Detailed Report

On

Fraud Detection in Financial Transactions Project

- By Anshika Tyagi

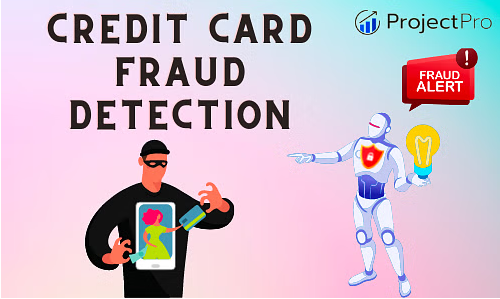
Date – 22 August ,2024

**Problem Description :** Our mission is to create a model capable of detecting fraudulent activities in financial transactions. The challenge lies in accurately identifying irregular patterns and anomalies that signify fraudulent behavior while minimizing false positives.

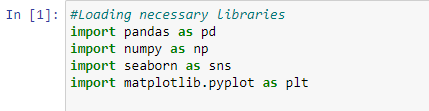
**Overview :** This report is about Fraud Detection in Financial Transactions Project , a machine learning model which is performed on Credit Card Fraud Detection dataset which is taken from website Kaggle in Jupyter Notebook.

The report presents entire process which involved in building the model,model selection criteria, challenges encountered and key insights obtained**.**

**Dataset link :** [**Credit Card Transactions Fraud Detection Dataset (kaggle.com)**](https://www.kaggle.com/datasets/kartik2112/fraud-detection)

****

* **Steps or Process which is used for building model :**
* Importing Necessary Libraries **:** Firstly ,import all python libraries such as numpy ,pandas ,matplotlib ,seaborn and sklearn etc.

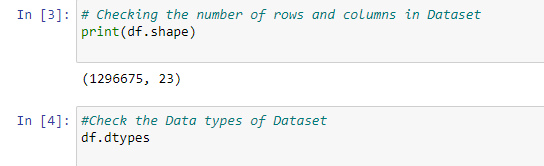


* Loading Dataset : Then load dataset in the jupyter notebook which is in csv file.

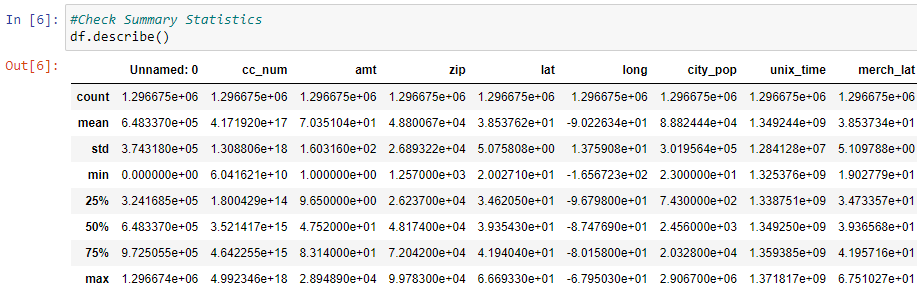
****

****

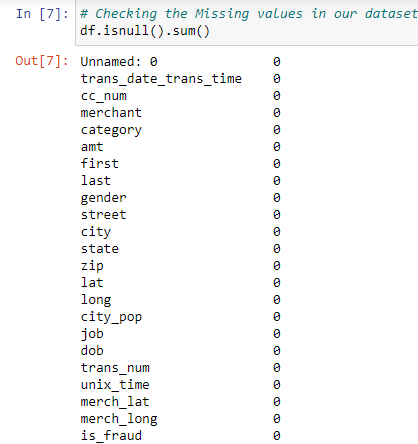
* Know about dataset : After loading , we try to know about the dataset such as number of rows and columns ,their datatypes and summary statistics.



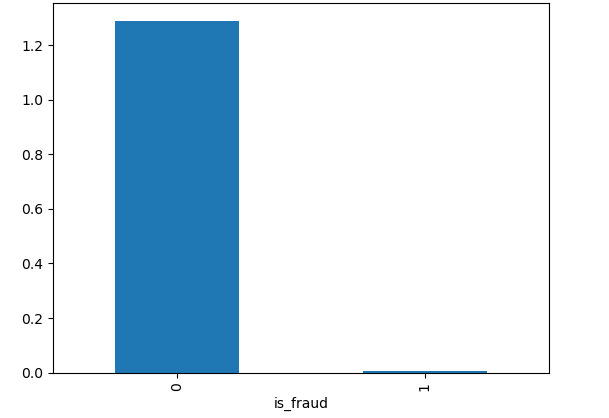




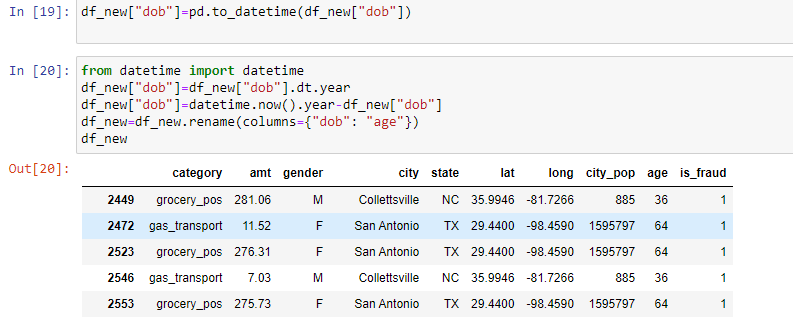
* Data Preprocessing and Visualization : In this step, clean and preprocess the data to make it suitable for analysis and modeling.This includes handling missing/null values,encoding categorical variables , data visualization and extract new features from existing ones.

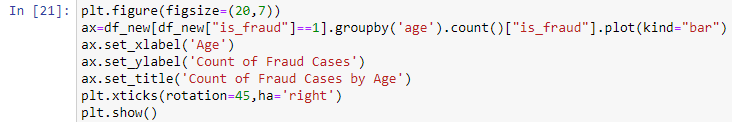


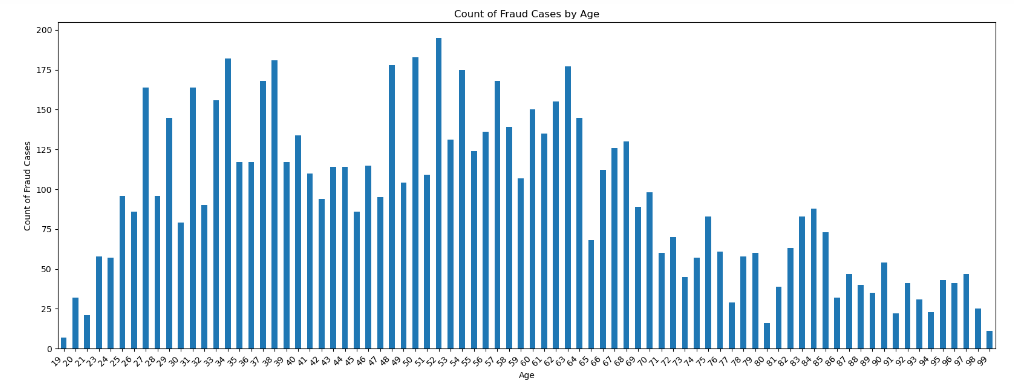


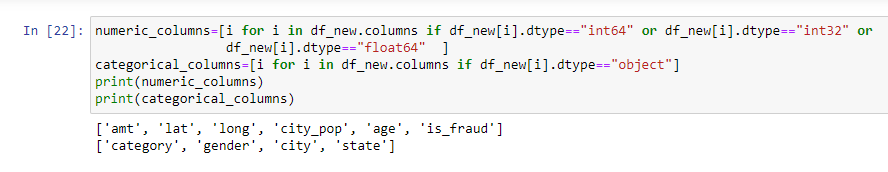


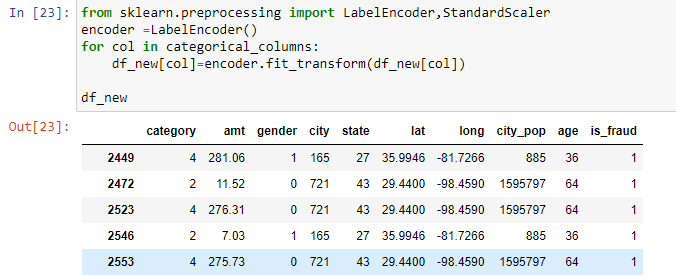
* Feature Engineering and Label Encoding : In this step ,create new freatures from the existing ones features that will assist to improve the performance of our models.Also we will convert categorical columns to numeric columns.



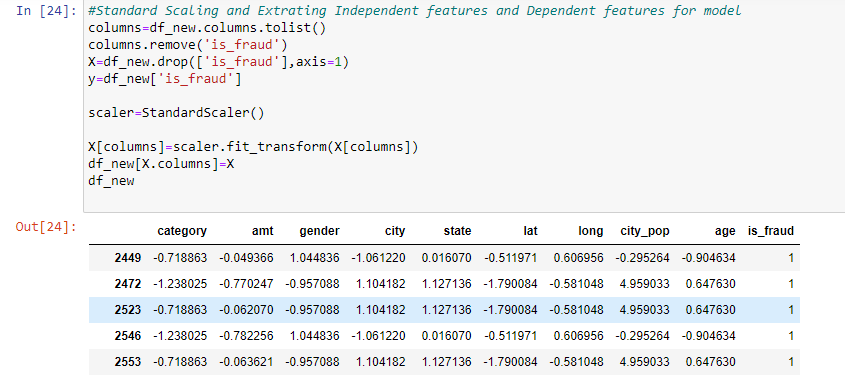




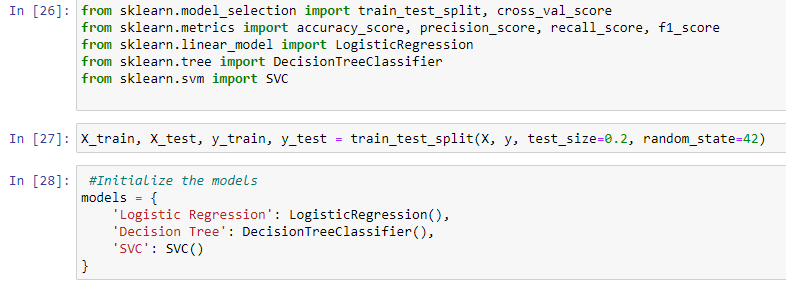


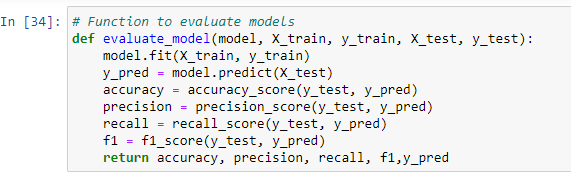


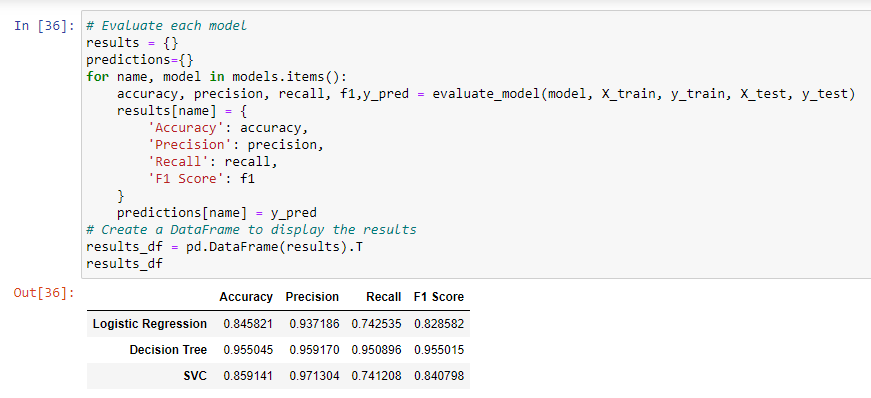
* Standard Scaling and Extracting Independent and Dependent Features: In this ,perform standard scaling and extract independent features which are category ,amt,gender, city ,state and lat etc. And target variable is is\_fraud which will be predicted by model.



* Building Models : After extracting features , implement various machine learning models ,as it classification problem .So we use Logistic Regression , Decision tree and Support Vector Machines etc to detect fraudulent transactions.







* Evaluation of Models : After the train the model ,we will evaluate the performance of our models using various metrics such as accuracy, precision, recall, and the F1 score.

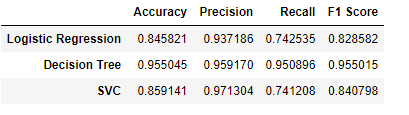
Evaluation Metrics:

1.Accuracy

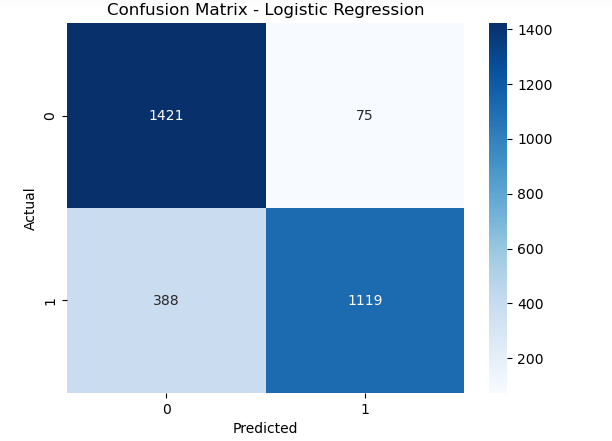
2.Precision

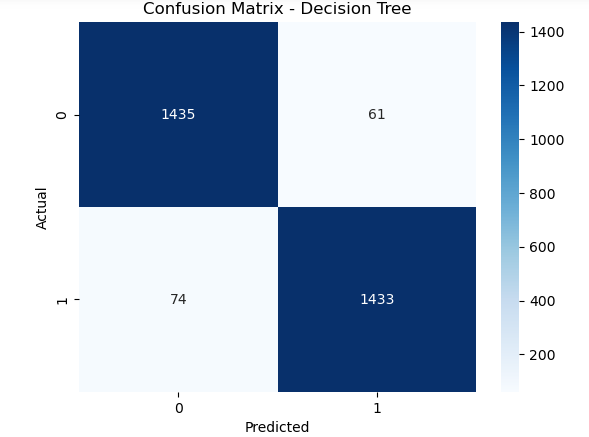
3.Recall

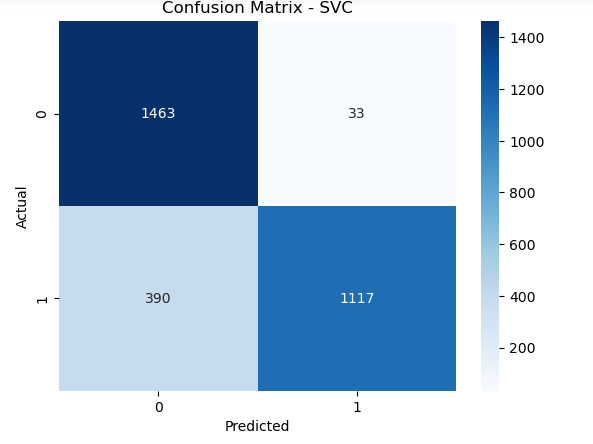
4.F1 Score



* Create Confusion Matrix : Create the Confusion matrix for each algo to determine the Fasle negatives and false positives







* **Model Selection Approach :** As it is Classification problem means we have to detect whether the it is fraud or not .For this we use these algo to implement this…

 **Logistic Regression -**The accuracy of LR is 84% obtained.

 **Decision Tree** - The accuracy of DT is 95% obtained.

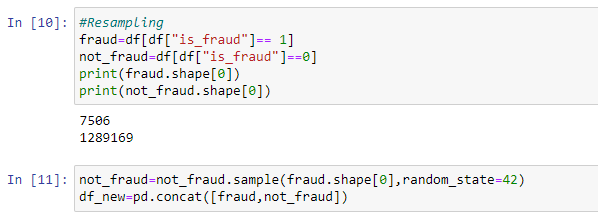
 **Support Vector Machine (SVM)** – The accuracy of SVM is 85% obtained.

So , from above Decision Tree has better accuracy, so it is more effective to predict/detect fraud .

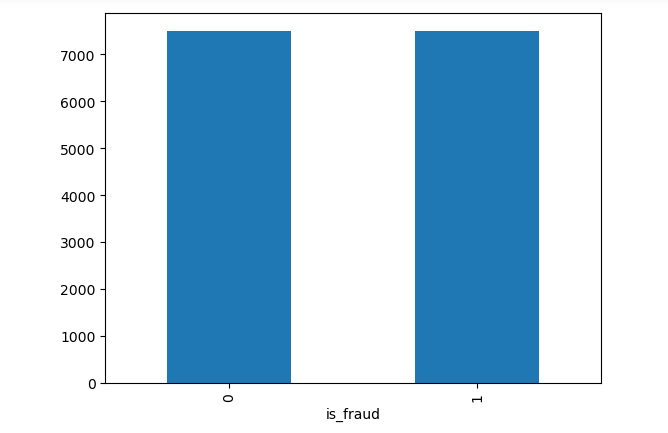
* **Challenges Occurred :**
* **Handling Large Dataset Challenge :** In the credit card fraud detection project, we faced a significant challenge due to the large size of the dataset. Addressing this challenge was crucial for ensuring efficient model development and accurate predictions**.**

Approach to Handle this Challenge :

We use data Sampling to manage the dataset size during this , we employed random sampling to create a smaller subset of the data .







* **Dealing with Categorical Columns :** It is also one challenge to deal with many categorical columns in the dataset , Because Categorical columns with a large number of unique values can lead to sparse representations and increase the dimensionality of the dataset, making it more complex to process.

Approach to Handle this Challenge :

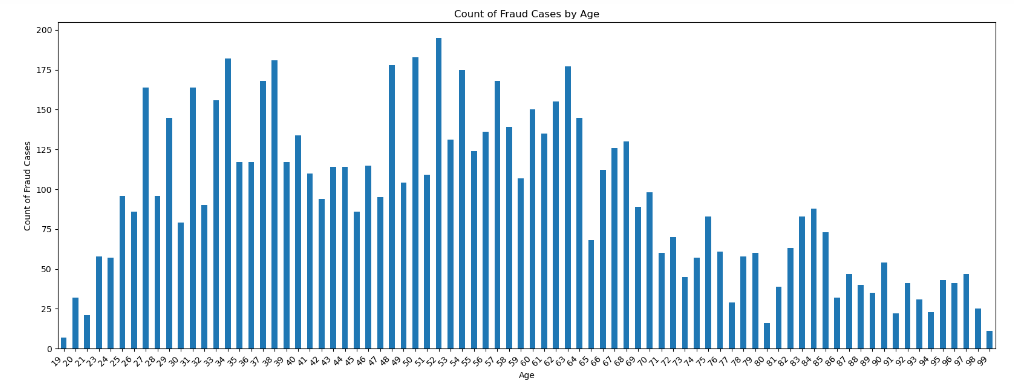
Techniques like Label Encoding or One hot Encoding applied on categorical columns with a manageable number of unique values. This method creates binary columns for each category, allowing the model to interpret the presence or absence of each category.

* **Memory Optimization Challenge** : There are many columns present in the dataset which need for memory for their storage ,that is also a challenge to optimize the memory ,the large dataset was problematic, leading to slow data retrieval and processing times.

Approach to Handle this Challenge :

We performed feature selection to identify and retain only the most relevant features for model training. By this only crucial features are taken from the dataset and rest of the features we have removed from the dataset which reduce the memory space or storage .

* **Insights obtained:**

****

 **High Fraud Occurrence in Middle Age Groups**: The age groups between 30 to 50 years old show the highest counts of fraud cases. These age ranges are often associated with individuals actively engaged in financial activities like loans, credit card usage, and investments, making them more susceptible targets.

 **Peak Fraud Cases Around Age 40**: There is a noticeable peak around the age of 40, with the count of fraud cases nearing 200. This suggests that individuals in this age range may be more exposed to risks or more actively involved in financial transactions.

 **Lower Fraud Incidents Among Younger and Older Age Groups**: The chart shows significantly fewer fraud cases in younger age groups (below 20) and older age groups (above 60). This could indicate less financial activity or different spending and financial management behaviors in these demographics.