CREDIT CARD FRAUD DETECTION USING THE MACHINE LEARNING ALGORITHMS & PYTHON

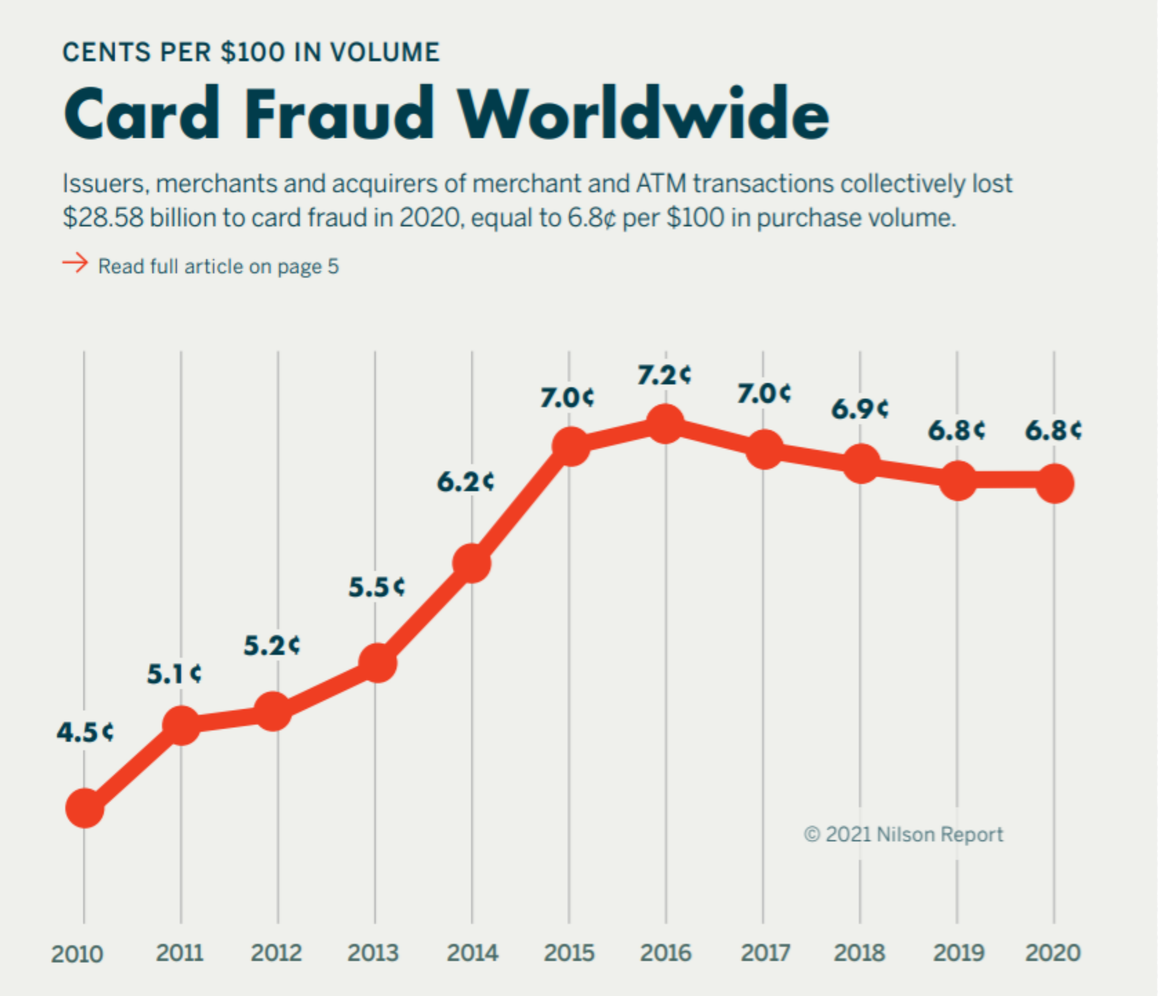
© Created by Iuri Catasov 2022 | Capstone Project

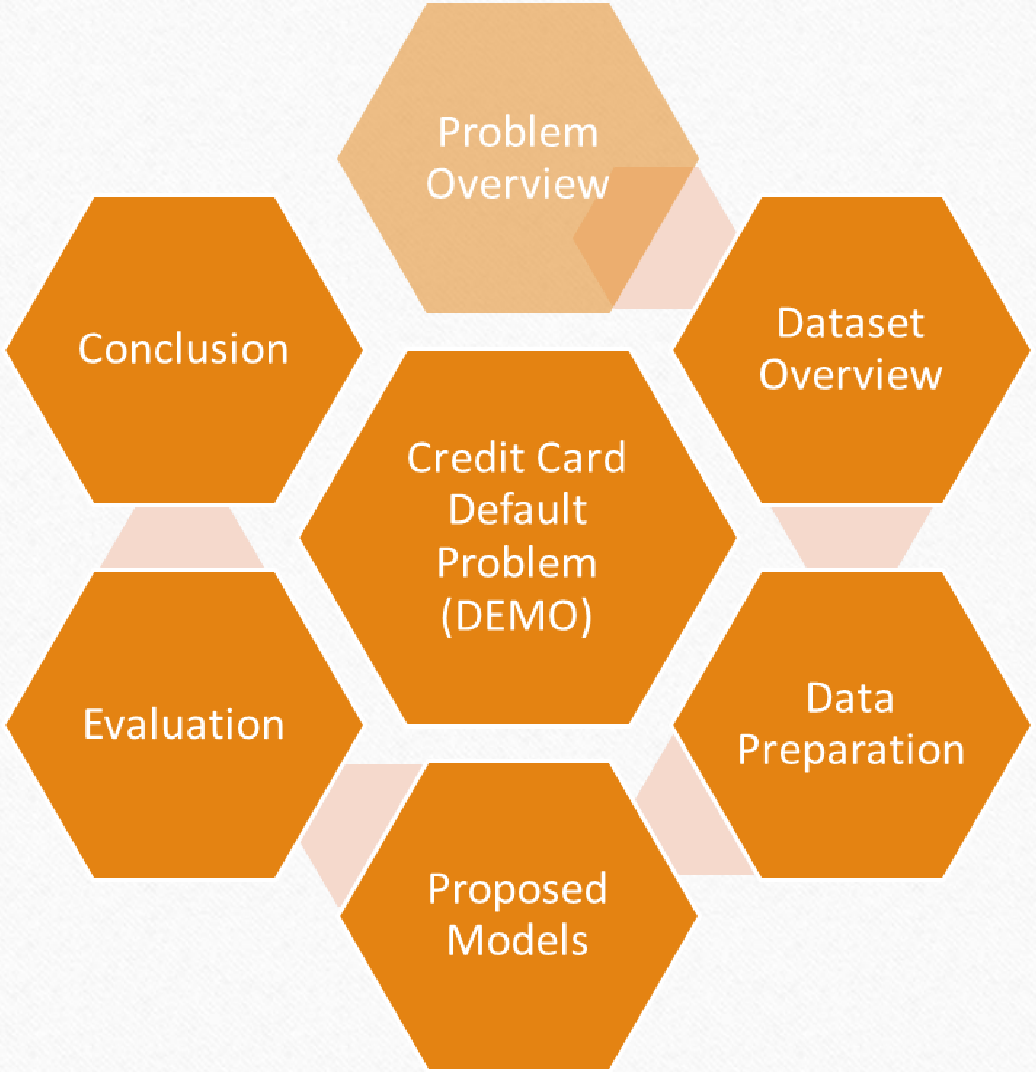
AGENDA

- Credit Card Default Problem (DEMO)
- Problem Overview
- Dataset Overview
- Data Preparation
- Proposed Models
- Evaluation
- Conclusion

PROBLEM OVERVIEW

- Fraudulent transaction is one of the most serious threats to online security nowadays.
- Payment card fraud losses reached $28.65 billion worldwide in 2019, according to the most Recent Nilson Report data.
- The coronavirus pandemic is also fueling explosive growth in card fraud activity.
- Companies that issue credit cards are looking to technological solutions to stop the fraud.




Problem Overview
Dataset Overview
Data Preparation
Proposed Models
Evaluation
Conclusion
Credit Card Default Problem (DEMO)

Dataset Overview

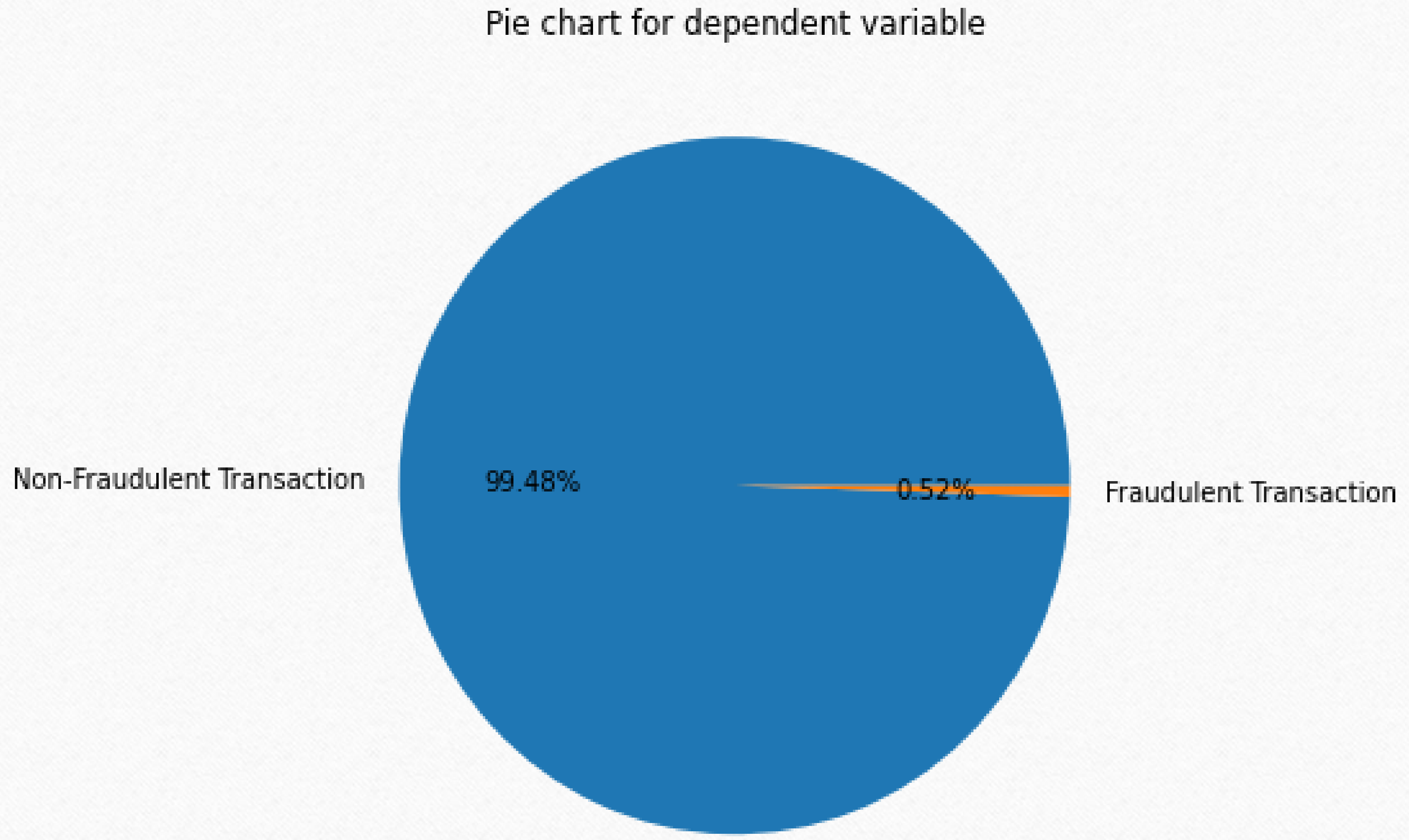
The Credit Card Fraud detection Data:
https://www.kaggle.com/kartik2112/fraud-detection?select=fraudTest.csv
https://www.kaggle.com/kartik2112/fraud-detection?select=fraudTrain.csv
This is a simulated credit card transaction dataset containing legitimate and fraud transactions. It covers credit cards of 1000 customers doing transactions with a pool of 800 merchants.
Data is collected for the period of 01/01/2019-12/31/2020 only inside the USA. There are 23 columns in the data and 1852394 rows of transaction records. The column 'is_fraud' can be considered as the entire data label/target, which I will be predicting.

DATA OVERVIEW

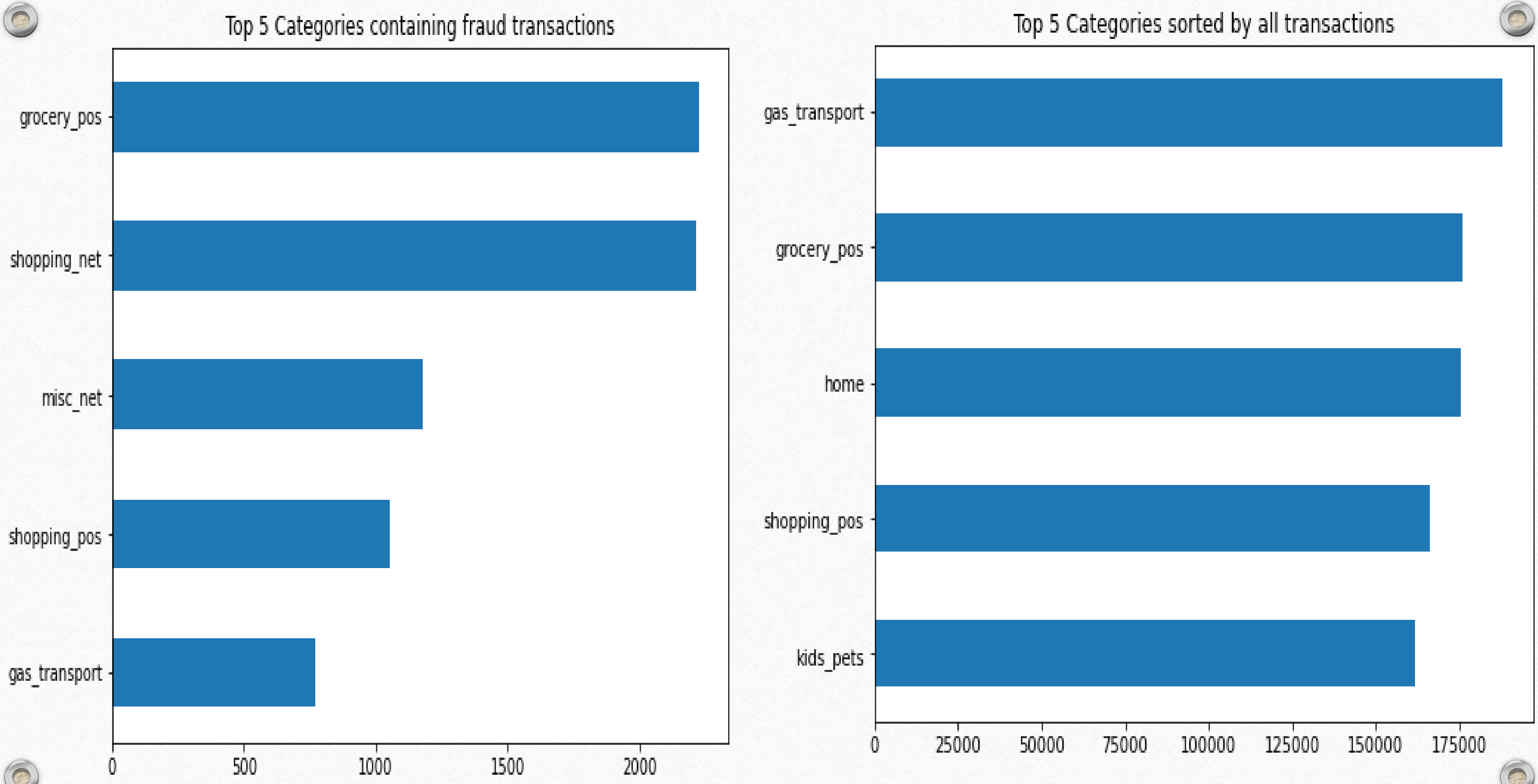
1	trans_date_trans_time	object	Transaction Date/Transaction Time
2	cc_num	int64	Customer's Credit Card Number
3	merchant	object	Merchant by whom the trade occurred
4	category	object	Type of Purchase
5	amt	float64	Amount of Transaction
6	first	object	First Name
7	last	object	Last Name
8	gender	object	Customer's Gender
9	street	object	Street Address
10	city	object	Home City
11	state	object	State
12	zip	int64	Zip Code
13	lat	float64	Latitude of the Customer
14	long	float64	Longitude of the Customer
15	city_pop	int64	Population of the City
16	job	object	Customers Job Title
17	dob	object	Customer's Date of Birth
18	trans_num	object	Unique Transaction Number for Each Transaction
19	unix_time	int64	Time of the Transaction in Unix
20	merch_lat	float64	Merchant Latitude
21	merch_long	float64	Merchant Longitude
22	is_fraud	int64	The Fraudulent Transaction /Not

dtypes: float64(5), int64(6), object(12)

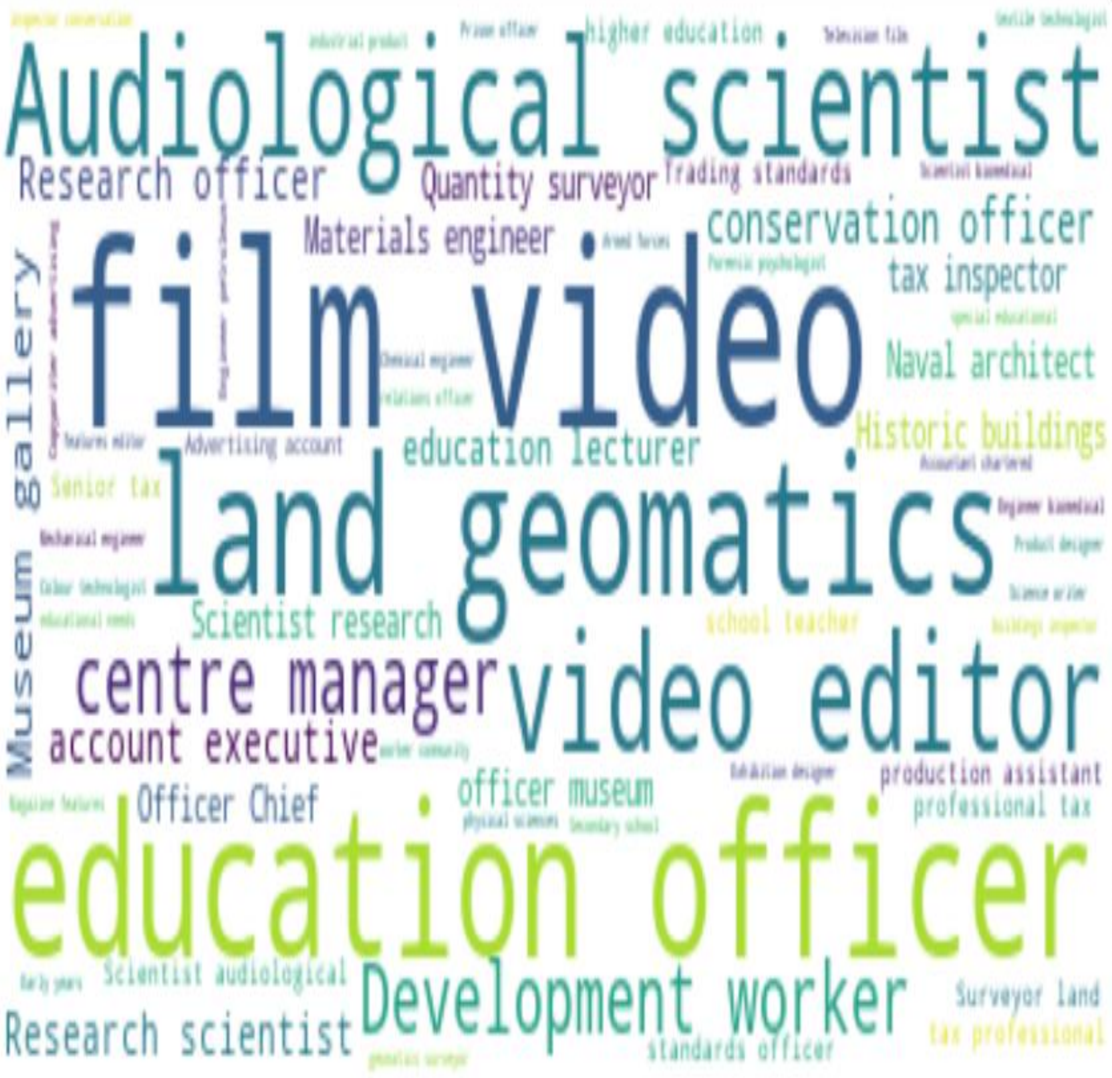
The data is imbalanced



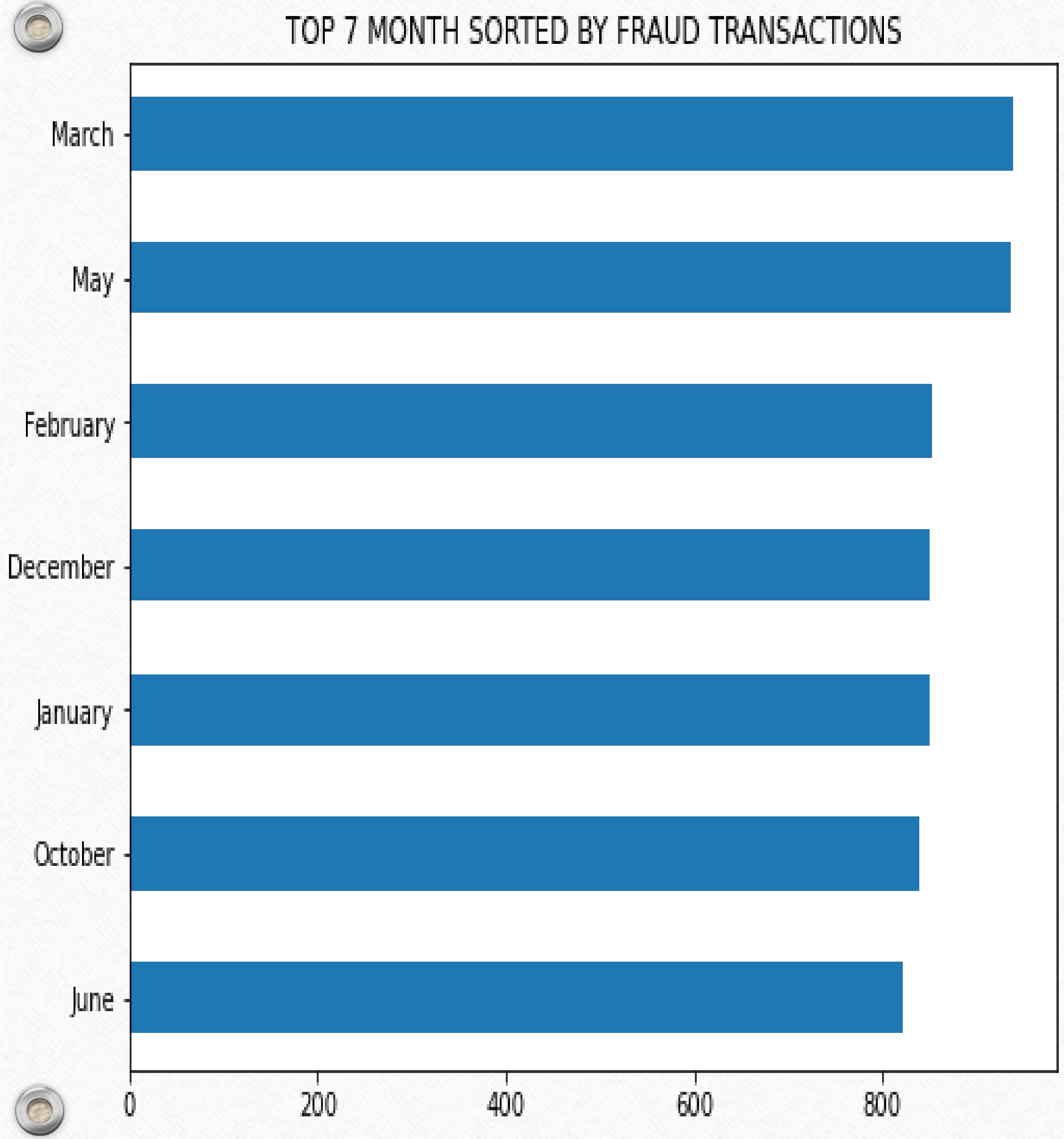
Non Fraud Transactions(0)	**1842743**
Fraud Transactions(1)	**9651**

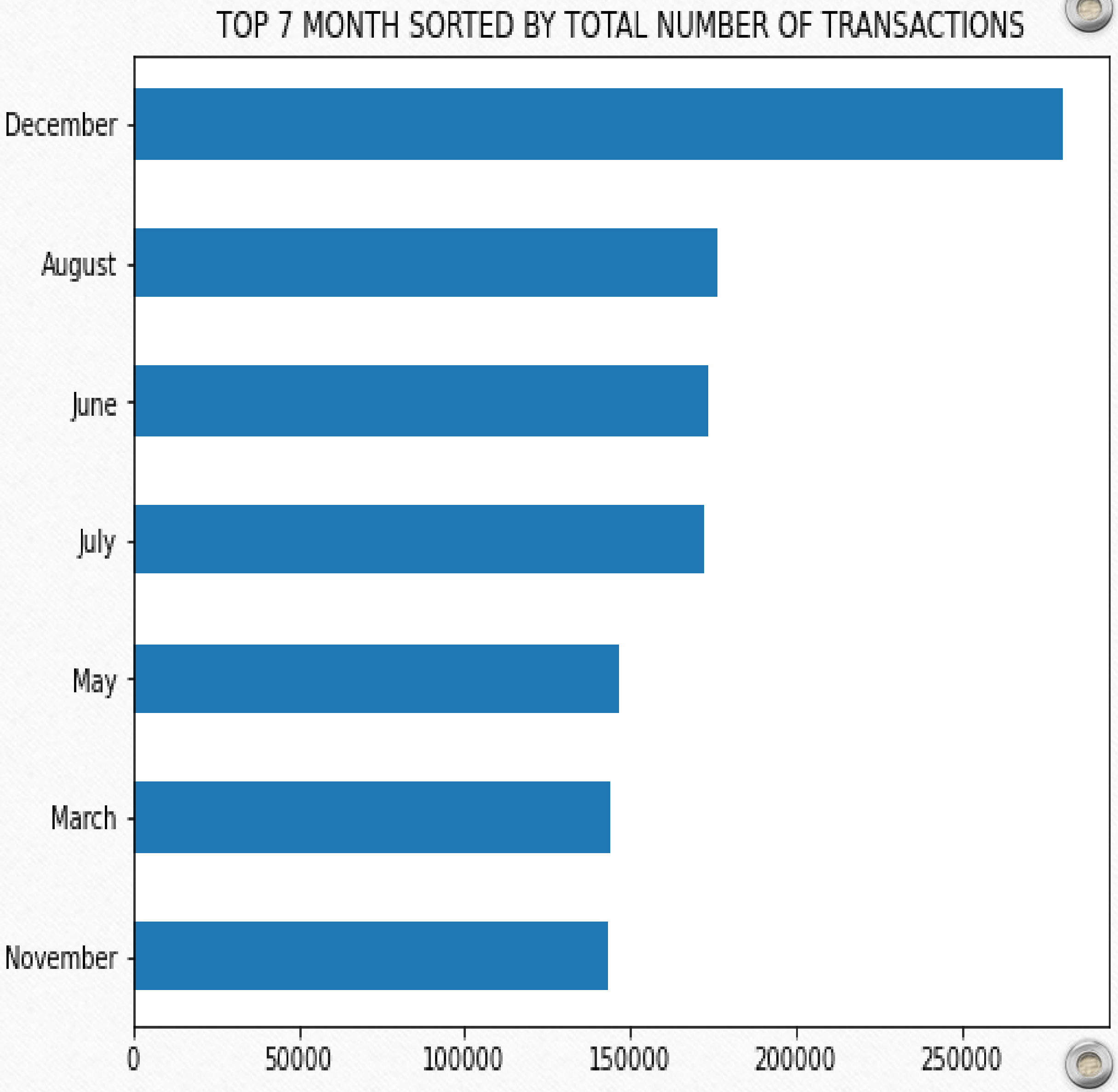

Top 5 Categories containing fraud transactions
grocery_pos
shopping_net
misc_net
shopping_pos
gas_transport
0
500
1000
1500
2000
Top 5 Categories sorted by all transactions
gas_transport
grocery_pos
home
shopping_pos
kids_pets
0
25000
50000
75000
100000
125000
150000
175000

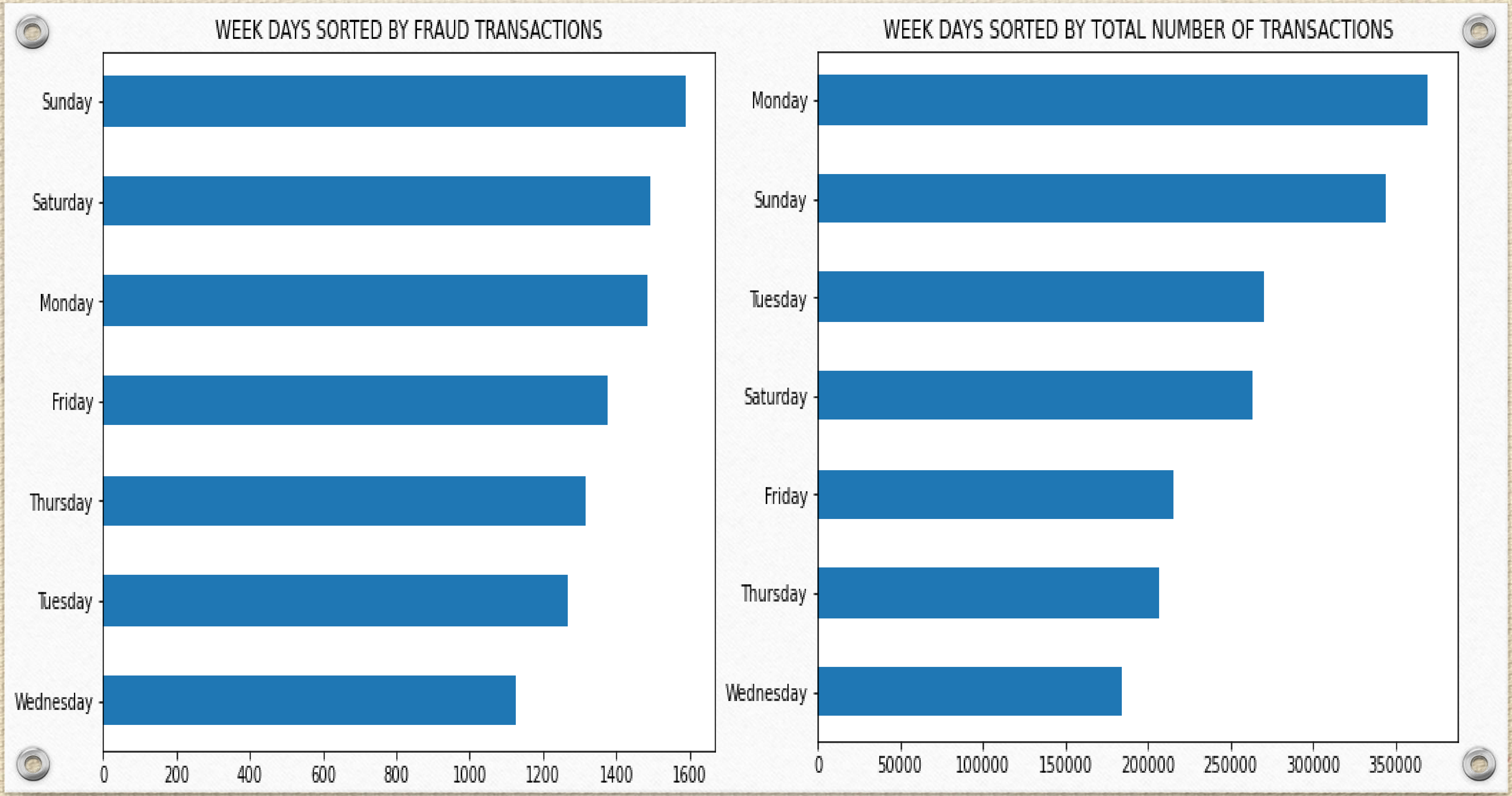
MOST FREQUENT JOB TITLES COMPARED BY FRAUD TRANSACTIONS

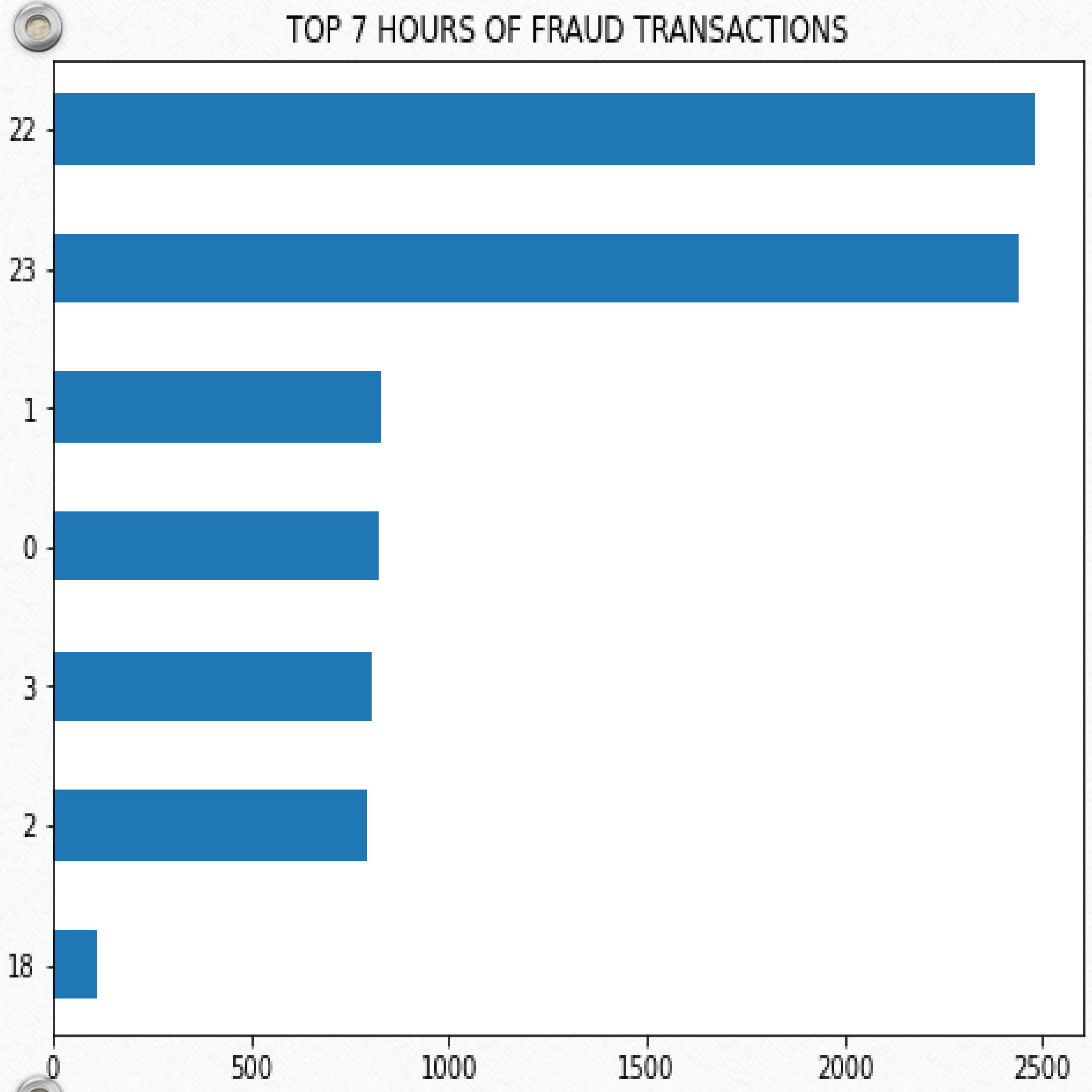


MOST FREQUENT JOB TITLES COMPARED BY TRANSACTIONS

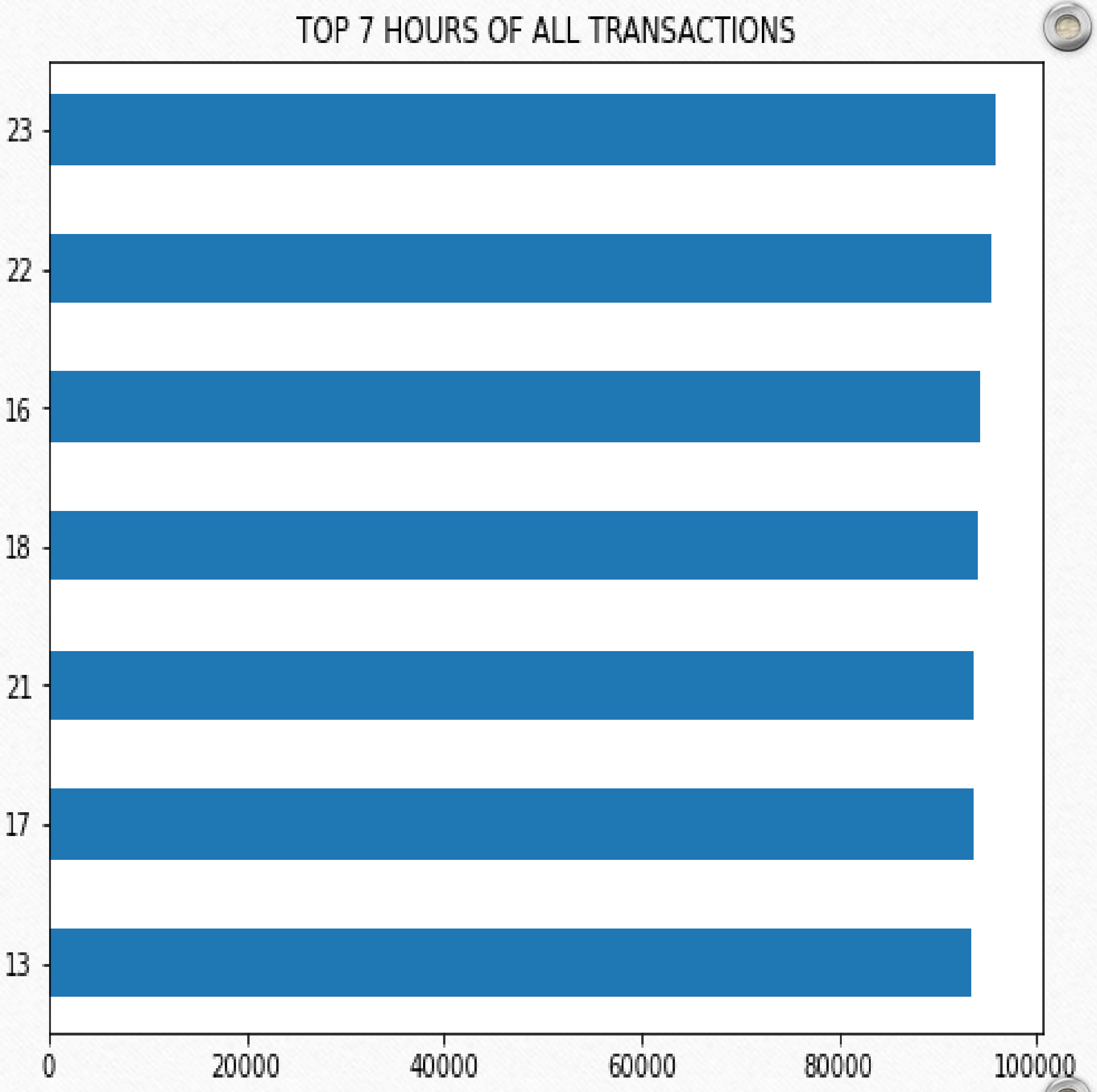

TOP 7 MONTH SORTED BY FRAUD TRANSACTIONS
March
May
February
December
January
October
June
0
200
400
600
800


TOP 7 MONTH SORTED BY TOTAL NUMBER OF TRANSACTIONS
December
August
June
July
May
March
November
0
50000
100000
150000
200000
250000

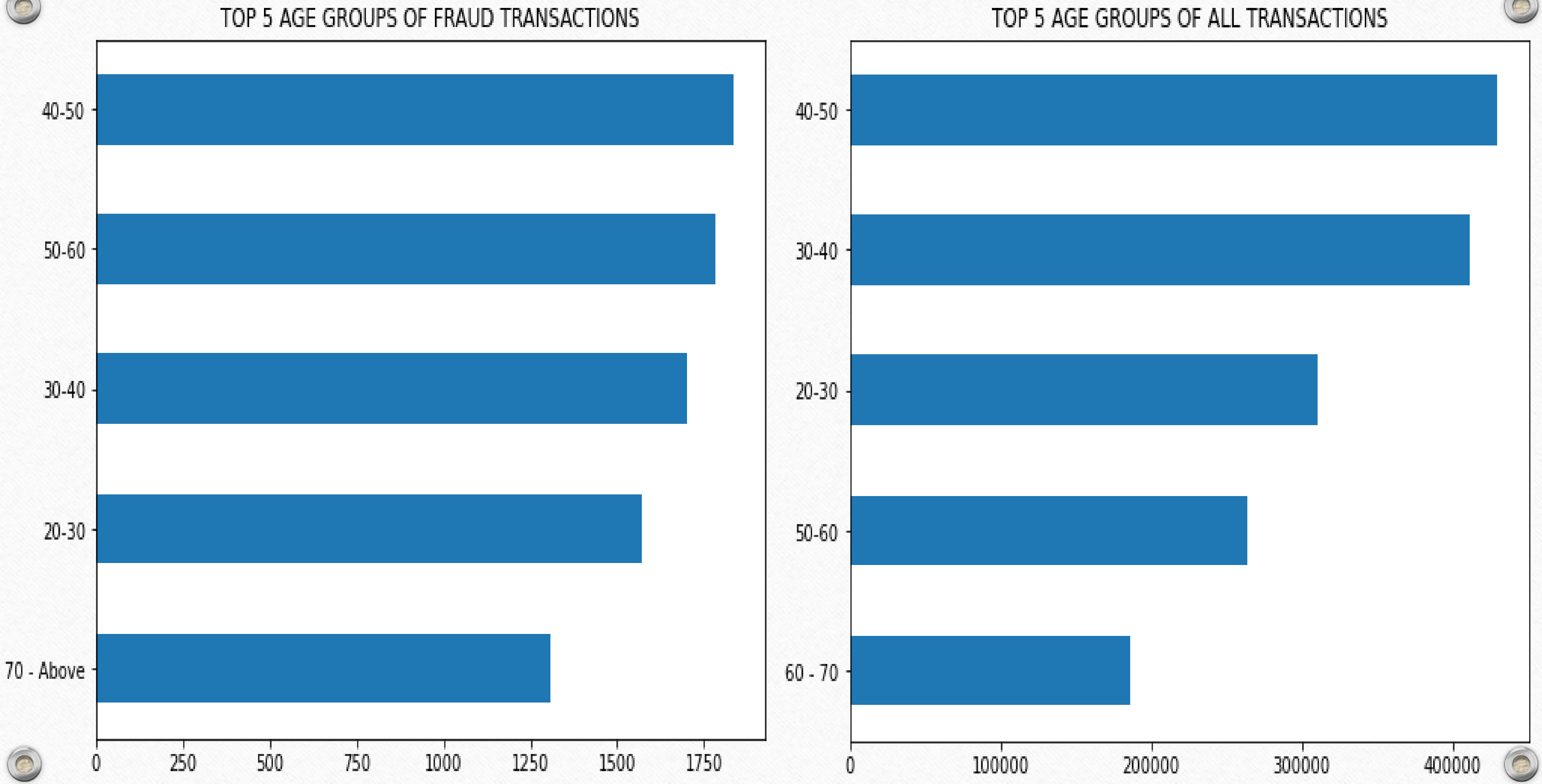

WEEK DAYS SORTED BY FRAUD TRANSACTIONS
Sunday
Saturday
Monday
Friday
Thursday
Tuesday
Wednesday
0
200
400
600
800
1000
1200
1400
1600
WEEK DAYS SORTED BY TOTAL NUMBER OF TRANSACTIONS
Monday
Sunday
Tuesday
Saturday
Friday
Thursday
Wednesday
0
50000
100000
150000
200000
250000
300000
350000

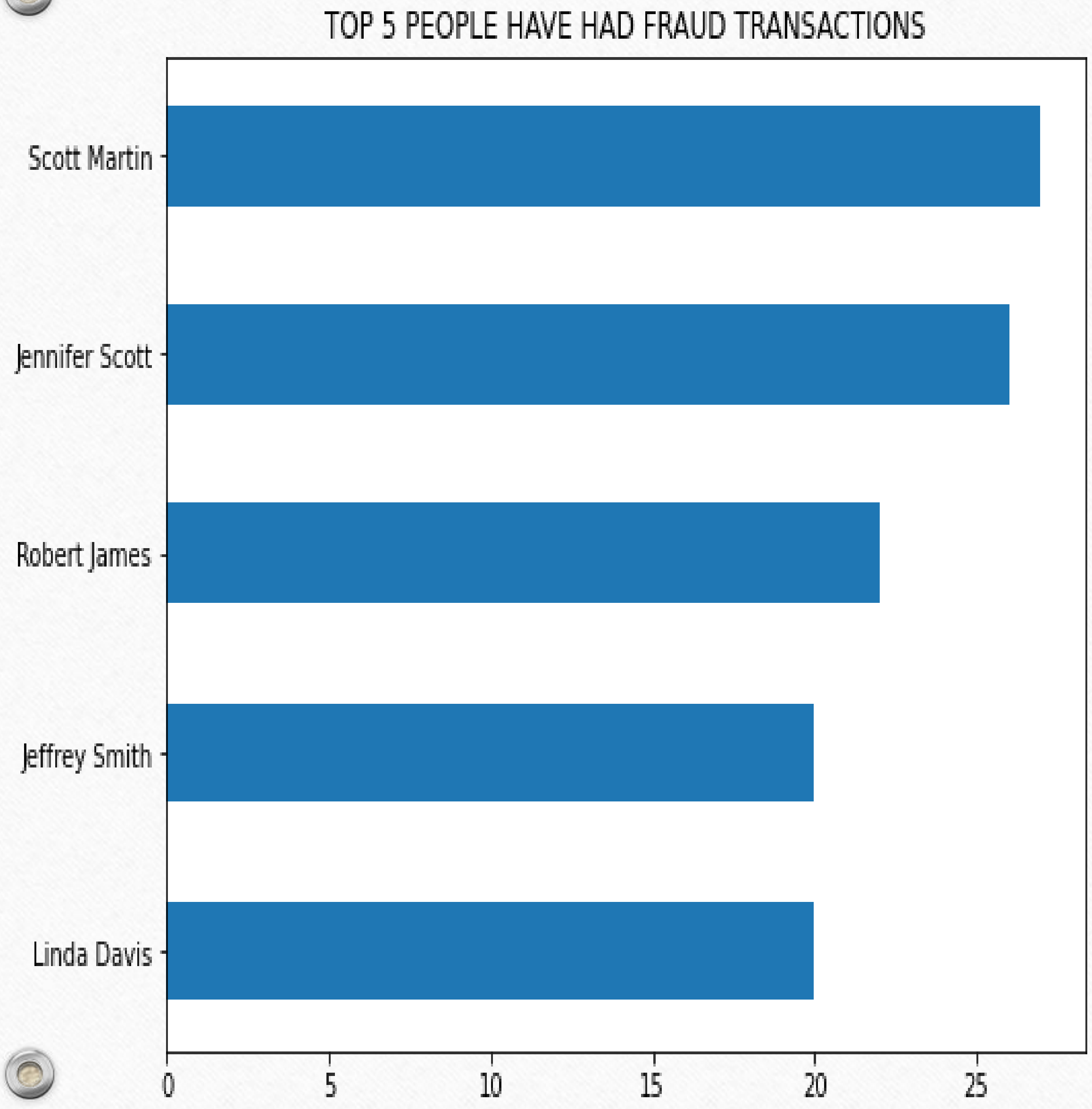


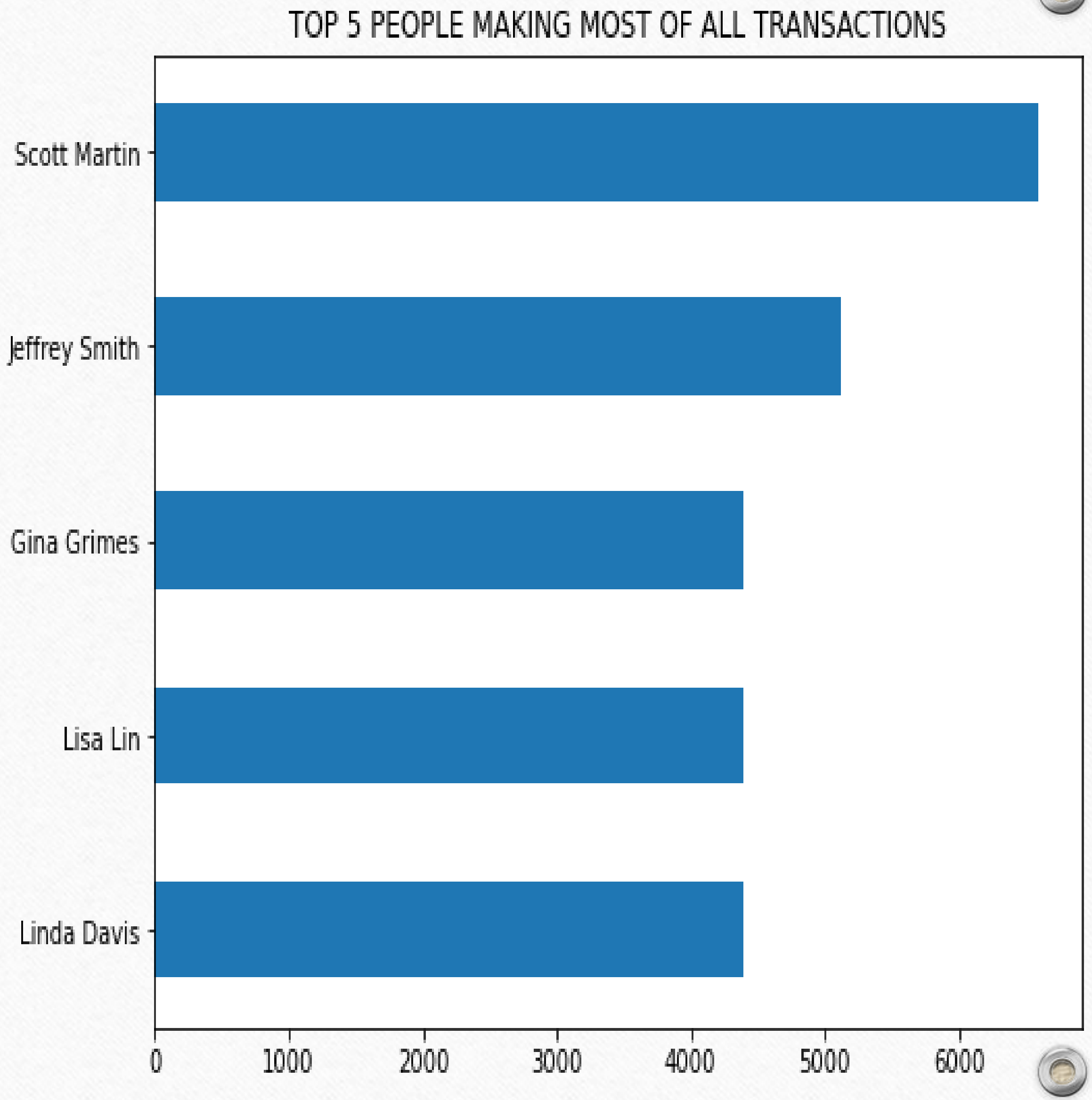
TOP 7 HOURS OF FRAUD TRANSACTIONS
22
23
1
0
3
2
18
0
500
1000
1500
2000
2500



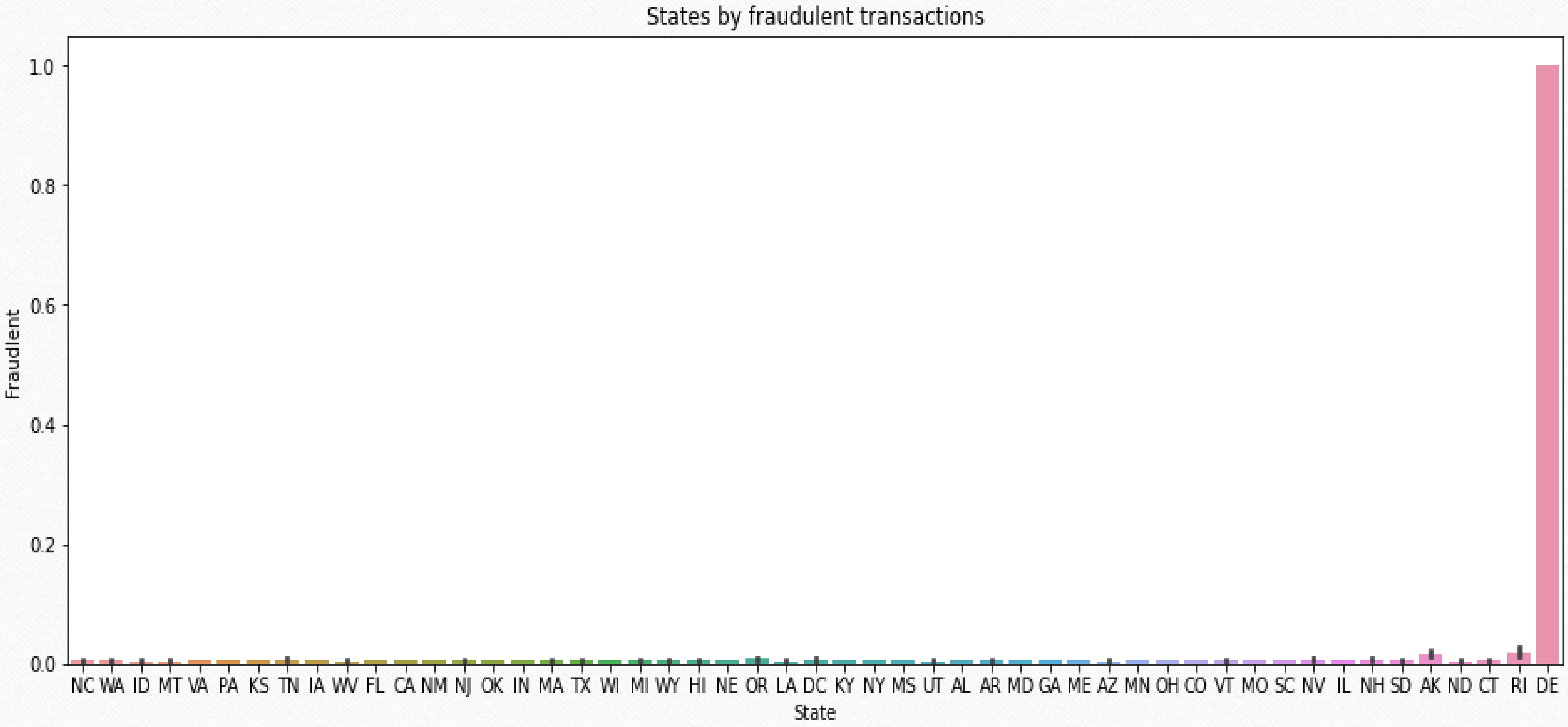
TOP 7 HOURS OF ALL TRANSACTIONS
23
22
16
18
21
17
13
0
20000
40000
60000
80000
100000


TOP 5 AGE GROUPS OF FRAUD TRANSACTIONS
40-50
50-60
30-40
20-30
70 - Above
0
250
500
750
1000
1250
1500
1750
TOP 5 AGE GROUPS OF ALL TRANSACTIONS
40-50
30-40
20-30
50-60
60 - 70
0
100000
200000
300000
400000

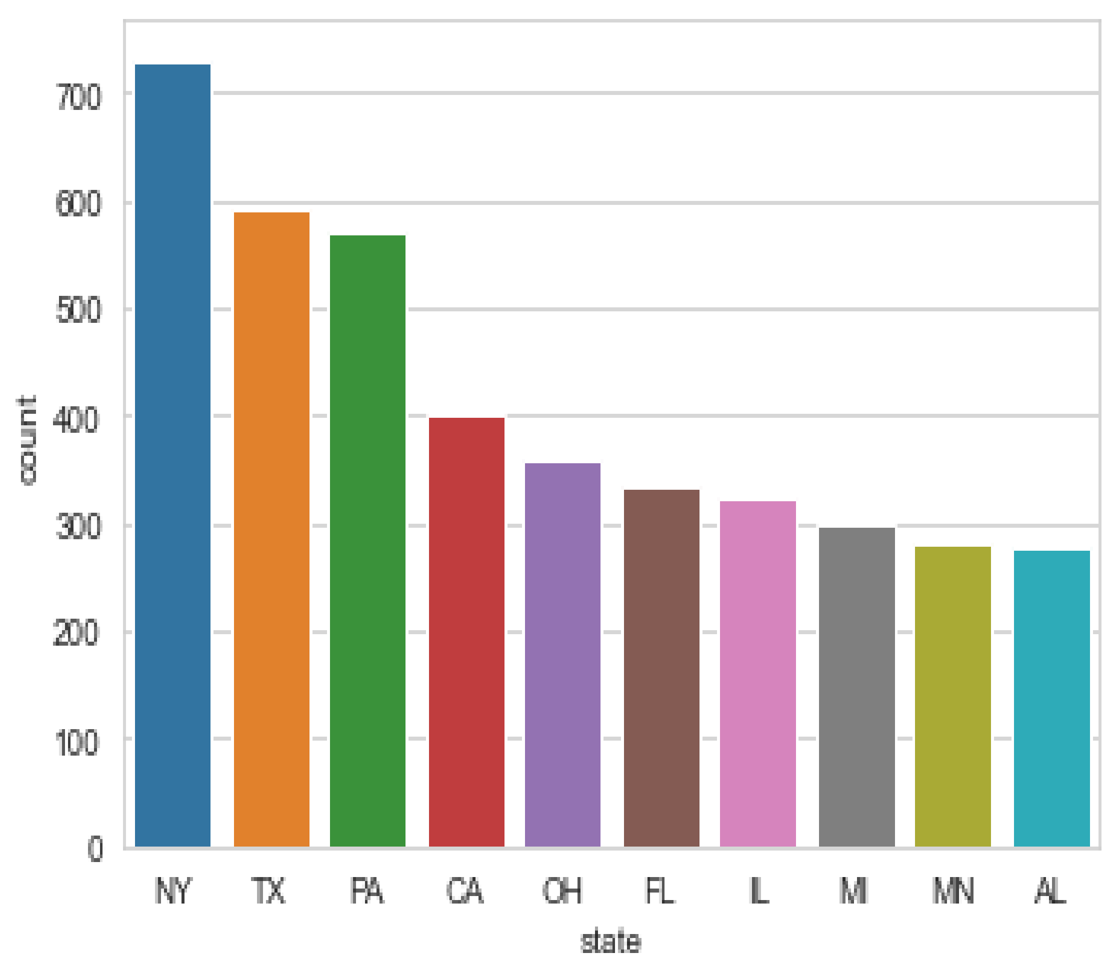

TOP 5 PEOPLE HAVE HAD FRAUD TRANSACTIONS
Scott Martin
Jennifer Scott
Robert James
Jeffrey Smith
Linda Davis
0
5
10
15
20
25


TOP 5 PEOPLE MAKING MOST OF ALL TRANSACTIONS
Scott Martin
Jeffrey Smith
Gina Grimes
Lisa Lin
Linda Davis
0
1000
2000
3000
4000
5000
6000

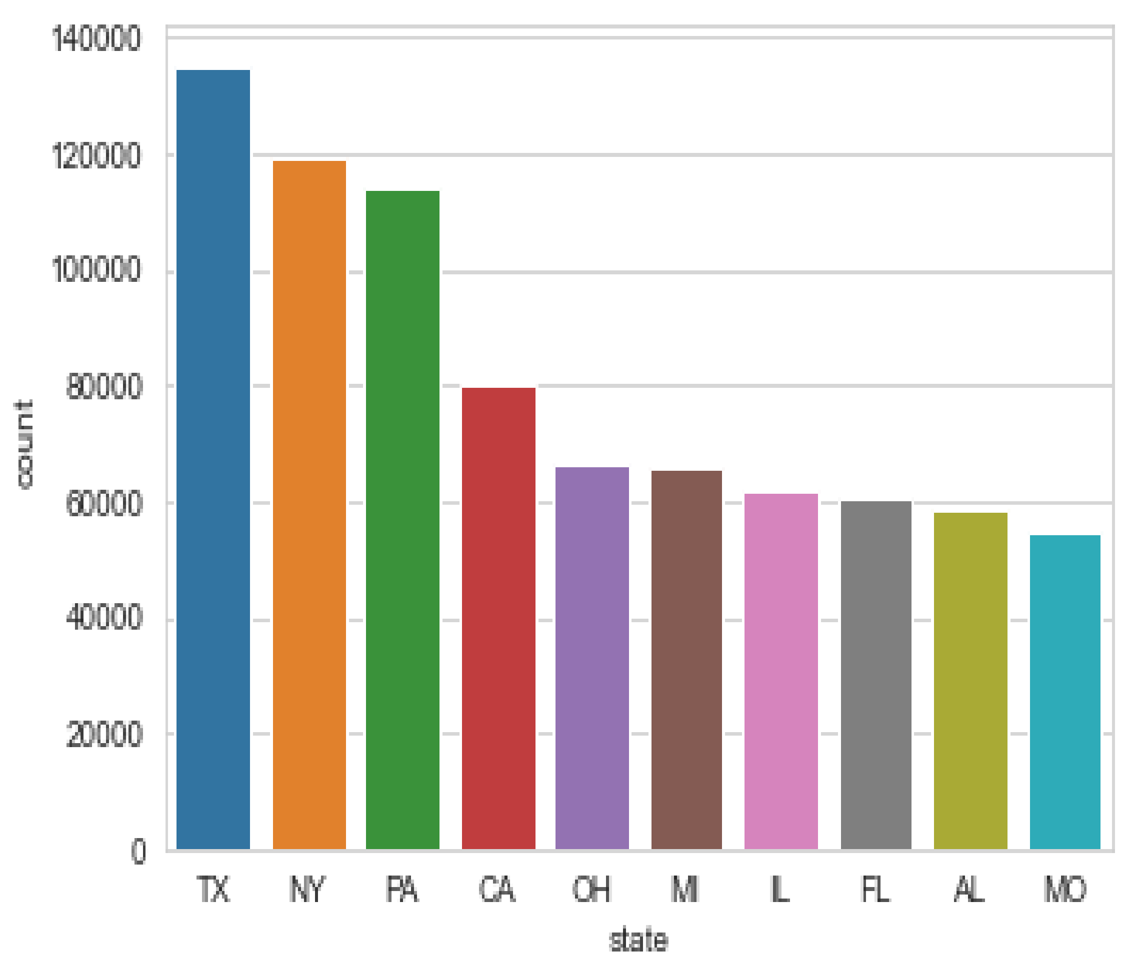
State Delaware is on top position by Fraud transactions.



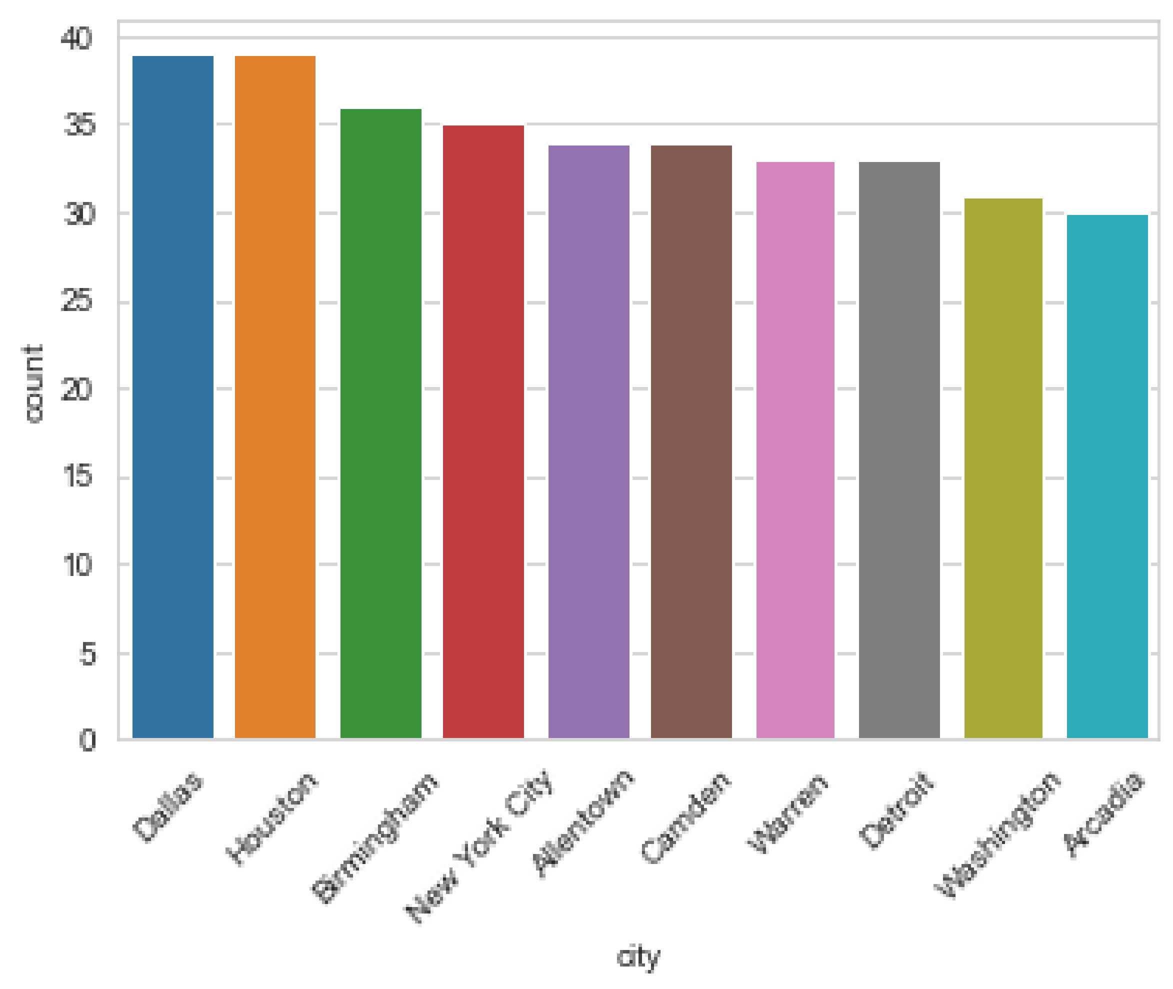
Top 10 States of Fraudulent Transactions



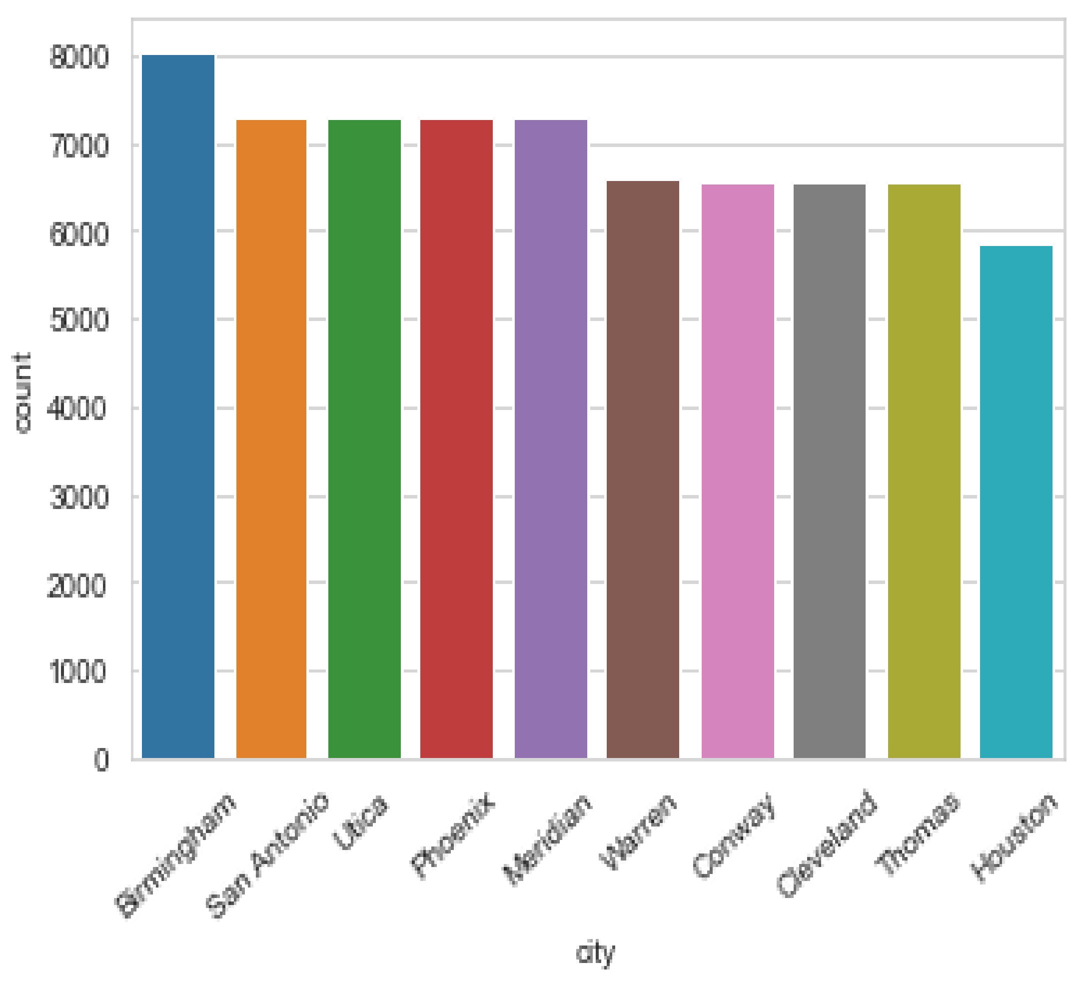
Top 10 States of All Transactions



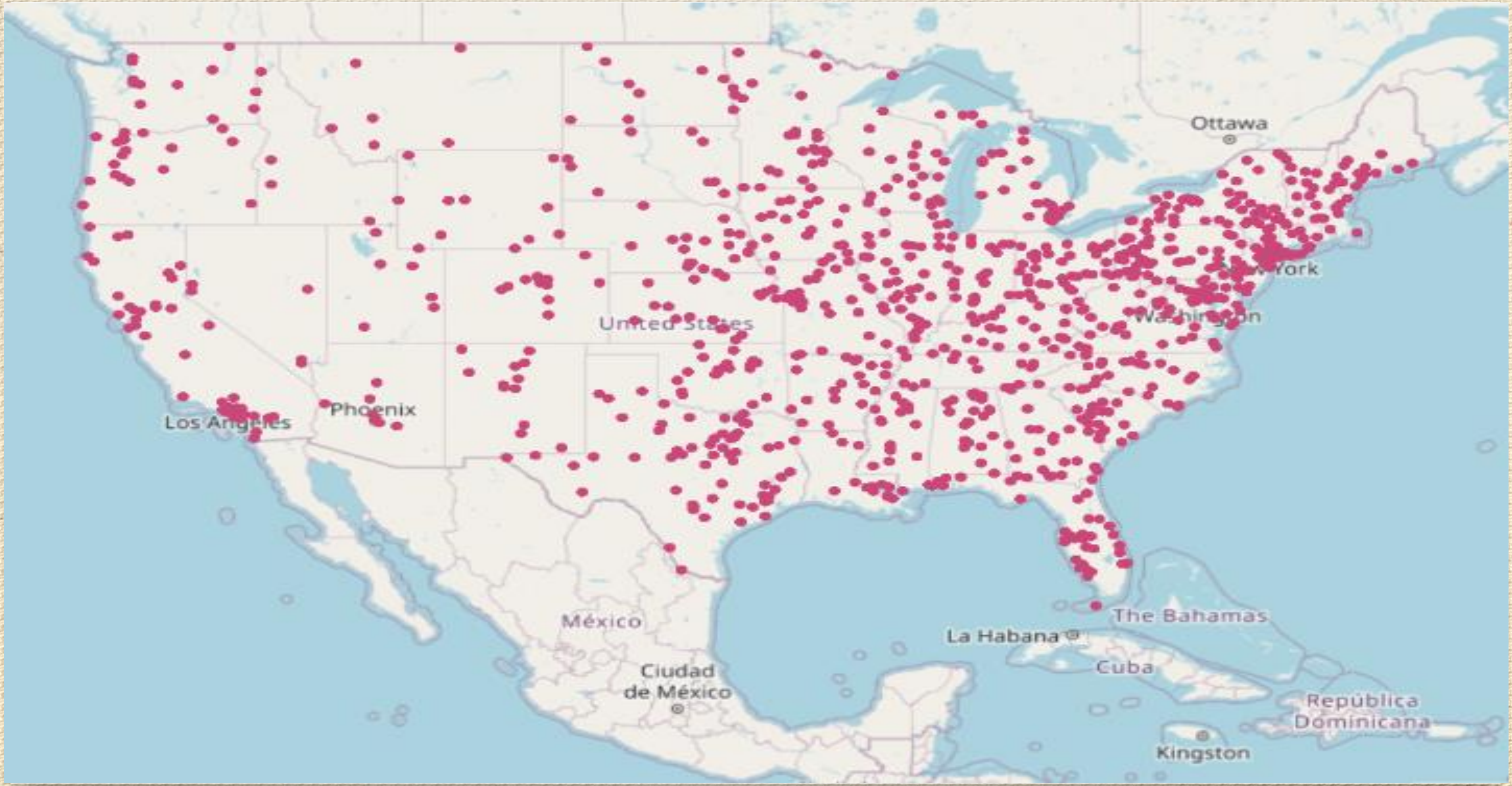
Top 10 Cities of Fraudulent Transactions

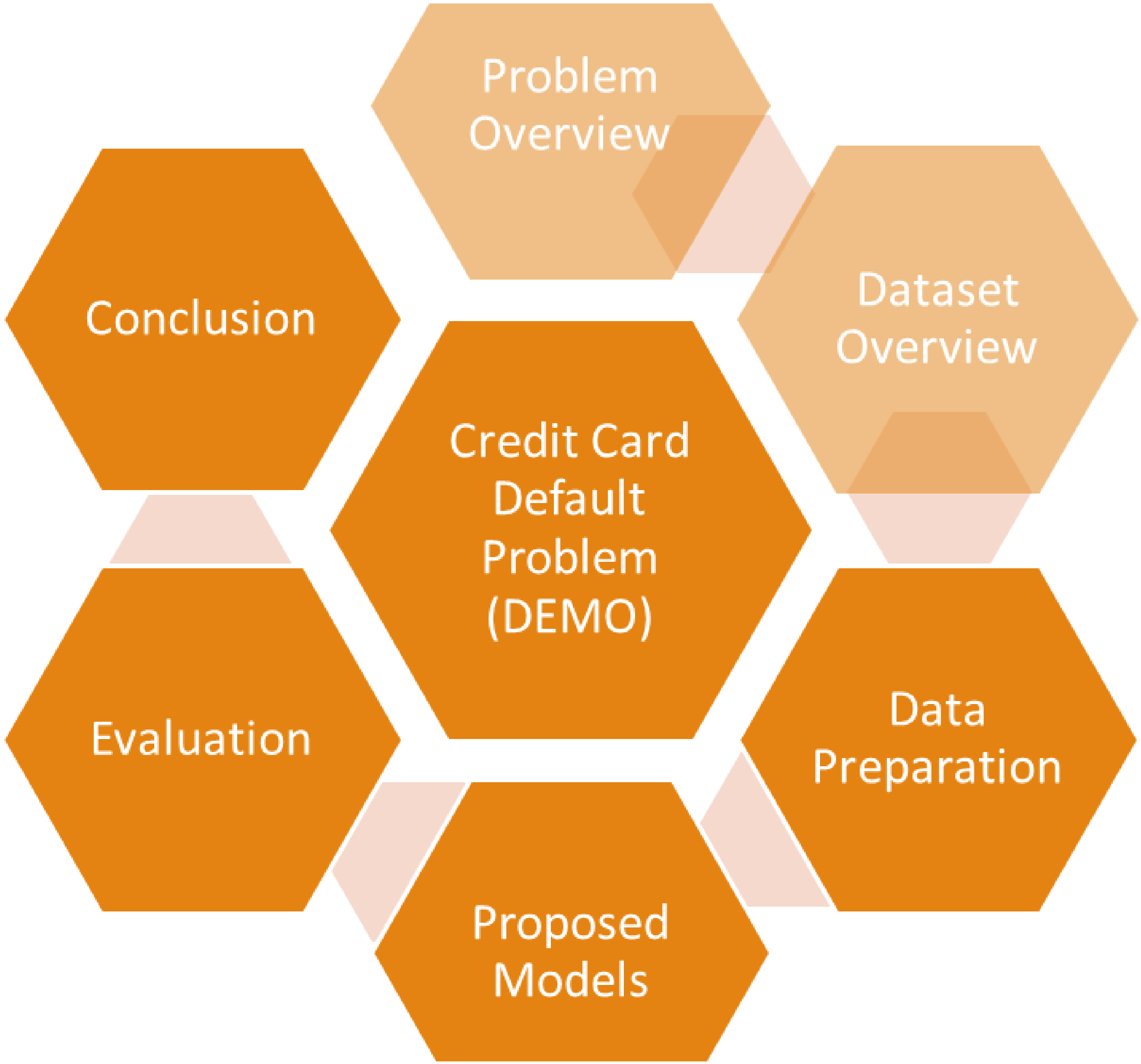


Top 10 Cities of All Transactions



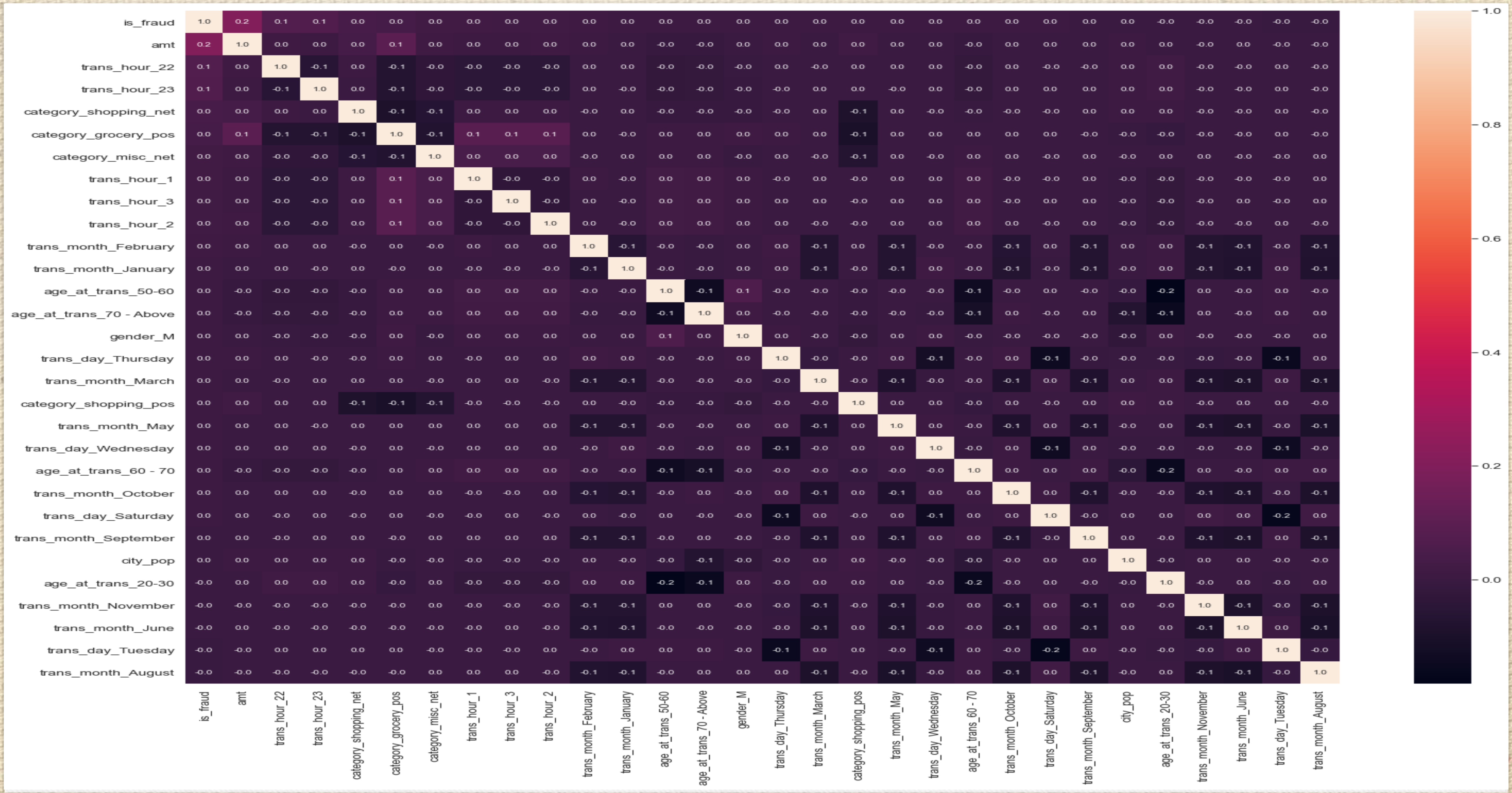
Fraud Transactions by location

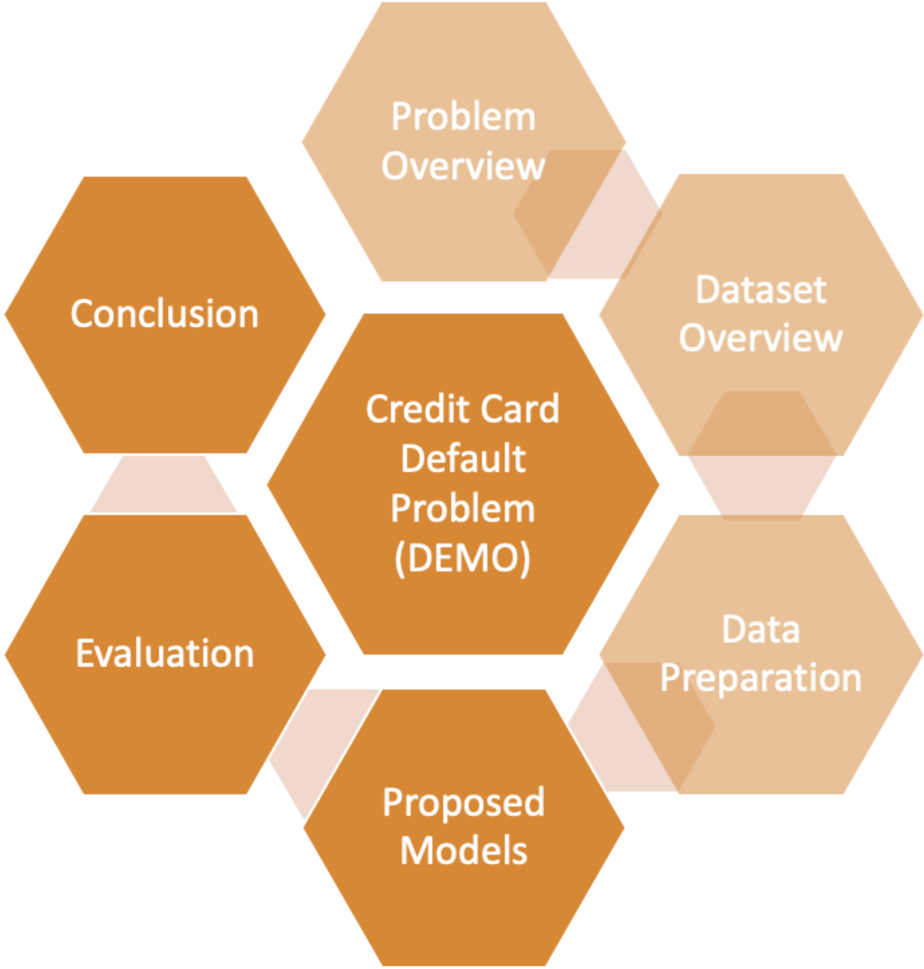



Problem Overview
Dataset Overview
Conclusion
Credit Card Default Problem (DEMO)
Evaluation
Data Preparation
Proposed Models

DATA PREPARATION

- Dropping unnecessary columns for modeling.
- Creating dummies for categorical variables.
- Split the data 70:30 ratio for train and test respectively.
- For proper Machine Learning results I've used SMOTE and Random Undersample techniques.
- On next slide I've create a correlation heatmap, which shows, that all variables are independent, which is good for modeling.

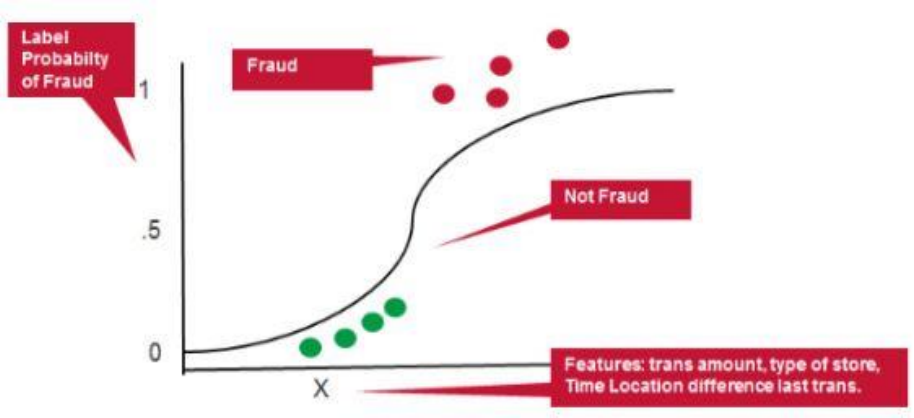

is_fraud
amt
trans_hour_22
trans_hour_23
category_shopping_net
category_grocery_pos
category_misc_net
trans_hour_1
trans_hour_3
trans_hour_2
trans_month_February
trans_month_January
age_at_trans_50-60
age_at_trans_70 - Above
gender_M
trans_day_Thursday
trans_month_March
category_shopping_pos
trans_month_May
trans_day_Wednesday
age_at_trans_60 - 70
trans_month_October
trans_day_Saturday
trans_month_September
city_pop
age_at_trans_20-30
trans_month_November
trans_month_June
trans_day_Tuesday
trans_month_August
1.0
0.8
0.6
0.4
0.2
0.0


Problem Overview
Dataset Overview
Data Preparation
Proposed Models
Evaluation
Conclusion
Credit Card Default Problem (DEMO)

Proposed Models

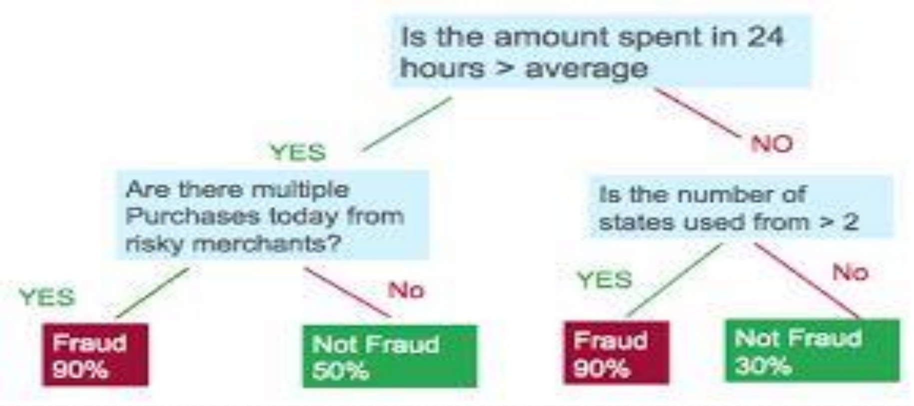
Logistic Regression:

- One of the most used ML algorithms in binary classification.
- Can be adjusted reasonably well to work on imbalanced data...useful for fraud detection.



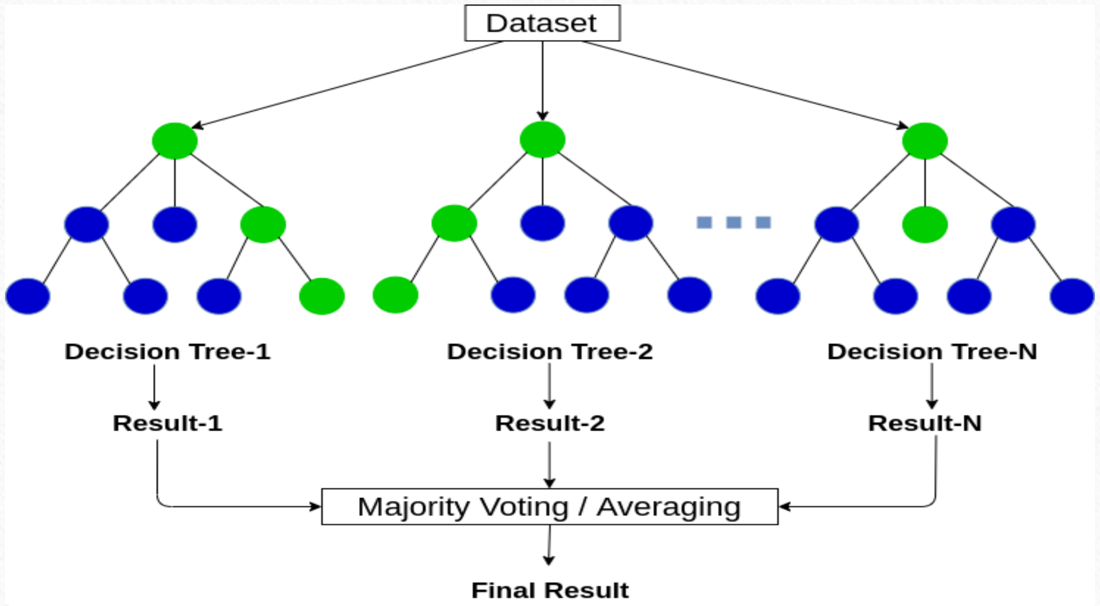
Decision Trees:

- Commonly used for fraud detection
- Transparent results, easily interpreted by analysts
- Decision trees are prone to overfit the data.



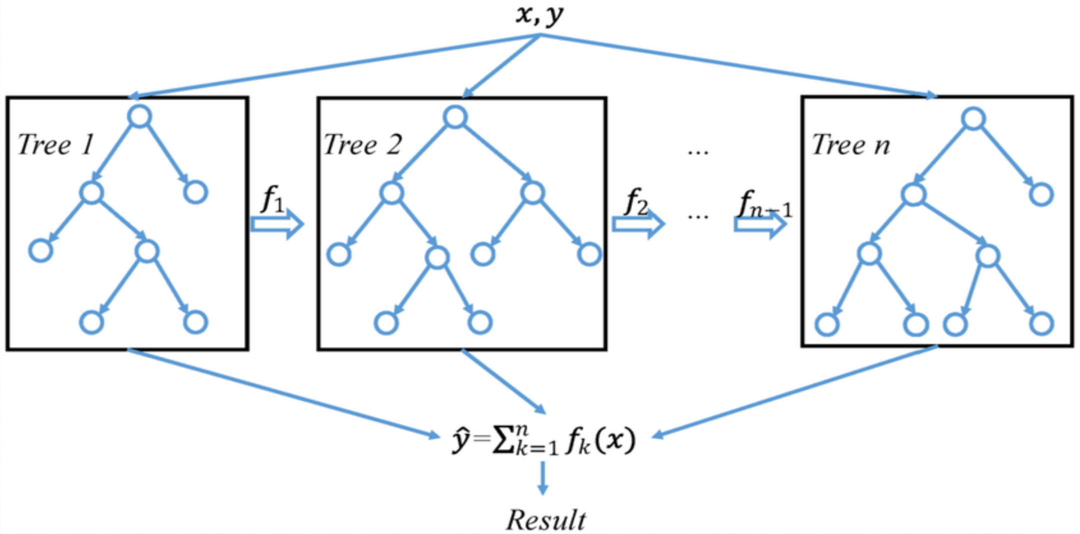
Random Forests:

- Are a more robust option than a single decision tree
- Construct a multitude of decision trees when training the model and outputting the class that is the mode or mean predicted class of the individual trees
- A random forest consists of a collection of trees on a random subset of features
- Final predictions are the combined results of those tree.
- Random forests can handle complex data and are not prone to overfit
- Very popular for fraud detection.



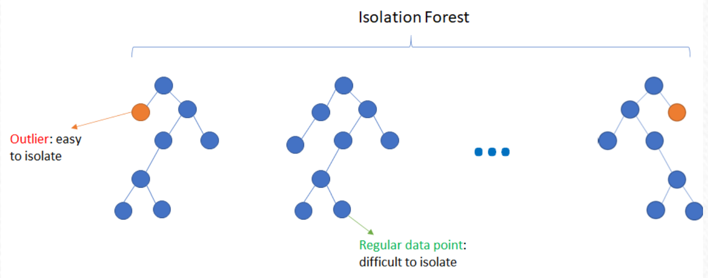
XGBoost Classifier:

- Is a popular and efficient open-source implementation of the gradient boosted trees algorithm.
- Gradient boosting is a supervised learning algorithm, which attempts to accurately predict a target variable by combining the estimates of a set of simpler, weaker models.



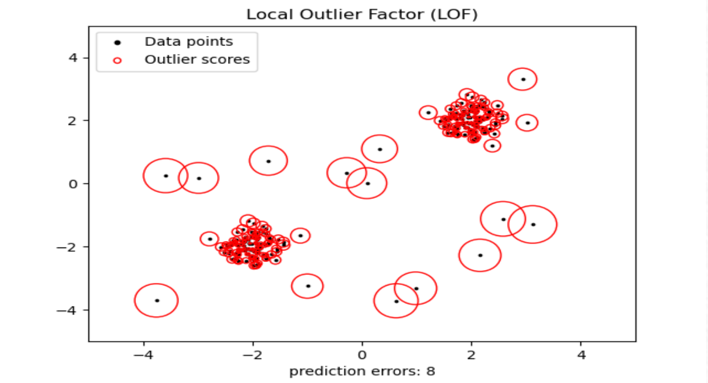
- **Isolation forest:**

Is an unsupervised algorithm for anomaly detection that works on principle of isolating anomalies. Instead of trying to build a model of normal instances, it explicitly isolates anomalous points in the dataset. It is a very fast algorithm with a low memory demand.



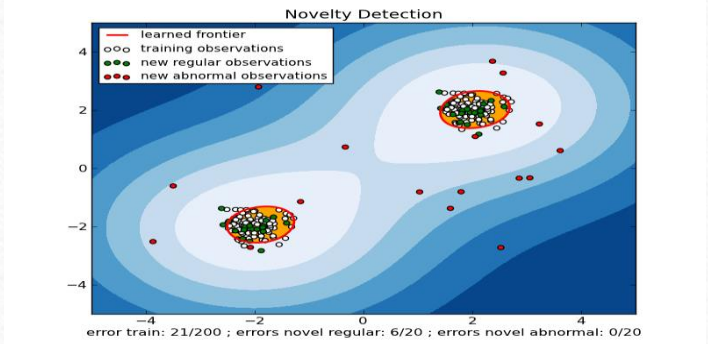
- **Local Outlier Factor (LOF):**

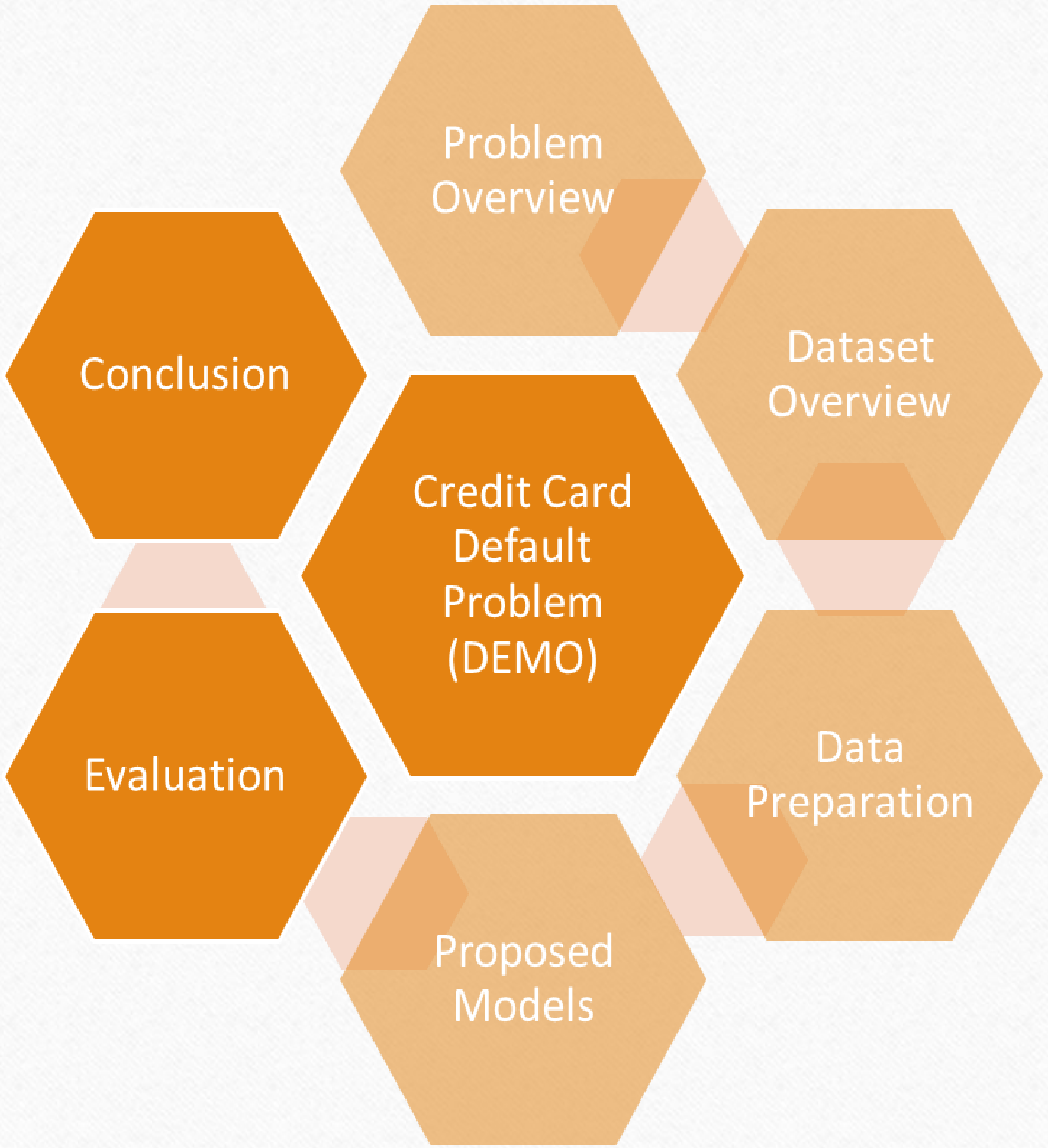
Is an unsupervised anomaly detection method which computes the local density deviation of a given data point with respect to its neighbors. It considers as outliers the samples that have a substantially lower density than their neighbors.



- **One-Class SVM:**

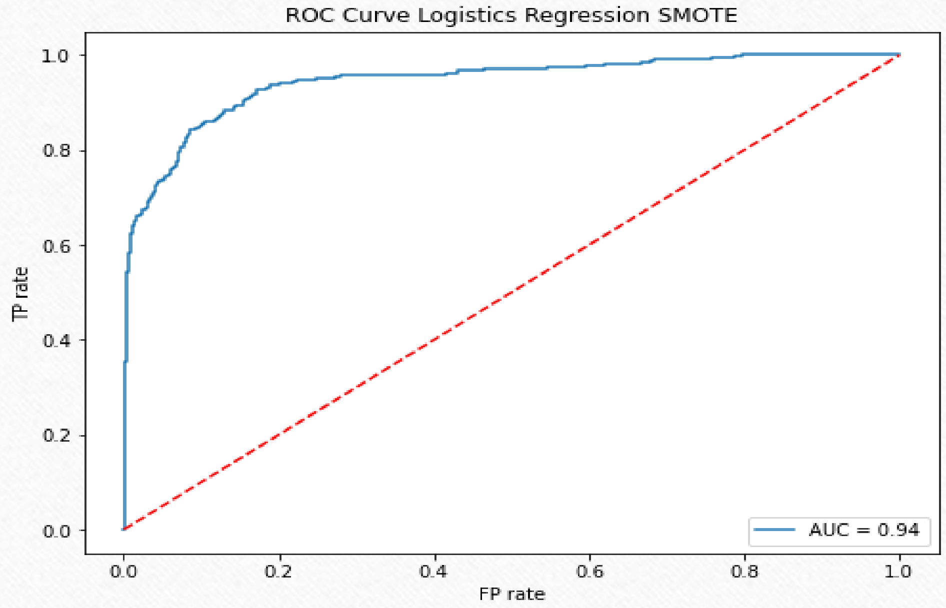
A classification method is used to detect the outliers and anomalies in a dataset. Based on Support Vector Machines (SVM) evaluation, the One-class SVM applies a One-class classification method for novelty detection.

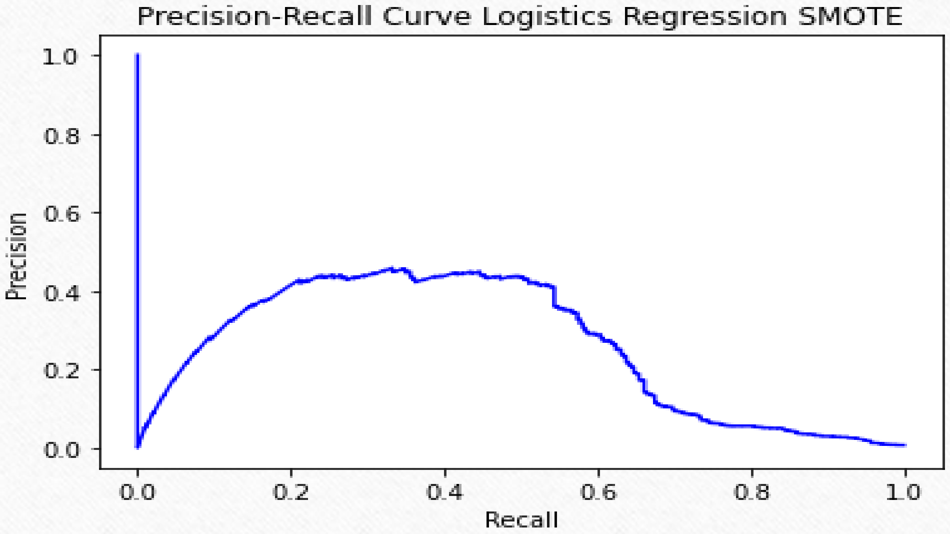



Problem Overview
Dataset Overview
Data Preparation
Proposed Models
Evaluation
Conclusion
Credit Card Default Problem (DEMO)

Logistic Regression using SMOTE technique.

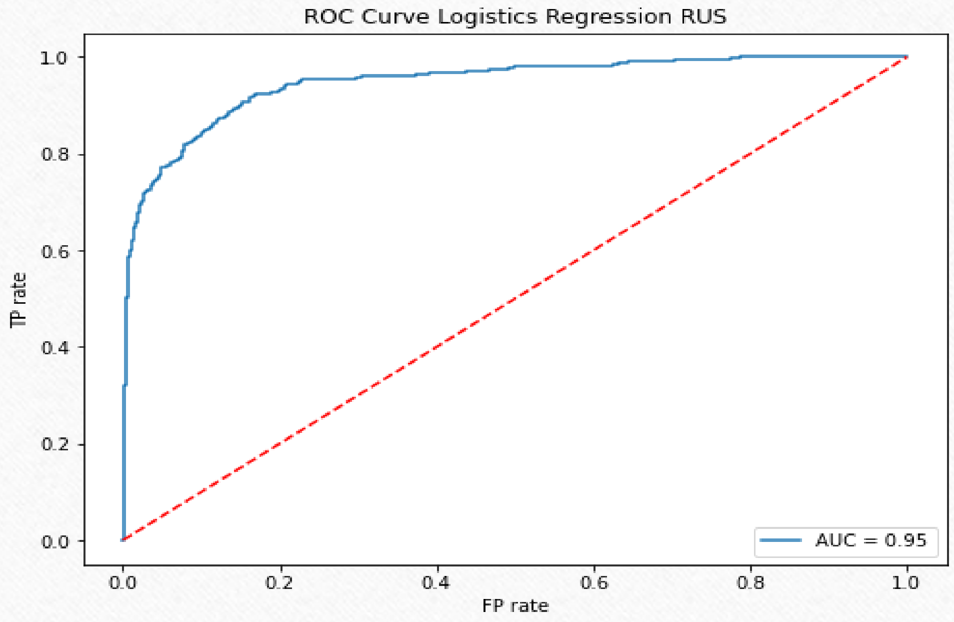
- Accuracy train score 0.91
- Accuracy test score 0.90
- Average Cross-Validation score 0.91
- Confusion Matrix :[49862 5412]
  [45 253]
- Precision 0.04
- Recall 0.85
- F1 score 0.08

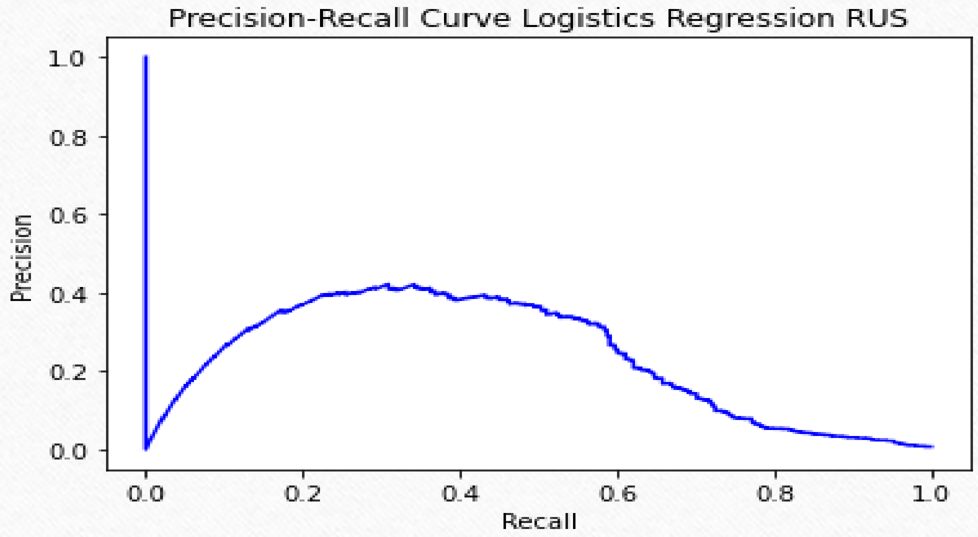




Logistic Regression using Random Under Sample technique.

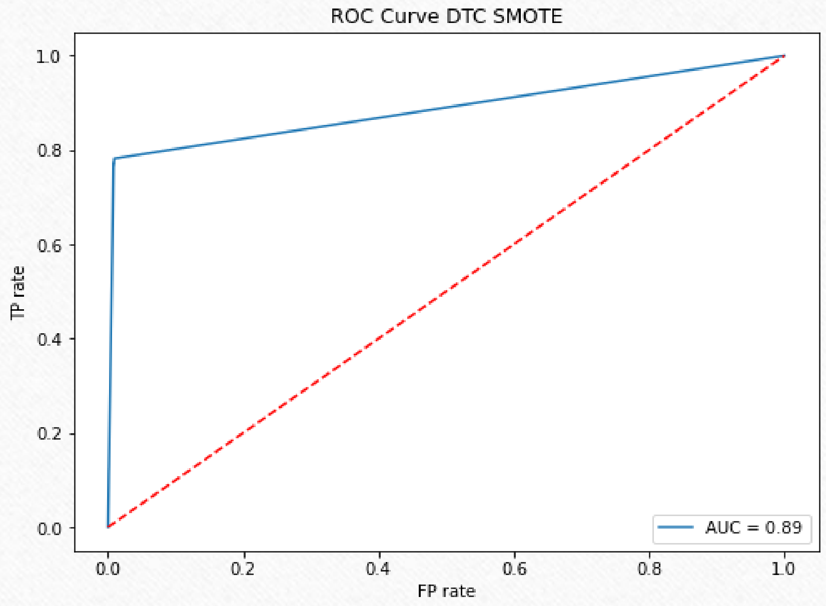
- Accuracy train score 0.90
- Accuracy test score 0.88
- Average Cross-Validation score 0.88
- Confusion Matrix :[48885 6389]
  [41 257]
- Precision 0.04
- Recall 0.86
- F1 score 0.07

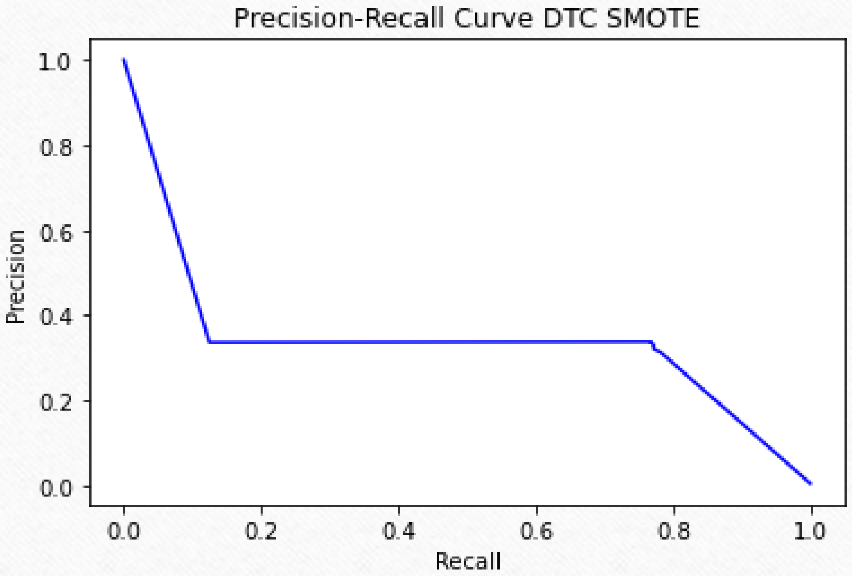




Decision Tree with SMOTE

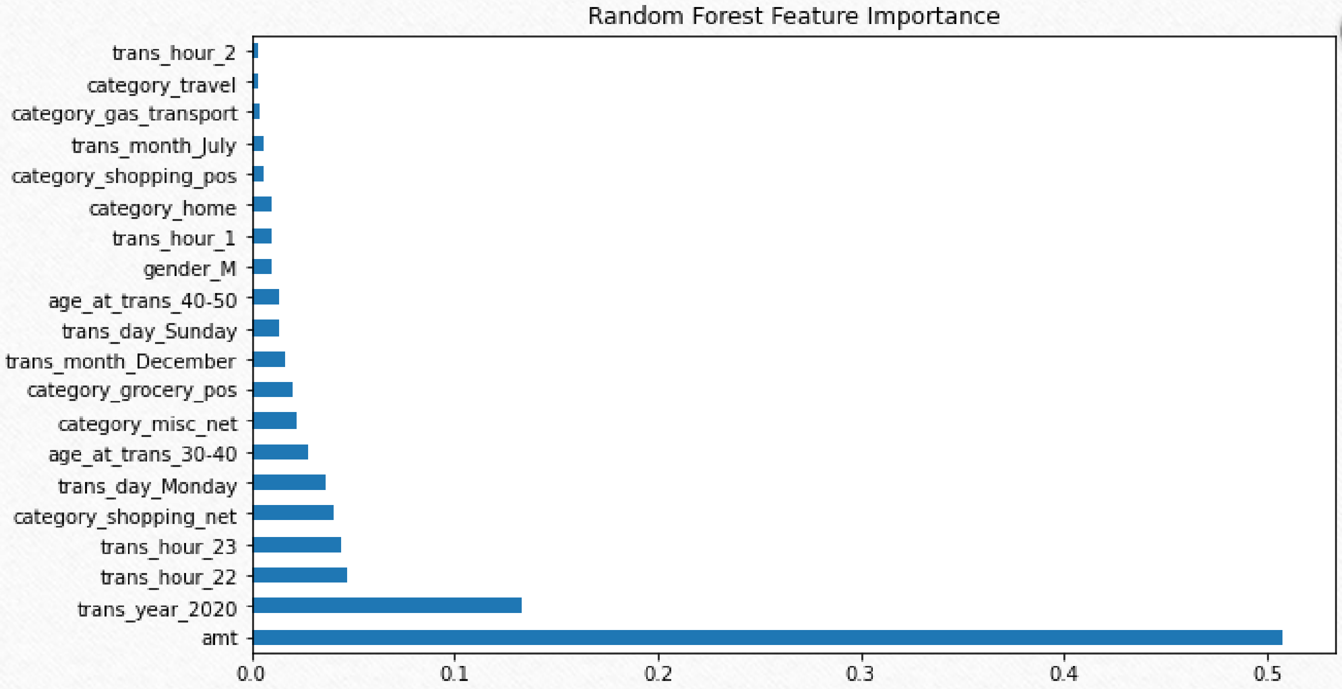
- Accuracy train score 0.99
- Accuracy test score 0.99
- Average Cross Validation score 0.99
- Confusion Matrix : [55170 104]
  [89   209]
- Precision 0.67
- Recall 0.70
- F1 score 0.68

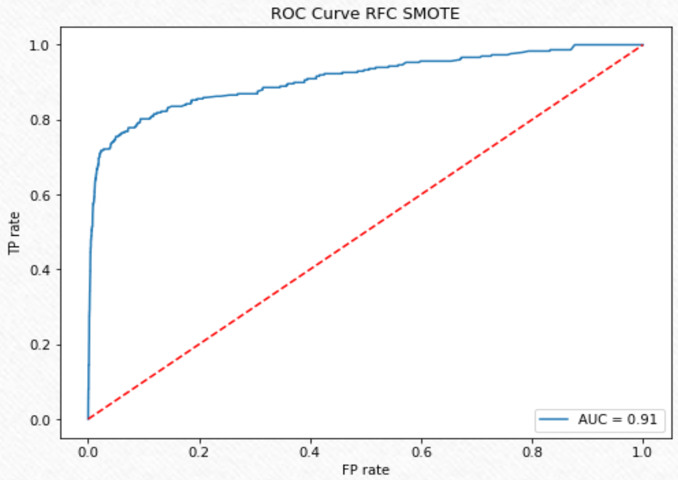


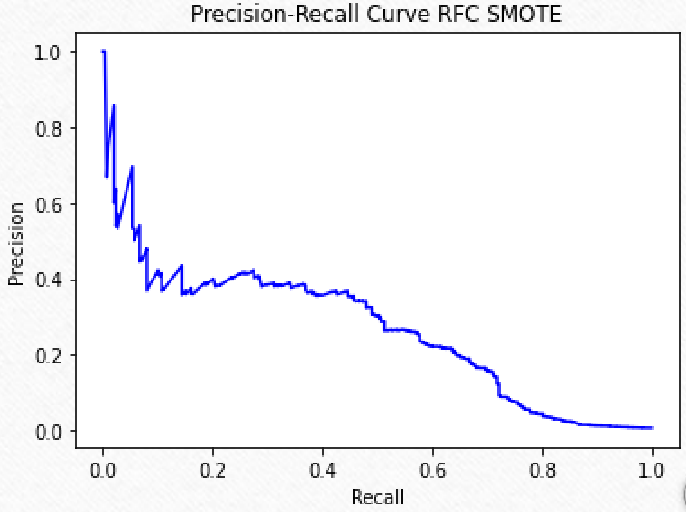


Random Forest Classification using SMOTE technique.

- Accuracy train score 0.92
- Accuracy test score 0.97
- Average Cross-Validation score 0.92
- Confusion Matrix : [53969 1305]
  [84 214]
- Precision 0.14
- Recall 0.72
- F1 score 0.24

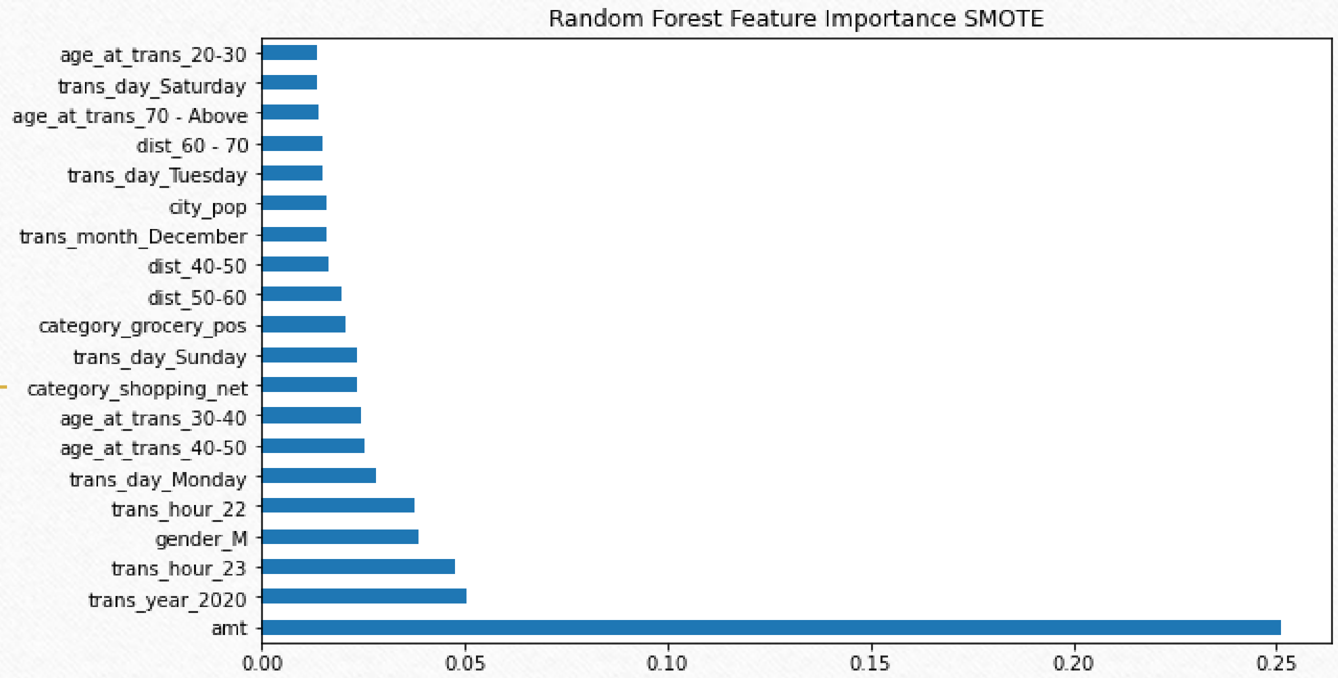


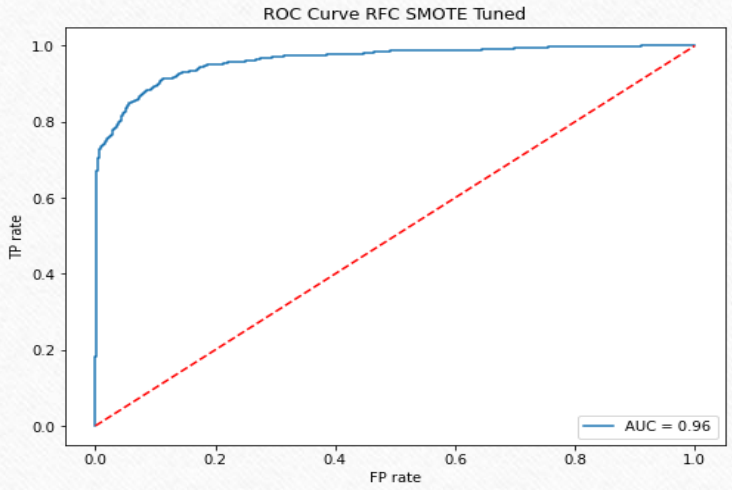


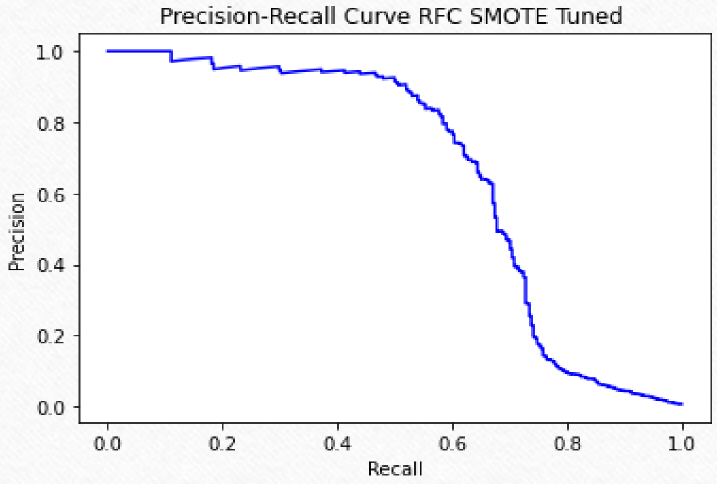


Random Forest optimized for best parameters.

- Accuracy train score: 0.99
- Accuracy test score 0.99
- Average Cross Validation score 0.99
- Confusion Matrix : [55268 6]
  [187 111]
- Precision 0.95
- Recall 0.37
- F1 score 0.53

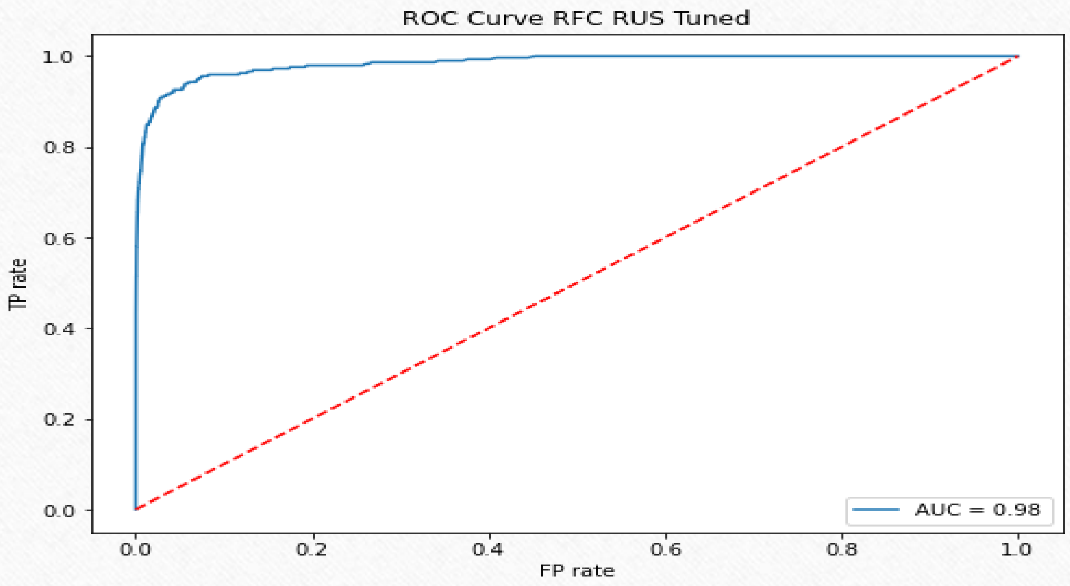


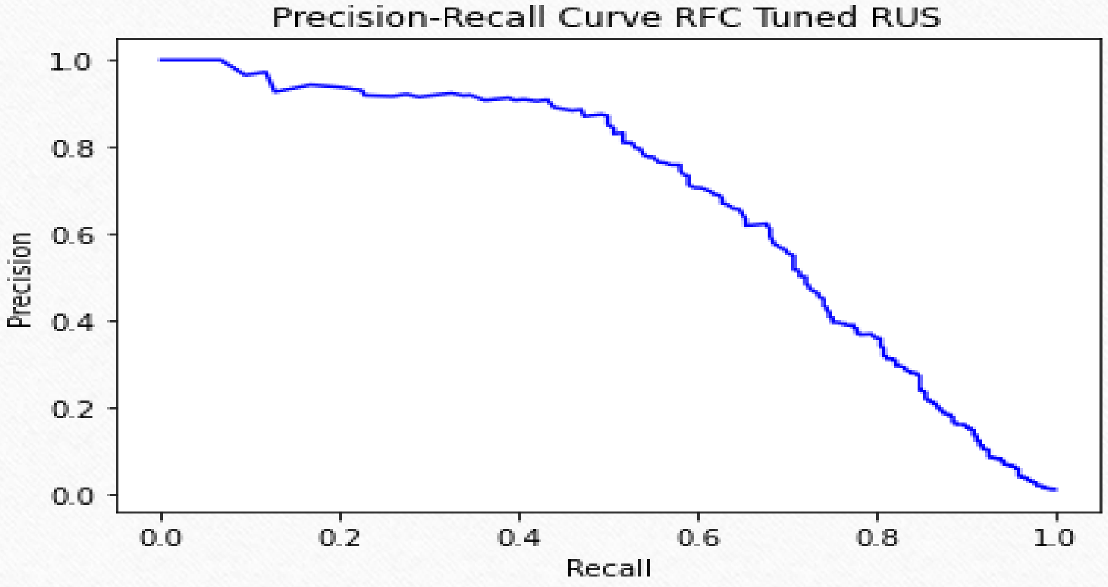




Random Forest Tuned with RUS

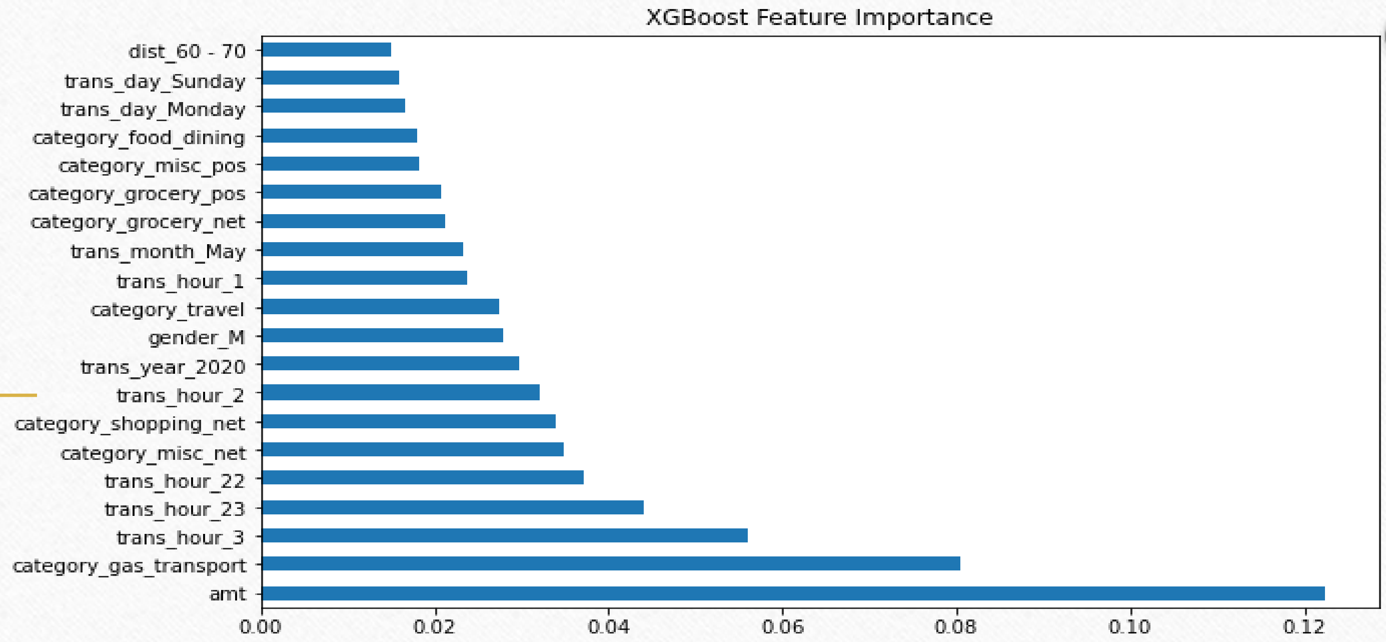
- Accuracy train score: 1.00
- Accuracy test score: 0.95
- Average Cross-Validation score: 0.95
- Confusion Matrix :[527784 2490]
  [22 276]
- Precision 0.10
- Recall 0.93
- F1 score 0.18

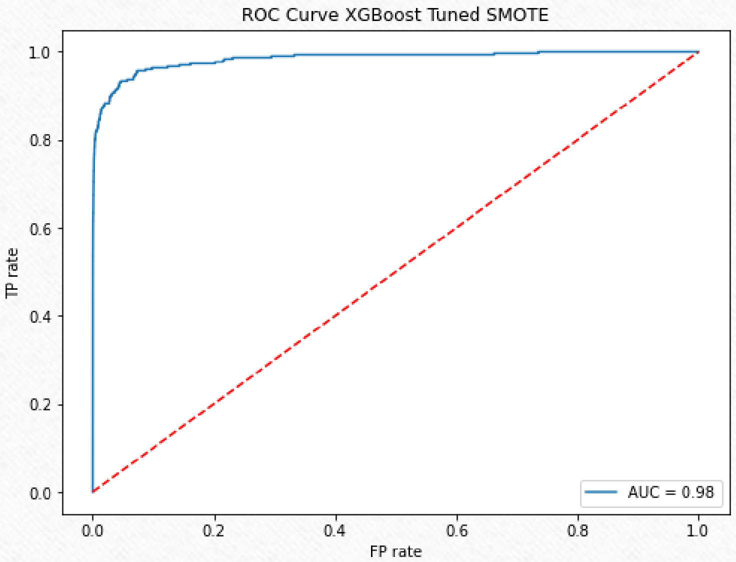


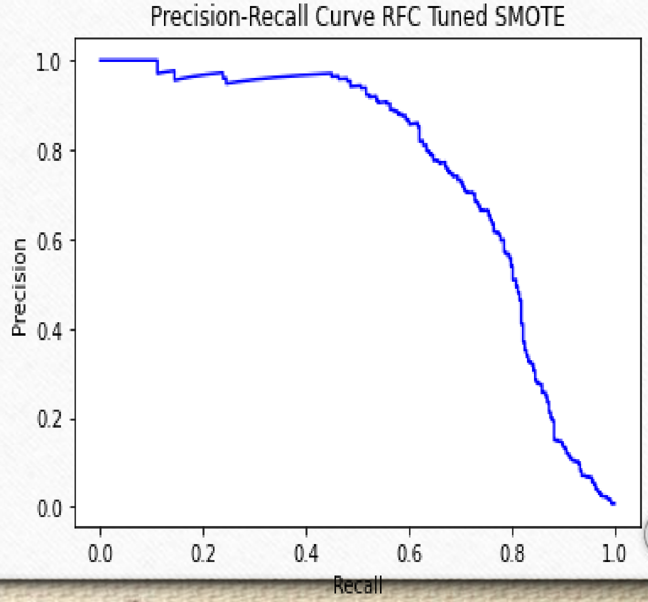


XGBoost Classifier using SMOTE technique and optimized for best parameters.

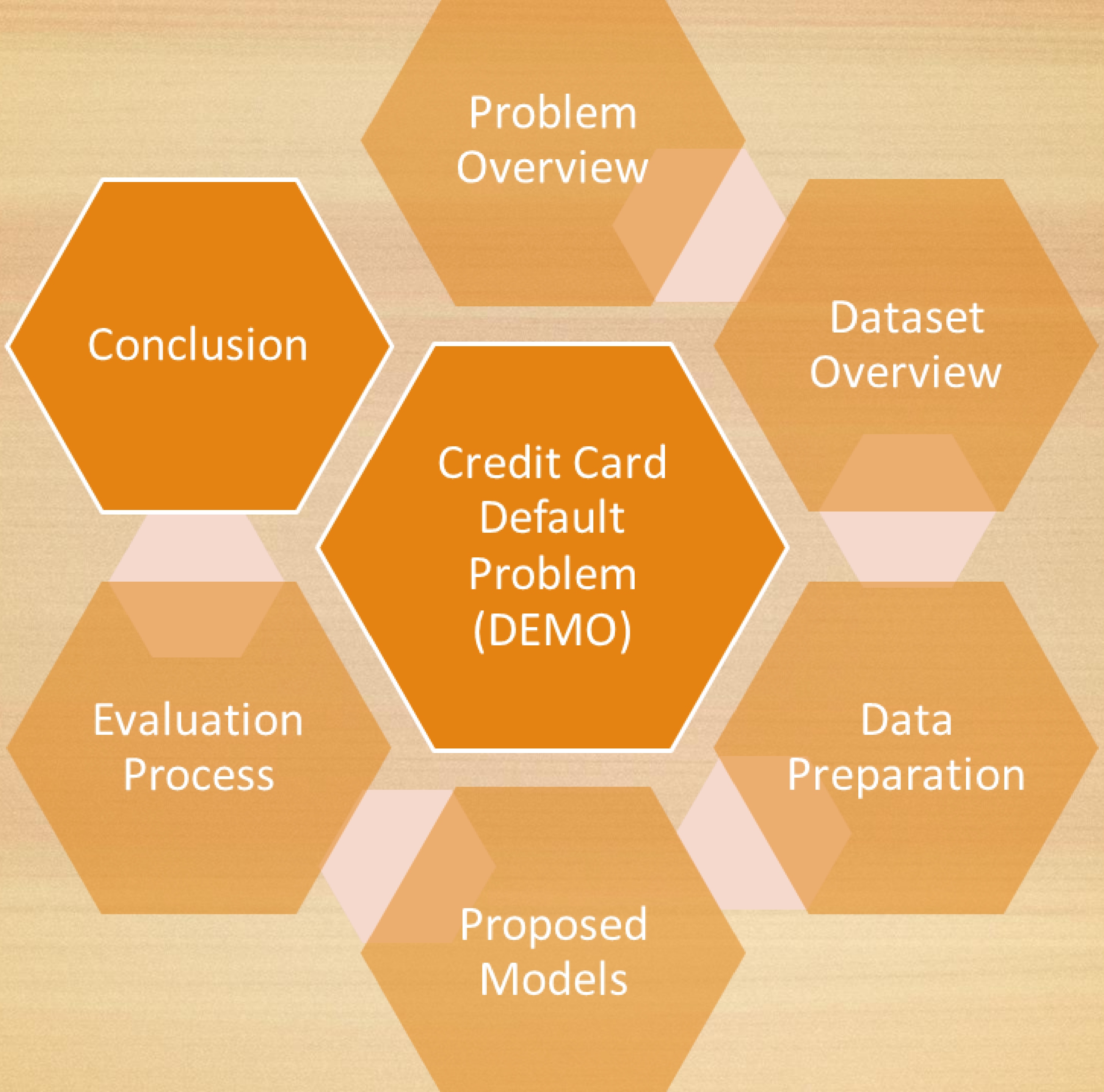
- Accuracy train score 0.99
- Accuracy test score 0.99
- Average Cross Validation score 0.99
- Confusion Matrix : [55143 131]
  [71 227]
- Precision 0.63
- Recall 0.76
- F1 score 0.69







Model	Accuracy	Precision	Recall	F1	Confusion Matrix	AUC Score
Logistic Regression With Smote	0.98	0.15	0.34	0.20	[9862 5412] [45 253]	0.92
Decision Trees With Smote	0.98	0.20	0.93	0.33	[542071 10753] [191 2704]	1.00
Random Forest With Smote	0.92	0.14	0.72	0.24	[53969 1305] [84 214]	0.91
Random Forest Tuned with SMOTE	0.99	0.95	0.37	0.53	[55268 6] [187 111]	0.96
Random Forest Tuned with RUS	1.00	0.10	0.93	0.18	[52784 2490] [22 276]	0.98
XGBoost Classifier Tuned SMOTE	0.99	0.63	0.76	0.69	[55143 131] [71 227]	0.98


Problem Overview
Dataset Overview
Data Preparation
Proposed Models
Evaluation Process
Conclusion
Credit Card Default Problem (DEMO)

CONCLUSION

I've investigated the data, checked for imbalance, visualized the features and understood the relationship between different features.

The data was split into 2 parts train and test sets . Four different Supervised Machine Learning algorithms have been used: Logistic Regression, Decision Tree Classifier, Random Forest Classifier and XGBoost Classifier as well as two techniques for imbalanced data The Random Under Sampled technique and SMOTE technique. The GridSearch was used to find optimal hyper parameters of Random Forest and XGBoost models . As a result of modeling , the best model for credit card fraud detection using supervised models is XGBoost Classifier with optimal parameters and imbalance technique SMOTE found 227 fraud cases from 298 but missed 71 fraud cases and only 131 identified as fraud but it's not. Other model which work better for financial organization is Random Forest with optimal parameters and RUS technique found 276 fraud cases from 298 and only 22 was recognized as not fraud but it's fraud but 2,490 False Positive.

Future Work.

One additional work that could have been achieved but could not be completed due to time crunch was using neural networks to see if it could further improve the model results. Also, if I could have time for each of the models, I would apply other techniques for imbalanced data and tune my models.

Happy Credit Card Holders!

Thank you !