CREDIT CARD FRAUD DETECTION

Himanshu Shukla

INDEX

PROBLEM STATEMENT

INCREASE IN FRAUD TRANSACTIONS IN CREDIT CARDS

Bank is losing High profitable customers due to loss of Trust in bank.

CUSTOMERS

Customers are not able to report fraud Transactions as transactions are happening at odd hours and No Mechanism is present in the bank to detect Fraud Transaction.

FINANCIALS

Bank is suffering financial loses as bank has to reimburse the lost amount to customers.

MECHANISM OF FRAUD DETECTION

Bank does not have any fraud detection mechanism to detect fraud transactions.

SOLUTION

CLOSE THE GAP

Given all possible hypotheses and considering the feasibility and customer time, the most suitable solution is to implement a fraud detection system.

TARGET AUDIENCE

Affected Customers are Target Audience.

COST SAVINGS

Fraud Detection System Will lead to Cost Savings to the bank by informing customers about the Fraud Transactions

EASY TO USE

Fraud Detection System will not affect the customer’s time with extra OTP checks on all transactions and is also quite feasible, as educating all customers on various fraudulent techniques is a challenging task.

MODEL OVERVIEW

CLASSIFICATION

Final Model is Classification Model and using Random Forest as Algorithm.

PARAMETERS

Model is based on different Parameters Such as age of Customer, Time of Transaction, Distance between Customer location and Merchant Location, Category of Purchase , day of purchase , Gender of Customer.

ANALYSIS

Exploratory Data Analysis done for the Data Set and Impact of Variables on final output is studied.

IMPLICATIONS

Cost Benefit Analysis done and calculated how much saving will be done by Model.

INSIGHTS BASED ON DATA

INSIGHTS ON TRANSACTION AMOUNT COLUMN

SUMMARY

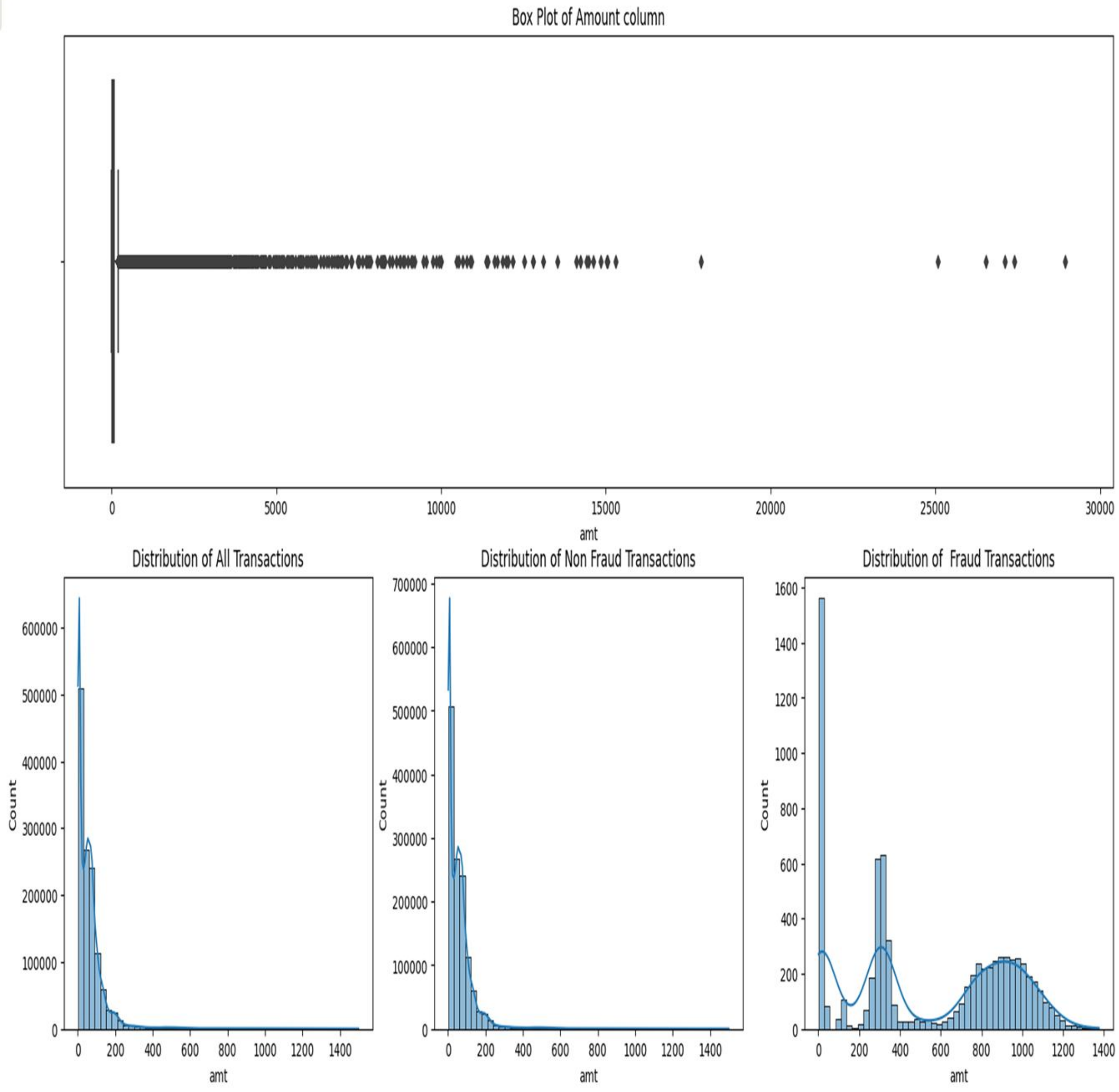
There are many Outliers in the Amount Column, but these can be genuine Transactions, so we have not capped this column for outlier Treatment.

SURPRISING RESULTS

Although Mean Transaction amount for Fraud Transaction is Higher Compared to Non-Fraud Transactions, Distribution of these Transactions Suggest that Transactions are of Small amount. So, Transactions of Small amount need to be monitored closely.

INFERENCE

Amount loss in Fraud Transaction is High but Transactions are of Small amounts. So, we need to monitor Transactions of Small amount Closely to Prevent Credit Card fraud



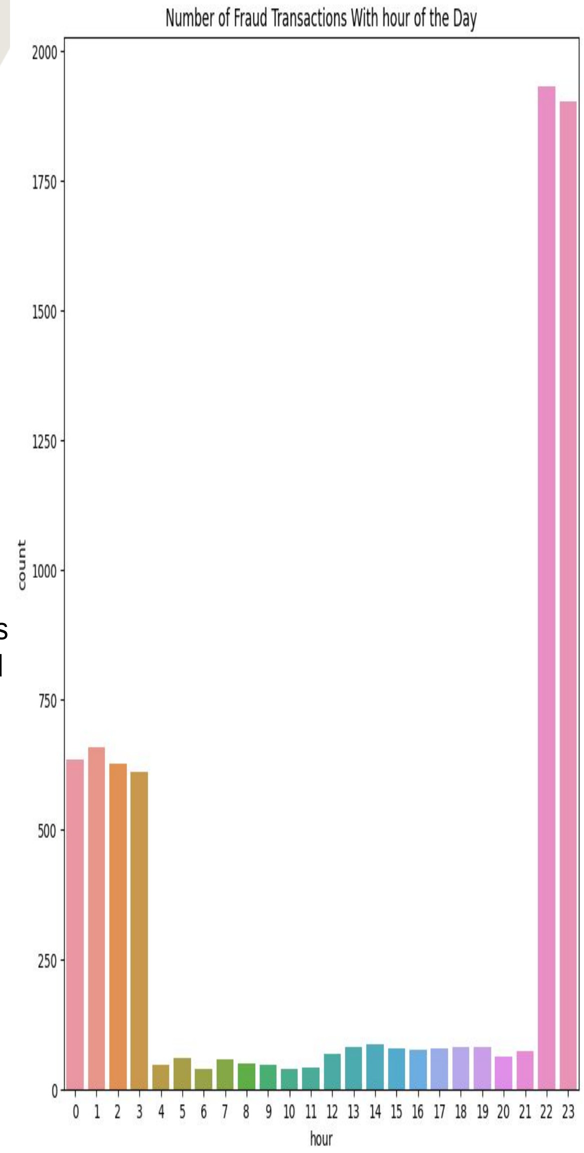
INSIGHTS ON TIME BASED COLUMNS (HOUR)

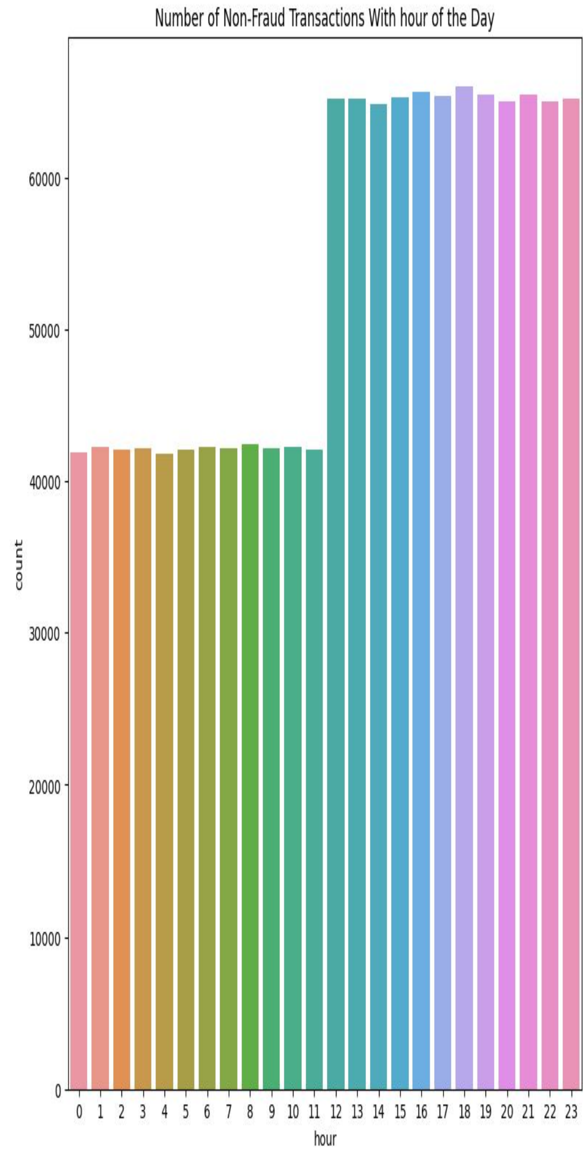
SUMMARY

Most Number of Transactions are happening after 12:00 hrs. So Survelience can be increased after 12:00 hrs.

INFERENCE

Most Number of Fraud Transactions are ahappening after 22:00 hrs till early morning 03:00 hrs which is not surprising as these are odd hours during which customer can not monitor the, Transactions So Close Survelience is required to monitor Transactions during this Time Frame.





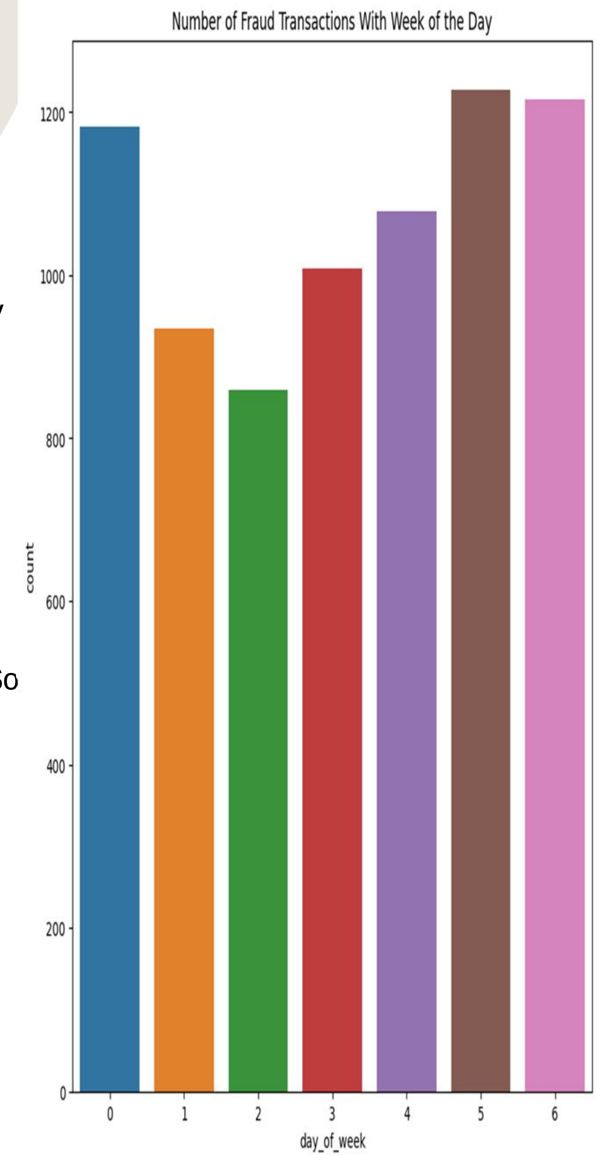
INSIGHTS ON TIME BASED COLUMNS (DAY OF WEEK)

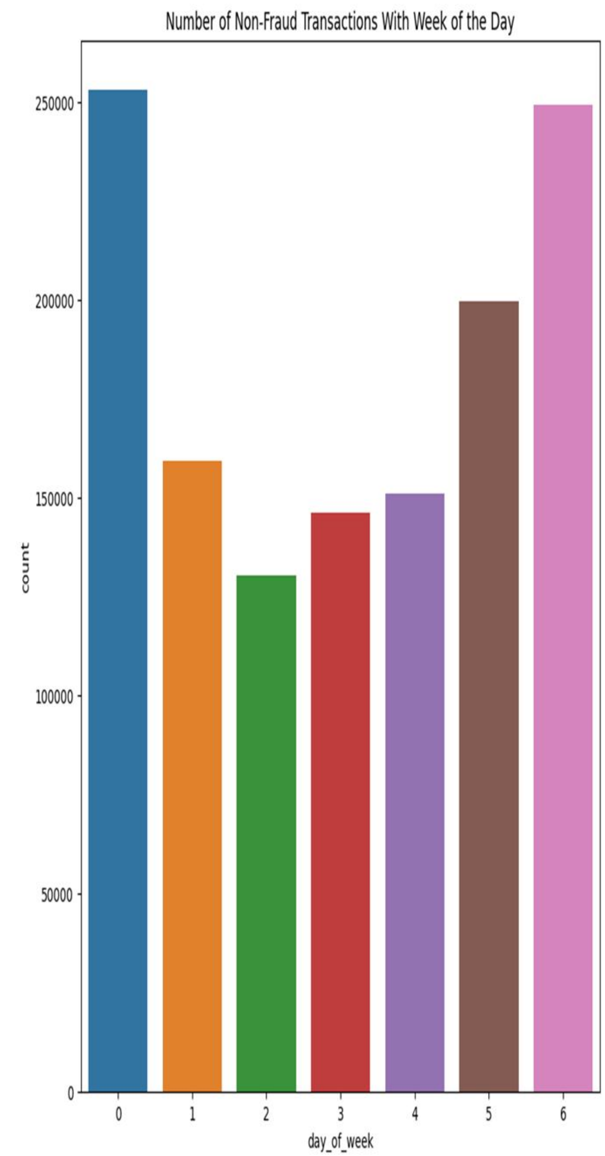
SUMMARY

Lot of Transactions are happening on Friday,Saturday and Sunday implying On Weekends and holiday lot of Transactions are happening. So Close Survelience is required on Weekends.

INFERENCE

Most of the Fraud Transactions are happening on the weekends. So Close Survelience is required on weekends.





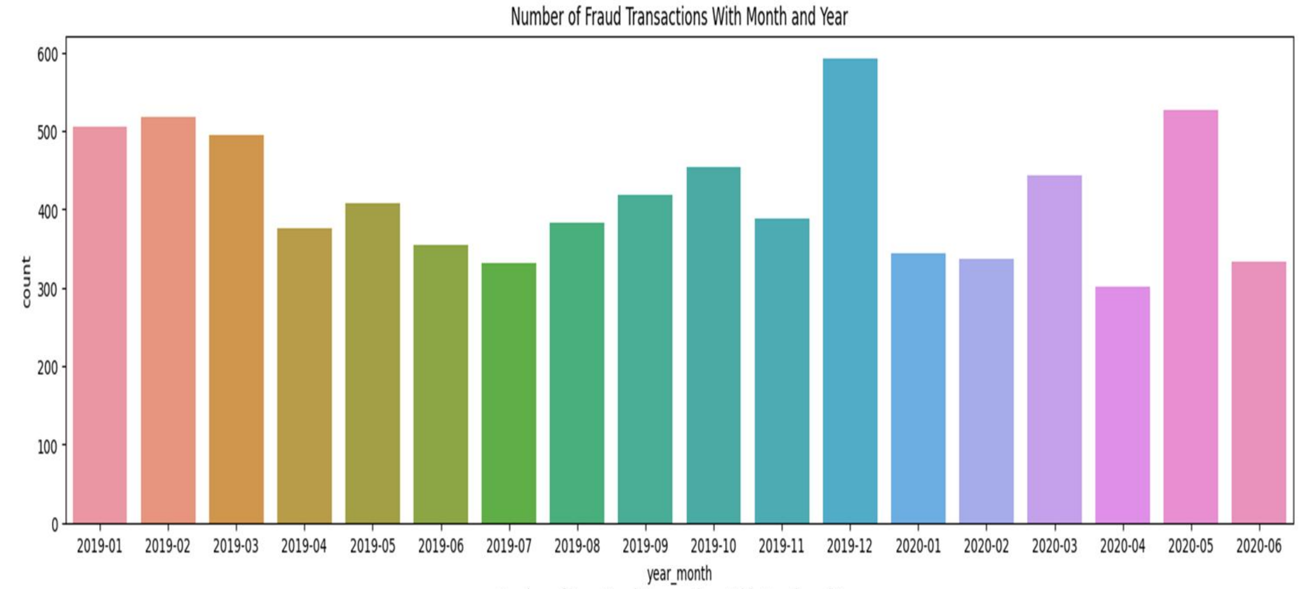
INSIGHTS ON TIME BASED COLUMNS (YEAR-MONTH)

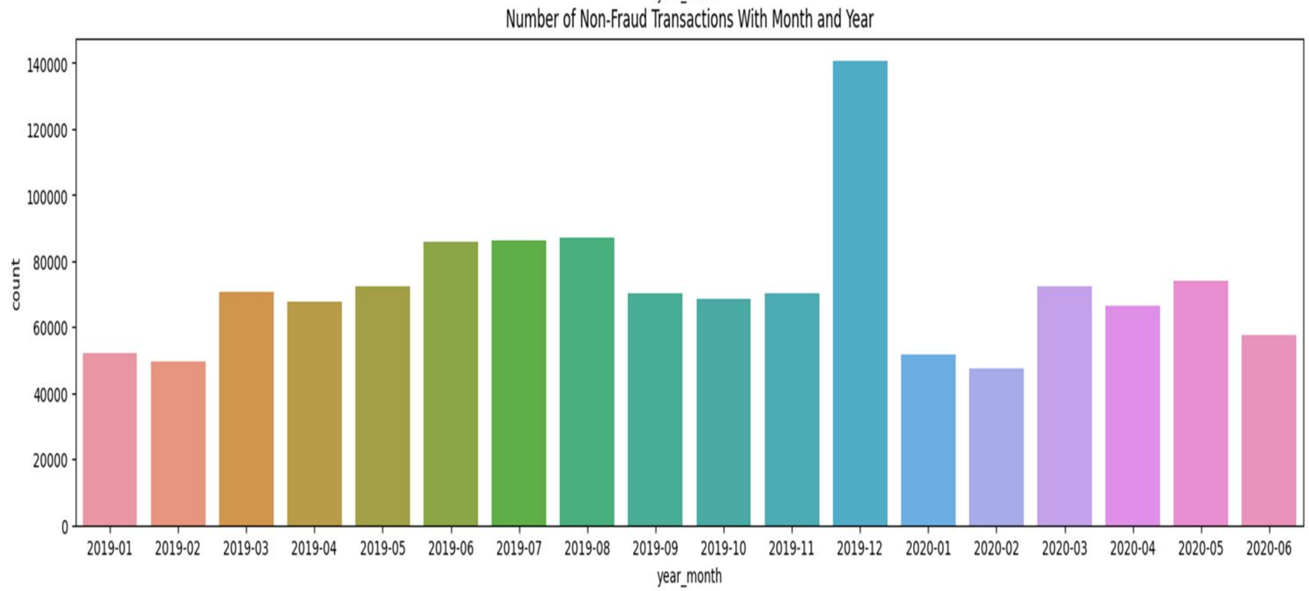
SUMMARY

Most Number of Transactions happened in January, February,March,December 2019,May 2020.

INFERENCE

Most Number of Fraud Transactions happened in December 2019 which is a Holiday Season So, Close Survelience is required during Holiday Season For Credit Card Fraud Prevention





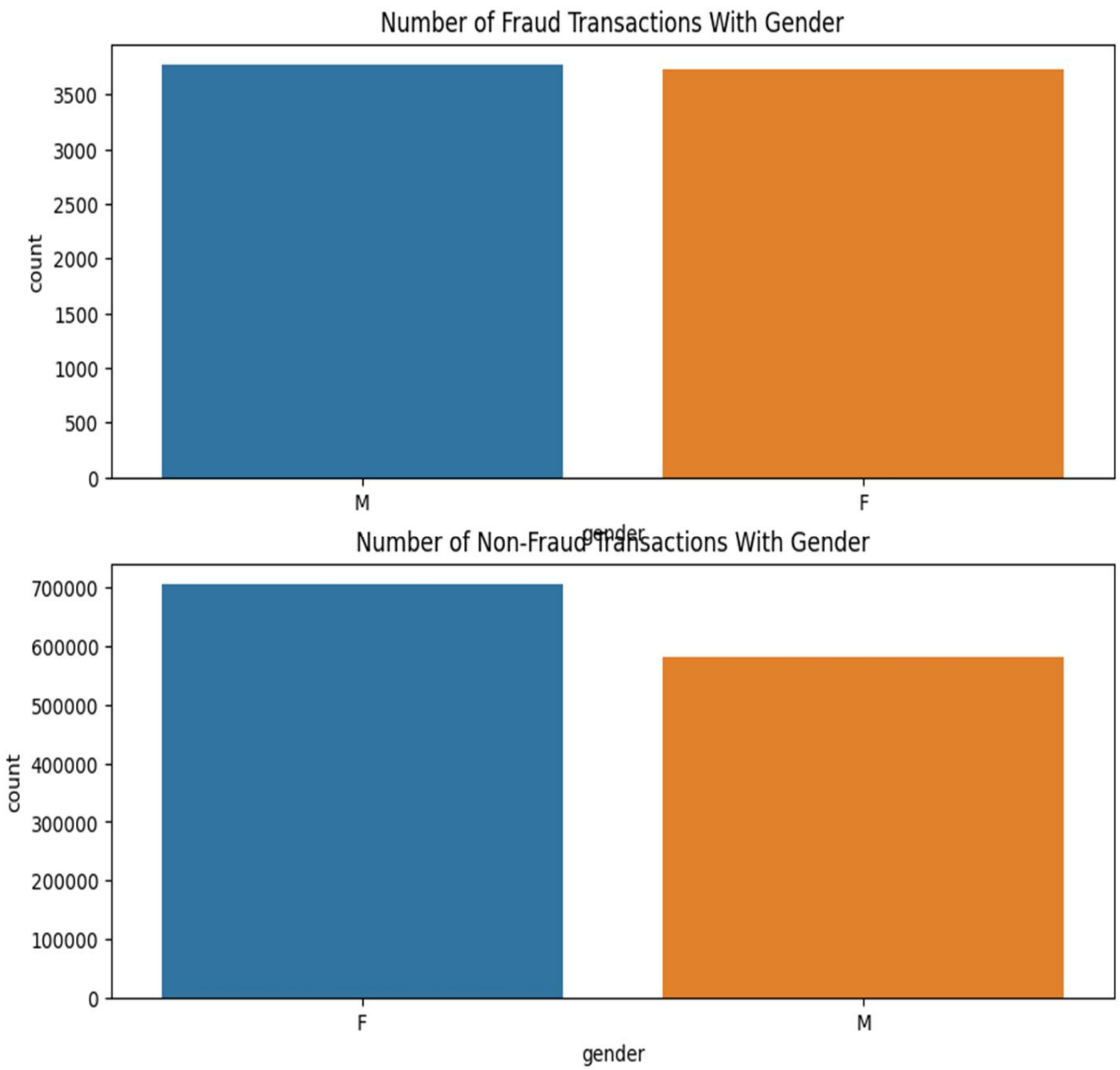
INSIGHTS ON GENDER COLUMN

SUMMARY

Number of Transactions Involving both sexes are almost equal so monitoring is required for both Sexes. However, Women are More involved in fraud Transactions Compared to Men.

INFERENCE

Women Customers need to be educated more for Fraud Transactions.



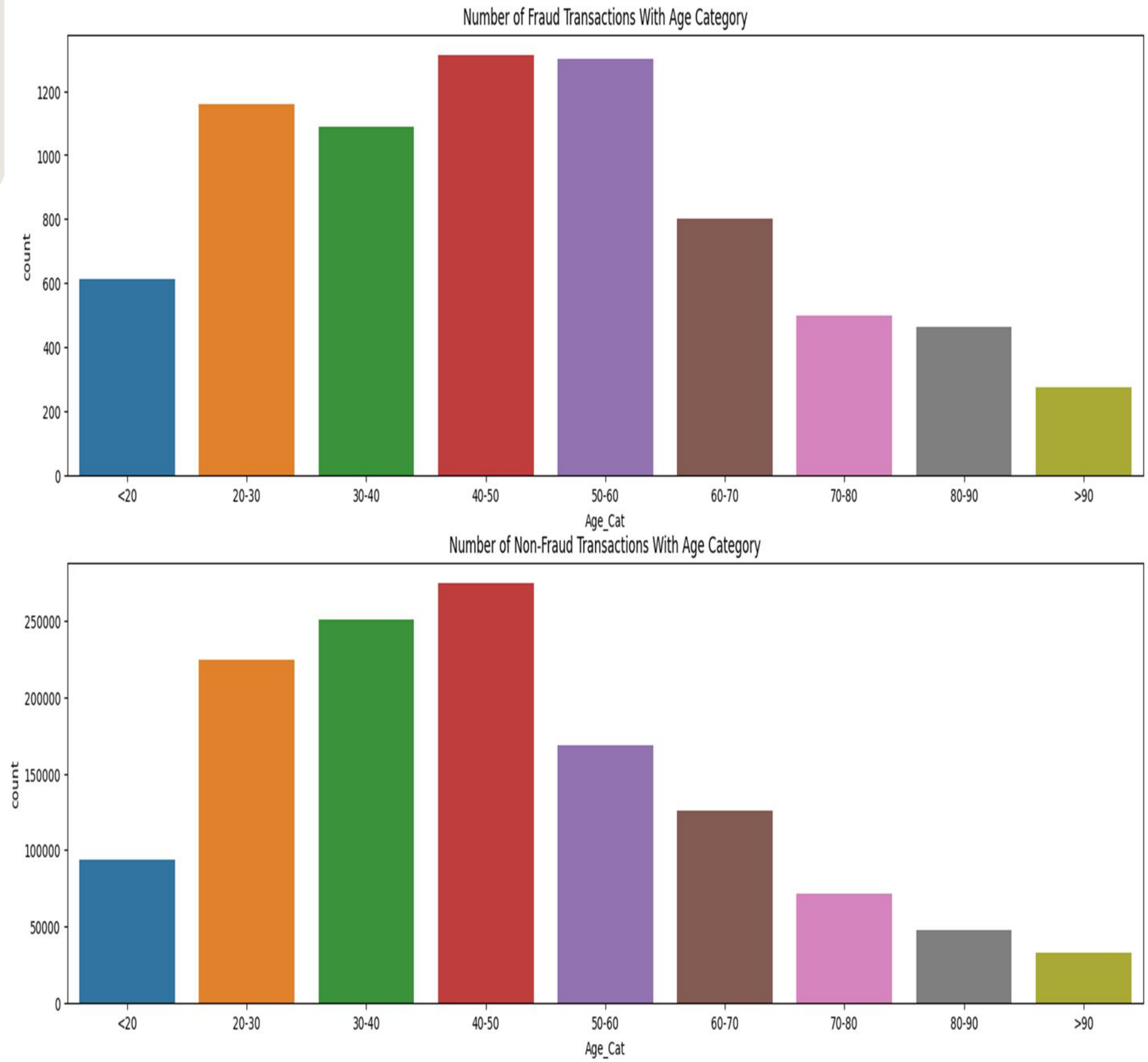
INSIGHTS ON AGE COLUMN

SUMMARY

People of Age Group 40-50 and 50-60 are more victims of Fraud Transactions Compared to all other Age Groups.

INFERENCE

Close Survelience is required for Customers with Age Group 40-50 and 50-60 as Fraud Transactions for these customers Is more as compared to other groups



INSIGHTS ON AGE COLUMN

SUMMARY

People of Age group 40-50 and 50-60 are more involved in Fraud Transactions. But In terms of Fraud Percentage Compared to Total Transactions done People with Age Group 80-90 and above 90 are having more Fraud percentage.

INFERENCE

Old Age Customers need to be educated more regarding Credit Card Fraud Transactions and Close Survelience is required for these Customers.

Age_Cat	Total Transactions	Fraud Transactions	Fraud_Percentage
<20	94520	611	0.646424
20-30	225516	1158	0.513489
30-40	252442	1090	0.431782
40-50	275871	1311	0.475222
50-60	169649	1300	0.766288
60-70	126387	800	0.632976
70-80	71707	500	0.697282
80-90	47801	463	0.968599
>90	32782	273	0.832774

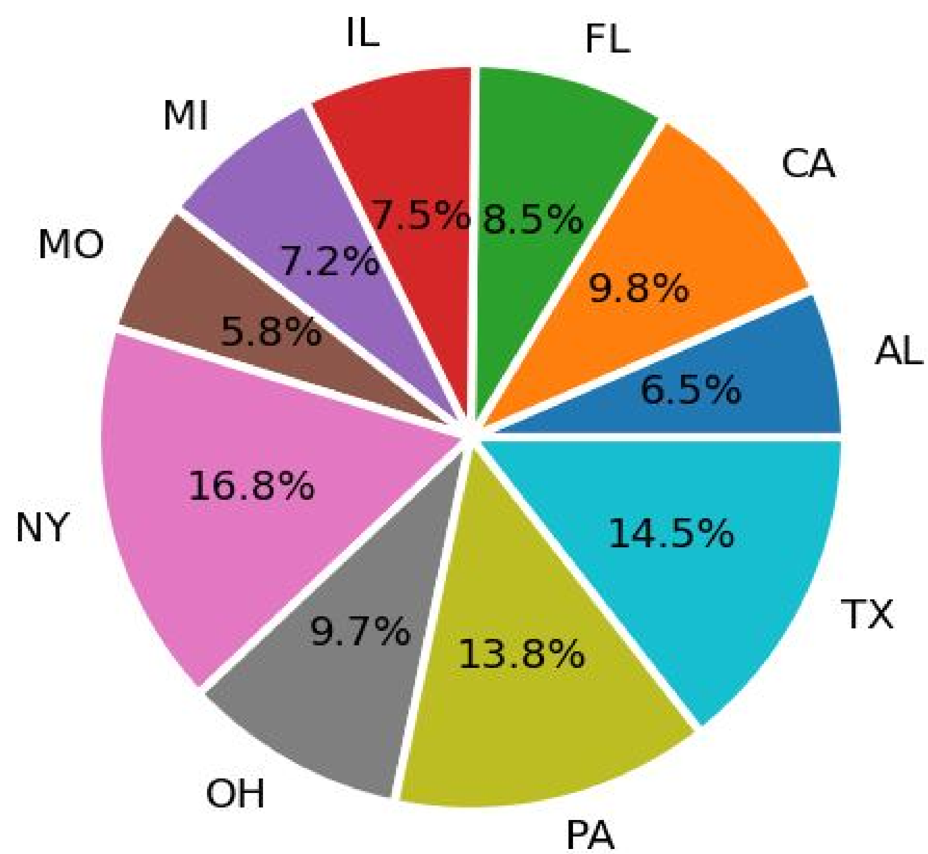
INSIGHTS ON STATE COLUMN

SUMMARY

Most Number of Fraud Transactions happened in the NY,PA,TX States.

INFERENCE

Close Survelience is required in NY,PA,TX States for Fraud Detection.



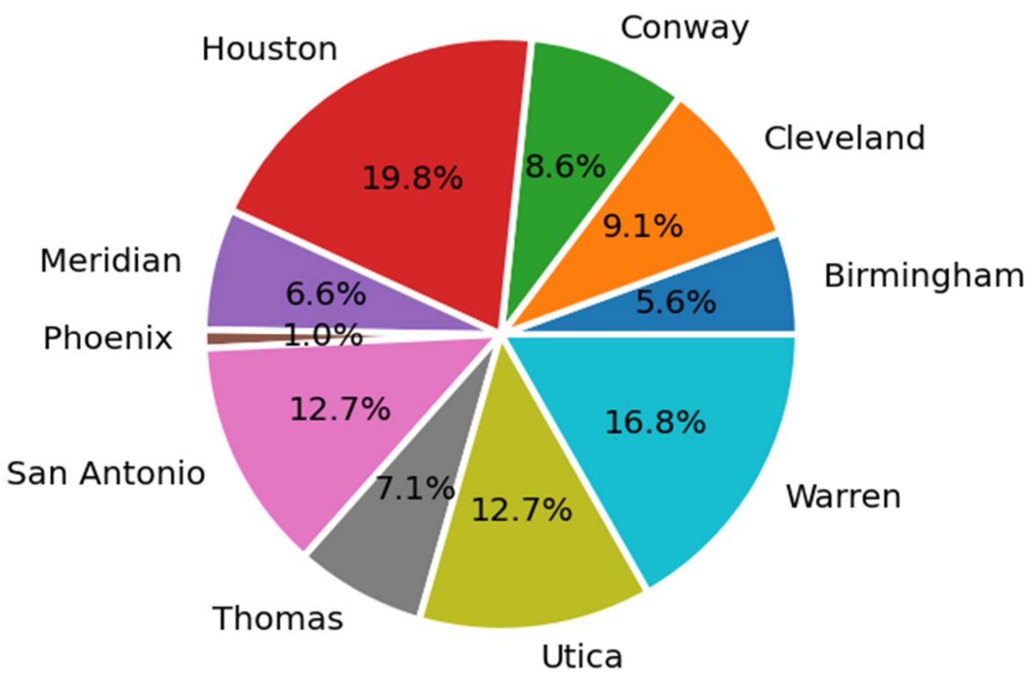
INSIGHTS ON CITY COLUMN

SUMMARY

Most Number of Fraud Transactions happened in the Houston,Warren,Utica.

INFERENCE

Close Survelience is required in Houston,Warren,Utica.



INSIGHTS ON CITY COLUMN

SUMMARY

Warren,Utica,San Antonio has high frequency of Fraud Transaction

INFERENCE

Warren,Utica,San Antonio require close Monitoring for Credit Card Fraud Detection.

city	Fraud_Transactions	Total Transactions	Fraud_Percentage
Birmingham	11	5617	0.195834
Cleveland	18	4604	0.390964
Conway	17	4613	0.368524
Houston	39	4168	0.935701
Meridian	13	5060	0.256917
Phoenix	2	5075	0.039409
San Antonio	25	5130	0.487329
Thomas	14	4634	0.302115
Utica	25	5105	0.489716
Warren	33	4599	0.717547

INSIGHTS ON JOB COLUMN

SUMMARY

Most Number of Fraud Transactions happened in the Materials Engineer, Naval architect,Exhibition Designer Job Categories.

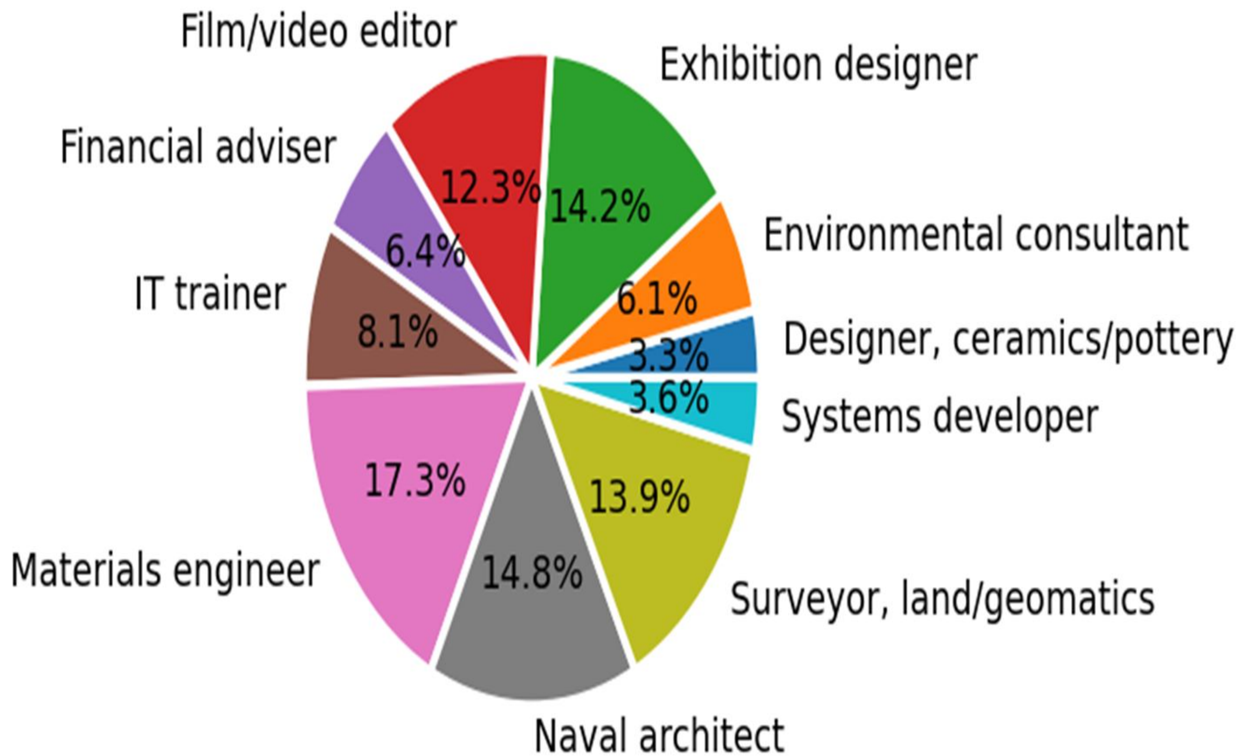
INFERENCE

Close Survelience is required for Materials engineer,

Naval architect, Exhibition Designer Job Categories for Credit

Card Fraud Detection.

,



INSIGHTS ON CATEGORY COLUMN

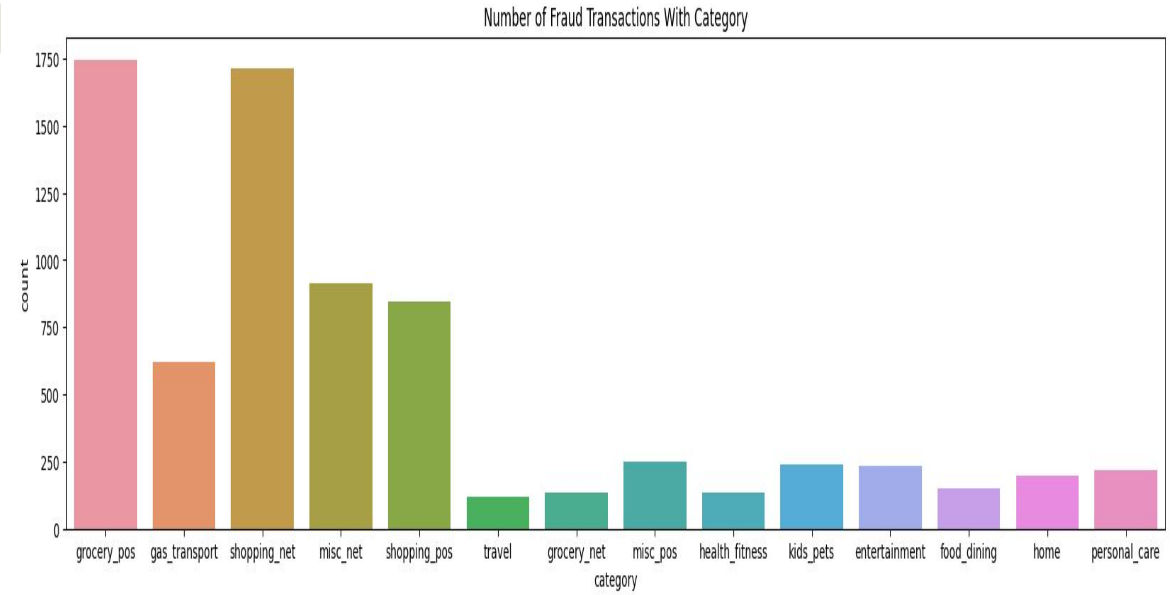
SUMMARY

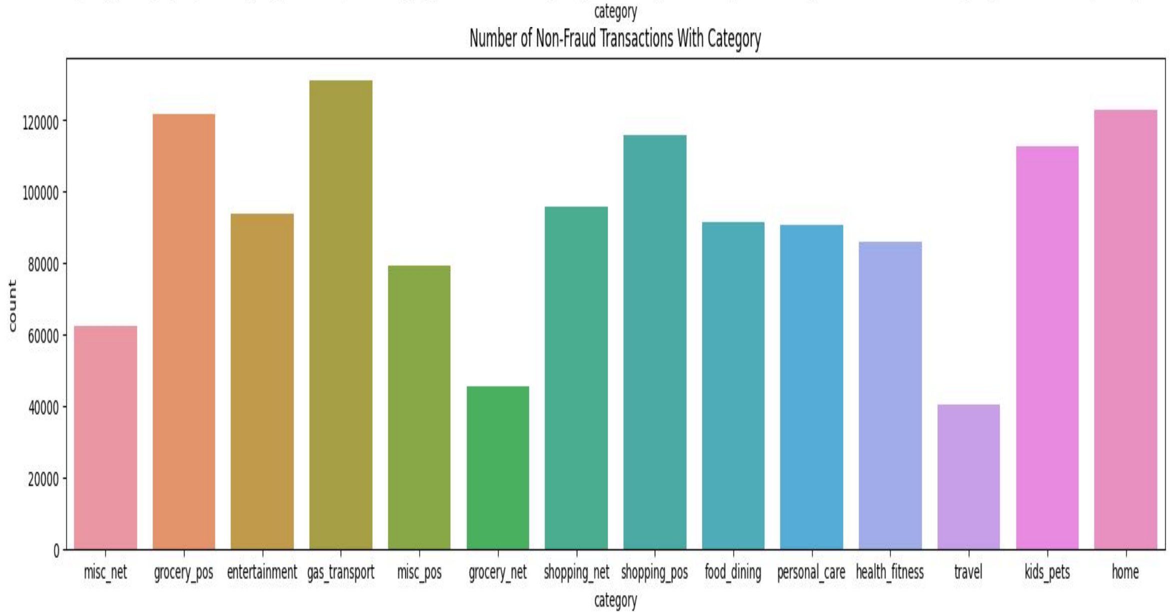
Most Number of Fraud Transactions happened in the grocery_pos ,shopping_net , misc_net Categories.

INFERENCE

Close Survelience is required for grocery_pos,shopping_net,misc_net Categories for Fraud Detection.

,





INSIGHTS ON DISTANCE COLUMN

SUMMARY

Most Number of Fraud Transactions happened when distance between Customer and Merchant is between 77-102 Km,51-77 Km

102-127 km.

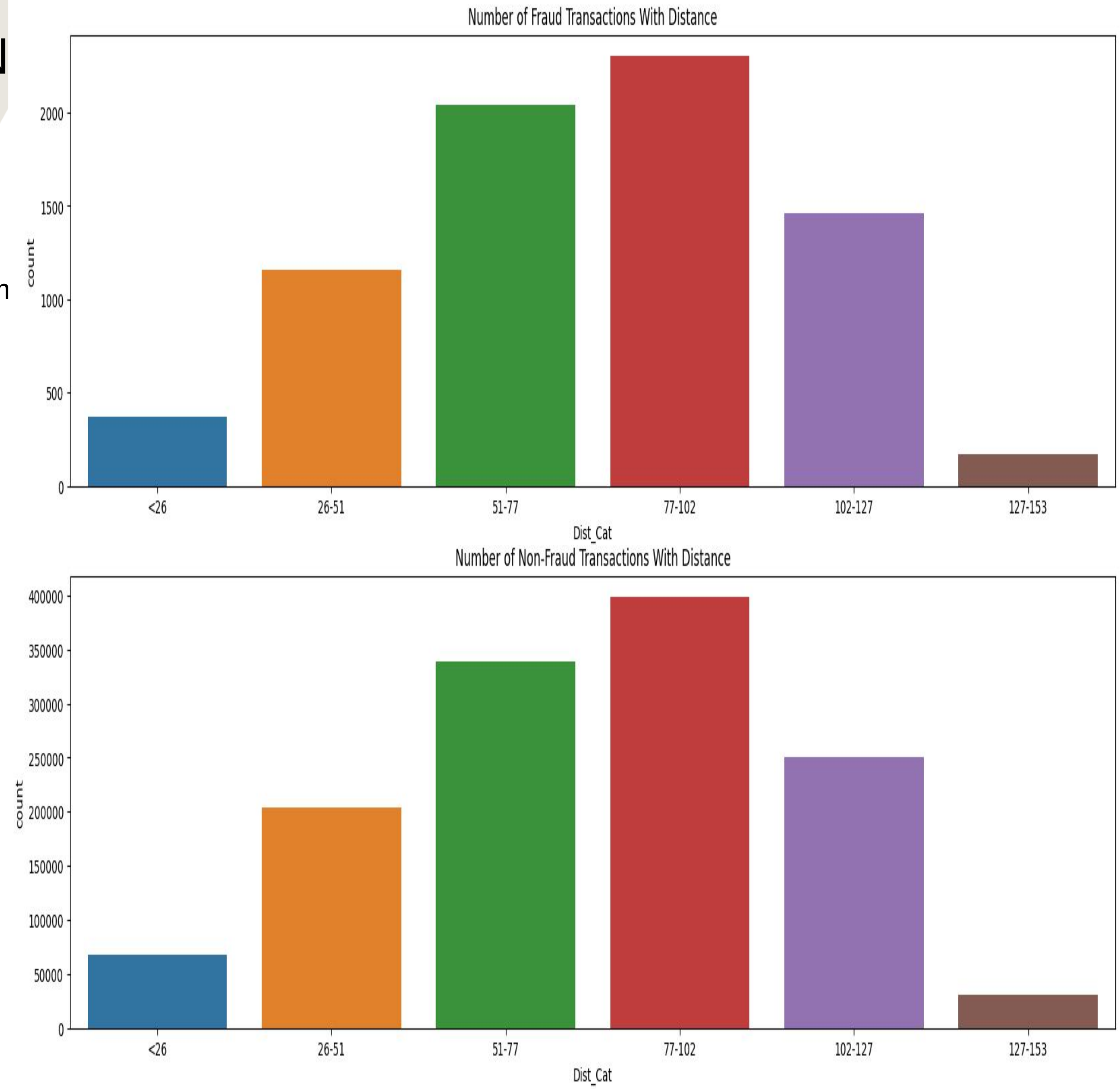
INFERENCE

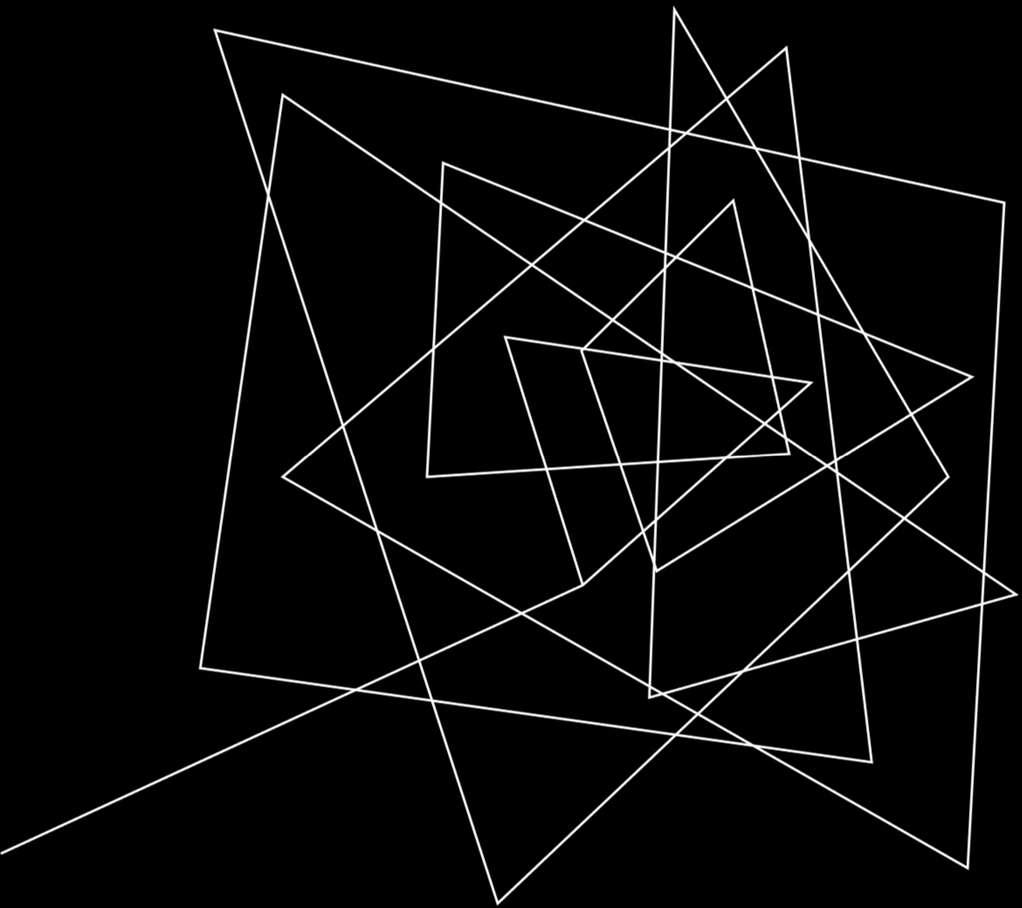
Close Survelience is required for transaction in which Distance

Between customer location and Merchant Location is between

77-102 Km,51-77 Km,102-127 Km.

,





MODEL BUILDING

MODEL BUILDING

1.Feature Encoding done for the Non-Numerical Variables.

2. Train and Test Data Prepared after Data Cleaning and Feature Encoding.

3. Scaling of Variables done for train and Test DataSet.

4. Since the Data is imbalanced, we used different imbalance techniques for Sampling so that Model will not overfit and give us correct results.

5. Credit Card Fraud Detection is the Classification Model. So, We used Logistic Regression, Decision Tree Classification, Random Forest Algorithms for building Models and Compared the results for different models.

6. We used Recall as Performance matrix for this Model because Cost of False Negative is more for the Application. If Model is predicting Fraud Transaction as Non-Fraud than it will be Costly for the Bank.

7. We Compared Different Models and Finally Choose Random Forest with SMOTE Technique as Final Model for Predicting Credit Card Fraud.

8. Finally we did the Cost Benefit Analysis after deployment of Model and found that bank will save Money after deployment of Model

COMPARISON OF MODELS

SUMMARY

Recall Value With Model Random Forest with RandomUnderSampling is 0.96 and Random Forest with SMOTE is 0.90.

INFERENCE

We choose Random Forest with SMOTE as our Final Model because RandomUnderSampling will lead to loss of Information while SMOTE uses neighbours for assigning new Value in the DataSet.

So, We used the Random Forest with SMOTE as Final Model and finally done the HyperParameter Tuning of the Model.

	Model_Name	Train Accuracy	Test Accuracy	Precision	Recall	F-1 Score
0	Logistic Regression With Random UnderSampling	0.818345	0.877407	0.023017	0.742191	0.044649
1	Logistic Regression With Random OverSampling	0.817793	0.878125	0.023192	0.743590	0.044982
2	Logistic Regression With SMOTE	0.829266	0.892939	0.026471	0.747319	0.051131
3	Logistic Regression With ADASYN	0.757107	0.759015	0.012893	0.813054	0.025384
4	Decision Tree With Random Undersampling	1.000000	0.972250	0.118593	0.962238	0.211162
5	Decision Tree With Random Oversampling	1.000000	0.998373	0.787935	0.791608	0.789767
6	Decision Tree With SMOTE	1.000000	0.901664	0.033945	0.891375	0.065400
7	Decision Tree With ADASYN	1.000000	0.882534	0.028795	0.899301	0.055803
8	Random Forest With RandomUnderSampling	1.000000	0.979628	0.155349	0.964103	0.267581
9	Random Forest With Random OverrSampling	1.000000	0.998872	0.902439	0.793473	0.844455
10	Random Forest With SMOTE	1.000000	0.982770	0.171297	0.902564	0.287945
11	Random Forest With ADASYN	1.000000	0.981482	0.160187	0.895105	0.271743

FINAL MODEL

SUMMARY

After doing HyperParameter Tuning We found following Parameters

For the Random Forest Model.

Max_depth=15

Max_features=20

Min_samples_leaf=11

INFERENCE

We choose Random Forest with SMOTE as our Final Model with Hyper Parameter Tuning and found Recall Value as 0.9547. We Choose this model for Final Prediction of Credit Card Fraud.

	Model_Name	Train Accuracy	Test Accuracy	Precision	Recall	F-1 Score
0	Random Forest With HyperParameters	0.992267	0.929261	0.049634	0.954779	0.094363

```
rf_best=RandomForestClassifier(max_depth=15,max_features=20,min_samples_leaf=11)
```

IMPORTANT VARIABLES FOR THE MODEL

SUMMARY

In the Image We Can find Important Variables with their Feature Importance for the Model.

INFERENCE

Amount of Transaction, hour at which Transaction happened,Gas Transport Category Transactions, Age of Customer, food Dining Category Transactions , Travel Category Transactions are Important Variables for Model.

	Varname	Imp
0	amt	0.688717
3	hour	0.151241
6	category_gas_transport	0.040229
2	age	0.023131
5	category_food_dining	0.011431
17	category_travel	0.010519
7	category_grocery_net	0.010387
13	category_misc_pos	0.009844
10	category_home	0.008271
1	city_pop	0.007858
8	category_grocery_pos	0.006362
15	category_shopping_net	0.006225
14	category_personal_care	0.005585
16	category_shopping_pos	0.004118

COST BENEFIT ANALYSIS

COMPARISON

$161568

COST BEFORE MODEL

Bank was losing this avg amount per month of money before deployment of Model

$9894

COST AFTER MODEL

Avg Cost incurred per month after deploying the Model for Fraud Detection

$151674

TOTAL SAVING

Total Average Saving Per Month After Deployment of Model

DETAILED COST BENEFIT ANALYSIS

Cost Benefit Analysis		
S. No	**Questions**	**Answer**
a	Average number of transactions per month	79388
b	Average number of fraudulent transaction per month	306
c	Average amount per fraud transaction	528

Cost Benefit Analysis		
S. No	**Questions**	**Answer**
1	Cost incurred per month before the model was deployed (b*c)	161568
2	Average number of transactions per month detected as fraudulent by the model (TF)	5894
3	Cost of providing customer executive support per fraudulent transaction detected by the model	$1.5
4	Total cost of providing customer support per month for fraudulent transactions detected by the model (TF*$1.5)	8841
5	Average number of transactions per month that are fraudulent but not detected by the model (FN)	13
6	Cost incurred due to fraudulent transactions left undetected by the model (FN*c)	1053
7	Cost incurred per month after the model is built and deployed (4+6)	9894
8	Final savings = Cost incurred before - Cost incurred after(1-7)	151674

MODEL BENEFITS

Cost Saving

Early Intimation to Customers regarding Fraud

Trust Improvement among Customers for Company

Improvement in Brand Value of Company

STRATEGY FOR FRAUD DETECTION

ODD HOUR SURVEILLANCE

Most of the Fraud Transactions happened during odd hours. So, Close Surveillance required during odd hours

AMOUNT OF TRANSACTION

Small Amount Transactions also need to be monitored Closely.

AGE OF CUSTOMERS

Old Age Customers are more Susceptible to Fraud so close Surveillance required for old Age People.

CATEGORY OF PURCHASE

Gas Transport Category, Grocery Category Online, Food Dining, Transport Category are having more Fraud Transactions. These Need to be monitored Closely.

DISTANCE BETWEEN MERCHANT AND CUSTOMER LOCATION

Distance also need to be monitored closely. Fraud Transactions are happening at large Distance from Customer Location.

SUMMARY

Credit Card Fraud Detection Model is providing Cost Benefit to the Company. It is providing Early Detection of Fraud Transaction which is saving money to the company and improving the Customer Trust.

Model is providing Important Feature Variables which need to be monitored Closely.

Identification of Right Customer Base which are susceptible to Fraud is also helpful. Customers Can be educated and made aware about the Fraud which will lead to less Fraud Transaction and Customer Awareness

THANK YOU

Himanshu Shukla

himanshushukla61@gmail.com