Credit Card Fraud Detection Capstone Project

By: Sanjay Tom Perayil

Introduction

- Finex has observed that a significantly large number of unauthorised transactions are being made, due to which the bank has been facing a huge revenue and profitability crisis.
- This has led to late complaint registration with Finex and by the time the case is flagged fraudulent, the bank incurs heavy losses and ends up paying the lost amount to the cardholders.

Objective

- To build a fraud detection model to help banks identify credit card frauds and be vigilant enough to reduce losses incurred due to such unauthorized transactions by the fraudsters.
- Also we will performing the cost benefit analysis to check for final savings with the help of cost incurred before and after model building.

Data Understanding

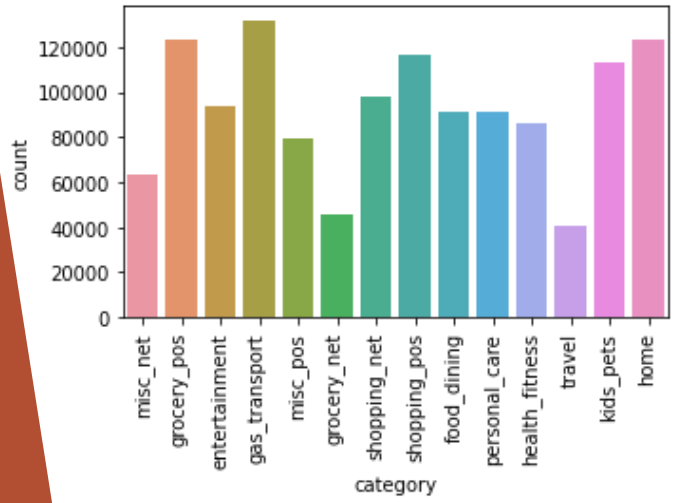
- Dataset contains legitimate and fraud transactions from the duration 1st Jan 2019 - 31st Dec 2020.
- Dataset contains 2 csv files named fraudtrain.csv and fraudtest.csv.
- It covers credit cards of 1000 customers doing transactions with a pool of 800 merchants. It contains a total of 18,52,394 transactions, out of which 9,651 are fraudulent transactions.
- The data has no null values in the given dataset.
- The data set is highly imbalanced.

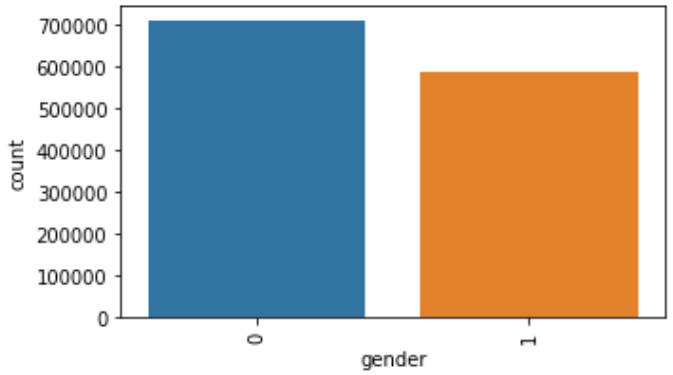
Data Preparation

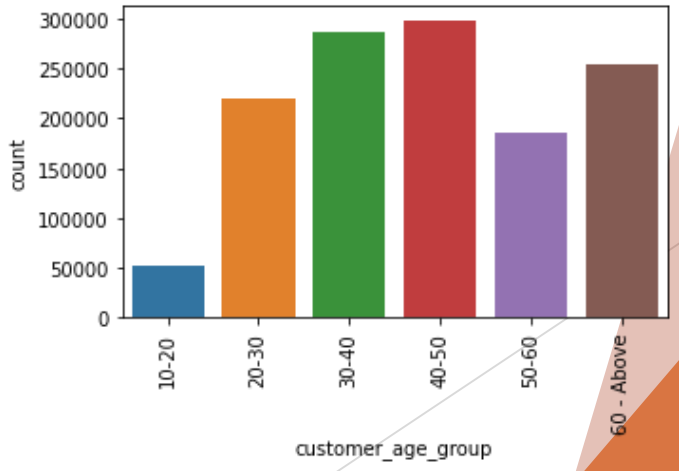
- converting date columns to datetime format on both the train and test dataset.
- Creating a column for customers making a transaction in the last 60 days for both the datasets.
- Binning the gender feature.
- Binning the customer age column.

Exploratory Data Analysis

- Categories mainly gas_transport and grocery_pos are the two main highest transaction dealing categories.
- There are more female customers compared to male customers.
- Maximum number of transactions occurs for both 30-40 and 40-50 age categories.

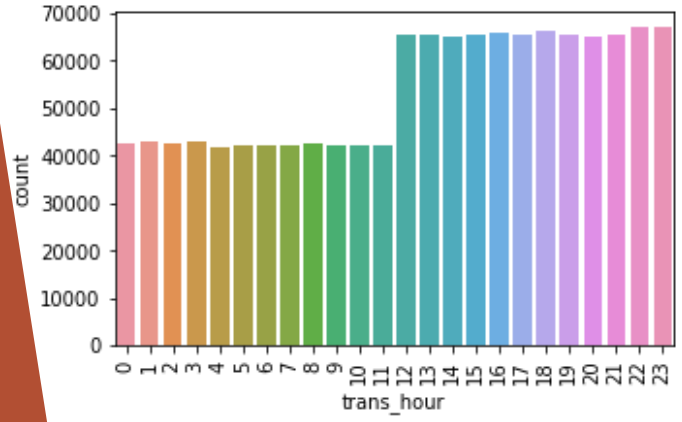


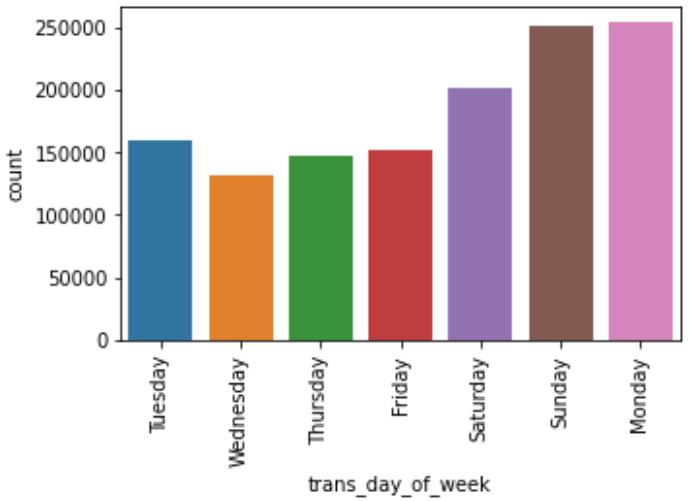


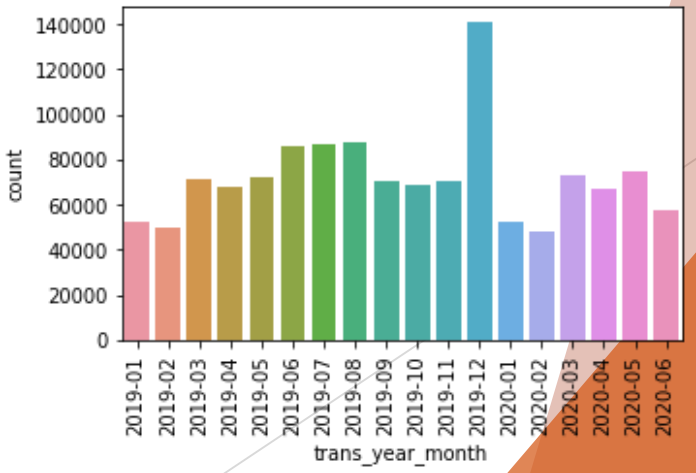


Exploratory Data Analysis

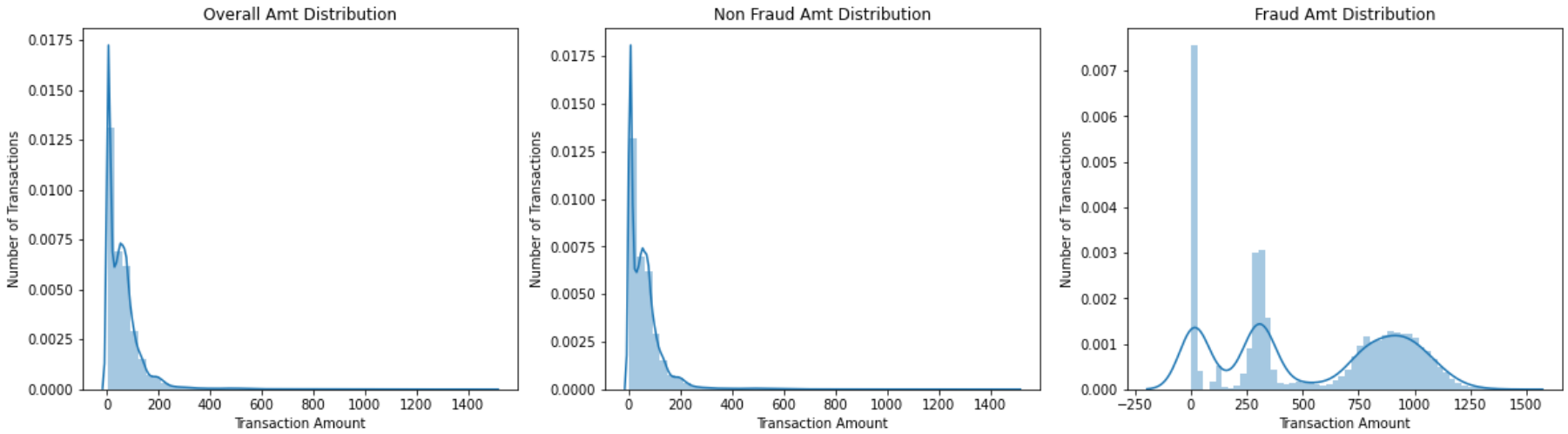
- Maximum Fraudulent transactions took place between 12pm and 11pm.
- Maximum number of transactions takes place on both Sunday and Monday.
- Maximum number of transactions occurs during December 2019 and minimum during February 2020.





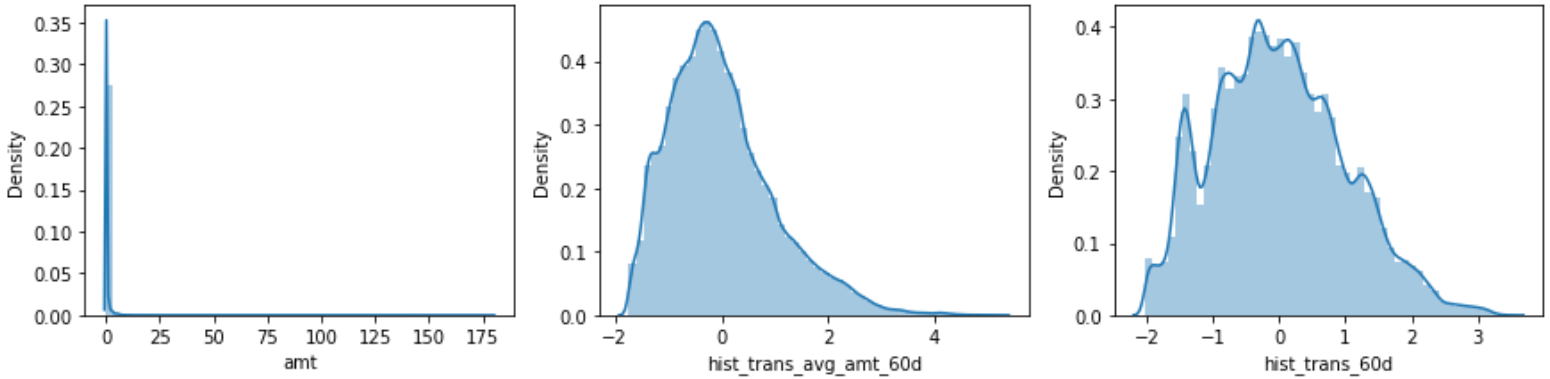


Bivariate Analysis

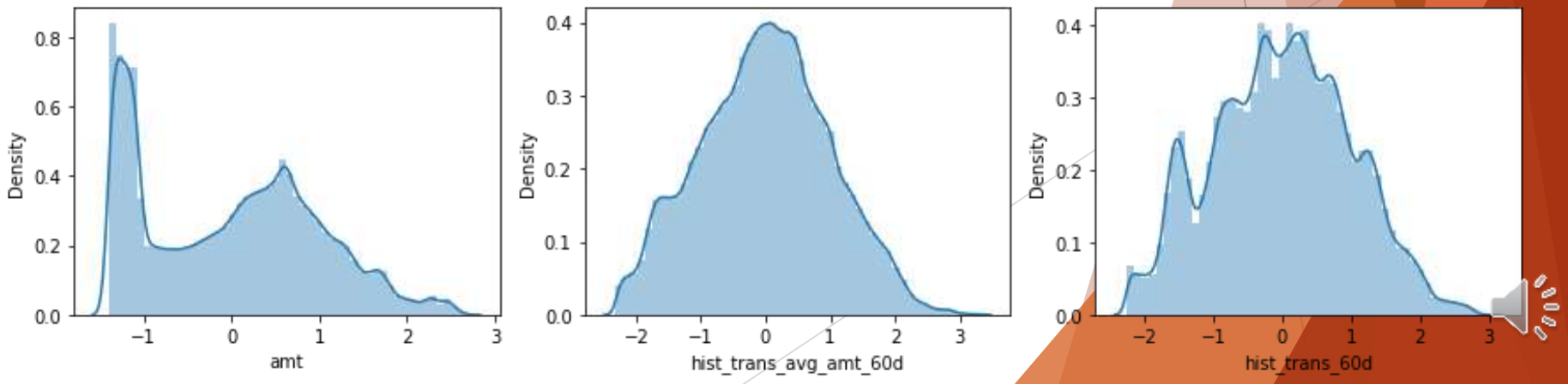


- The amount of fraudulent transactions majorly distributed around the range 750-1250.
- The mean of fraud transactions are higher than the non- fraudulent transactions.

Skewness Treatment



- As the data was properly evenly distributed as there was skewness on the dataset.
- The dataset is properly treated for skewness and now it is ready for model building.



Model Building Methodology

- As the class is not properly balanced, ADASYN (Adaptive Synthetic Sampling) is used for the oversampling.
- After the dataset is balanced, we could start building models for the further exploration and computation.
- We start building a Logistic Regression model, Decision Tree model and finally XGBoost on balanced dataset.
- Evaluation on both the train and test datasets are performed.

Logistic Regression Model

Accuracy:

0.91

Recall:

0.89

***Precision*:**

0.93

Train Dataset

Accuracy:
0.93

Recall:
0.93

***Precision*:**
1.0

Test Dataset

Model summary

Train set

AUC-ROC: 0.97

Test set

AUC-ROC: 0.73

Decision Tree Model

Accuracy:

0.95

Recall:

0.94

Precision:

0.96

Train Dataset

Accuracy:

0.93

Recall:

0.93

Precision:

1

Test Dataset

Model summary

Train set

AUC-ROC: 0.99

Test set

AUC-ROC: 0.97

KGBoost Model

Accuracy:

0.96

Recall:

0.96

***Precision*:**

0.96

Train Dataset

Accuracy:
0.96

Recall:
0.96

***Precision*:**
1

Test Dataset

Model summary

Train set

- AUC-ROC: 0.99

Test set

- AUC-ROC: 0.99

Model Evaluation metrics

- ADASYN techniques were used to balance the dataset.
- Logistic regression, decision tree and XGBoost algorithms were used to build models on each sampling method. While selecting the best model, few factors must be considered such as infrastructure, resources or computational power to run the model.
- COST OF DEPLOYMENT
- MONETARY LOSS
- After conducting the experiment, we observe that XGBoost model is performing well on the dataset which is balanced with ADASYN

Cost Benefit Analysis

- The costs associated with the implementation of the model are compared with the costs before and after it was deployed. Before the model was deployed, the bank would pay the entire amount of the fraudulent transaction to the customer.
- The cost incurred after the model is deployed is due to the left out fraudulent transactions that the model fails to detect. Hence, the total savings for the bank would be the difference of costs incurred after and before the model deployment.

Cost Benefit Analysis

Cost Benefit Analysis		
S. No	Questions	Answer
a	Average number of transactions per month	77183.08
b	Average number of fraudulent transaction per month	402.12
c	Average amount per fraud transaction	530.66

Cost Benefit Analysis		
S. No	Questions	Answer
1	Cost incurred per month before the model was deployed (b*c)	213389
2	Average number of transactions per month detected as fraudulent by the model (TF)	5211.04
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)	7816.56
5	Average number of transactions per month that are fraudulent but not detected by the model (FN)	10.71
6	Cost incurred due to fraudulent transactions left undetected by the model (FN*c)	5683.37
7	Cost incurred per month after the model is built and deployed (4+6)	13499.93
8	Final savings = Cost incurred before - Cost incurred after(1-7)	199889.07

Conclusions

- The model has detected more fraudulent transactions than the undetected fraudulent transactions.
- We recommend to apply the KGBoost model which has 99% accuracy for both the train and test datasets.
- For models such as XGBoost, the computational resources required to build and deploy the infrastructure are very high. On the other hand, simpler models such as Logistic Regression or Decision Tree require less computational resources. This makes the cost of deploying the model less.

Thank You