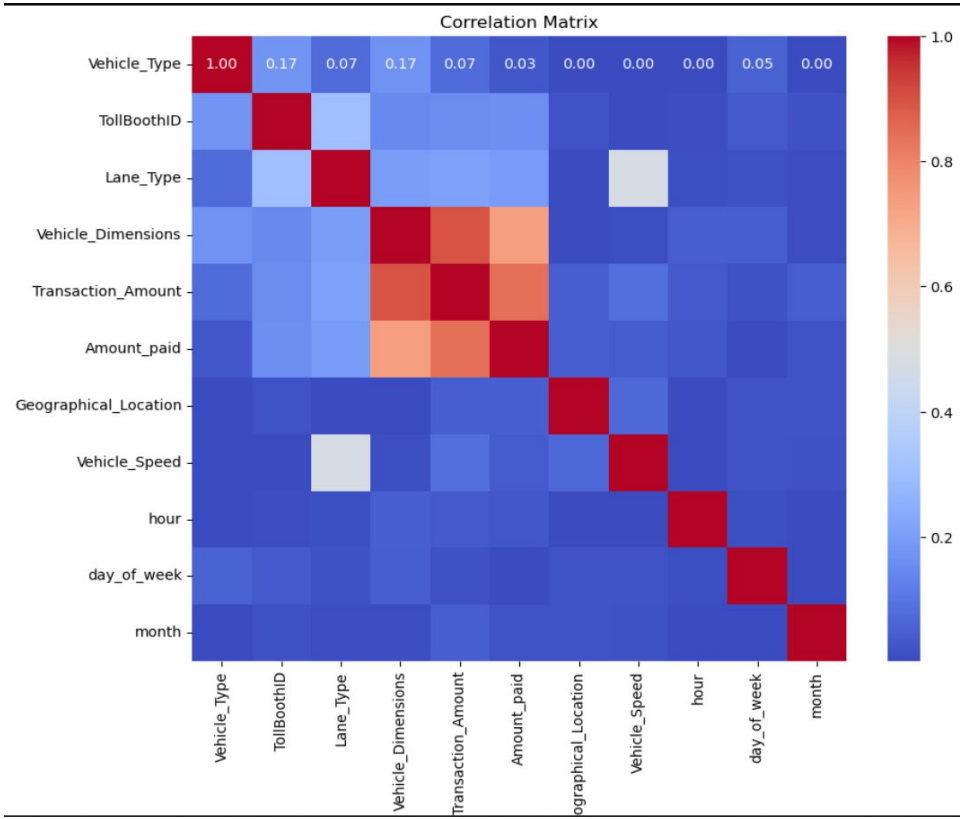


The relationship between the features and the label.

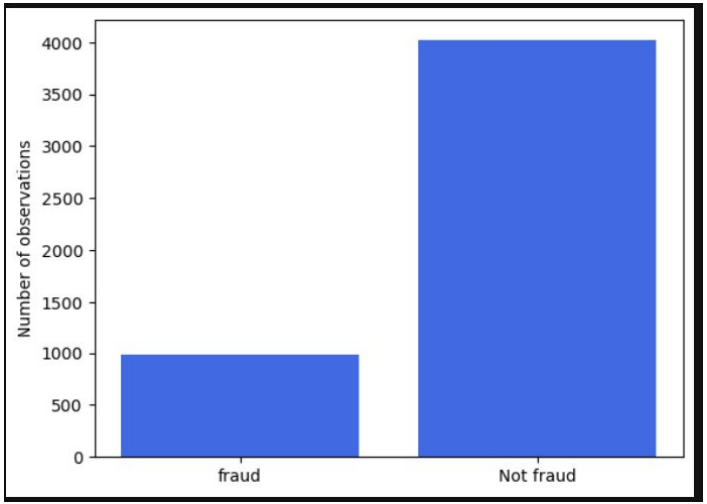
RELATIONSHIP BETWEEN THE LABEL AND EACH FEATURE:

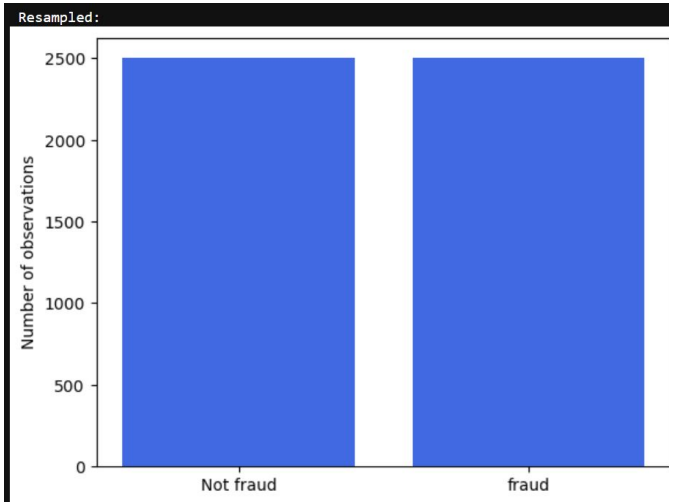


In training the model, one feature that correlated up to **90%** with another was dropped. Hence reducing the features to **11.**



The imbalanced issue of the target variable was handled by creating a class size of **2500** (the average of the total observationin the dataset), and upsampling the low observation(“Fraud”) and down-sampling the high observation(“Not fraud”) to the class size.





After training the model with several classifiers, the **AdaBoostClassifier** with **number of estimator 40,** is finally used for my prediction.

Classifier	Train Accuracy	Train Recall	Train Precision	Train F1 Score	Test Accuracy	Test Recall	Test Precision	Test F1 Score	Train Time
Logistic Regression	0.92675	0.9655	0.896056	0.929483	0.931	0.964	0.904315	0.933204	0.125149
K Neighbors	0.96150	0.9580	0.964753	0.961365	0.929	0.940	0.919765	0.929773	0.018245
Linera SVM	0.93100	0.9770	0.894689	0.934034	0.935	0.976	0.902033	0.937560	1.106525
RBF SVM	1.00000	1.0000	1.000000	1.000000	0.935	1.000	0.884956	0.938967	2.629628
Decision Tree	0.93400	1.0000	0.883392	0.938086	0.929	1.000	0.875657	0.933707	0.010998
Random Forest	0.84800	0.8340	0.858025	0.845842	0.817	0.790	0.835095	0.811922	0.043343
Ada Boost	0.95900	0.9815	0.939234	0.959902	0.958	0.970	0.947266	0.958498	0.293143

FEATURE IMPORTANCES

The diagram shows the relative importance of the features used in prediting the model.

Vehicle Type, Vehicle Dimensions, and **Day of Week** are not contributing to the model and therefore will be dropped.

The most most important features are arraged from topo to bottom.

	Feature Importances (%)
Amount_paid	45.0
Transaction_Amount	32.5
hour	7.5
Vehicle_Speed	5.0
TollBoothID	2.5
Lane_Type	2.5
Geographical_Location	2.5
month	2.5
Vehicle_Type	0.0
Vehicle_Dimensions	0.0
day_of_week	0.0