Capstone Project : Credit Card Fraud Detection

- Angad Gothe
- Anurag Bhardwaj
- Devanshi Gupta

Agenda

- Objective / Problem statement
- Background
- Model analysis
  - Exploratory data analysis
  - Key features
  - Model evaluation
- Cost benefit analysis
- Conclusions
- Appendix

Objective / Problem Statement

- Objective is the identify and stop the fraudulent transactions before they happen.
- Catching fraudsters or trapping them is not something that a banks AI / ML model can achieve.
- The increasing fraudulent transactions and customers coming late to the bank to highlight the fraud transactions is the issue. Due to this the credibility of bank is on stakes. This impacts the banks reputation.
- Identifying these fraud transactions before they happen is problem statement.
- Finally build a model that will be cost effective and ultimately result in revenue by controlling the cost incurred due to fraud transactions.

Background

- The number of fraud transactions has increased drastically due to which credit card companies are facing a lot of challenges.
- Retaining these customers is prime importance, frauds can be a threat in maintaining these high profile customers.
- Financial loss, trust and credibility due to banking fraud is a concerning issue for both banks and customers.
- With growing digital payments and increasing online shopping the fraud transactions are increasing,
- Hence stopping such frauds is not in scope for a bank, but identifying such transactions and stopping in time can help retaining these customers.

Model analysis

- Machine learning model is built to detect frauds before they happen, add an additional verification as to mitigate the losses.
- Understand the data: The data contains both legitimate and fraud transactions. This data has details of 1000 credit card customers and multiple merchants where the transactions were made from 1st Jan'19 to 31st Dec'20.
- Out of total 18,52,394 transactions provided only 9,651 are fraudulent. Only 0.52% which creates imbalance in the data hence to build a model we will need to handle this imbalance.
- Fine tune the hyperparameters to get good level of model performance.
- As a final step a cost benefit analysis is required to assess the benefits to business and its stakeholders after model implementation.

EDA

Perform univariate analysis to get insights on data pattern and distribution.

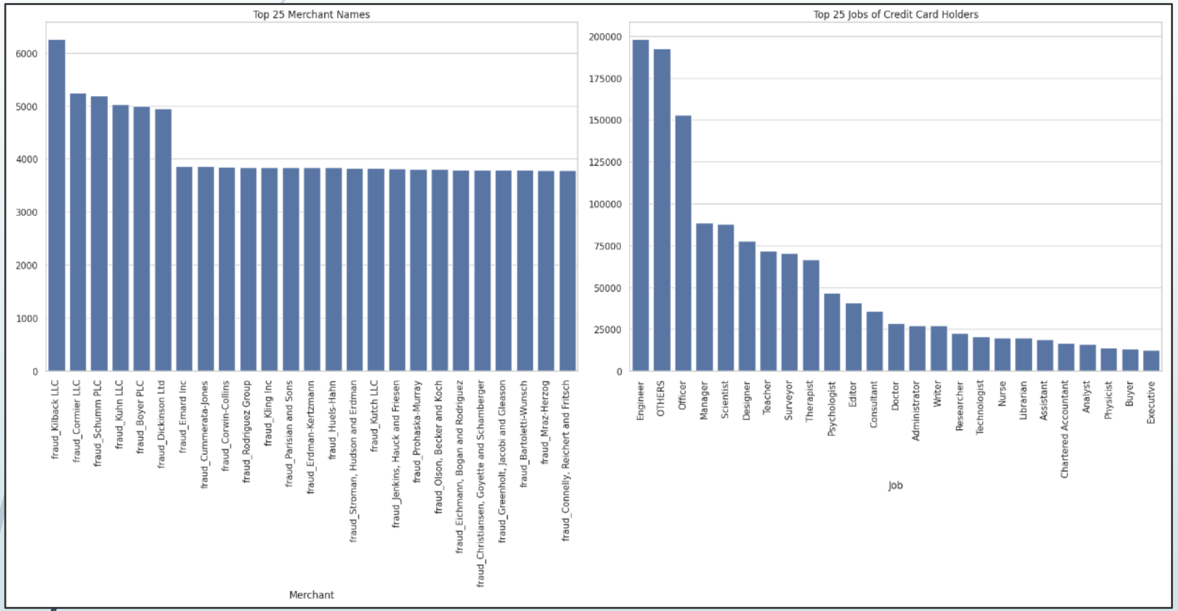
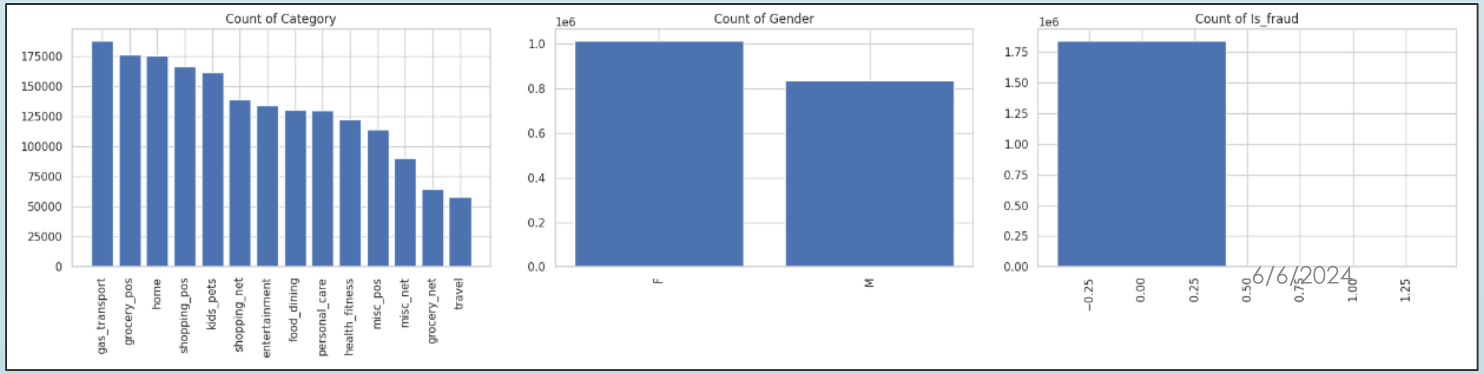


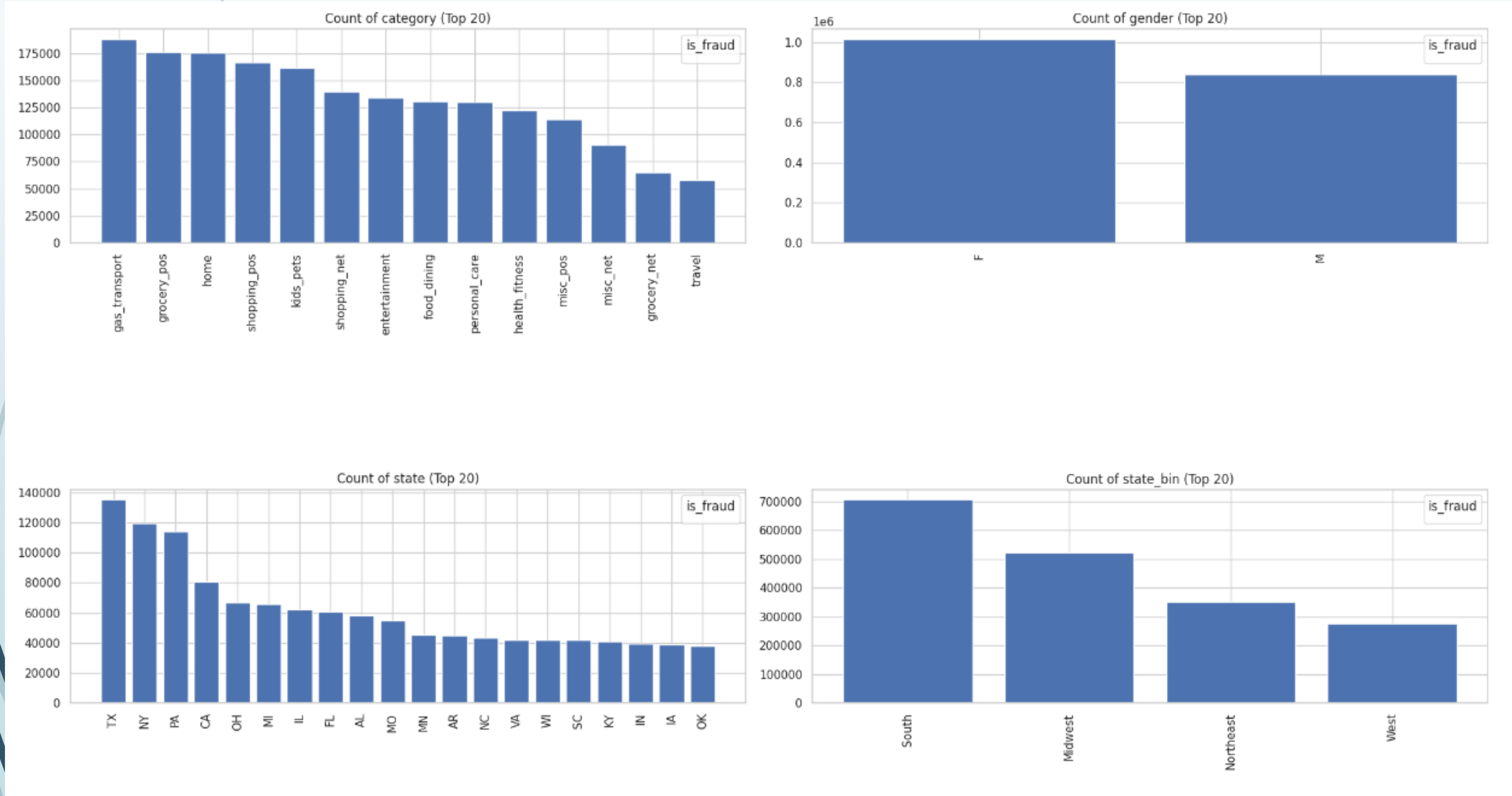
Chart of display top 25 Merchants and credit card holders.

Chart to display count of categorical variables like category, Gender, if fraudulent.



EDA

Perform bivariate analysis to get insights on data pattern and distribution.



Count of category, gender where the data is fraudulent.

Statewise fraudulent transactions.

Also region wise the number of frauds encountered.

Model evaluation

- In comparison with 2020, fraud transactions in 2019 are on higher side.
- Count of fraud transactions is slightly more with female's credit card holders. Hence gender can be considered as one feature.
- Model should handle fraud transactions specifically during 12th month of the year as regular and fraud transactions both are at peak. Days of the weeks can play critical role here.
- Category plays an important feature where the category like shopping_net, misc_net, shopping_pos, gas_transport and grocery_posries have high number of fraud transactions.
- Evaluation of Frauds transactions with catefory == shopping_net where count of normal transaction is less but fraud is more needs to be handled. As for others like gas_transport and shopping_pos regular transactions exceed number of fraud transaction.
- Count of Frauds transactions are done more in people with occupations as Engineer, officer, others, scientist, Manager and Teacher. Hence occupation is feature to be considered by model.
- Region and States play an important role as states like NY, TX and PA have higher fraud. Accordingly South and Midwest region have higher frauds.
- Amount spend needs to be monitored to understand any abnormal transactions to identify fraud transactions.

Numbers for Cost benefit analysis

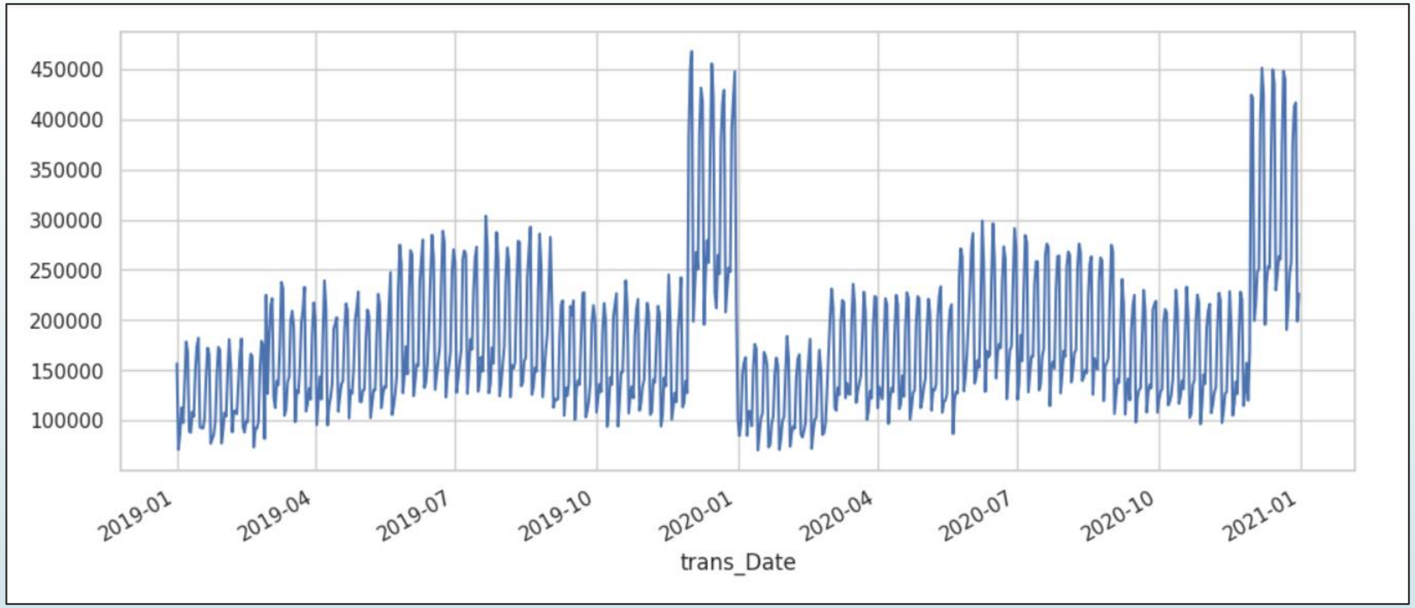
- Average number of transactions per month detected as fraudulent by the model (TF) : ***405.13***
- Total cost of providing customer support per month for fraudulent transactions detected by the model(TF*$1.5): ***607.69***
- Average number of transactions per month that are fraudulent but not detected by the model (FN) : ***16.25***
- Cost incurred due to fraudulent transactions left undetected by the model (FN*c) : ***8623.25***
- Cost incurred per month after the model is built and deployed (2+4): ***9230.94***
- Final savings = Cost incurred before - Cost incurred after(1-5): ***204161.28***

Conclusion – based on model

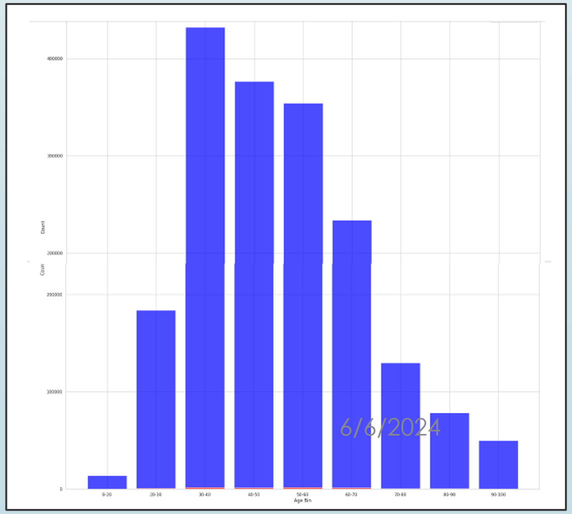
- Based on average amount spent against last 24 hours transaction by the credit card if there is a great increase, then its ideal for the Bank to send an SMS ALERT! to customer confirming the transactions.
- Pattern shows that probability of fraud transactions increases on Thursday, Saturday, Sunday and Monday, when the banks needs to be extra cautious and alert on these specific days to avoid frauds.
- Trend shows the fraud transaction amounts are high, hence any transaction where the spend is high than regular can be flagged and necessary alerts can be sent to card holder.
-
- Probability of frauds increases for categories like home, shopping, grocery, health, fitness and gas for transport and these are regular transactions hence here banks will have to keep card holders updated with SMS Alerts mentioning the transaction details. This can help identify the frauds early.
- The fraud transactions are majorly done during odd hours of the day i.e. between 22 - 3 Hr, So banks needs to ensure to send an SMS ALERT during such odd hours.
- Additional security and some upper limit alert for transactions online and offline shopping can help control the large amount of frauds that are happening in shopping sectors.

Appendix –

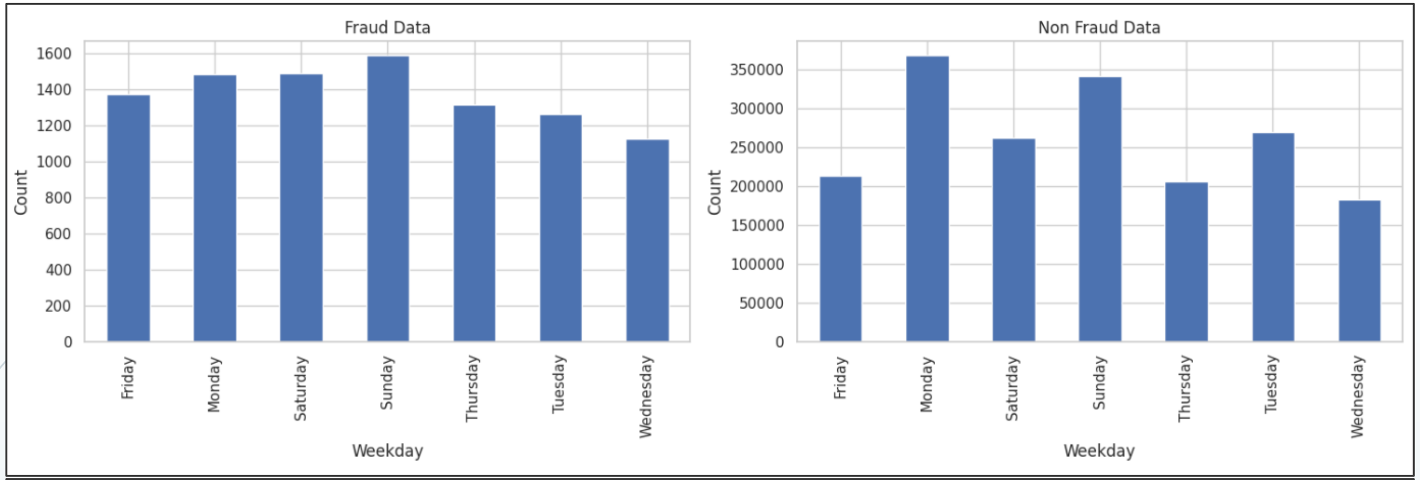
We have added some visualization for reflecting flavor of fraud patterns



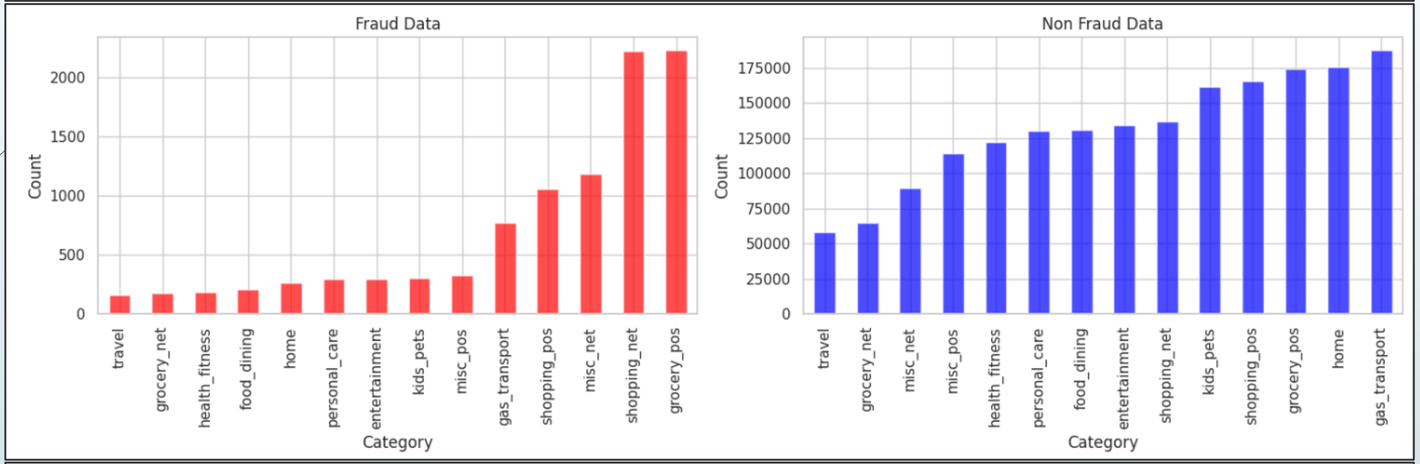
We have 2 years data and this shows trend during which months of the year when frauds are high. (towards end of year)



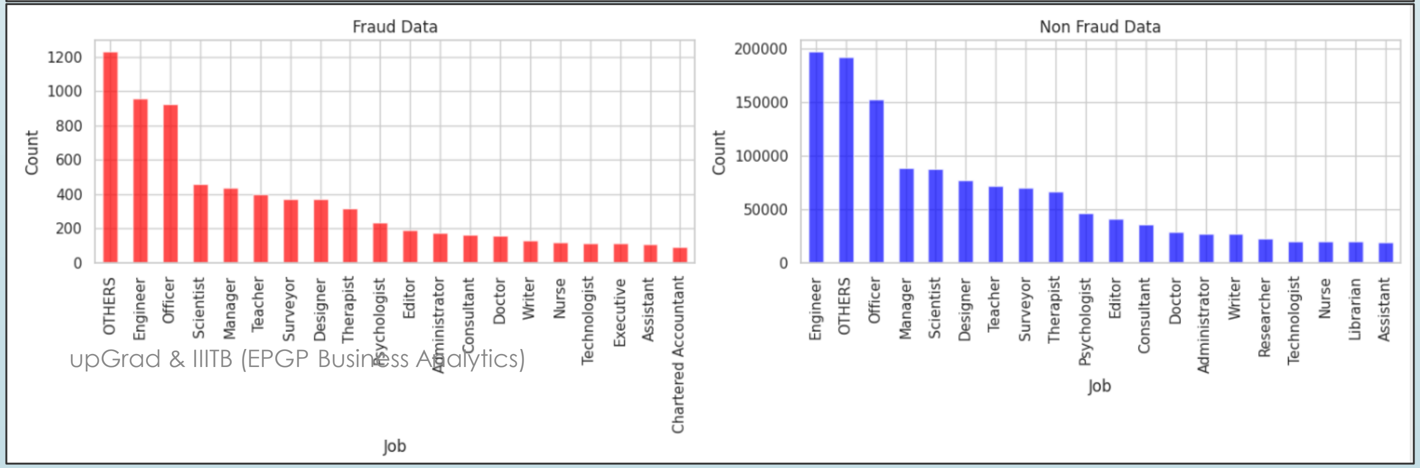
Shows the bucket where maximum frauds are happening.
(30-40 age group has max frauds)



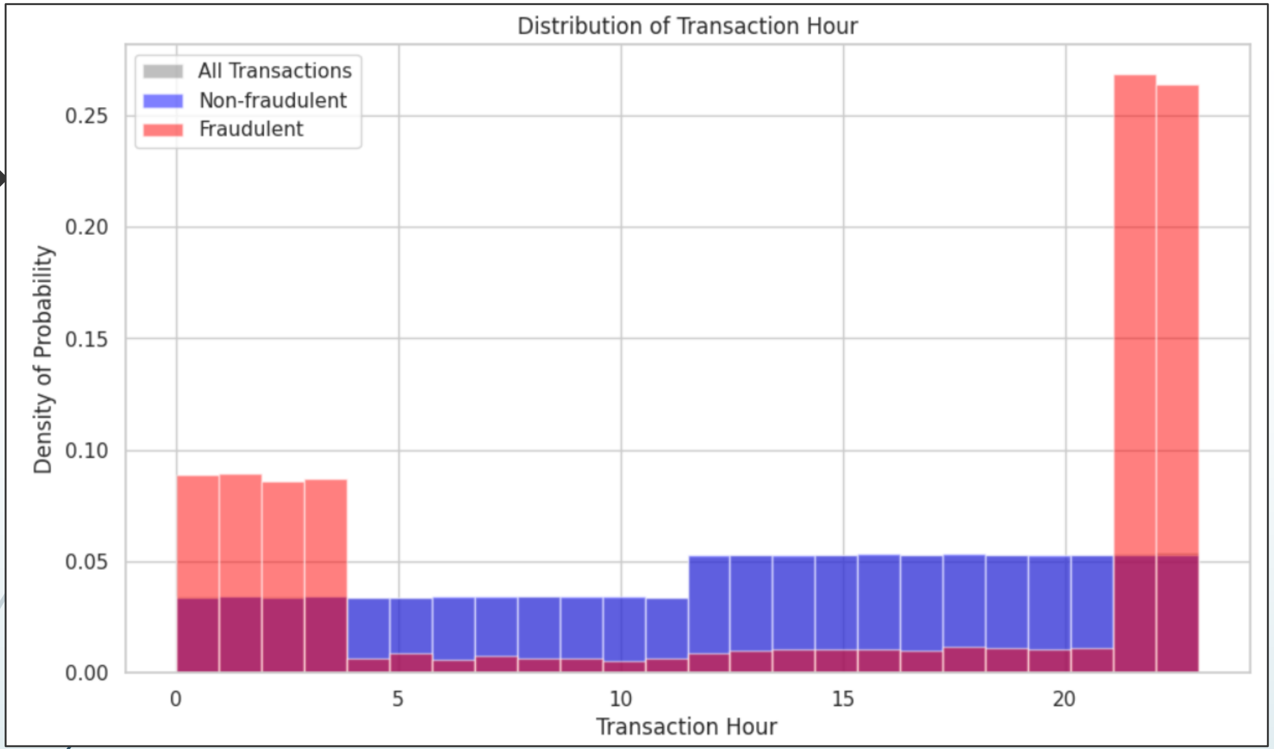
Weekdays wise pattern for fraudulent and non-fraudulent transactions



Category pattern for fraudulent and non-fraudulent transactions

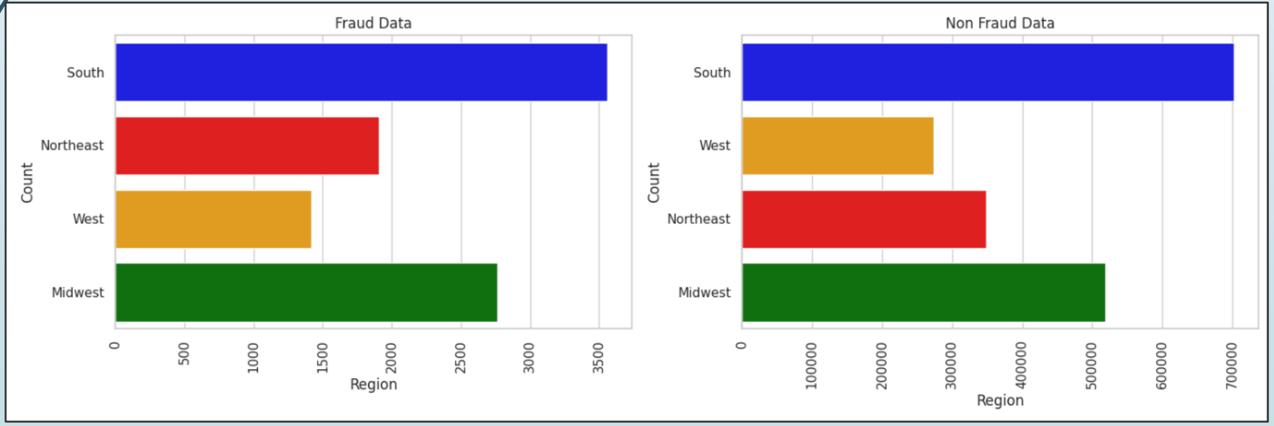


Job wise pattern for fraudulent and non-fraudulent transactions



Time of the day when the fraud transactions are high.

Late night and early morning 22hrs to 3hrs is the time of the day when frauds are high.



Regionwise fraud and non-fraud transactions.