Capstone Project (DA):-
Credit Card Fraud Detection

Submitted By:- (DSC 43)

- Prashant Singh
- Vaibhav Singh
- Shreya Mandlesha

AGENDA

- Objective /Problem Statement
- Background
- Problem-solving approach
- Key Insights /Visualization
- Cost Benefit Analysis
- **Appendix:-**
  - Data Attributes
  - Data Methodology
  - Attached Files

OBJECTIVE / PROBLEM STATEMENT

- As a part of the analytics team working on a fraud detection model and its cost-benefit analysis. We need to develop a machine learning model to detect fraudulent transactions based on the historical transactional data of customers with a pool of merchants.
- We have to analyze the business impact of these fraudulent transactions and recommend the optimal ways that the bank can adopt to mitigate the fraud risks.
- We need to put proactive monitoring and fraud prevention mechanisms in place.
- Machine learning helps these institutions reduce time-consuming manual reviews, costly chargebacks and fees, and denial of legitimate transactions.

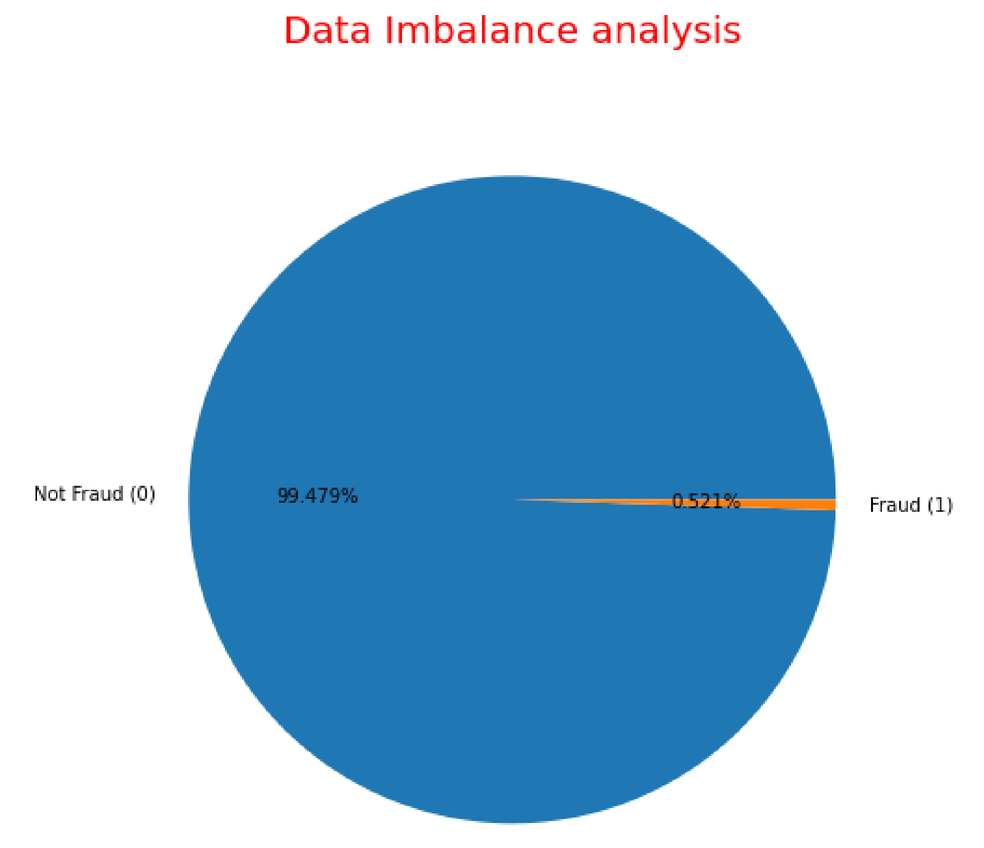
BACKGROUND

- The number of fraudulent transactions has increased drastically, due to which credit card companies are facing a lot of challenges. For many banks, retaining highly profitable customers is the most important business goal. Banking fraud, however, poses a significant threat to this goal.
- In terms of substantial financial loss, trust and credibility, banking fraud is a concerning issue for both banks and customers alike.
- With the rise in digital payment channels, the number of fraudulent transactions is also increasing as fraudsters are finding new and different ways to commit such crimes.
- We have performed the root cause analysis for the increasing number of frauds and high revenue loss, and you realized that building a fraud detection system using different machine learning techniques is quite important to identify such fraudulent activities at the right time and prevent them from happening.

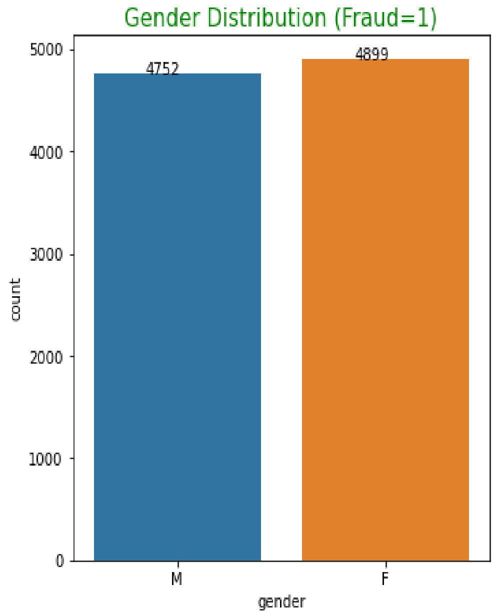
PROBLEM-SOLVING APPROACH

1. Reading and Understanding the Data
2. Data Inspection/ Cleaning / Transformation
3. EDA (Univariate and Bivariate analysis)
4. Data Preparation (Train/Test Data Splitting)
5. Multiple Model Building or Hyperparameter Tuning
6. Model Evaluation
7. Business Impact: - Cost Benefit Analysis (Before and After Model deployment)

KEY INSIGHTS /VISUALIZATION



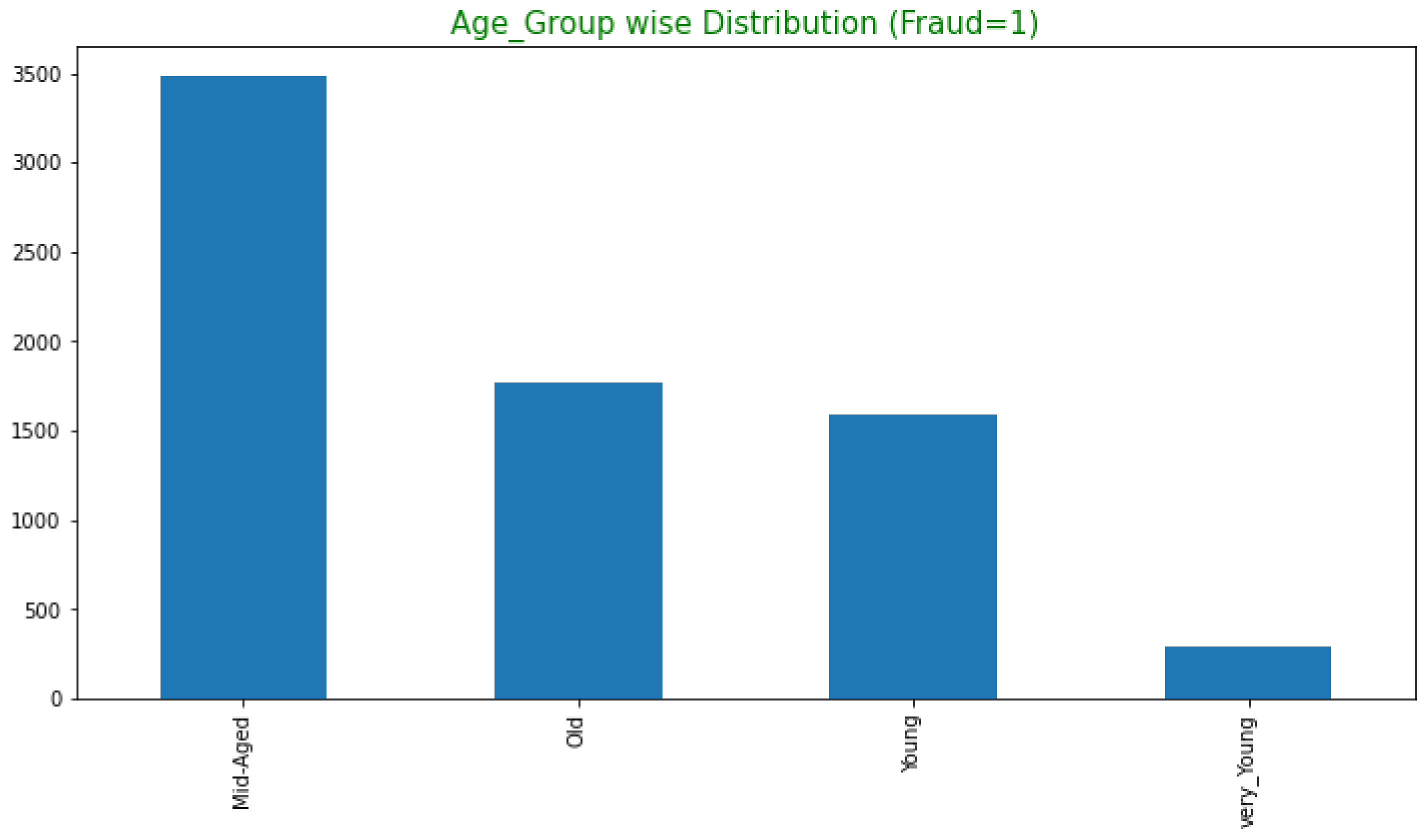
- Data set is highly imbalanced, Out of a total of 18,52,394 transactions, 9651 are fraudulent, with the positive class (frauds) accounting for 0.521% of the total transactions. Class Not fraud accounting for 99.479% of total transactions.



Comments:-

-Female having more fradulent transaction count over man the difference is not huge

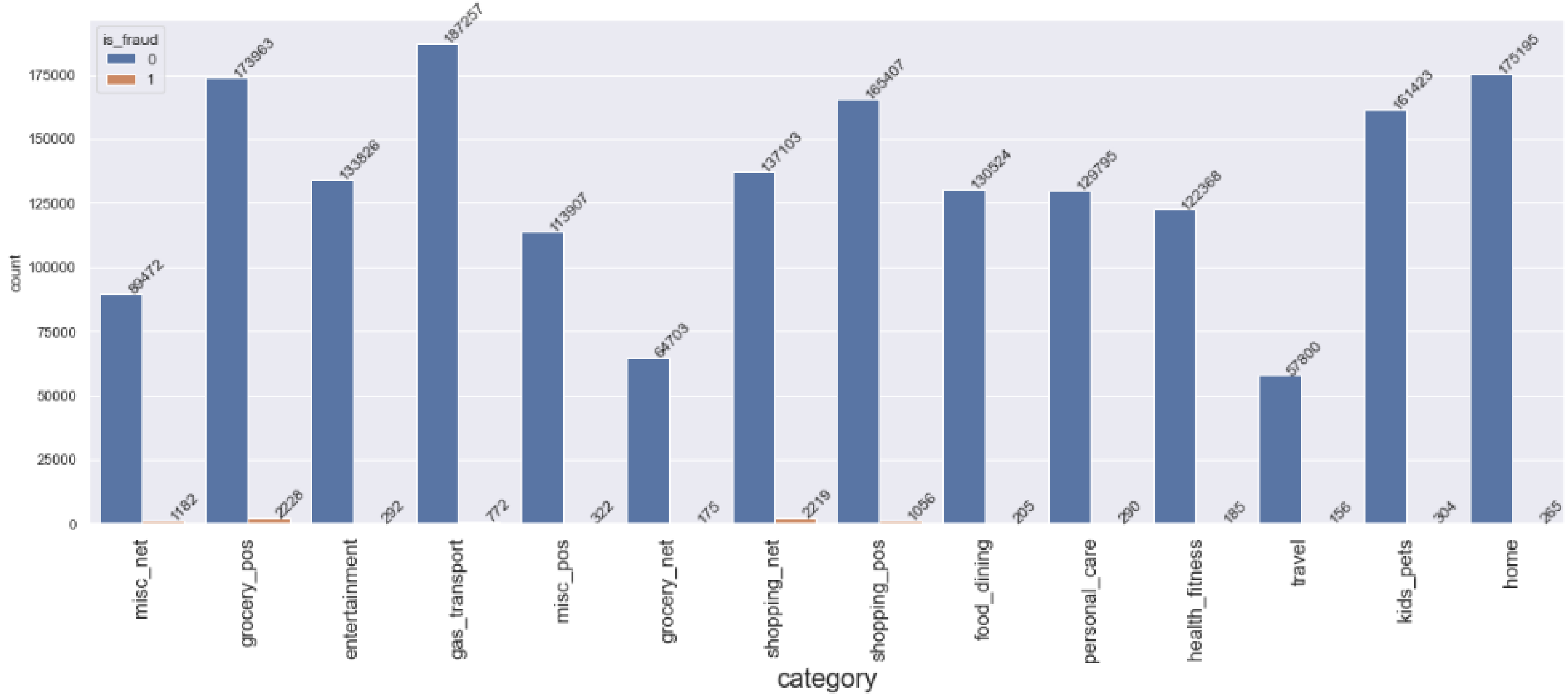
- Need to focus on both gender equally to check fraudulent



Comment:-

- Mid Aged people between 30 to 50 are most no of fraudulent activity

Transaction amount, category and gender are the most important variables

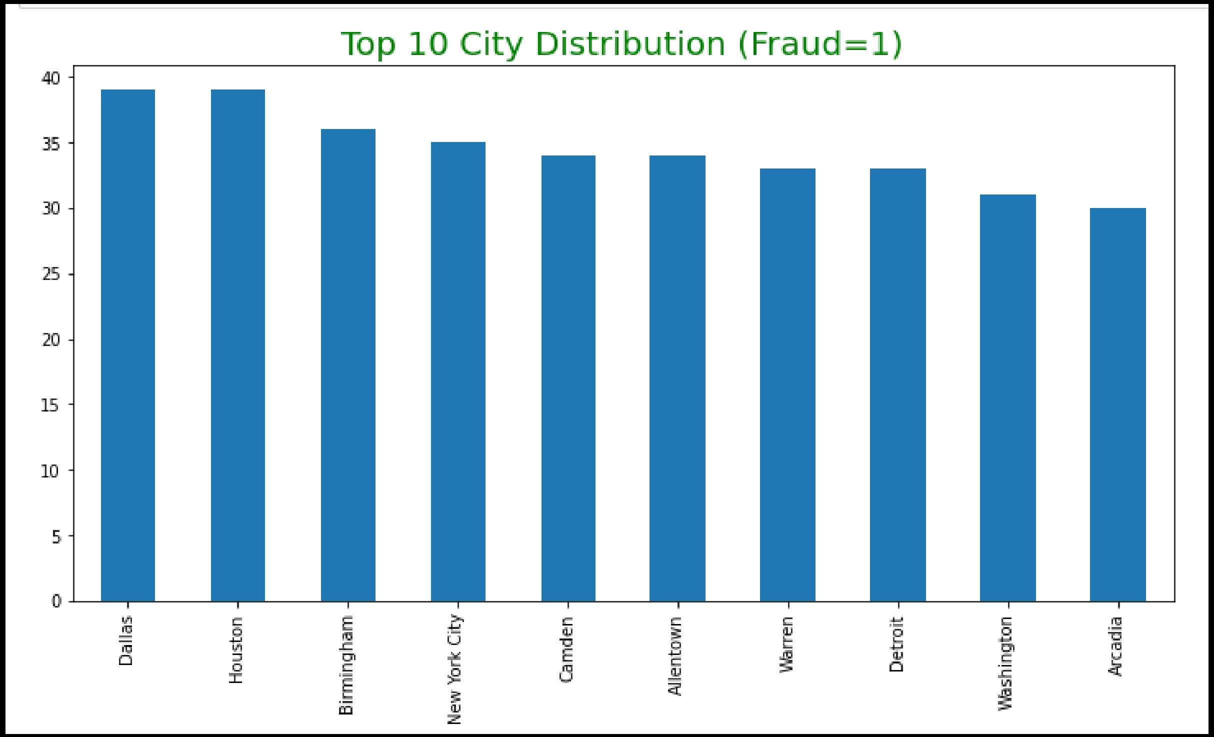


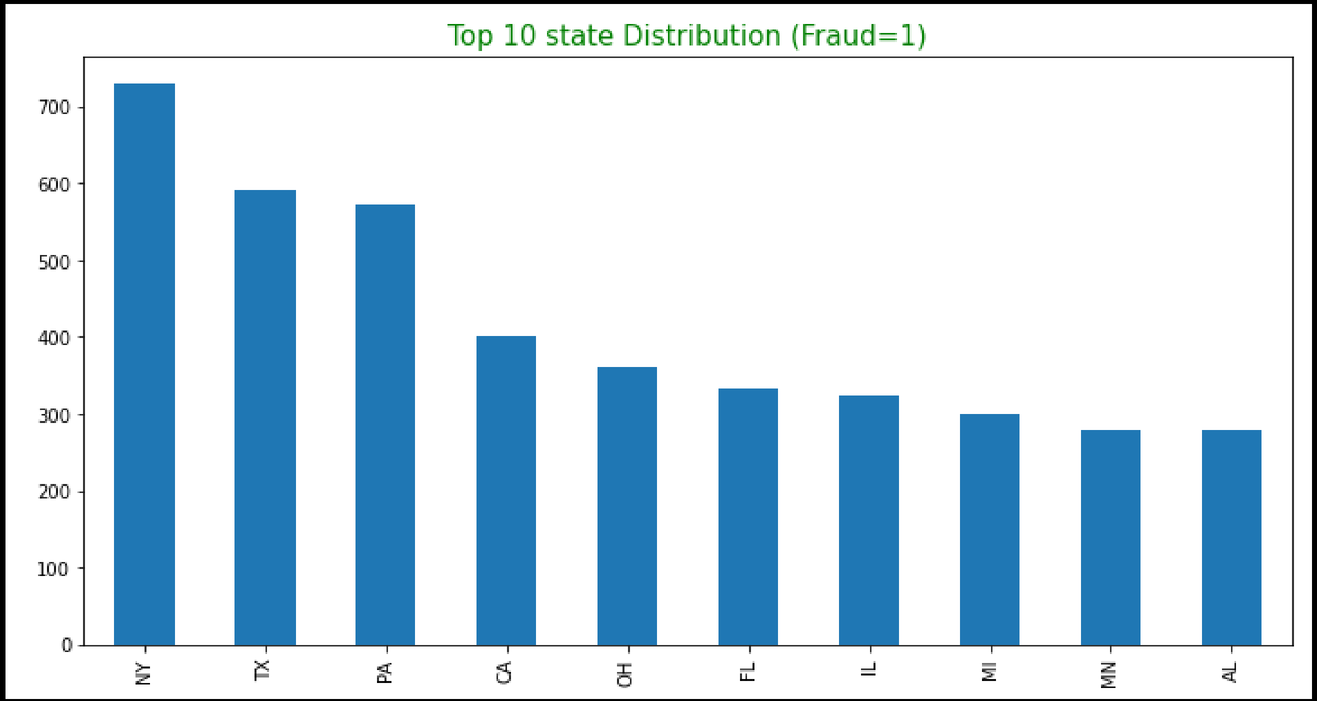
Observation:-

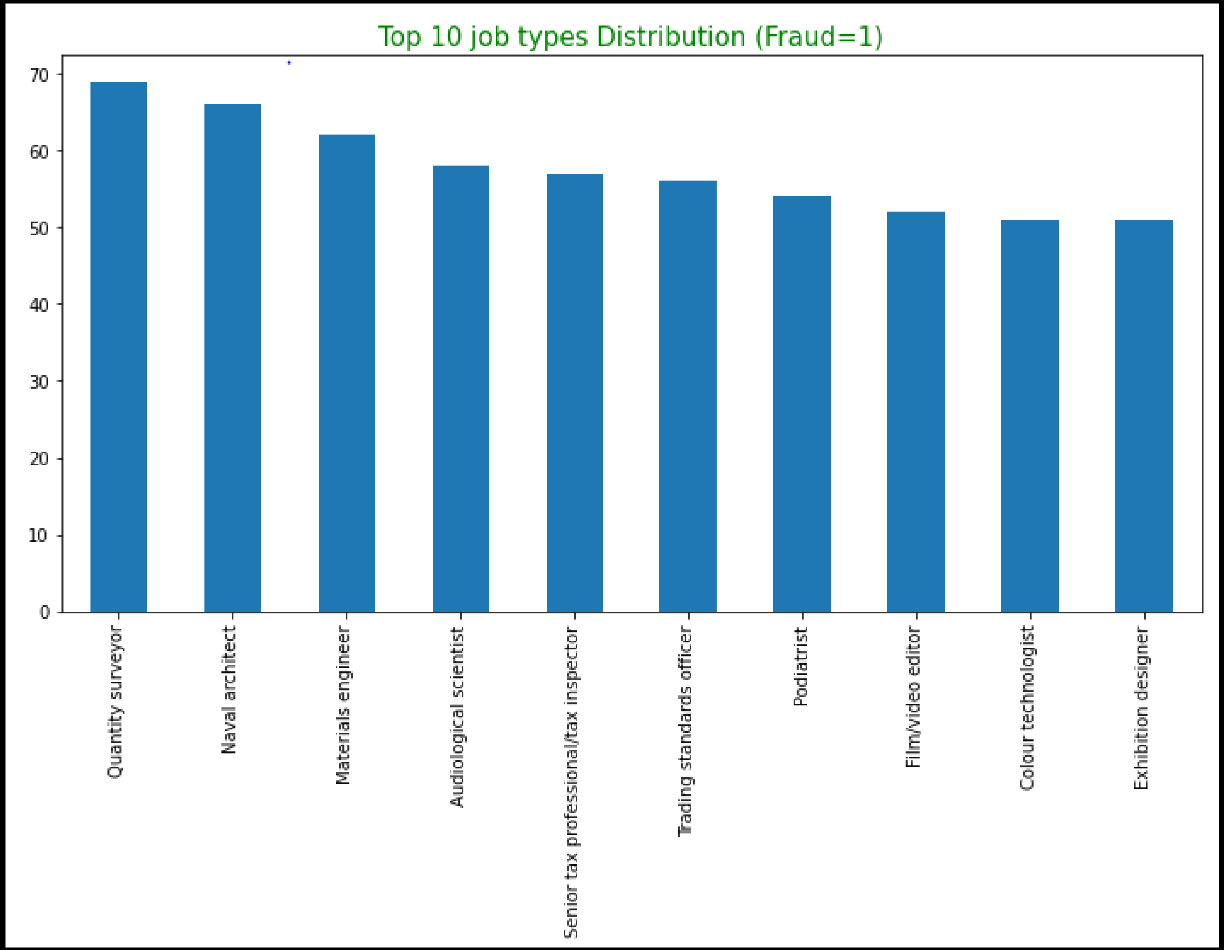
In column "category" (Is_fraud =1)

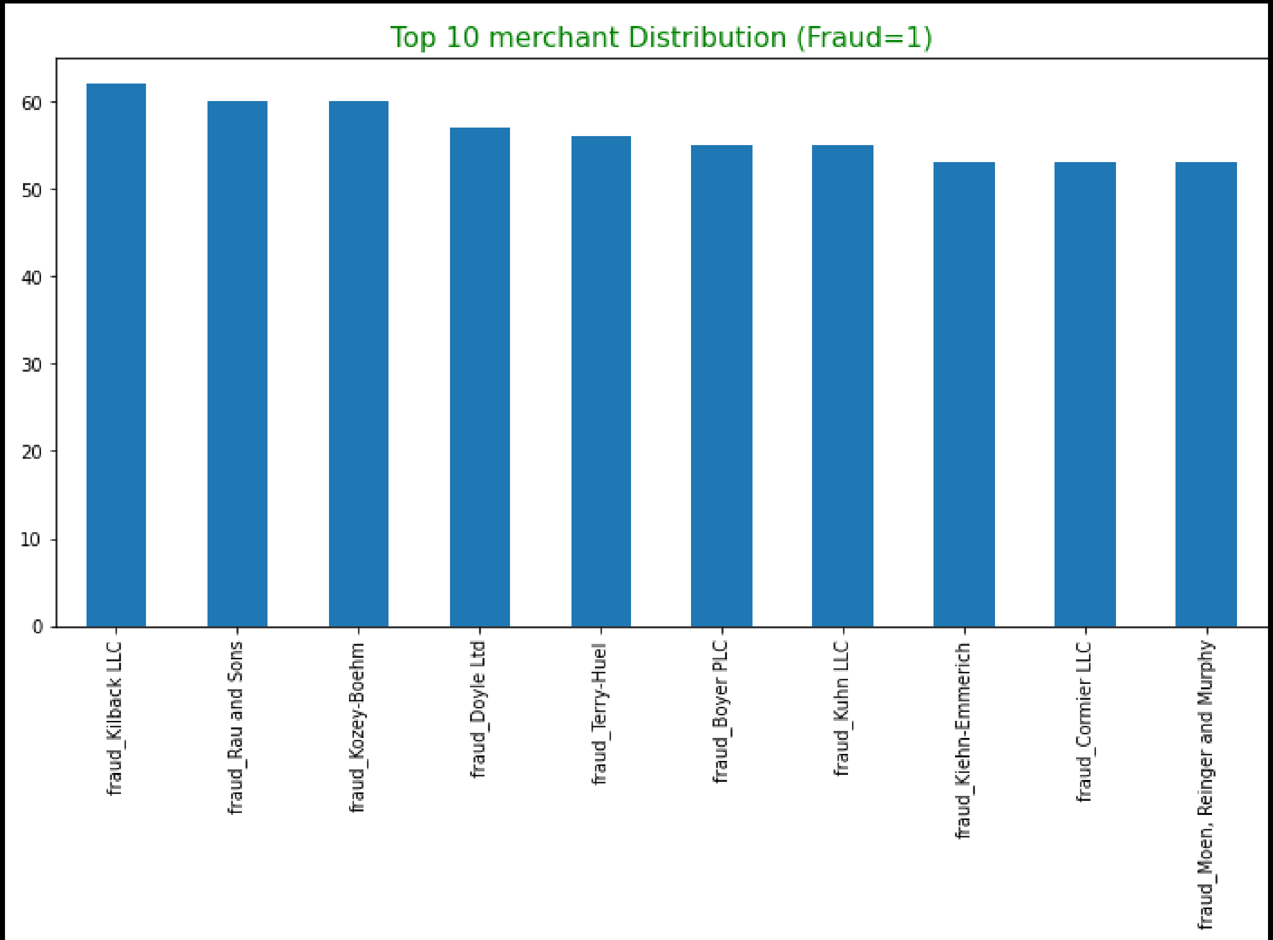
- grocery_pos, shopping_net, misc_net, shopping_pos, gas_transport these are top 5 category which are having more chances of having fraudulent transaction.

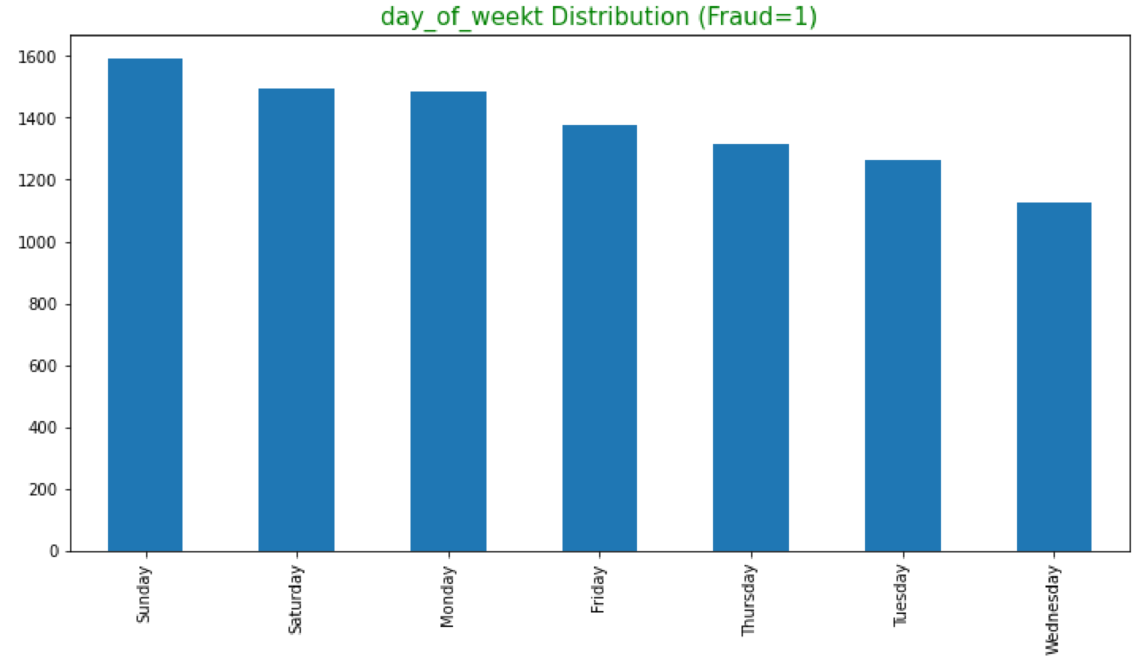
- food_dining, health_fitness, grocery_net, travel having least fraud transaction.


Top 10 City Distribution (Fraud=1)
0
5
10
15
20
25
30
35
40
Dallas
Houston
Birmingham
New York City
Camden
Allentown
Warren
Detroit
Washington
Arcadia


Top 10 state Distribution (Fraud=1)
0
100
200
300
400
500
600
700
NY
TX
PA
CA
OH
FL
IL
MI
MN
AL

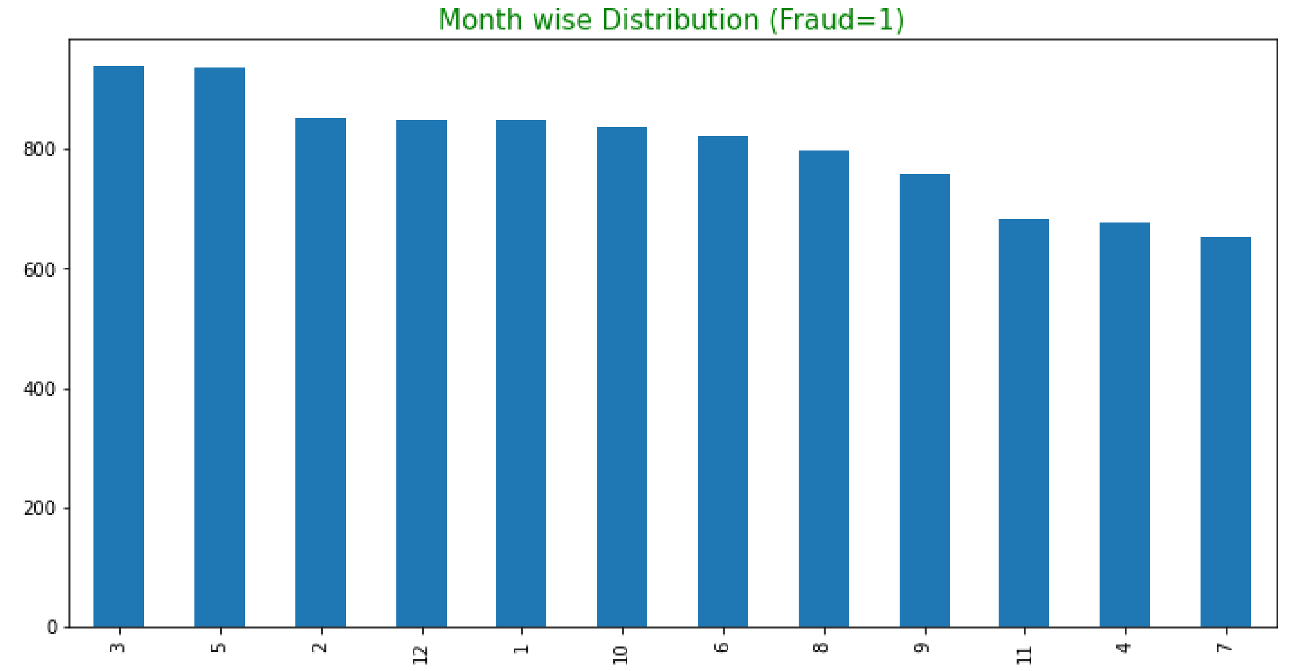

Top 10 job types Distribution (Fraud=1)
0
10
20
30
40
50
60
70
Quantity surveyor
Naval architect
Materials engineer
Audiological scientist
Senior tax professional/tax inspector
Trading standards officer
Podiatrist
Film/video editor
Colour technologist
Exhibition designer


Top 10 merchant Distribution (Fraud=1)
0
10
20
30
40
50
60
fraud_Kilback LLC
fraud_Rau and Sons
fraud_Kozey-Boehm
fraud_Doyle Ltd
fraud_Terry-Huel
fraud_Boyer PLC
fraud_Kuhn LLC
fraud_Kiehn-Emmerich
fraud_Cormier LLC
fraud_Moen, Reinger and Murphy



Comments:-

-Saturday, sunday having highest fraud transactions



Comments

-March & May showing highest no of fradulent transaction.

- April & July having least fraud transactions

COST BENEFIT ANALYSIS

- Part 1 – Cost Benefit Analysis

Sr. No.	Questions	Results
1	Average number of transactions per month	77,183
2	Average number of fraudulent transaction per month	402.12
3	Average amount per fraud transaction	530.661

- Part 2 – Cost Benefit Analysis
  - Reduction in losses by ~82%

Sr. No.	Questions	Results
1	Cost incurred per month before the model was deployed (b*c)	213392.22
2	Average number of transactions per month detected as fraudulent by the model (TF)	237
3	Cost of providing customer executive support per fraudulent transaction detected by model	1.5
4	Total cost of providing customer support per month for fraudulent transactions detected by the model (TF*$1.5)	355.50
5	Average number of transactions per month that are fraudulent but not detected by the model (FN)	68
6	Cost incurred due to fraudulent transactions left undetected by the model (FN*c)	35908.09
7	Cost incurred per month after the model is built and deployed (4+6)	36263.59
8	Final savings = Cost incurred before - Cost incurred after(1-7)	**177128.63**

APPENDIX: DATA ATTRIBUTES

Snapshot of the data:

- index - Unique Identifier for each row
- trans_date_trans_time- Transaction DateTime
- cc_num - Credit Card Number of Customer
- merchant - Merchant Name
- category - Category of Merchant
- amt - Amount of Transaction
- first - First Name of Credit Card Holder
- last - Last Name of Credit Card Holder
- gender - Gender of Credit Card Holder
- street - Street Address of Credit Card Holder
- city - City of Credit Card Holder
- state - State of Credit Card Holder
- zip - Zip of Credit Card Holder
- lat - Latitude Location of Credit Card Holder
- long - Longitude Location of Credit Card Holder
- city_pop - Credit Card Holder's City Population
- job - Job of Credit Card Holder
- dob - Date of Birth of Credit Card Holder
- trans_num - Transaction Number
- unix_time - UNIX Time of transaction
- merch_lat - Latitude Location of Merchant
- merch_long - Longitude Location of Merchant
- is_fraud - Fraud Flag <--- Target Class

APPENDIX: DATA METHODOLOGY

- Multiple ML Model classifier built on top of a Kaggle-simulated dataset - screenshot
- Class imbalance adjusted using Adaptive Synthetic (ADASYN)/SMOTE sampling method
- Manual hyperparameter tuning is done due to extensive computational times when using Grid Search Cross Validation

	Model	Recall on Train	Recall on Test	AUC Score
9	XGBoost - Unsampled	0.650	0.460	0.98
6	Random Forest - Unsampled	0.270	0.230	0.96
10	XGBoost - SMOTE	0.870	0.810	0.95
11	XGBoost - ADASYN	0.860	0.810	0.95
4	Decision Trees - SMOTE	0.850	0.830	0.93
5	Decision Trees - ADASYN	0.850	0.850	0.93
7	Random Forest - SMOTE	0.820	0.780	0.93
8	Random Forest - ADASYN	0.820	0.790	0.93
3	Decision Trees - Unsampled	0.330	0.300	0.73
2	Logistic Regression - ADASYN	0.729	1.000	0.63
1	Logistic Regression - SMOTE	0.780	1.000	0.56
0	Logistic Regression - Unsampled	0.000	0.988	0.52

APPENDIX: ATTACHED FILES

Cost Benefit Analysis:

- Cost Benefit Analysis_FRAUD_PS_VS_SJ.xlsx

Multiple ML Model deployments:

- credit card fraud_DA_capstone_PS_VS_SJ.ipynb

video submission Link:-