Credit Card Fraud Detection Capstone Project

Ujwal kesharwani

Manish Dabhade

Kyatham Nikhil

Introduction

Following are the goals of this Credit Fraud Detection Model

To build the most accurate model to detect the maximum credit card fraud transaction. We will also detect the different type of fraud transaction and their trends

- Also we will performing the cost benefit analysis to check for final savings with the help of cost incurred before and after model building.

Problem Statement

To help Finex the detect fraud transaction and business impact of these fraud transaction

Build the most accurate model to detect the maximum credit card fraud transaction so as to reduce the fraud transaction

Identify the driver variables and understand their significance which are strong indicators of fraud transaction

Identify the outliers, if any, in the dataset and justify the same

Check and fix the imbalance and skewness in the data

Consider both technical and business aspects while building the model

Summarize the fraud detection predictions by using evaluation metrics like accuracy, sensitivity, specificity and precision. And also perform the cost benefit analysis check business impact of the fraud transactions

Business Goal

01

Finex company want to develop a machine learning model to detect fraudulent transactions based on the historical transactional data of customers with a pool of merchants.

02

Finex company to analyze the business impact of these fraudulent transactions and recommend the optimal ways that the bank can adopt to mitigate the fraud risks.

03

The company want to know the benefit of the model using cost benefit analysis

Solution Methodology : Data Exploration

fraudTrain.csv & fraudTest.csv contains all the information about the fraud transaction generated through various sources and their activities

The train file contains 1296675 rows and 23 columns

The test file contains 555719 rows and 23 columns

Out of 23 columns, 10 are numeric columns and 13 are non-numeric or categorical columns

We merged the train and test data into a single file credit_fraud.csv for model building.

The merged dataset contains 1852394 rows and 22 columns.
Out of 22 columns, 10 are numeric columns and 12 are non-numeric or categorical columns

Solution Methodology : Data Cleaning and Preparation

```
--  ------                 -----
0   trans_date_trans_time  object
1   cc_num                 int64
2   merchant               object
3   category               object
4   amt                    float64
5   first                  object
6   last                   object
7   gender                 object
8   street                 object
9   city                   object
10  state                  object
11  zip                    int64
12  lat                    float64
13  long                   float64
14  city_pop               int64
15  job                    object
16  dob                    object
17  trans_num              object
18  unix_time              int64
19  merch_lat              float64
20  merch_long             float64
```

- Check for the shape and datatypes in the dataset.
  credit_fraud.csv:
- Check for the null values.
- Convert the incorrect datatypes to correct datatypes for better results.
- Binning the columns for better analysis.
- Creation of new column
- Drop columns that are not useful for the analysis.
- Check and handle outliers in the dataset.

Solution Methodology : Data Cleaning and Preparation

Check for the null values

Few columns that are not required will delete those.

Check and handle outliers in data.

Solution Methodology : EDA and Findings

Univariate data analysis: value count, distribution of variable etc.

Bivariate data analysis: correlation coefficients and pattern between the variables etC.

Oversampling technique to fix data imbalance

Feature Scaling & Dummy Variables and encoding of the data.

Fixing the skewness in the data

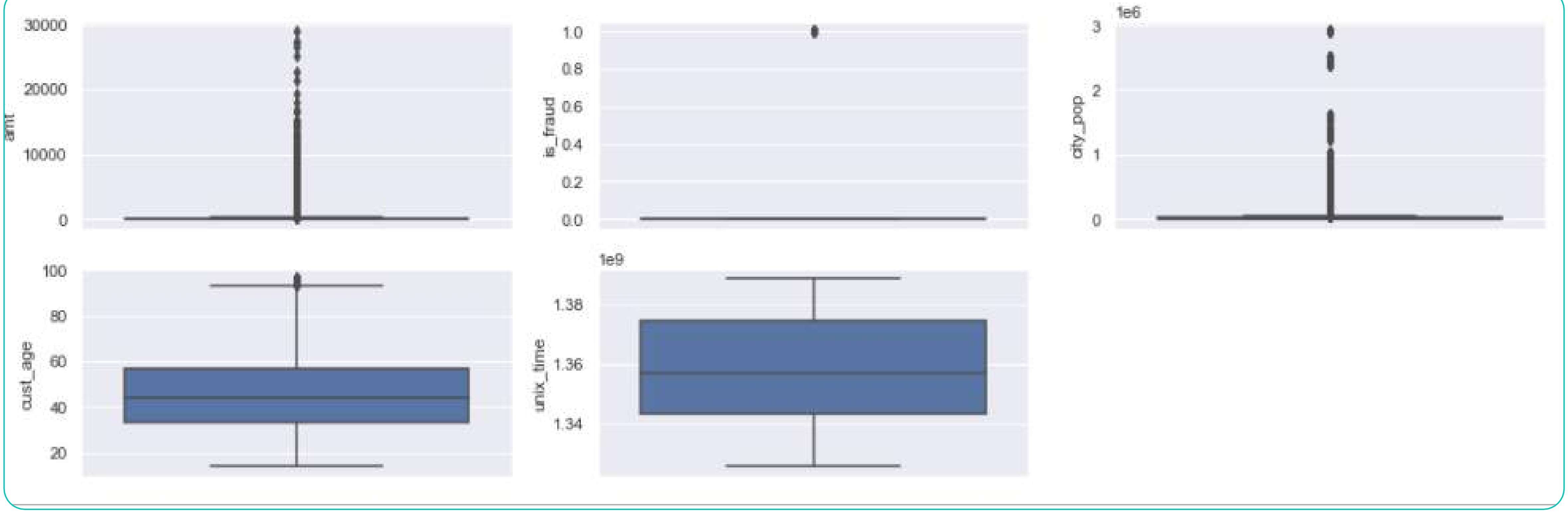
Classification technique: **Logistic Regression, Decision Tree & Random Forest** used for the model making and prediction.

Validation of the model.

Model presentation.

Conclusions and recommendations.

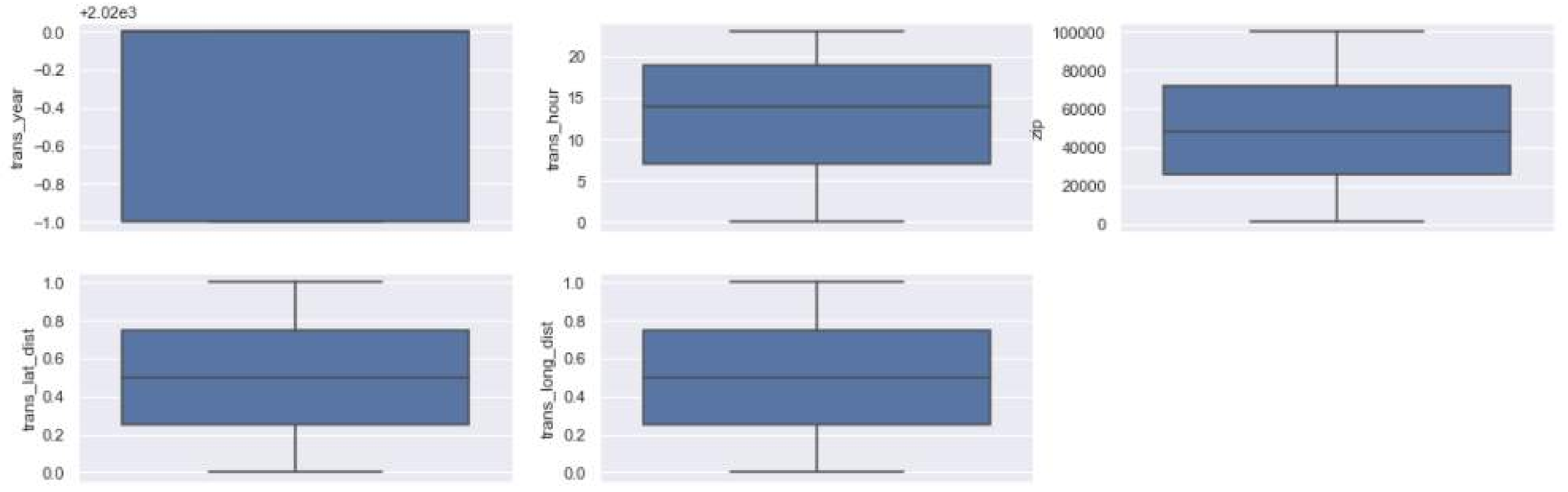
Performing Cost Benefit Analysis



Check the outliers in the data set

Analysis:-

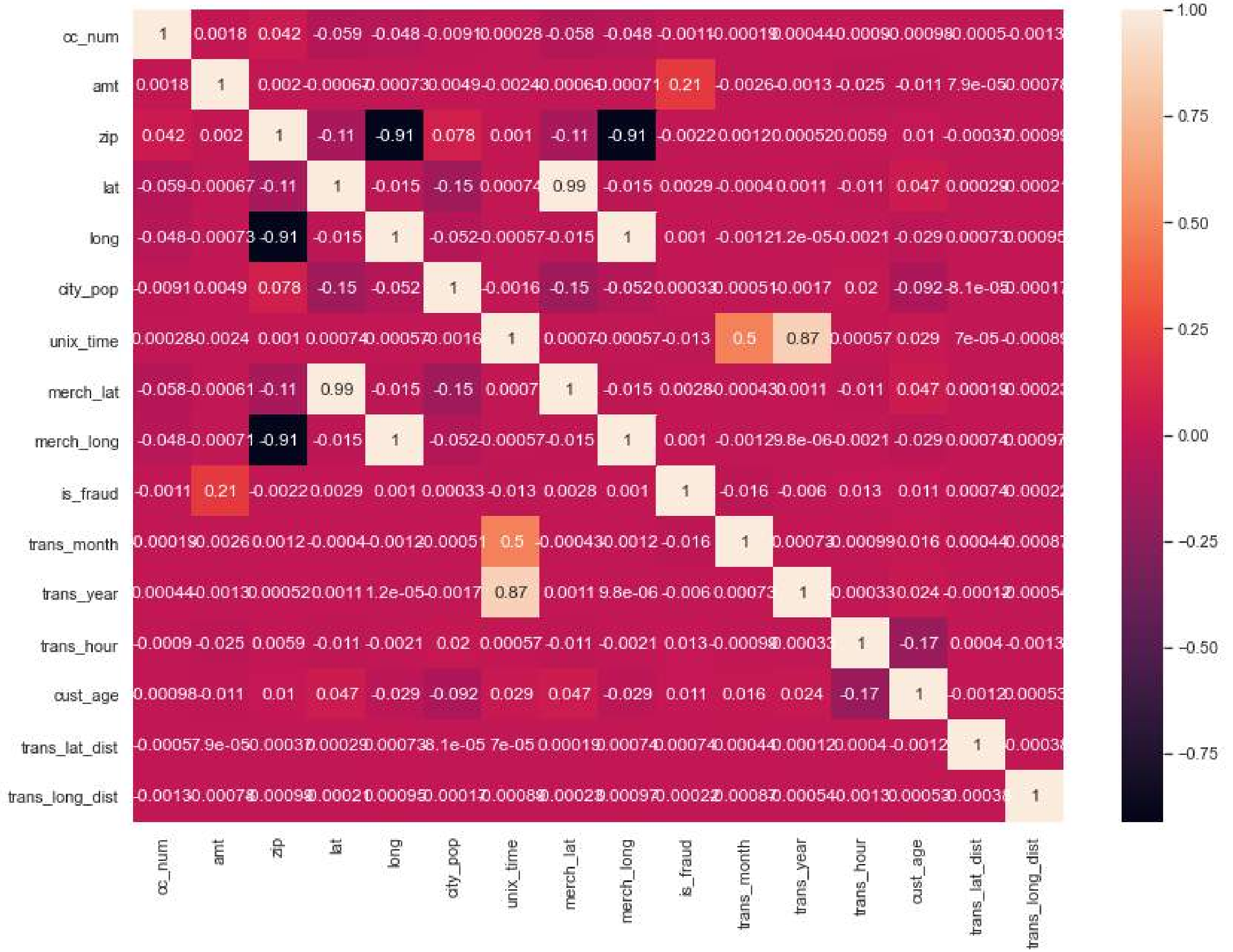
From the boxplot we can infer that There are no outliers in the dataset



Check the outliers in the data set

Analysis:-

From the boxplot we can infer that There are no outliers in the dataset



Visualizing correlation in Dataset

Analysis:

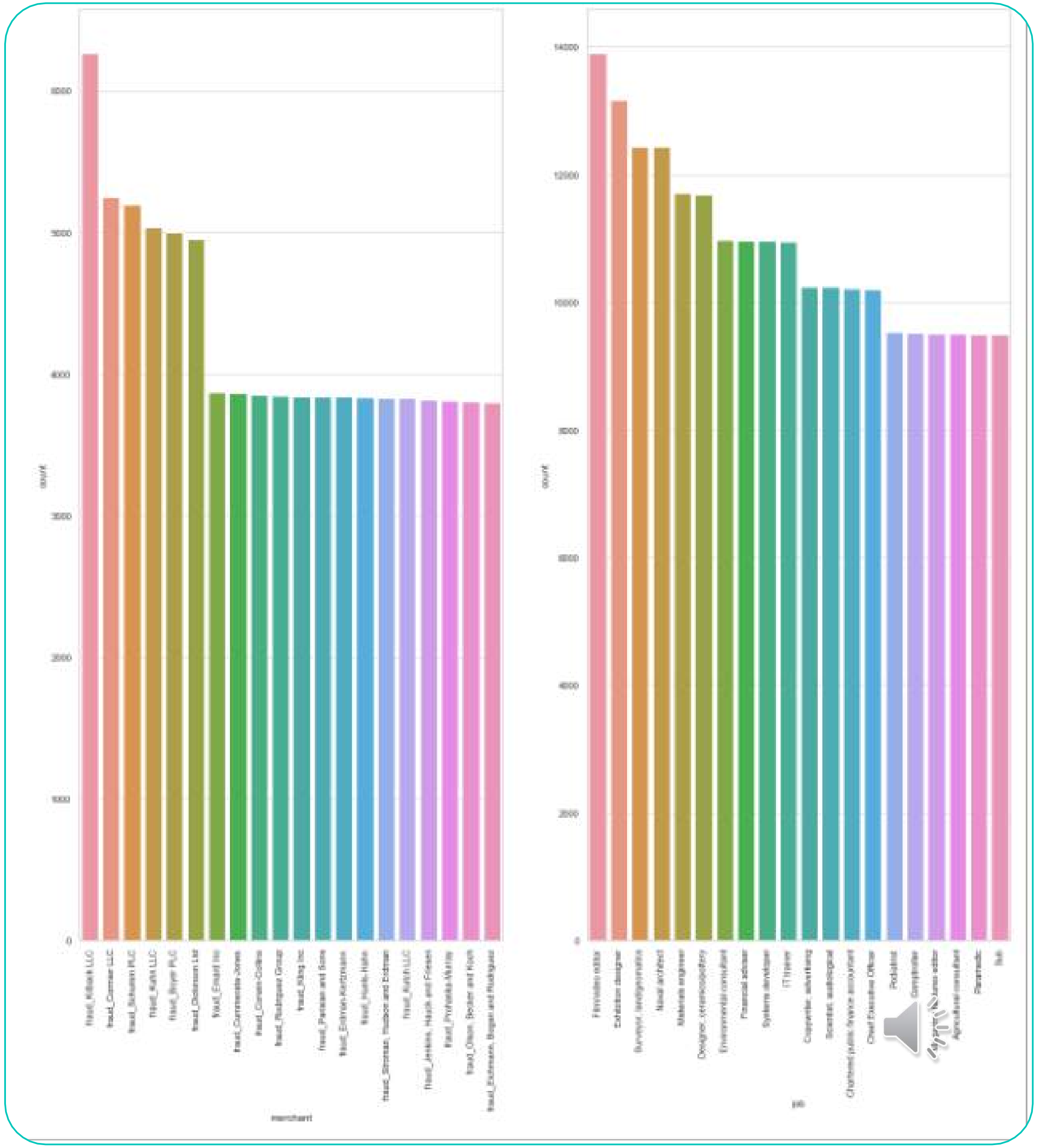
It can be seen that merch_lat and lat, merch_long and long are positively correlated whereas merch_long and zip are negatively correlated. we can remove one of the correlated columns

Univariate Analysis For Categorical Columns and Numerical Columns

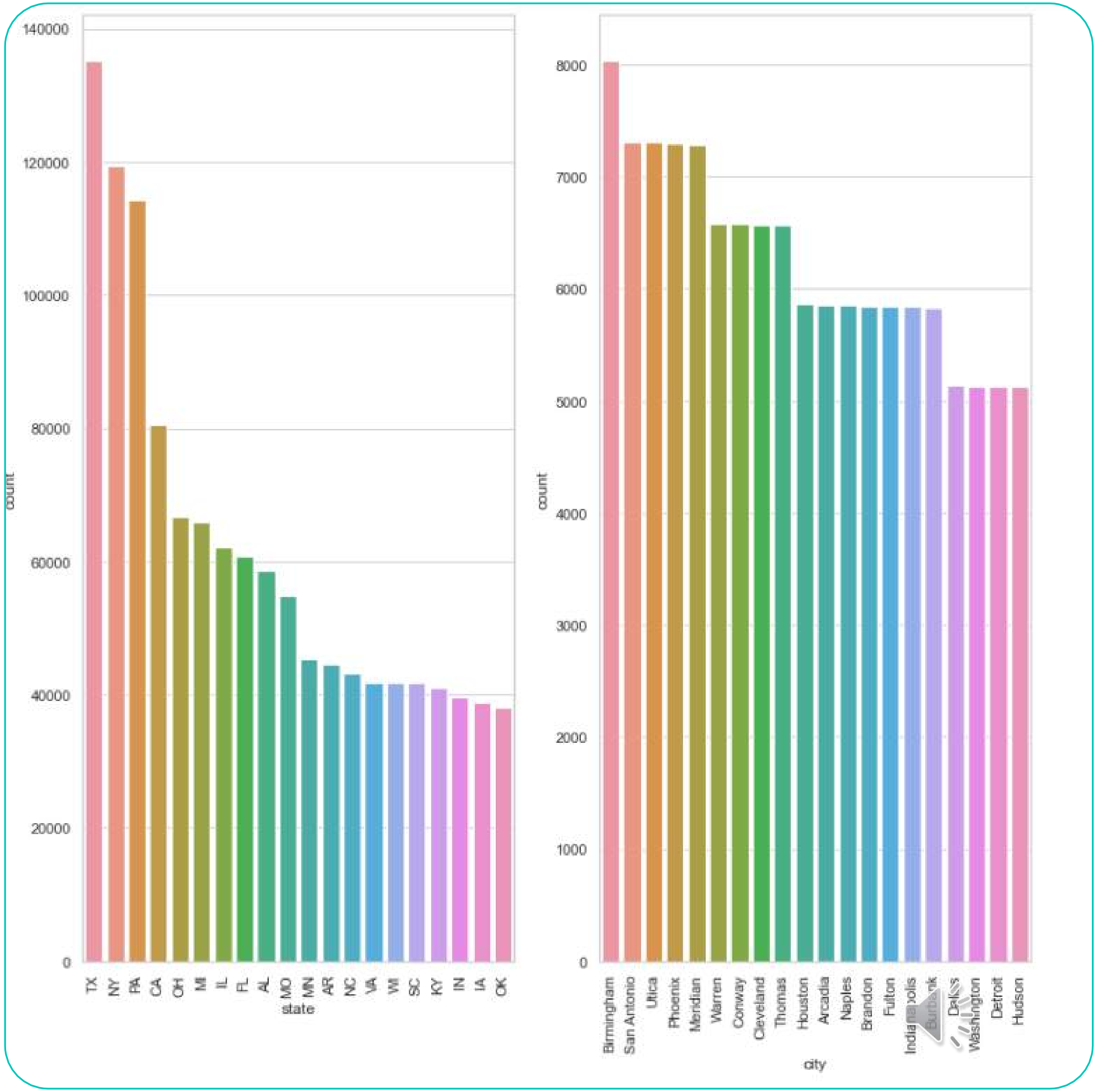
Univariate Analysis for Categorical Columns

Analysis:-

- The merchant fraud_Killback_LLC mostly accepts credit card for transactions.
- The people in Film/Video Editor & Exibition Desinger makes maximum use of transaction via credit card.

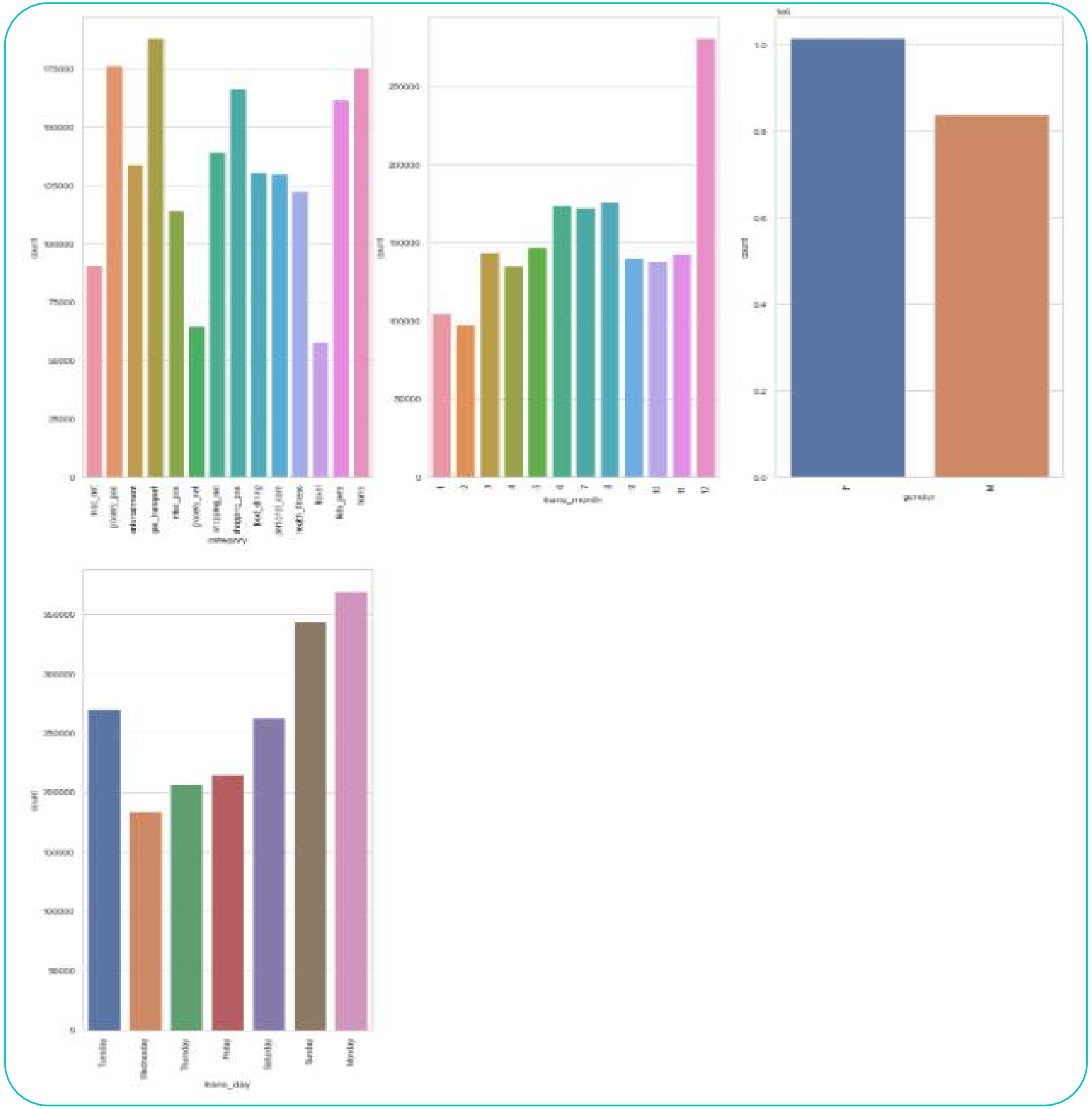


Univariate Analysis for Categorical Columns



__Analysis:-__

- The Texas state use credit card for performing transaction.
- The people living in Birmingham do highest transaction among other cities via credit card.



Univariate Analysis for Categorical Columns

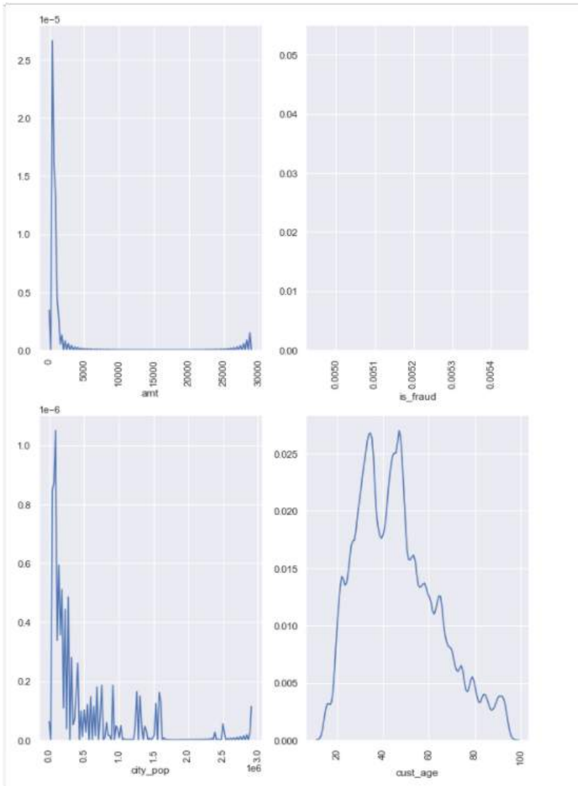
Analysis:-

- Gas_transport and grocery_pos categories are the highest transaction dealing categories.
- The maximum transactions are done on Monday and Sunday.
- The month of December reported the maximum number of transaction.
- The females do more transaction than males.

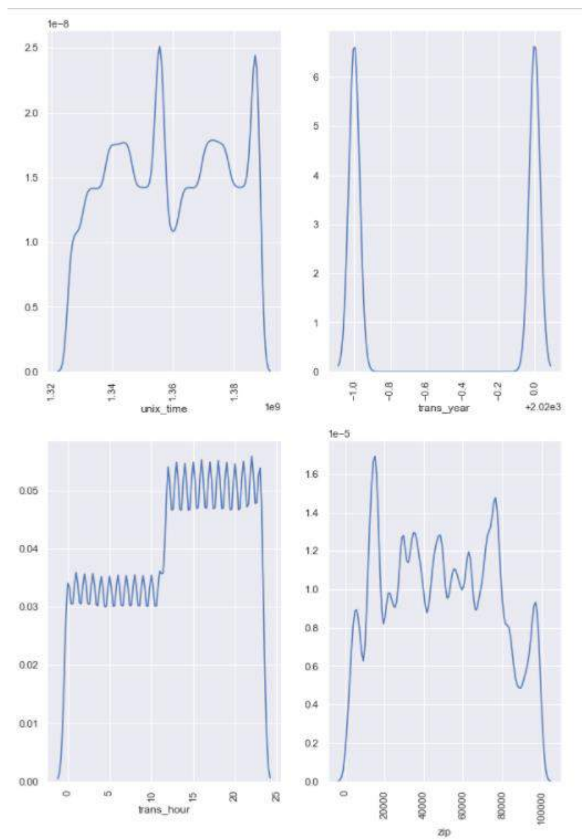
Univariate Analysis for Numeric Columns

***Analysis*:-**

- The most of transaction done by credit card is of 1000 and very few transaction of above 5000
- The people of city with a population of 0.2 are doing maximum transaction via credit card.
- The people with a age of 35 and 45 are doing maximum number of transaction.



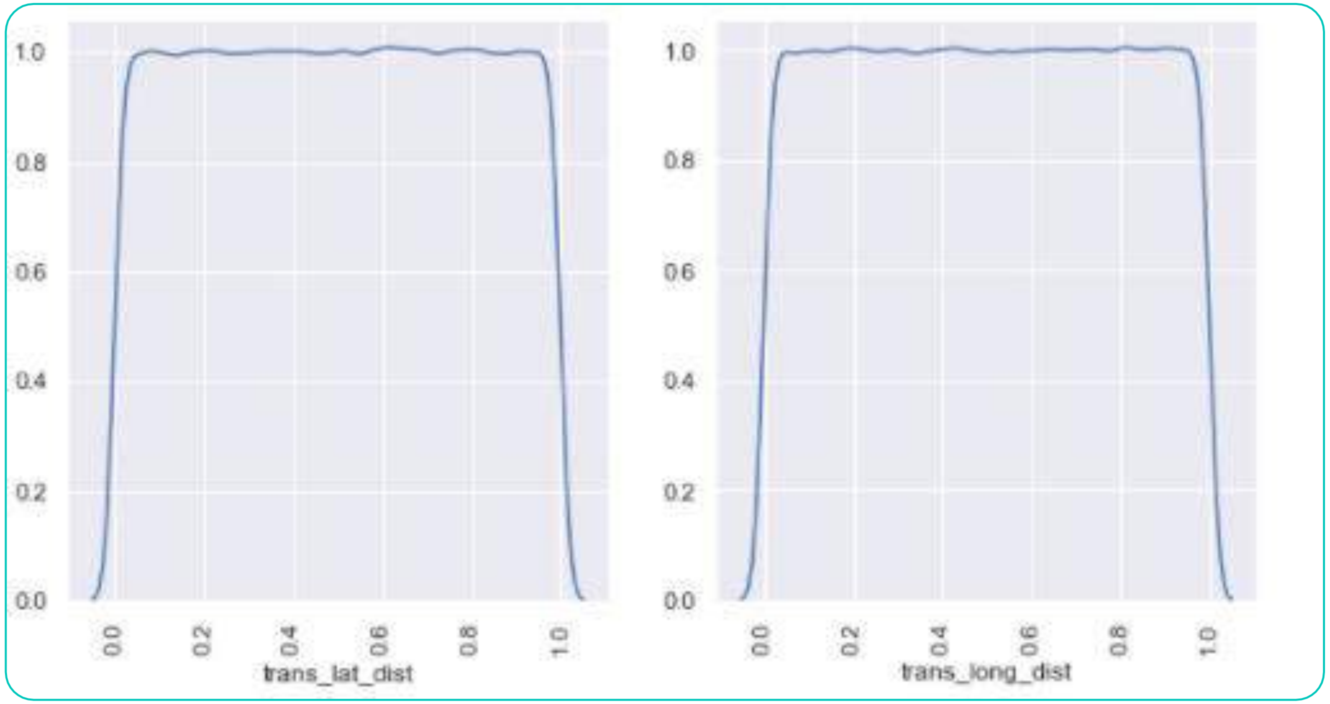
Univariate Analysis for Numeric Columns



***Analysis*:-**

- The most of the transaction are at between unix time between 1.35-1.36 and at 1.39.
- The majority transaction are happening between 11:00 AM - 12:00 Midnight and the minimum transaction are done between 12:00 AM -11:00AM.
- The people do less transaction in the morning and more transaction in afternoon and night.
- The city with the zip code of 17000 has the maximum number of transactions
- There are similar number of transactions in both the years

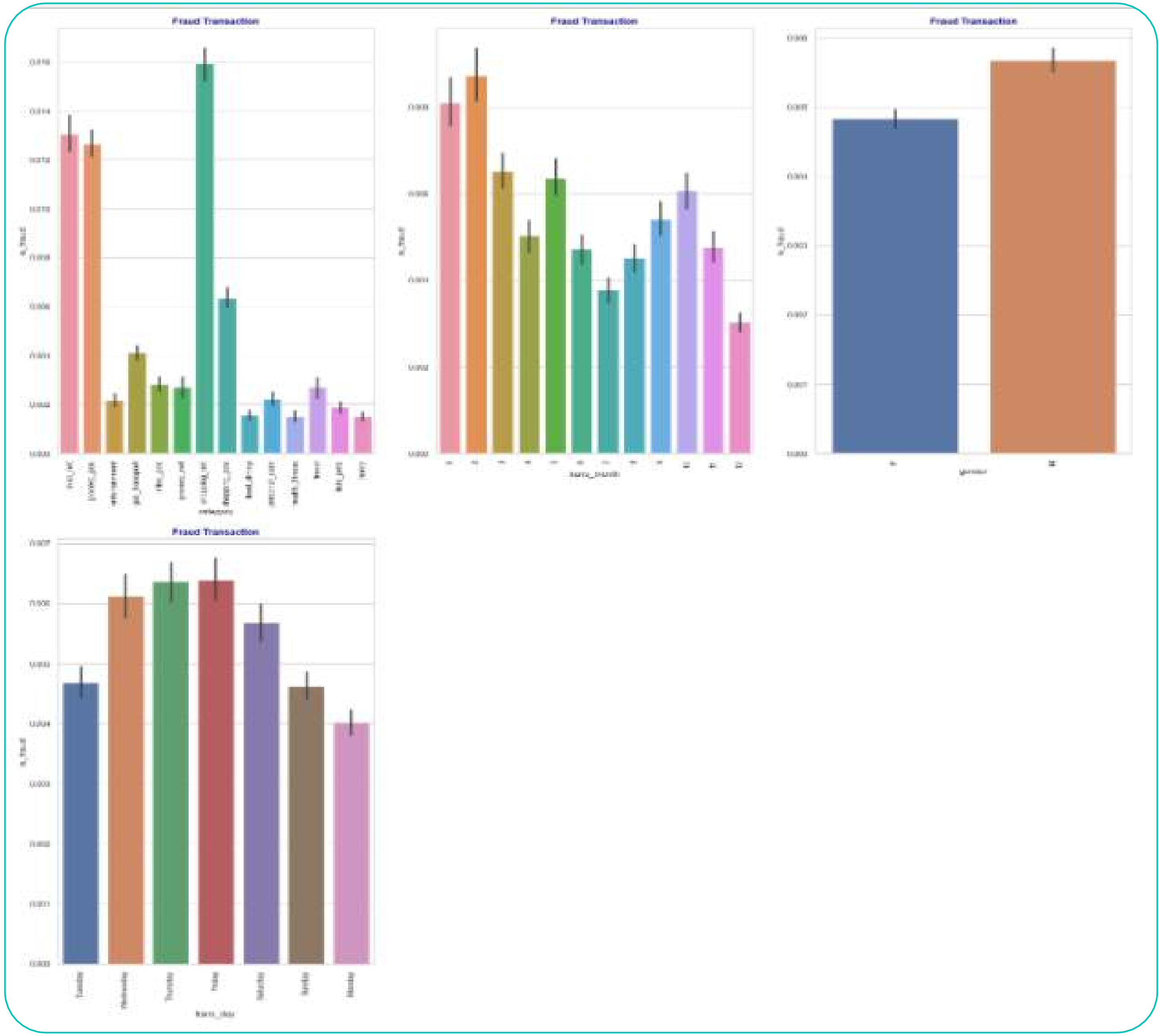
Univariate Analysis for Numeric Columns



***<u>Analysis</u>*:-**

- There are constant number of transaction at a latitude distance between 0.1-1.0
- There are constant number of transaction with a latitude distance between 0.1-1.0

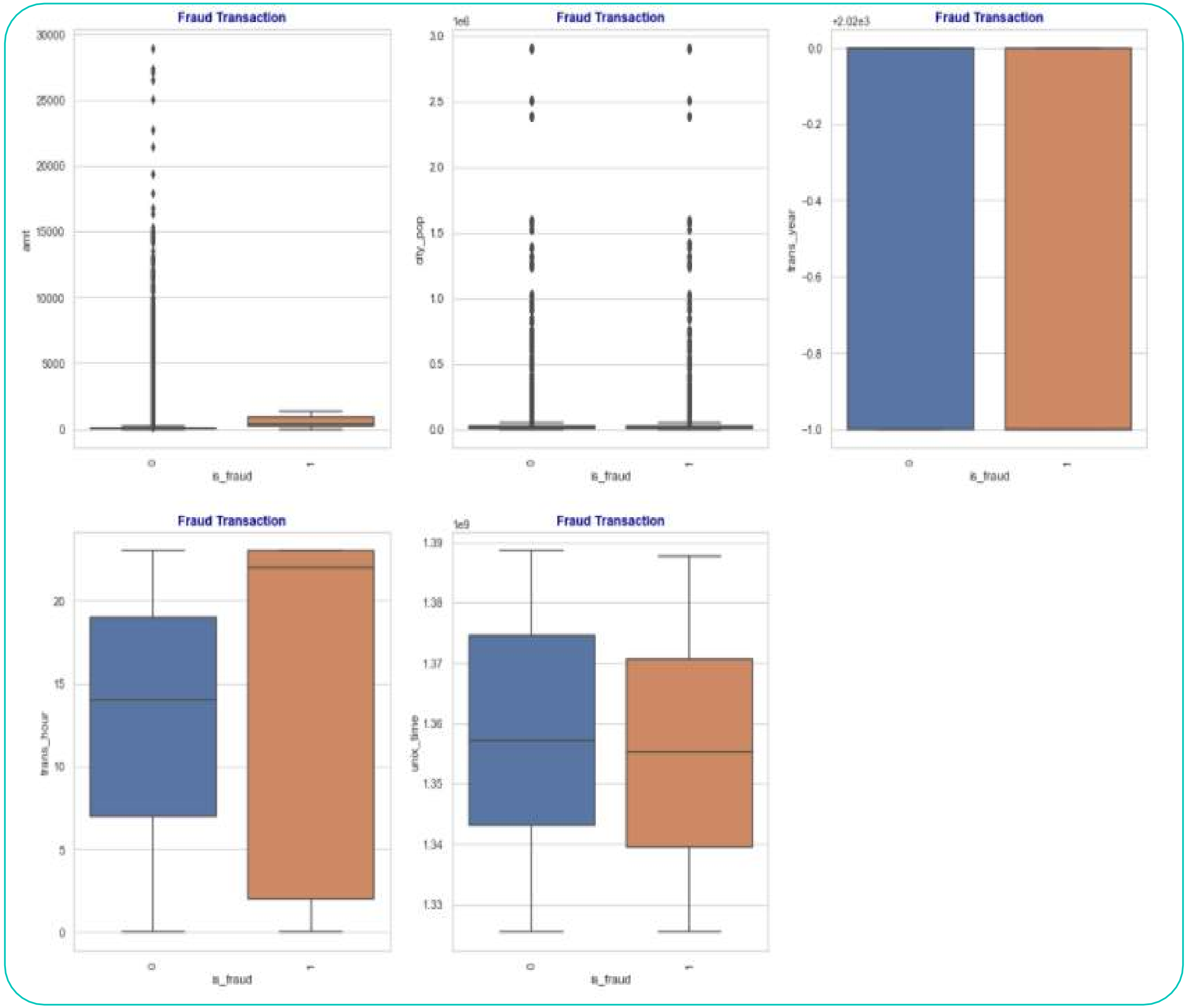
Bivariate Analysis For Categorical and Numeric Variables



Bivariate Analysis For Categorical variables

Analysis:-

- The maximum fraud transaction is done for the category 'shopping_net'
- The month of February has reported the maximum transaction that were fraud.
- The fraud transaction reported for male are more than female.
- The maximum number of fraud transaction are done on Thursday and Friday.



Bivariate Analysis For Numerical variables

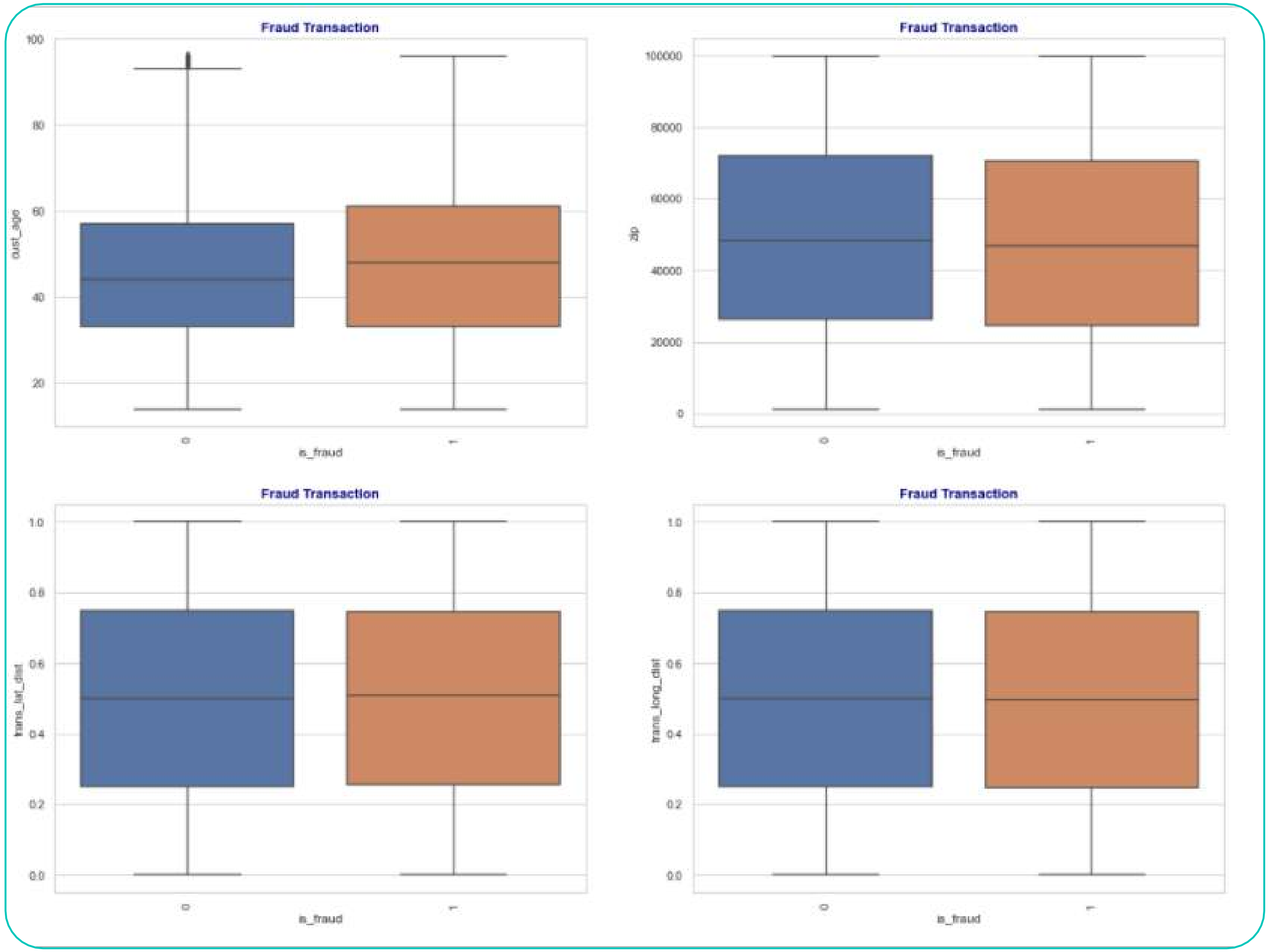
Analysis:-

- The fraud transaction ammount is more than non-fraud transaction.
- The fraud transaction is same in both the year.
- The fraud transaction are more in the early morning and in late night than non-fraud transaction.
- However in the afternoon fraud and non-fraud transaction are same.
- The maximum fraud transaction are reported at unic time 1.34.

Bivariate Analysis For Numerical variables

Analysis:-

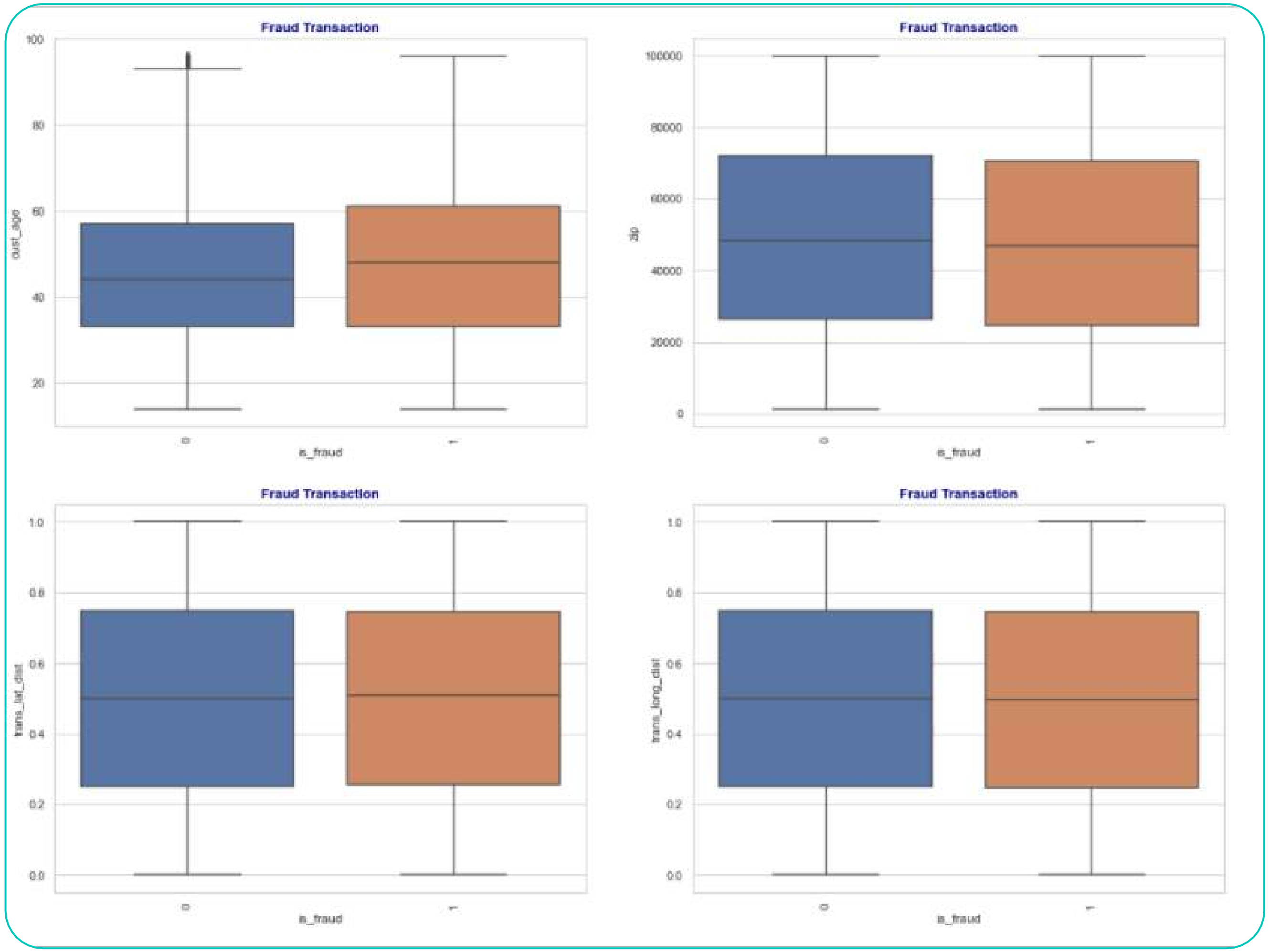
- The fraud transaction and non-transaction are constant with respect too the city population
- The zip of the fraud and non-fraud transaction is similar
- The people with a age of 60 have reported maximum fraud transaction.
- The latitude and longitude distance of transaction and merchant are same for both fraud and non-fraud transaction.



Bivariate Analysis For Numerical variables

Analysis:-

- The fraud transaction and non-transaction are constant with respect too the city population
- The zip of the fraud and non-fraud transaction is similar
- The people with a age of 60 have reported maximum fraud transaction.
- The latitude and longitude distance of transaction and merchant are same for both fraud and non-fraud transaction.

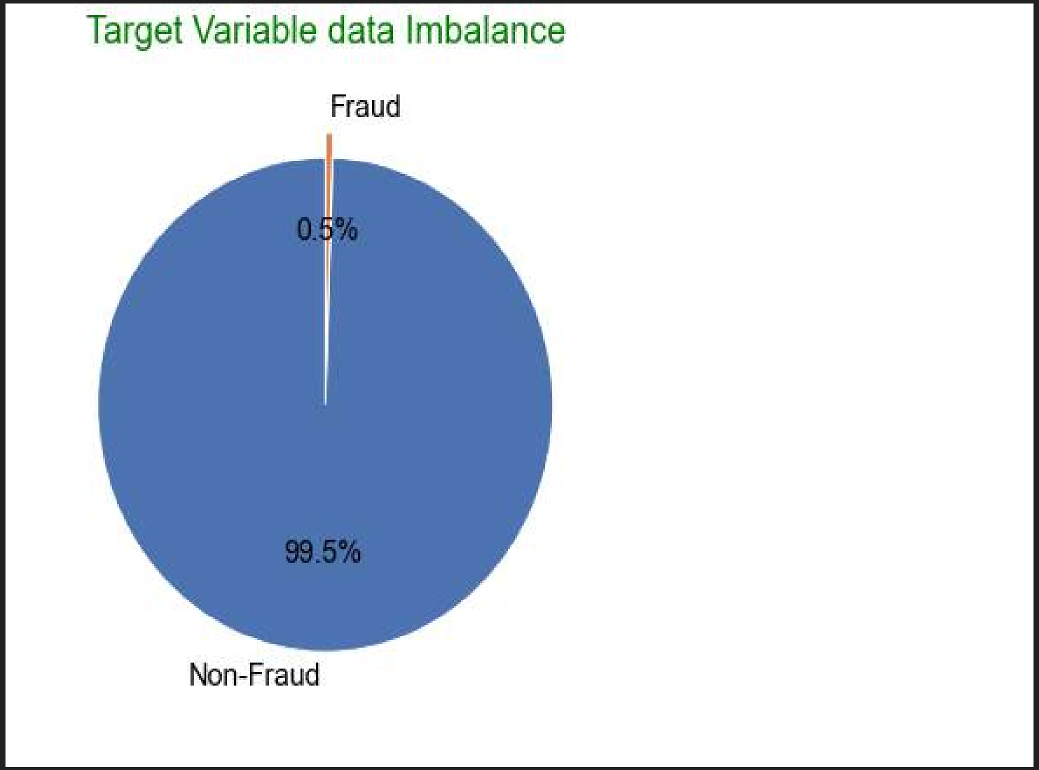


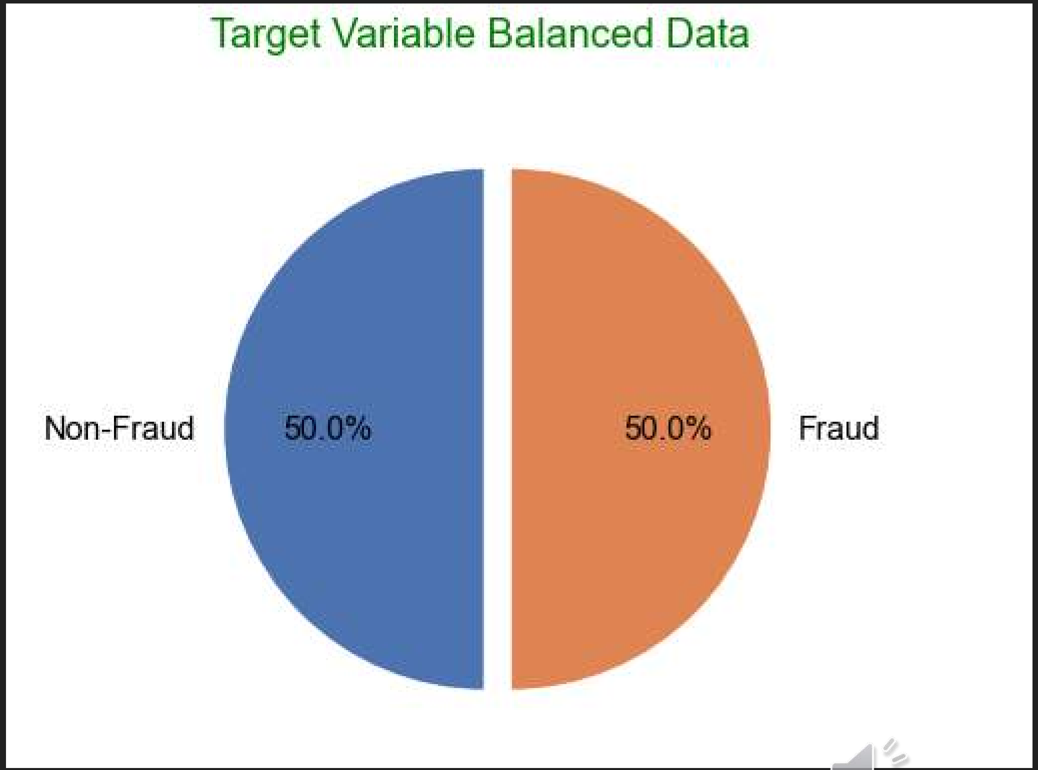
Data Preparation for Modeling

Create Dummy Variables:

- Independent variables as dummy variables allows easy interpretation and calculation of the odds ratios, which increases the stability and significance of the coefficients.

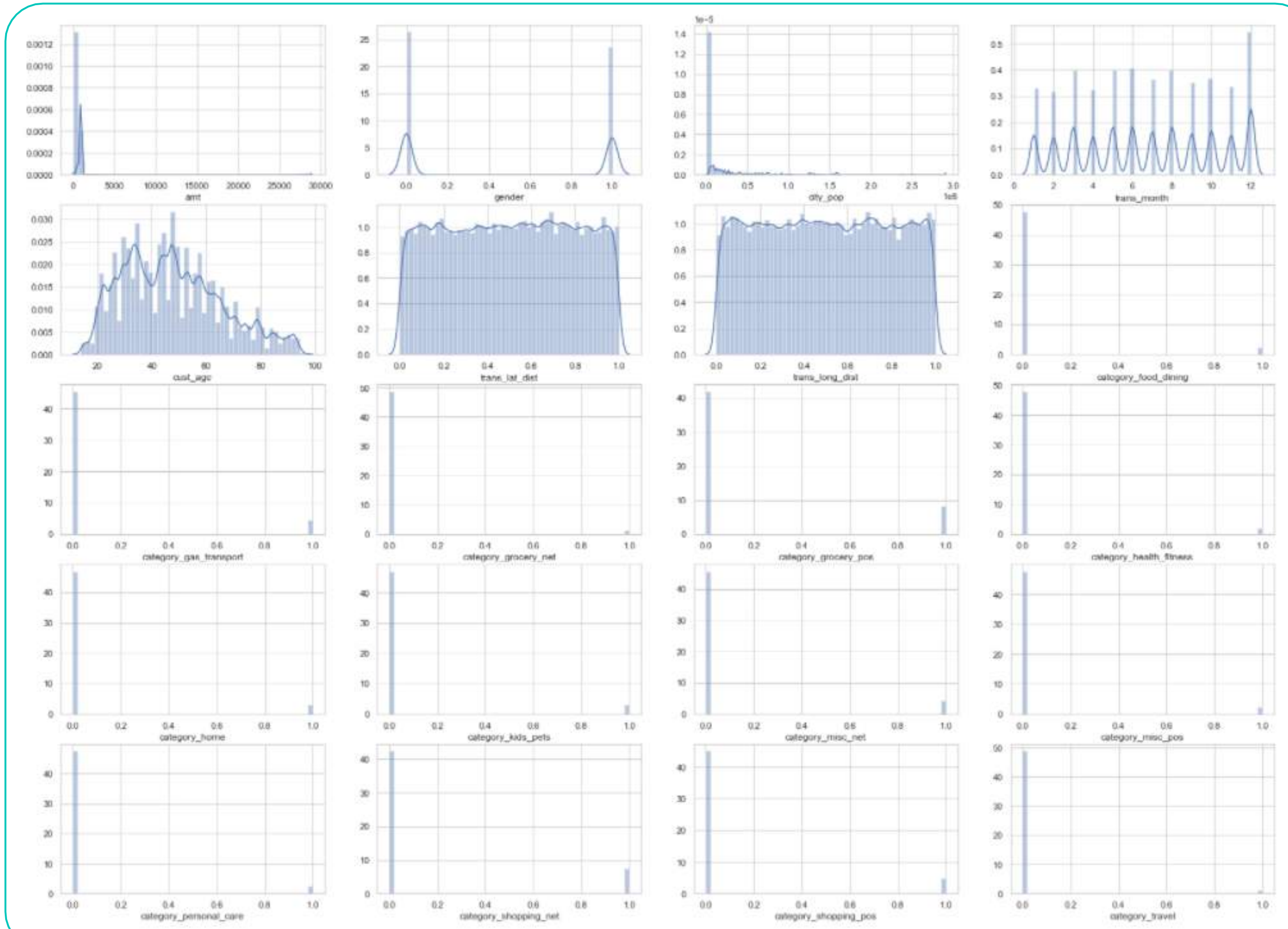
Data Imbalance





Data Preparation for Modeling

- Train- Test Split
- The modified ' dataset has been split into Train and test dataset in the ratio 70 30
- Train dataset has been used to train the model whereas Test dataset has been used to evaluate the model

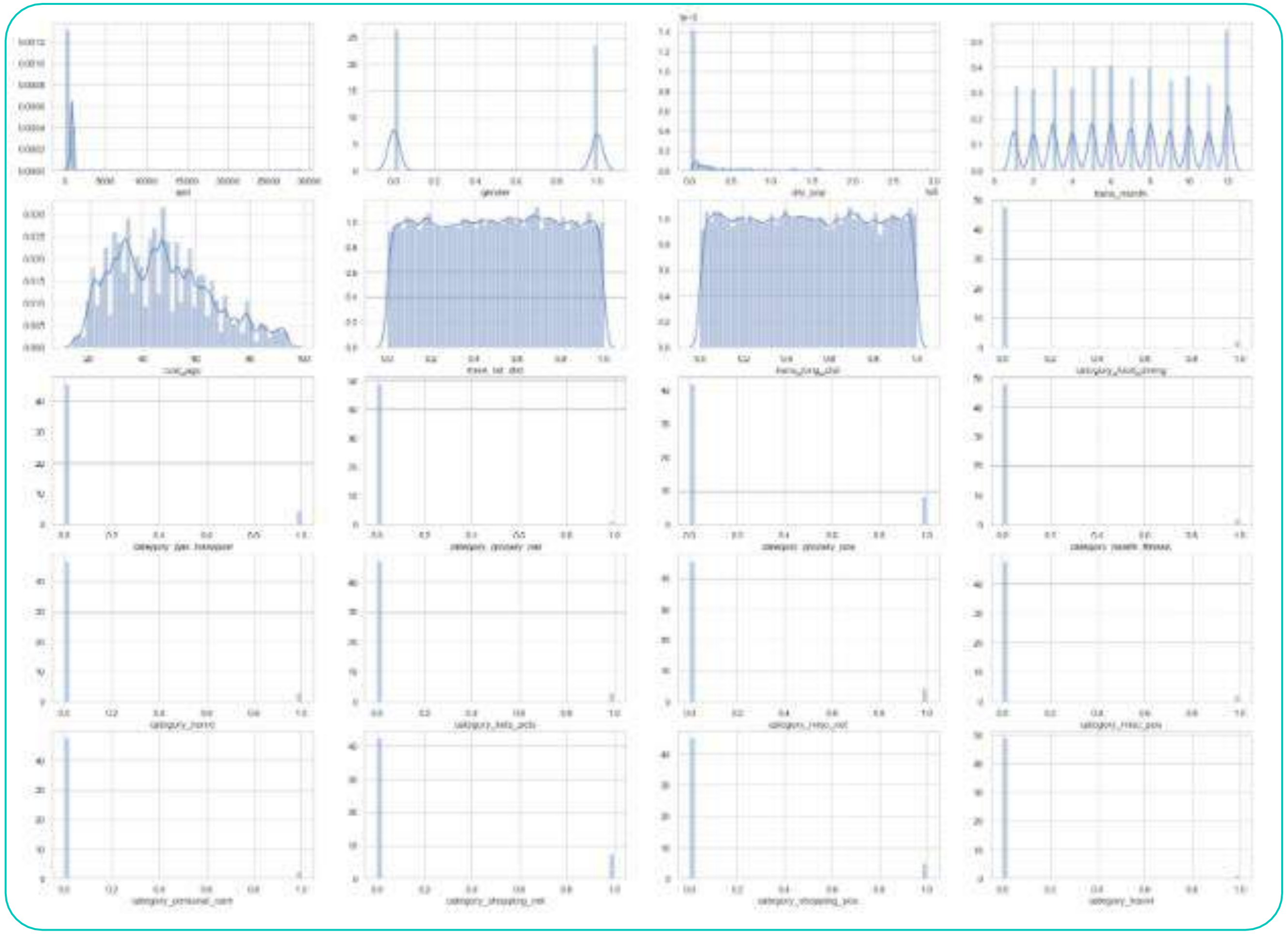


Skewed Data

<u>Analysis:-</u>

- ***<u>We see that the data is not evenly distributed as there is skewness in the data that will not give good result during model building.</u>***

Unskewed Data



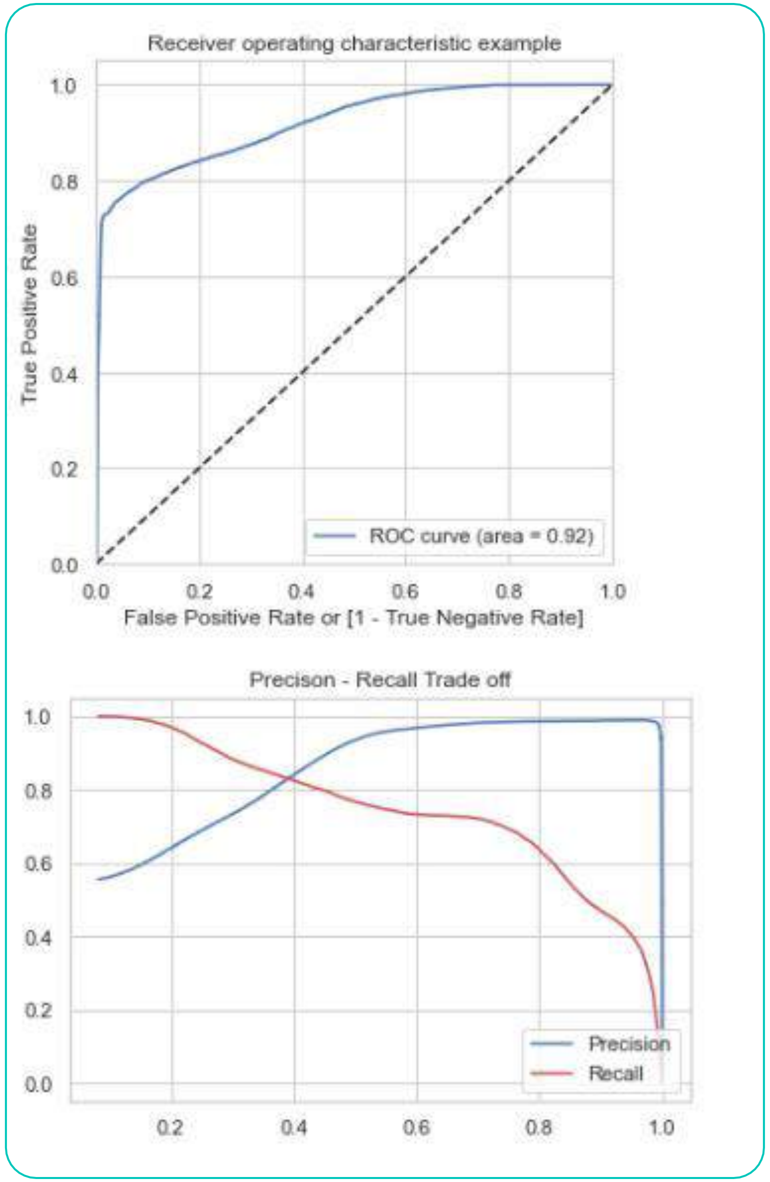
Analysis:-

We can see the data is unskewed so the now is best suitable for model building.

Model Building: Using Logistic Regression

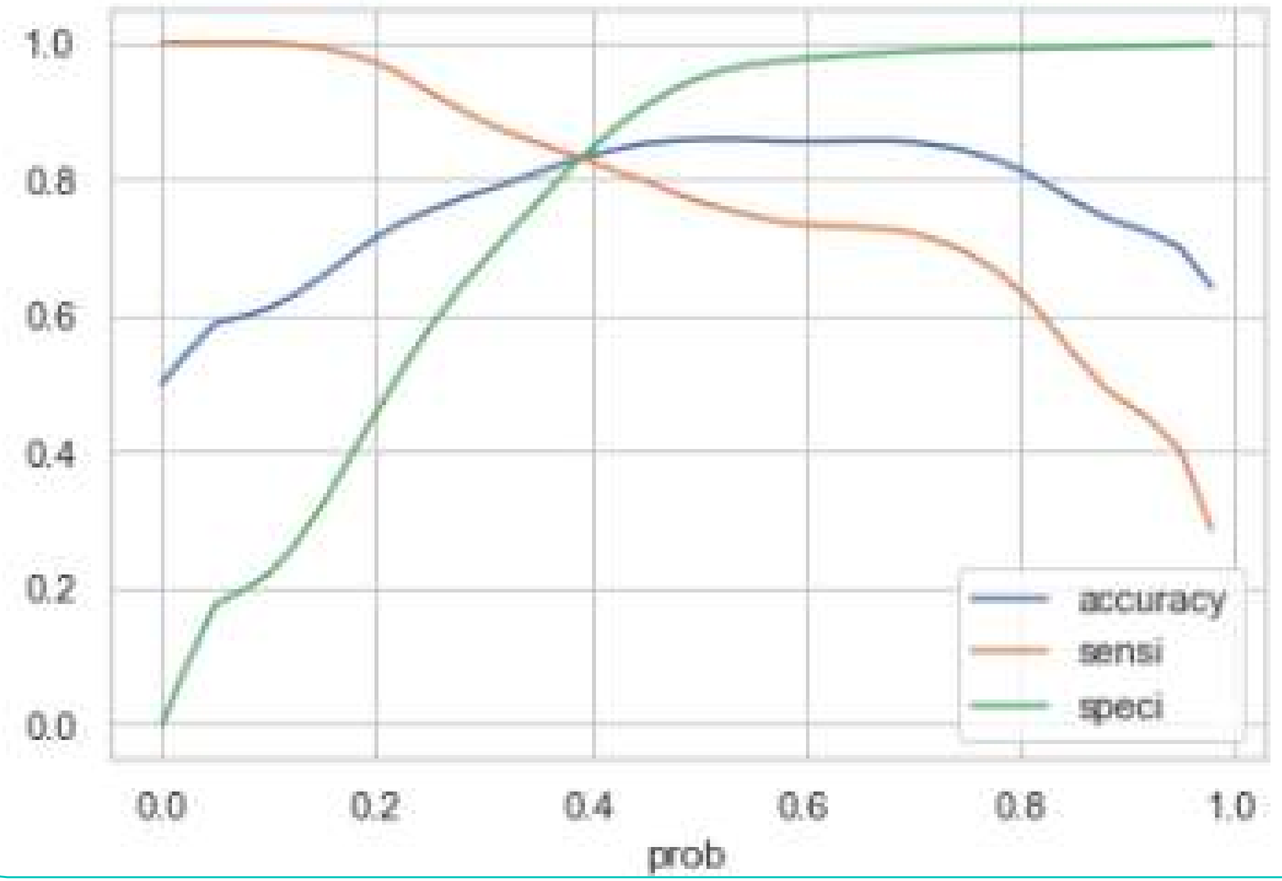
Feature Selection using Recursive Feature Elimination (RFE)

- RFE is an optimization technique for finding the best performing subset of features It is based on the idea of repeatedly constructing a model and choosing either the best (based on coefficients), setting the feature aside and then repeating the process with the rest of the features This process is applied until all the features in the dataset are exhausted Features are then ranked according to when they were eliminated.
- We ran RFE with 20 variables for further model building process
- Insignificant features were dropped one by one after checking the P value and Variance Inflation Factor



Measuring Model Performance

Sensitivity (Recall): 0.7672231101823273	***Specificity***: 0.9471088016052823
Precision: 0.935524177960679	***F-Score***: 0.843056129895273



Finding Optimal Cut-off

Optimum cut-off value is: 0.4

Measuring Performance on Train Set

Accuracy: 0.8343598827834284

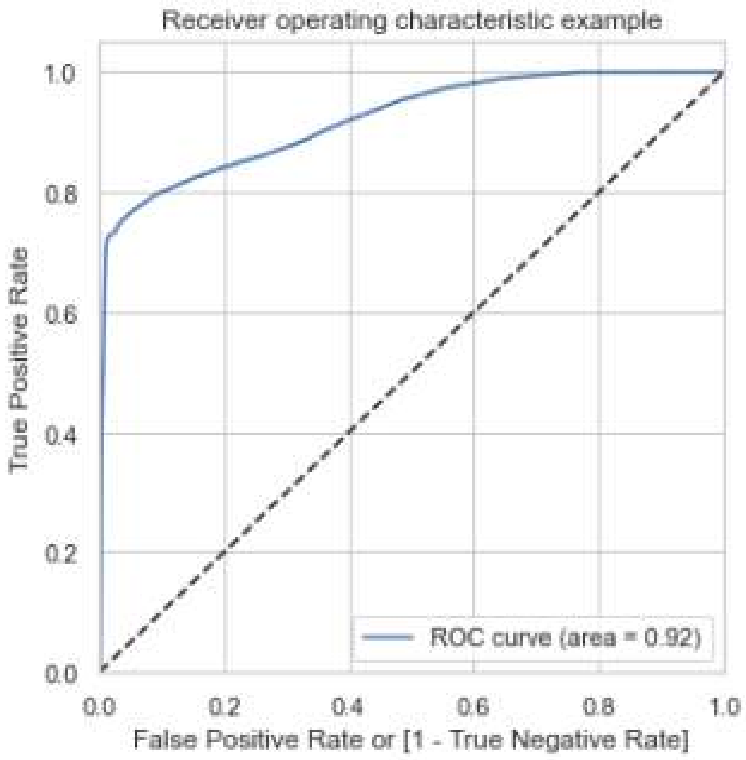
Sensitivity (Recall): 0.8255605388117802

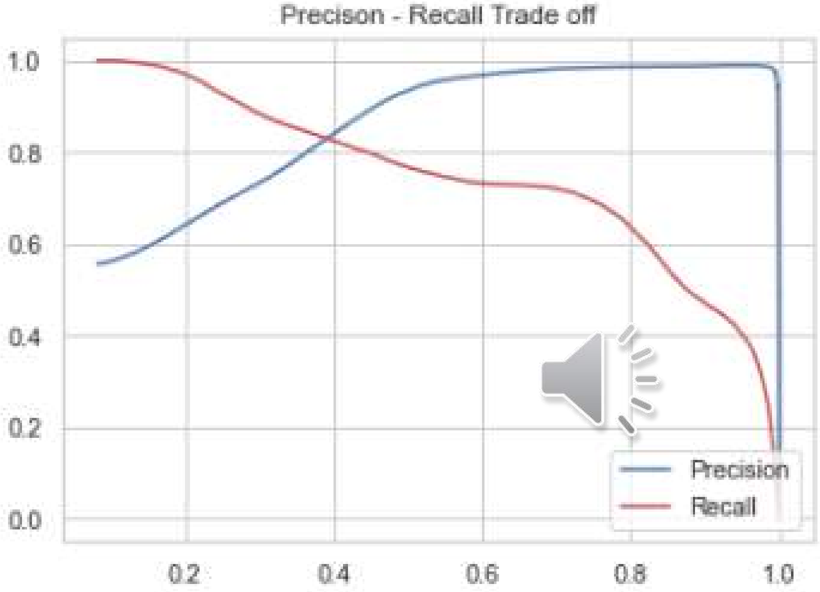
Specificity: 0.8431616555869313

Precision: 0.8403886896519074

F-Score: 0.8329086236081259

Finally, we have an overall accuracy of approx. 0.8 on our Logistic Regression model.





Measuring Performance on Test Set

Accuracy:
0.8090084891547565

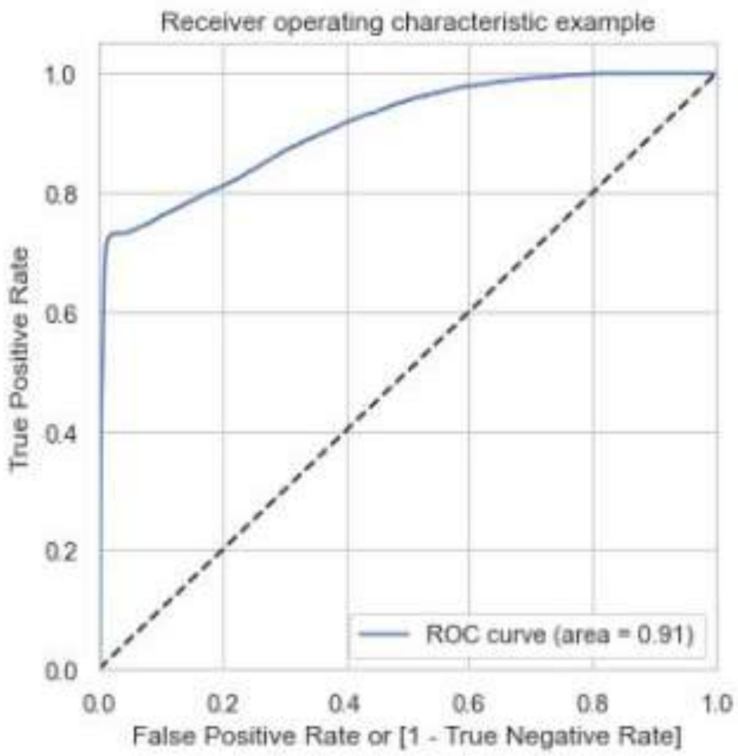
Sensitivity (Recall):
0.8051533986555565

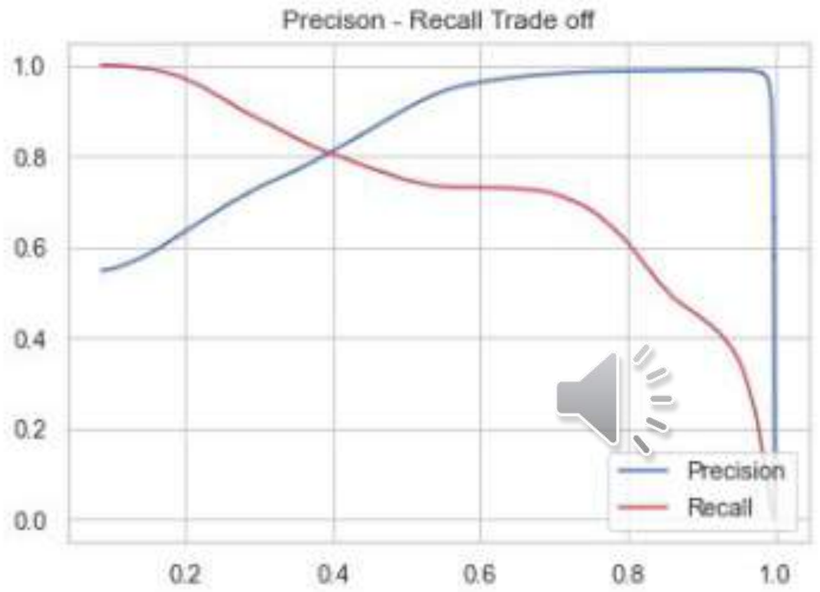
Specificity:
0.8128610979003654

Precision:
0.8113089203795409

F-Score:
0.8082194393409482

We have an overall accuracy of approx. 0.80 on our Logistic Regression model.





Model Building: Using Decision Tree

Measuring Performance on Train Set

Accuracy: 0.9888454322748698

Sensitivity (Recall): *0.9961700583986642*

Precision: 0.981790582531763

Measuring Performance on Test Set

Accuracy: 0.9886419342176429

Sensitivity (Recall): *0.9814253242412088*

Precision: 0.9961295225687377

Analysis:-

- We are getting accuracy of 98 % on both train and test data set

Model Building: Using Random Forest

Measuring Performance on Train Set

Accuracy: 0.9021435437856611

Sensitivity (Recall): 0.8314833446761409

Precision: 0.9683582698560862

Measuring Performance on Test Set

Accuracy: 0.9023557268782232

Sensitivity (Recall): 0.9961700583986642

Precision: 0.8315663762451483

Analysis:-

- We are getting accuracy of 90 % on both train and test data set

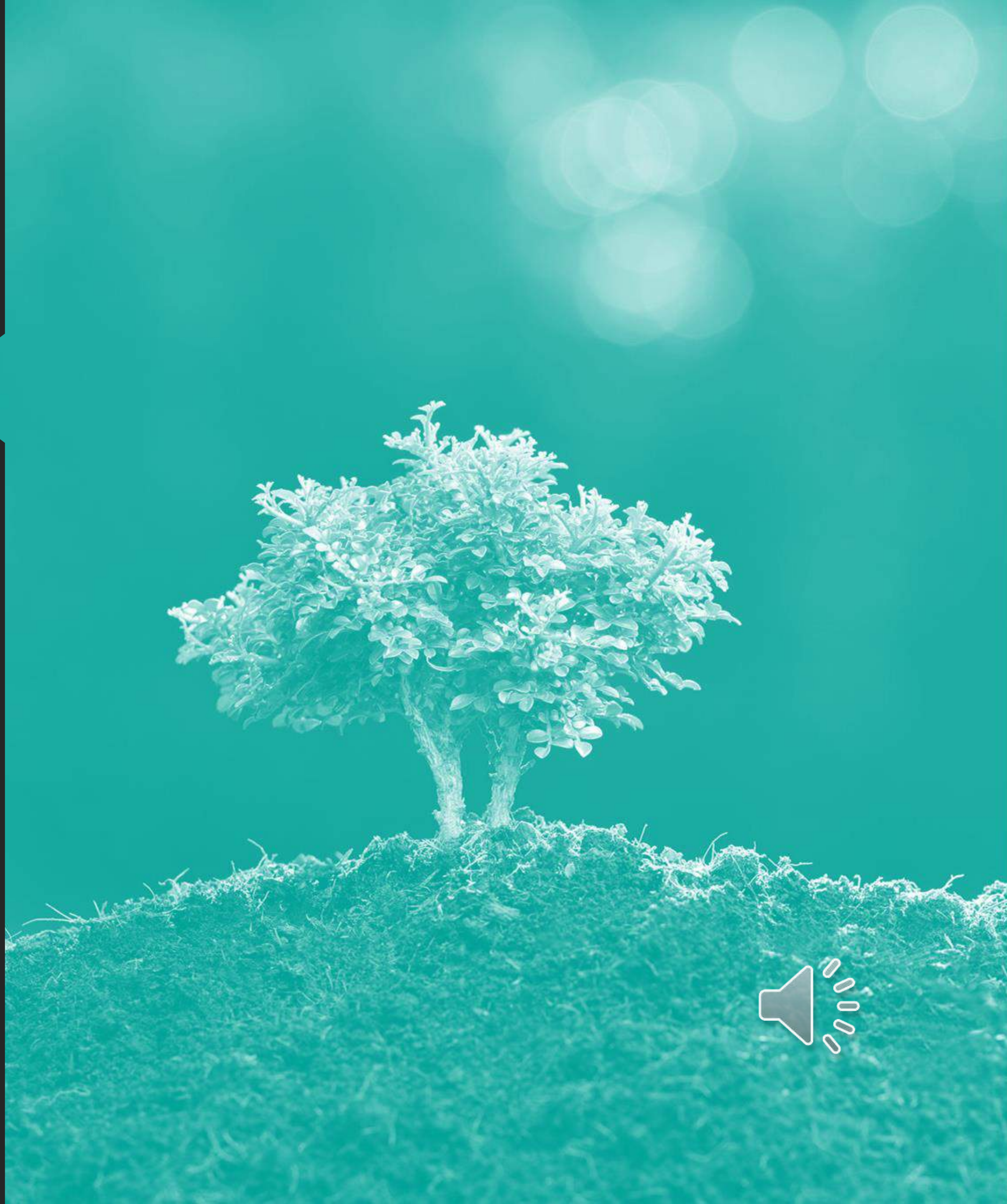
Evaluation Metrics

After evaluating all the three models we will using the Decision Tree model for the cost benefit analysis. As the predicted data we get from the Decision Tree Model because its accuracy, precision and recall percentage is highest among all the models.

Accuracy: 98%

Precision: 98%

Recall: 99%



Cost Benefit Analysis

- After the model has been built and evaluated with the appropriate metrics, we need to demonstrate its potential benefits by performing a cost-benefit analysis which can then be presented to the relevant business stakeholders.
- To perform this analysis, you need to compare the costs incurred before and after the model is deployed. Earlier, the bank paid the entire transaction amount to the customer for every fraudulent transaction which accounted for a heavy loss to the bank.
- We will perform the following calculations sequentially to arrive at the final savings that your model can potentially provide to Finex.

Current Loss Incurred

Average number of transactions per month	77183.1
Average number of fraudulent transaction per month	402.12
Average amount per fraud transaction	530.66
Cost incurred per month before the model was deployed	213389

Analysis After Model Building

Average number of transactions per month detected as fraudulent by the model (TF)	31890.62
Cost of providing customer executive support per fraudulent transaction detected by the model	$1.5
Total cost of providing customer support per month for fraudulent transactions detected by the model (TF*$1.5)	47997.12
Average number of transactions per month that are fraudulent but not detected by the model (FN)	233.29
Cost incurred due to fraudulent transactions left undetected by the model (FN*c)	123797.67
Cost incurred per month after the model is built and deployed (4+6)	171794.79
Final savings = Cost incurred before - Cost incurred after(1-7)	41467.07

SUMMARY

Findings

- The males are more prone to fraud transaction
- People do more transaction from 11:00 AM to 11:00 PM
- The fraud transaction generally happens in the later night
- Maximum fraud transaction happens on Sunday and Monday.
- December reported maximum number of fraud transaction.
- People between the age of 50-60 are more prone to fraud transaction.

Findings

Following three variables are contributing the most towards the probability of a lead getting converted:

- Amount
- Category
- Gender

Again, based on the coefficient values the following are the top three categorical/dummy variables that should be focused the most in regarding the fraud transaction:

  - gas_transport,
  - grocery_pos
  - shopping_pos

233.29

Conclusion and Recommendations

- The fraud transaction detected by model is more than fraud transaction undetected by the model.
- Average number of transactions per month detected as fraud by the model is **31890.62**
- Average number of transactions per month not detected as fraud by the model is **233.29**
- The cost incurred after the model is built is less than cost before the model is built.
- Cost incurred per month before the model was deployed is **213389**
- Cost incurred per month after the model was deployed is **171794.79**
- Final saving of **41594.21**

Conclusion and Recommendations

- We have use Decision Tree Model with an Accuracy of 98%, Precision of 98% and Recall of 99%.
- We will using this model because of its high accuracy among all the three models.
- To perform the cost benefit analysis this model is best.
- This model detects high number of fraud cases.
- This model is cost effective.
- Hence overall this model seems to be good.

Attached Files

- Raising Fraud Root Cause Analysis
- Structured Problem Solving
- Cost Benefit Analysis
- Credit Card Fraud Detection Capstone Project File
- YouTube link for video of the project - https://youtu.be/gaLQqKYBNnQ
- fraudTrain and fraudTest Dataset.

THANK YOU