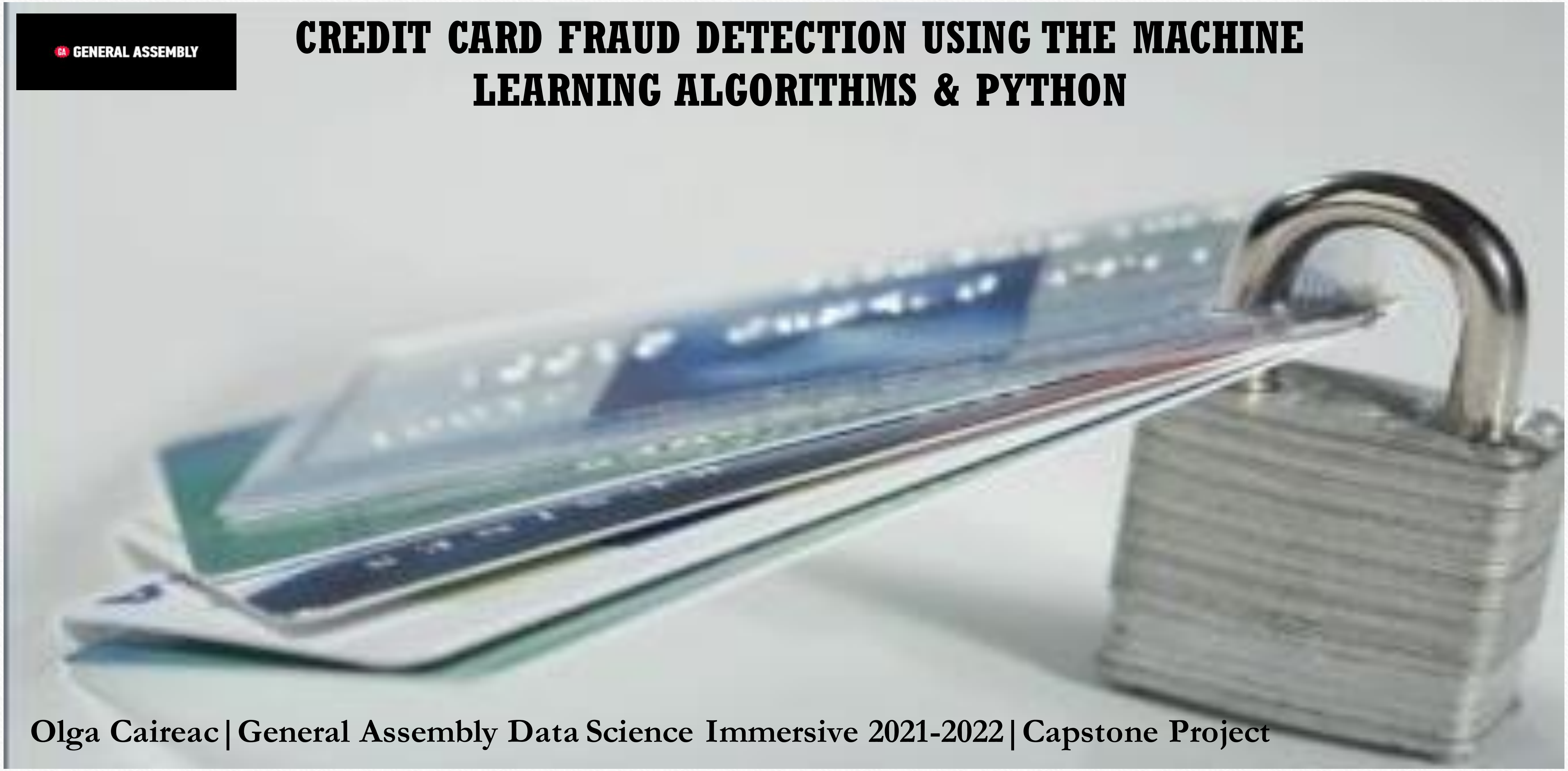
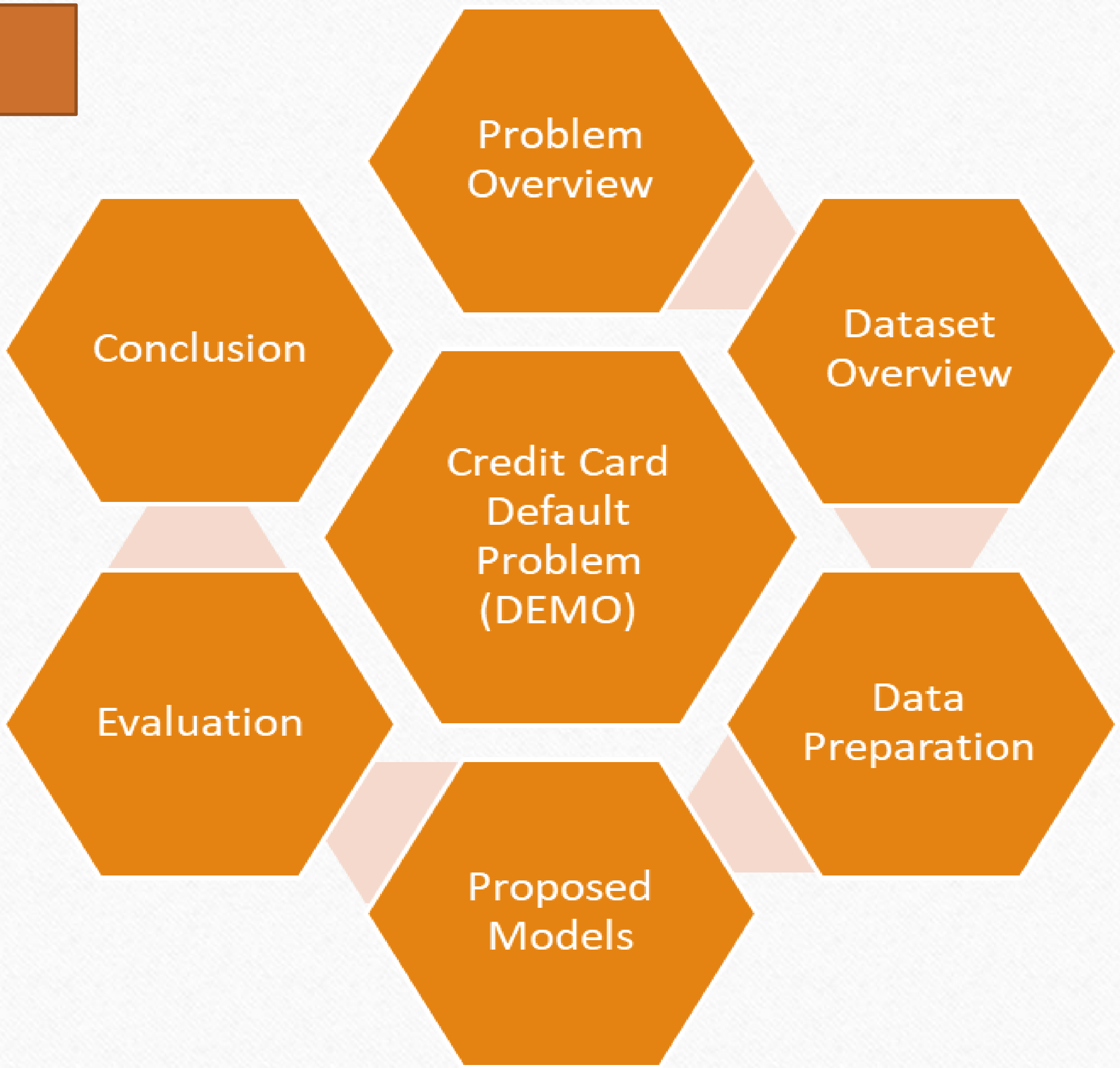
GENERAL ASSEMBLY

CREDIT CARD FRAUD DETECTION USING THE MACHINE LEARNING ALGORITHMS & PYTHON



Olga Caireac|General Assembly Data Science Immersive 2021-2022|Capstone Project

AGENDA



Credit Card Default Problem (DEMO)

Problem Overview

Dataset Overview

Data Preparation

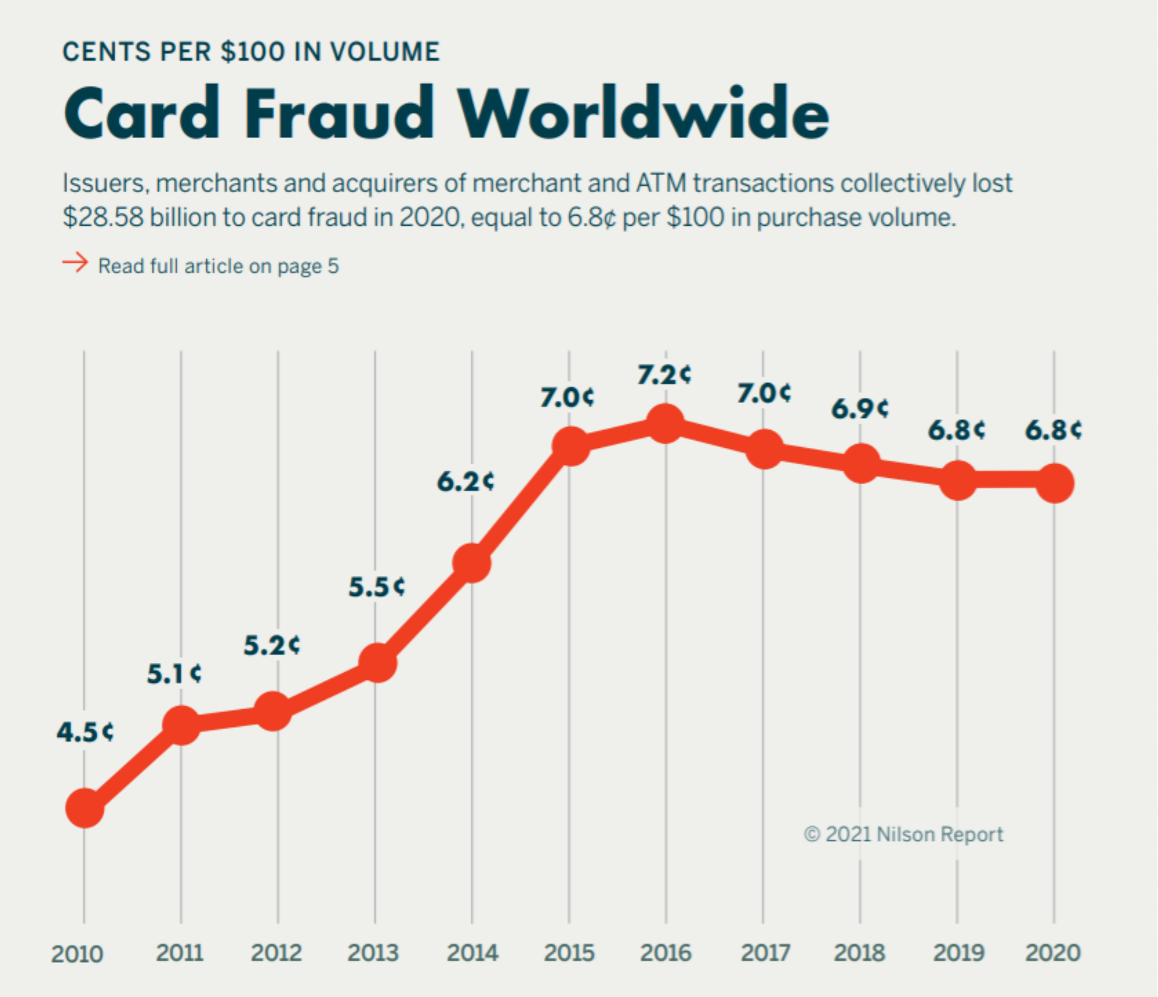
Proposed Models

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.



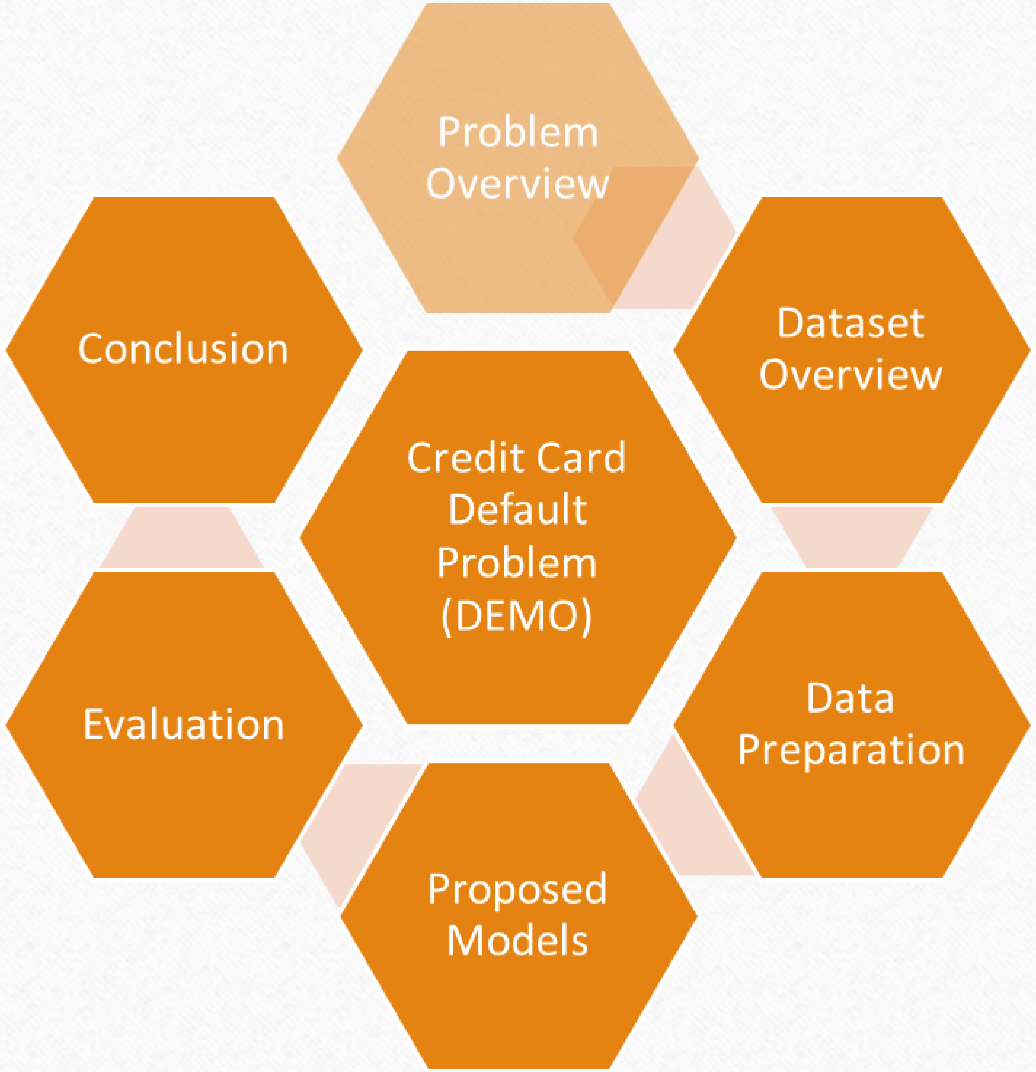
CENTS PER $100 IN VOLUME

Card Fraud Worldwide

Issuers, merchants and acquirers of merchant and ATM transactions collectively lost $28.58 billion to card fraud in 2020, equal to 6.8¢ per $100 in purchase volume.

→ Read full article on page 5

© 2021 Nilson Report


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

Dataset Overview

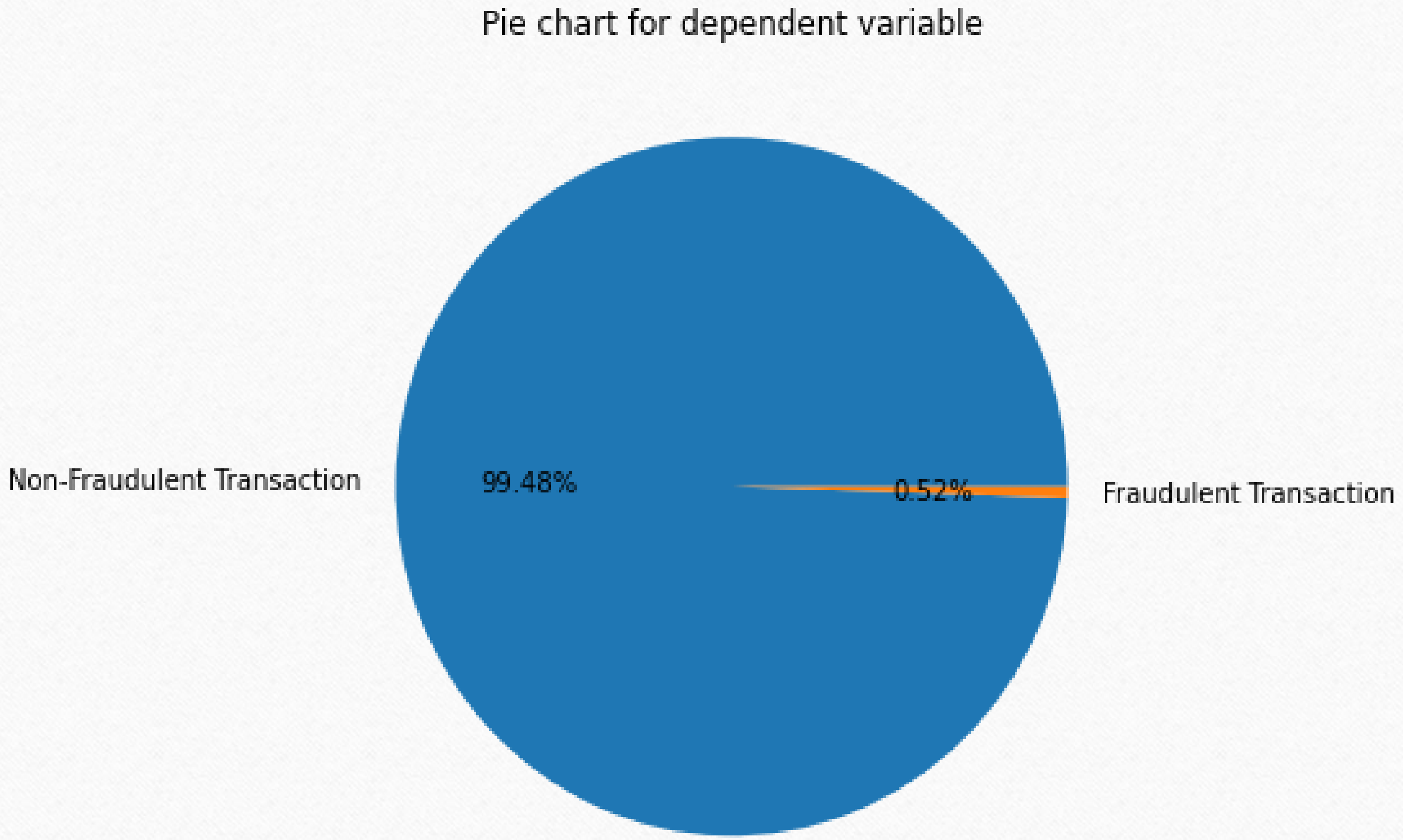
The Credit Card Fraud detection Data:
https://www.kaggle.com/kartik2112/fraud-detection?select=fraudTest.csv
https://www.kaggle.com/kartik2112/fraud-detection?select=fraudTrain.csv
This is a simulated credit card transaction dataset containing legitimate and fraud transactions. It covers credit cards 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

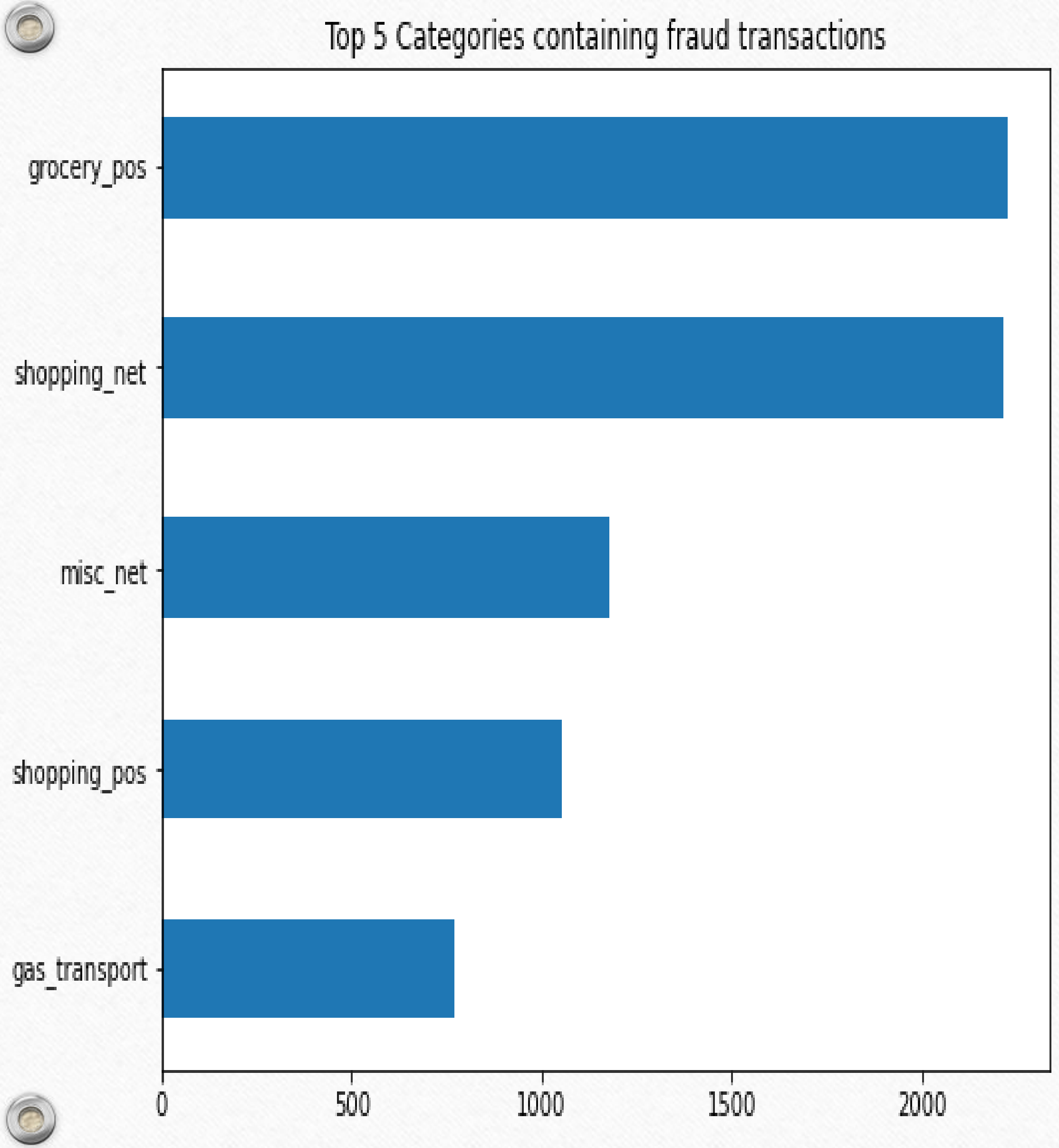
1	trans_date_trans_time	object	Transaction Date/Transaction Time
2	cc_num	int64	Customer's Credit Card Number
3	merchant	object	Merchant by whom the trade occurred
4	category	object	Type of Purchase
5	amt	float64	Amount of Transaction
6	first	object	First Name
7	last	object	Last Name
8	gender	object	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)
```

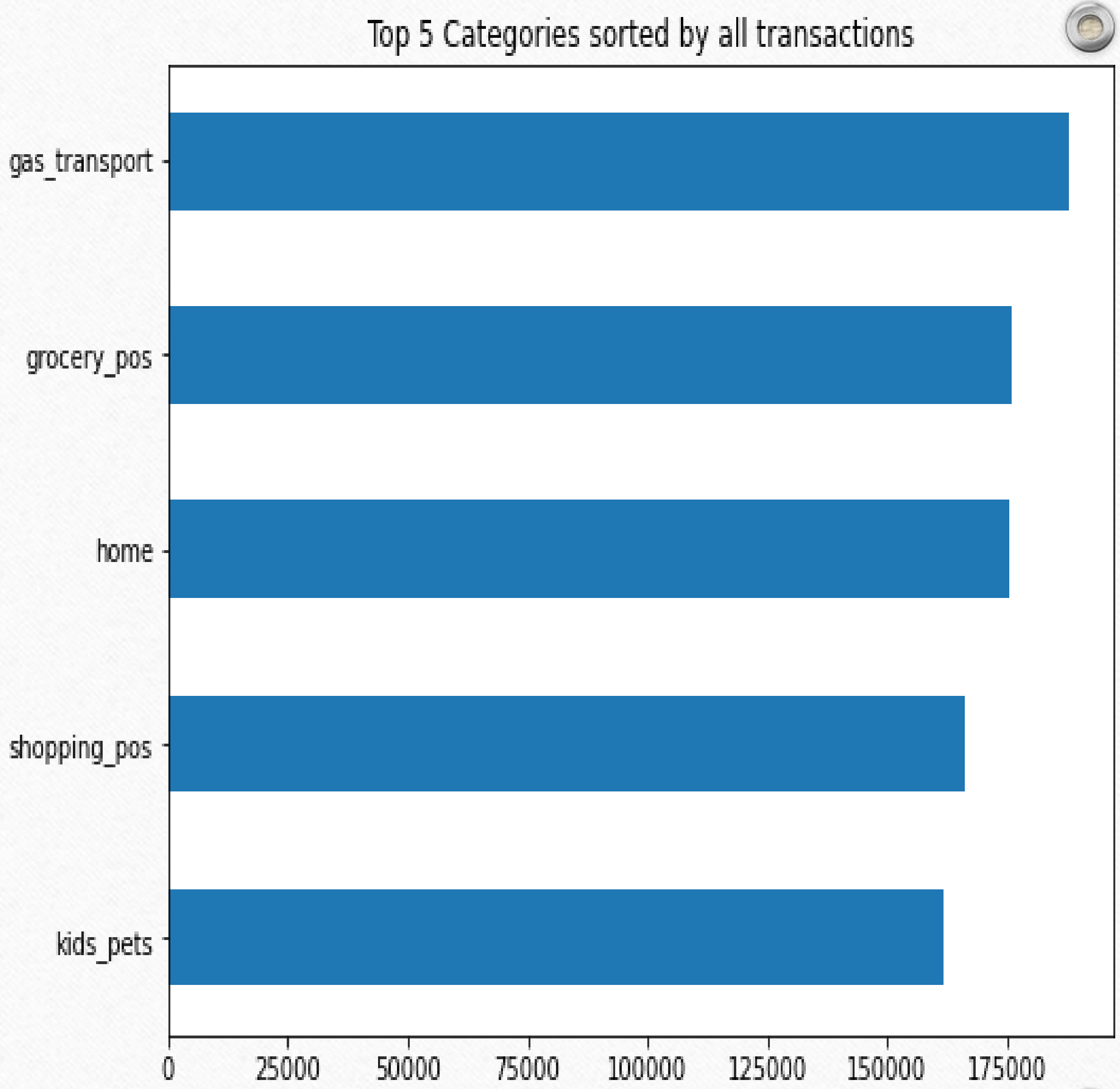
The data is imbalanced



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

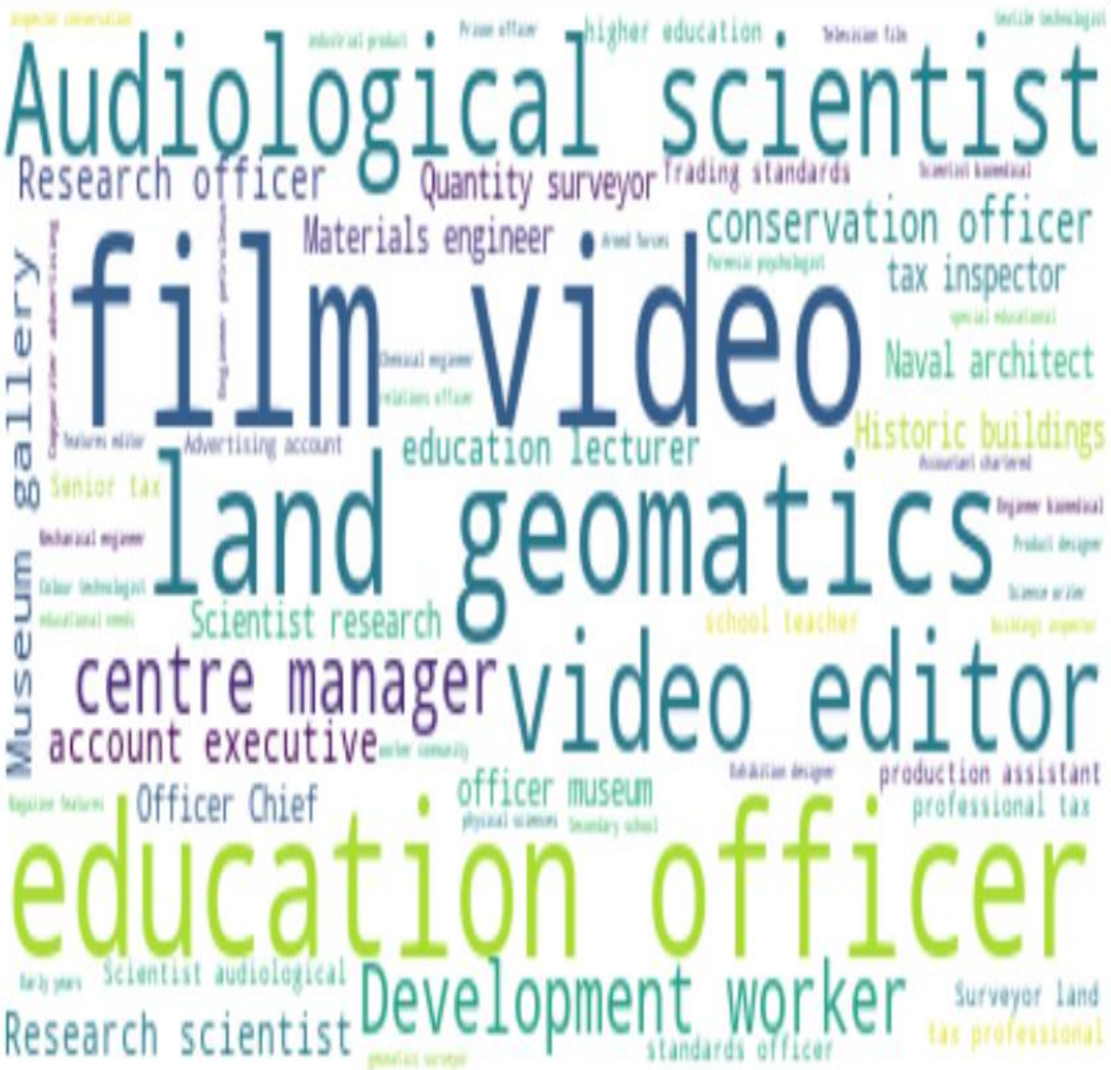


Top 5 Categories containing fraud transactions

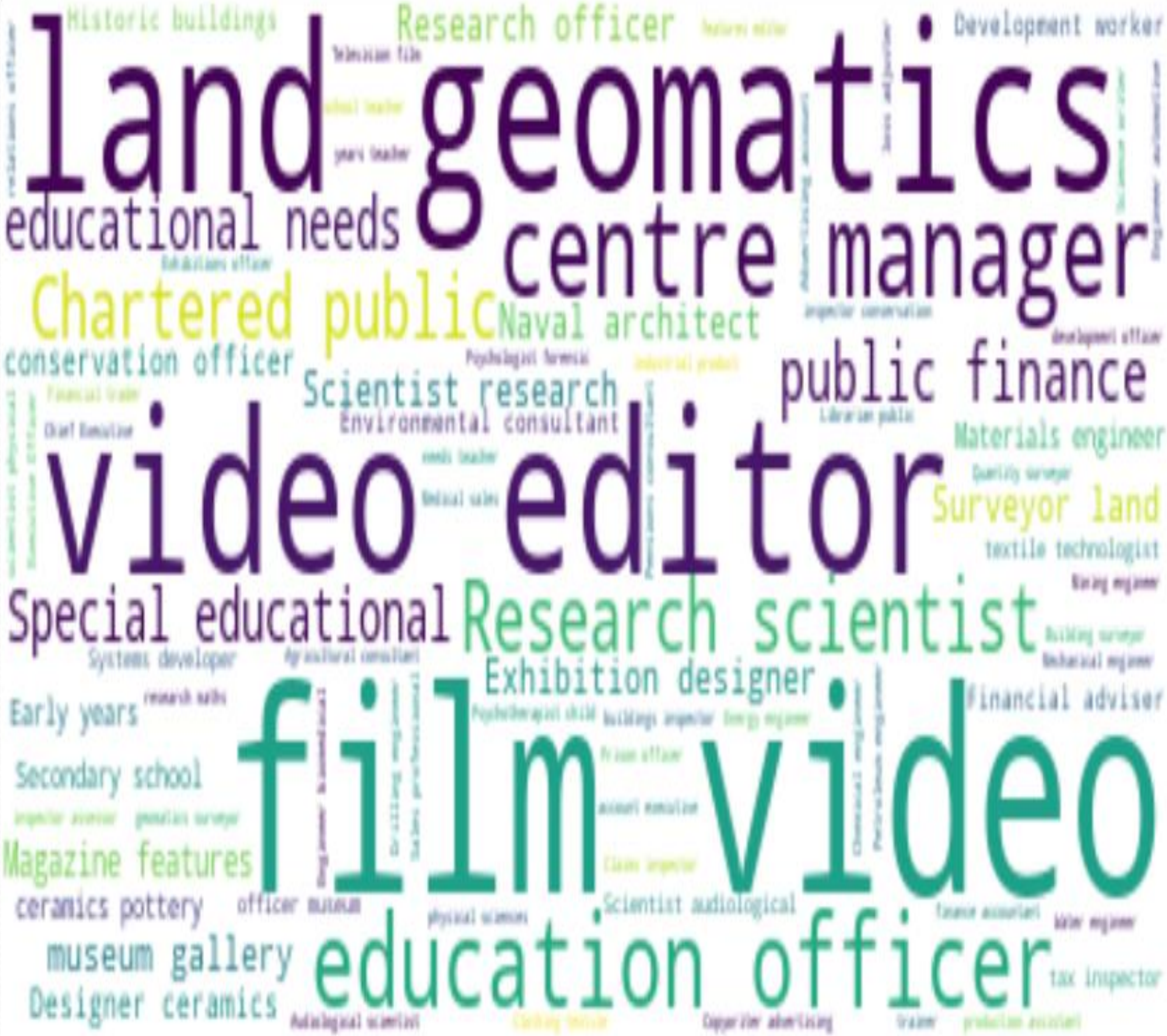


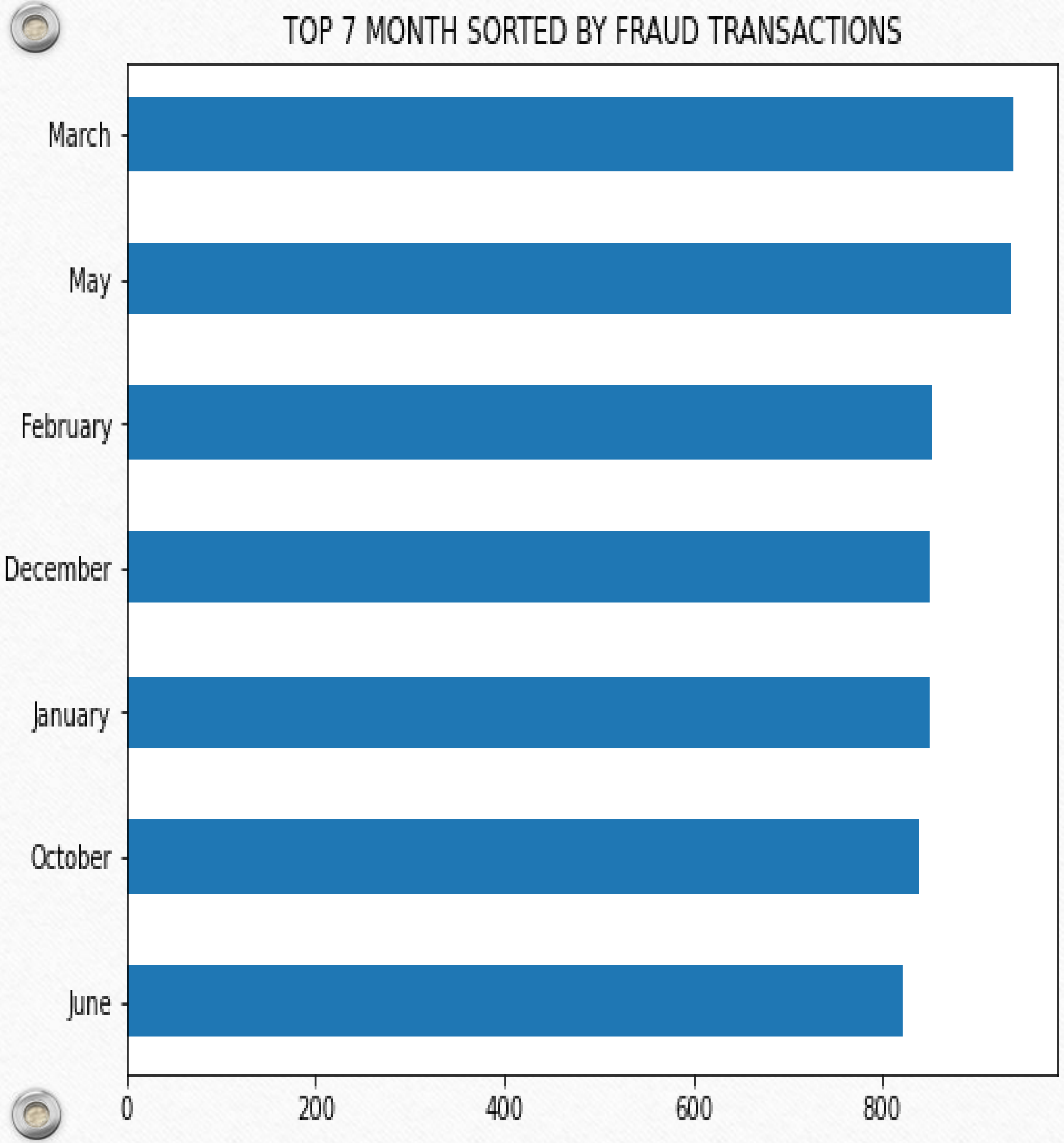
Top 5 Categories sorted by all transactions

MOST FREQUENT JOB TITLES COMPARED BY FRAUD TRANSACTIONS

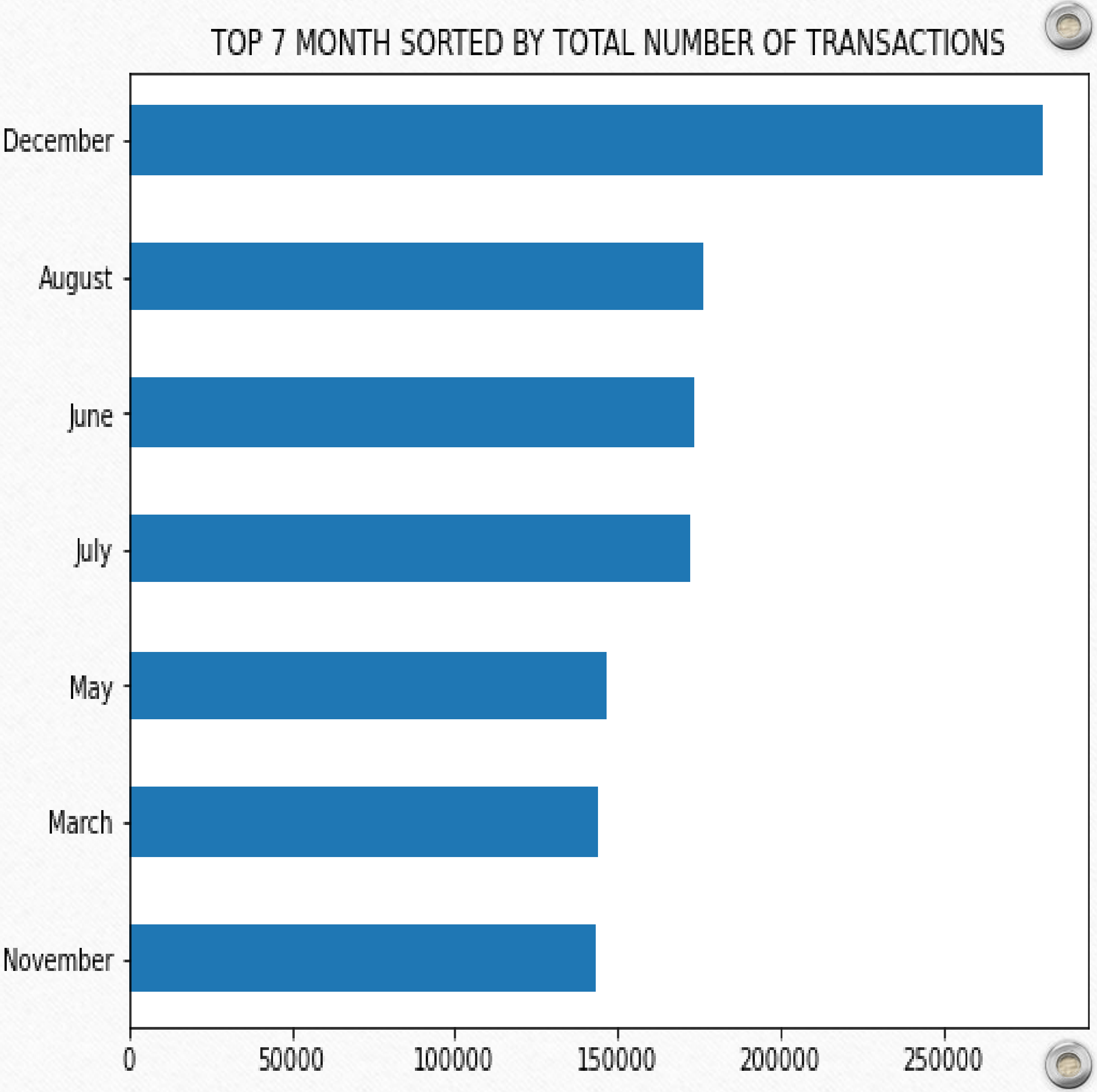


MOST FREQUENT JOB TITLES COMPARED BY TRANSACTIONS

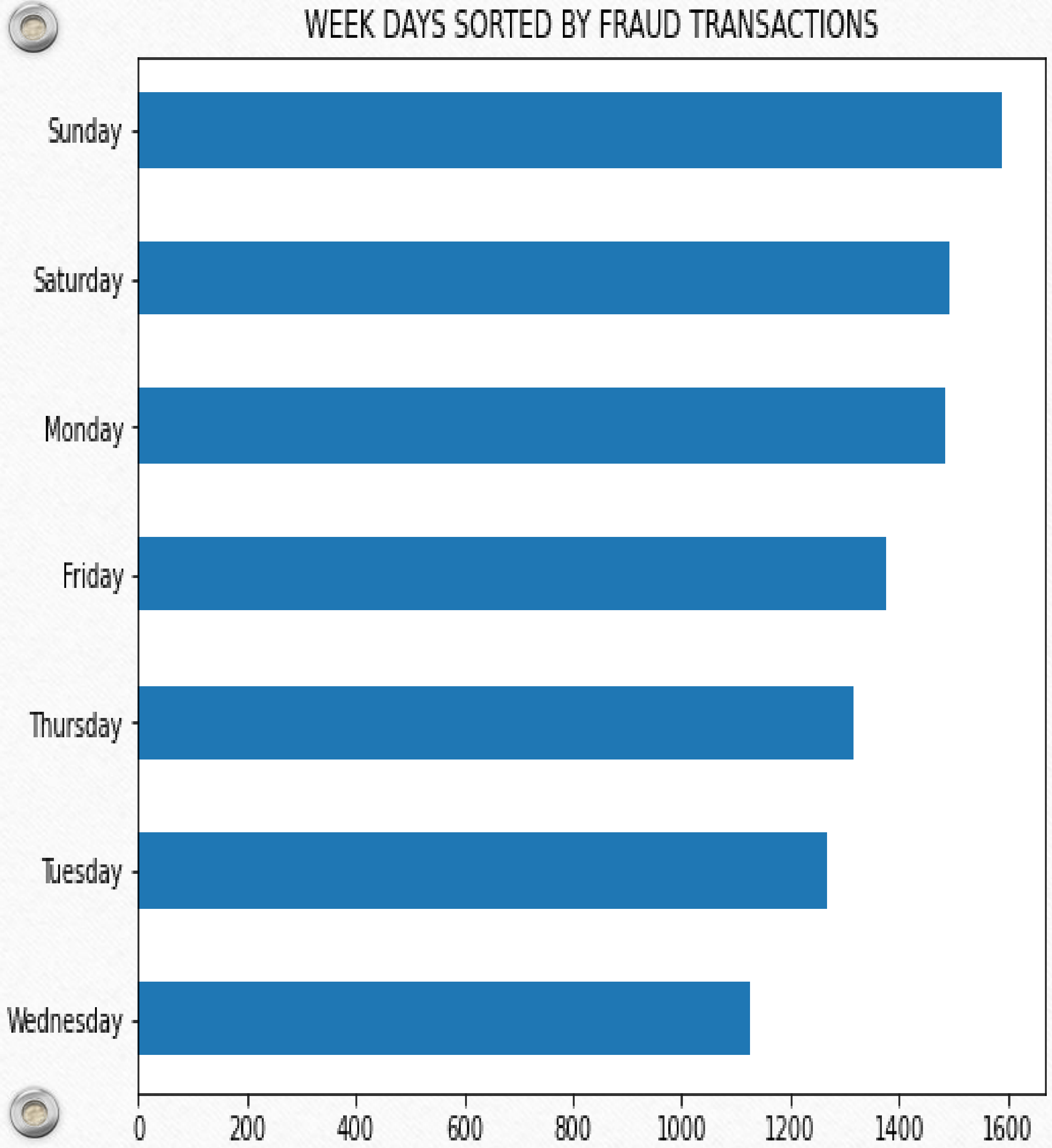




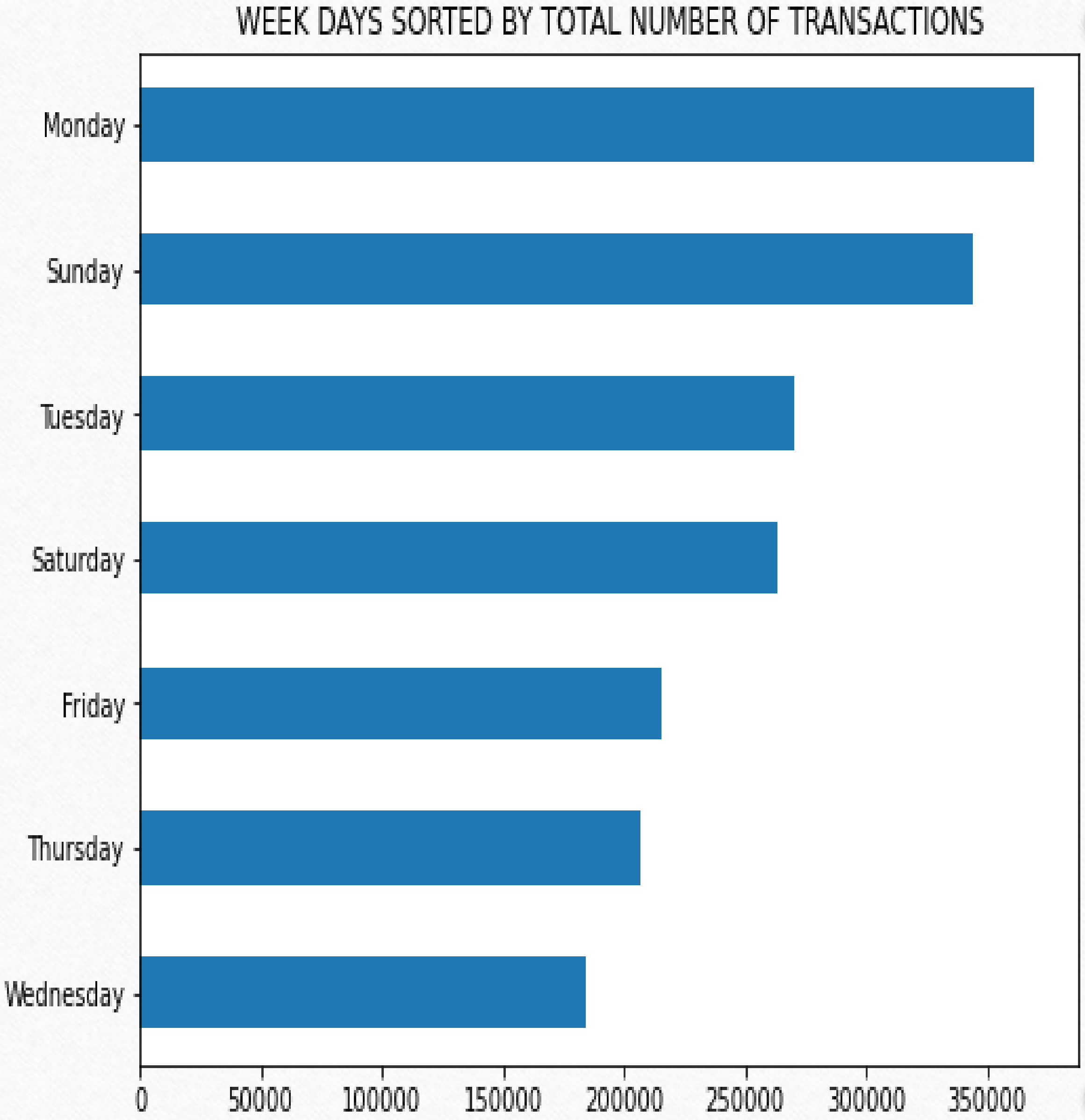
TOP 7 MONTH SORTED BY FRAUD TRANSACTIONS



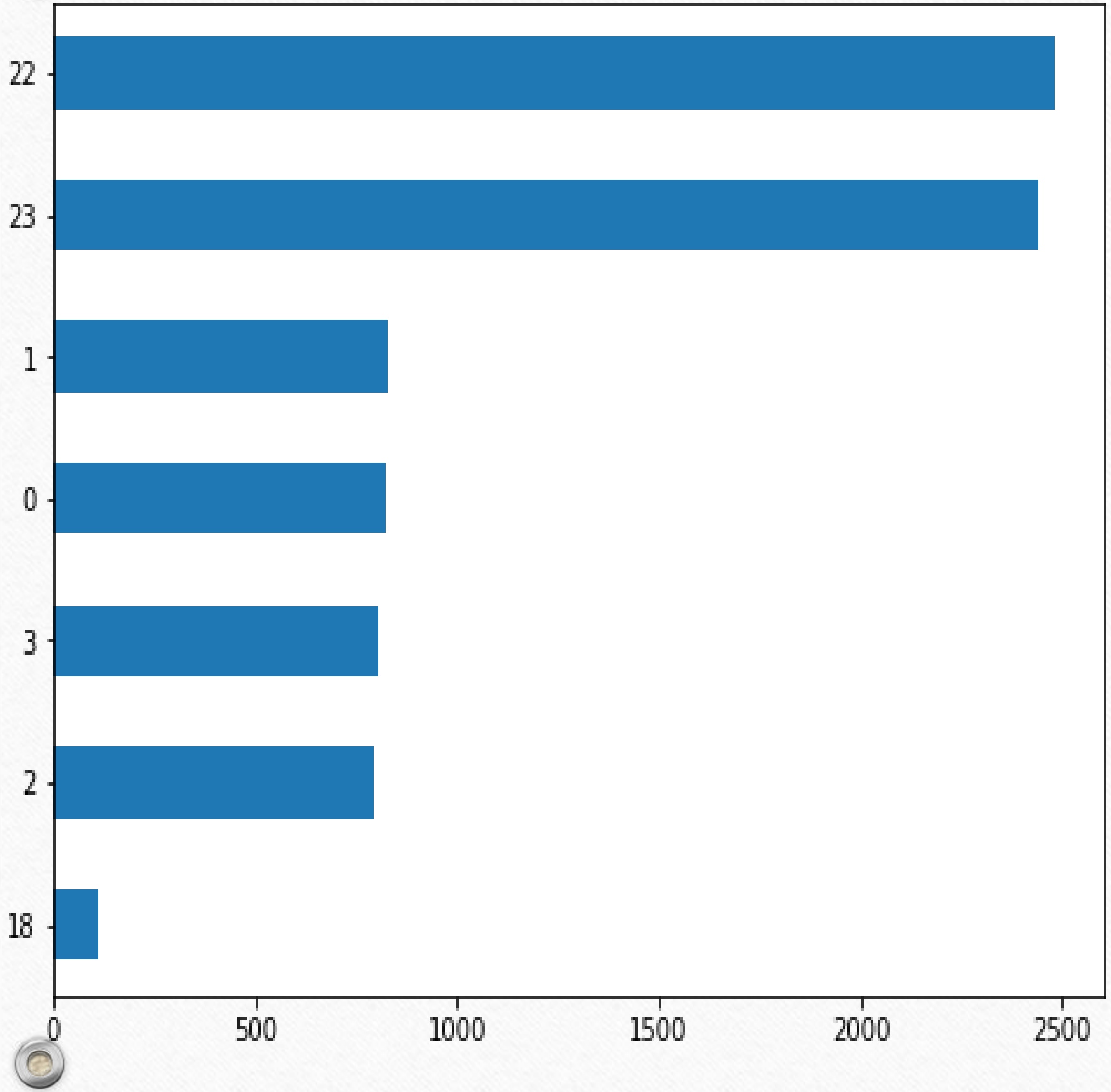
TOP 7 MONTH SORTED BY TOTAL NUMBER OF TRANSACTIONS



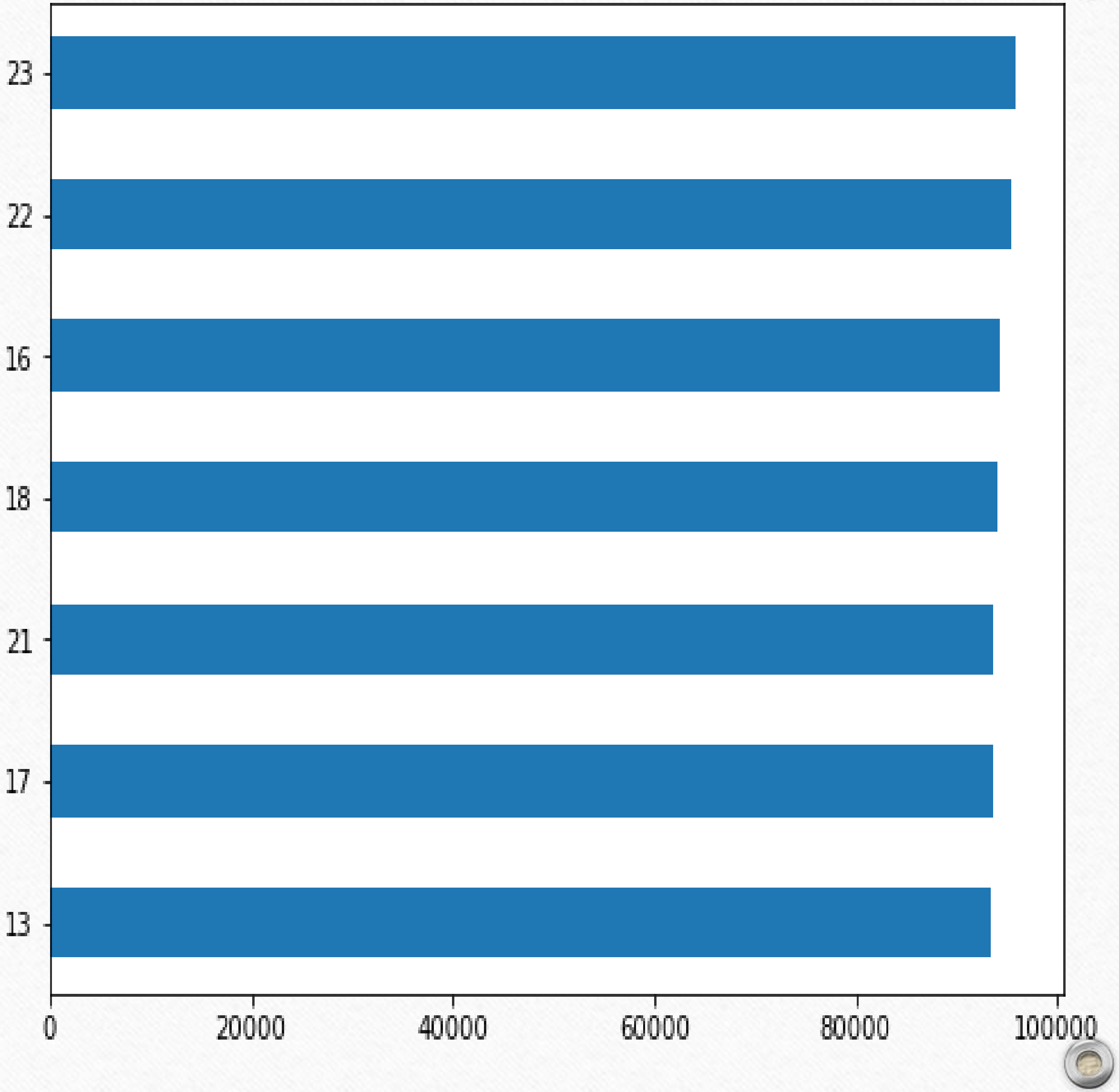
WEEK DAYS SORTED BY FRAUD TRANSACTIONS



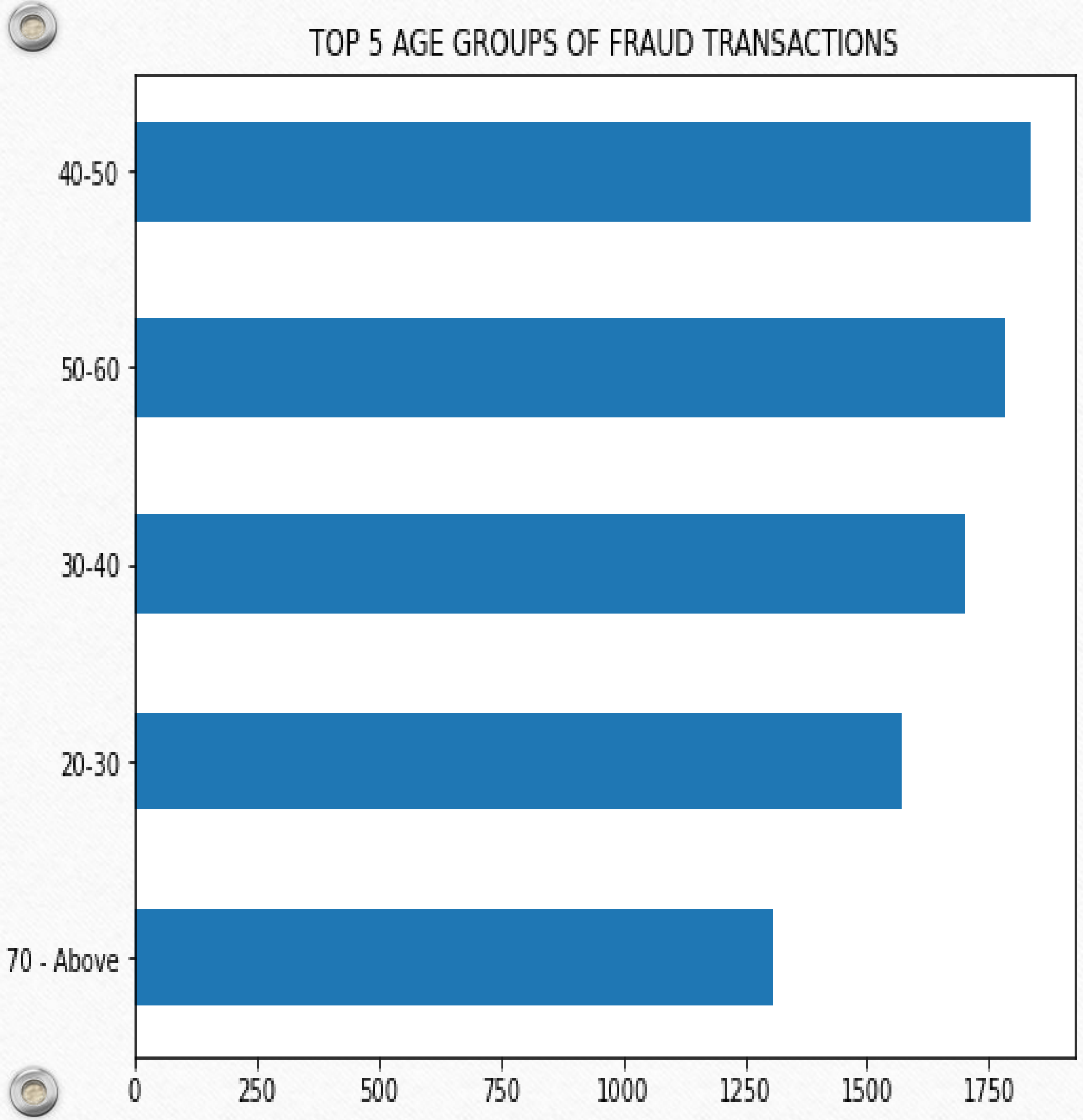
WEEK DAYS SORTED BY TOTAL NUMBER OF TRANSACTIONS



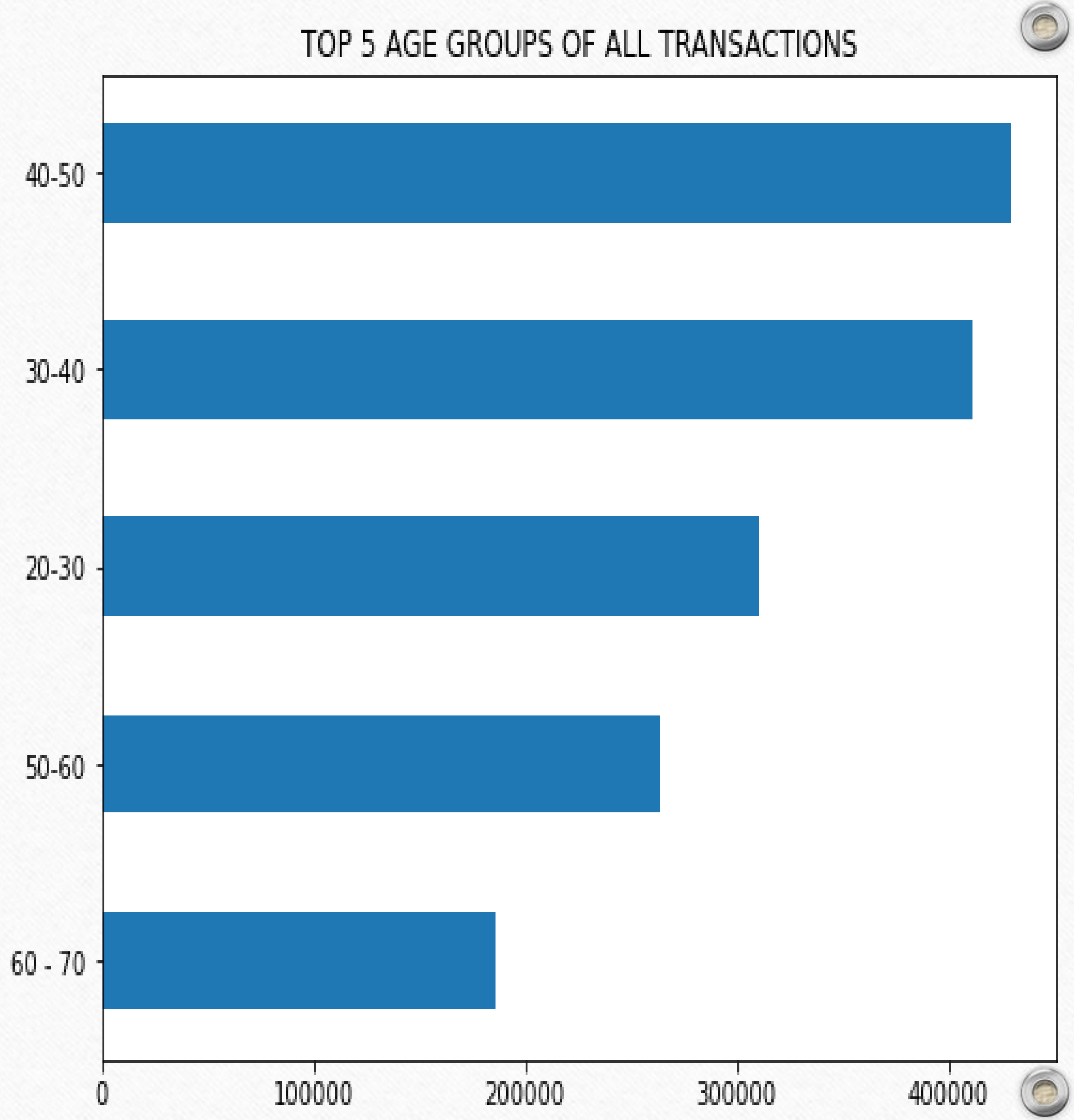
TOP 7 HOURS OF FRAUD TRANSACTIONS



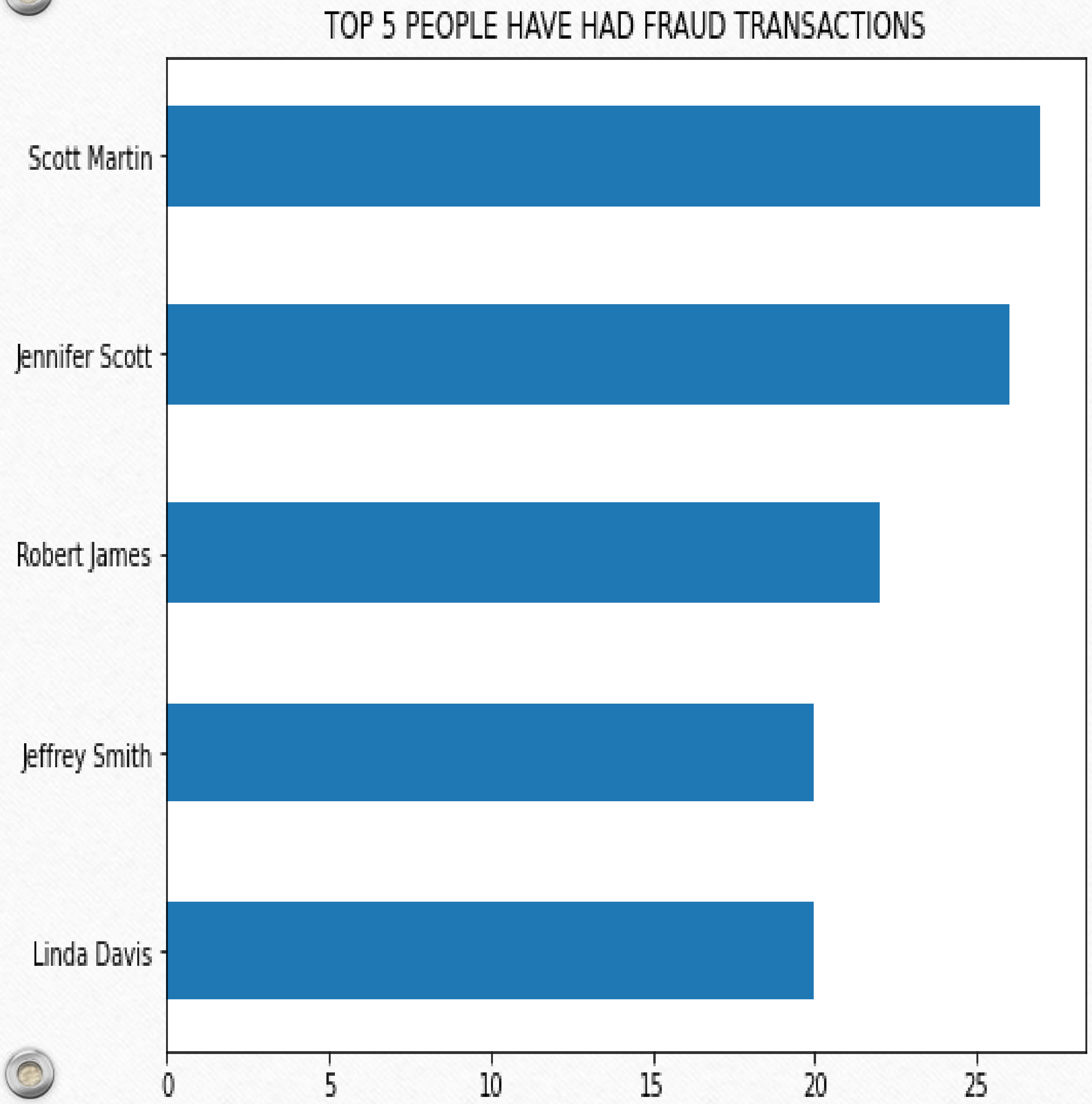
TOP 7 HOURS OF ALL TRANSACTIONS



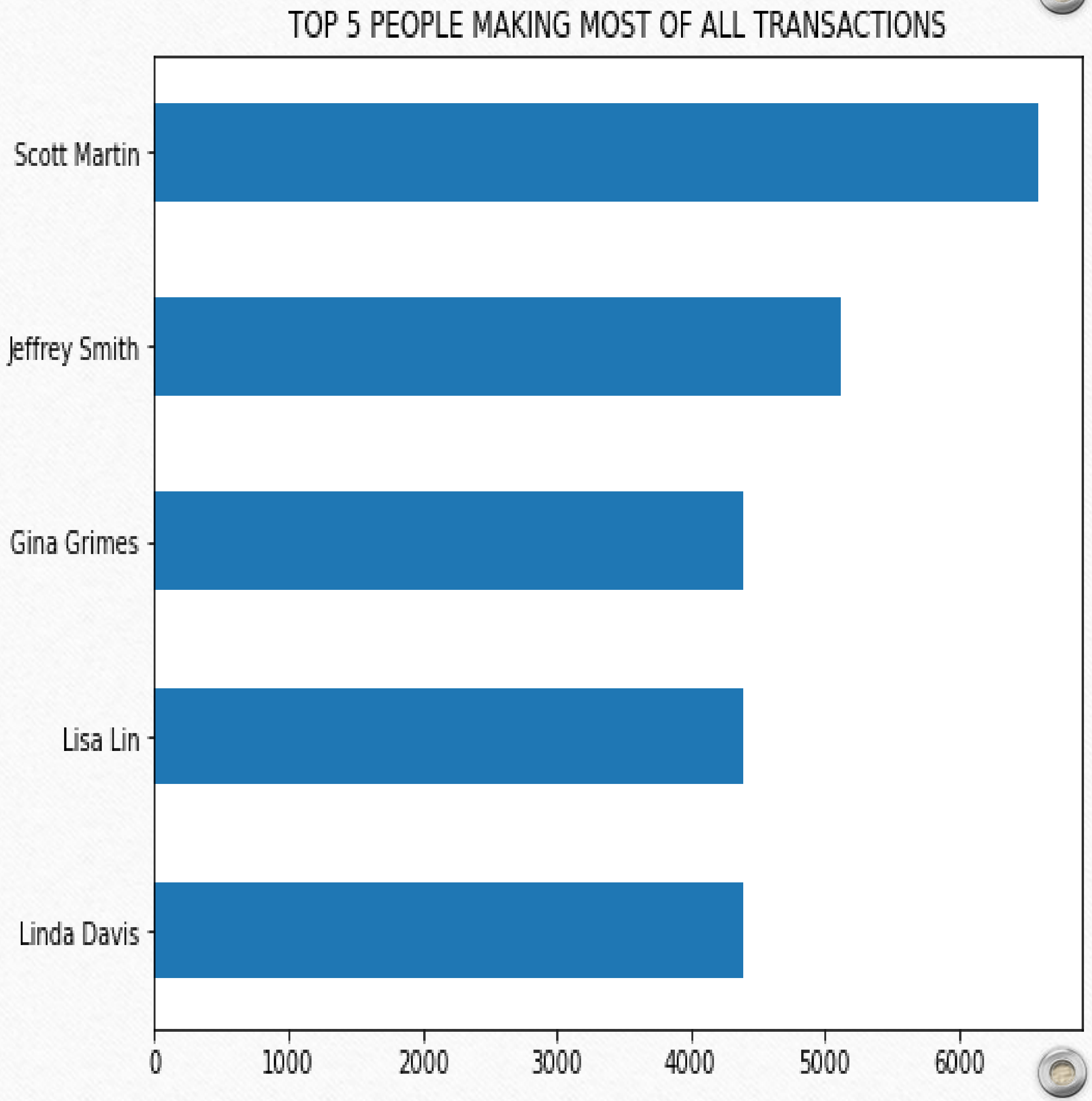
TOP 5 AGE GROUPS OF FRAUD TRANSACTIONS



TOP 5 AGE GROUPS OF ALL TRANSACTIONS

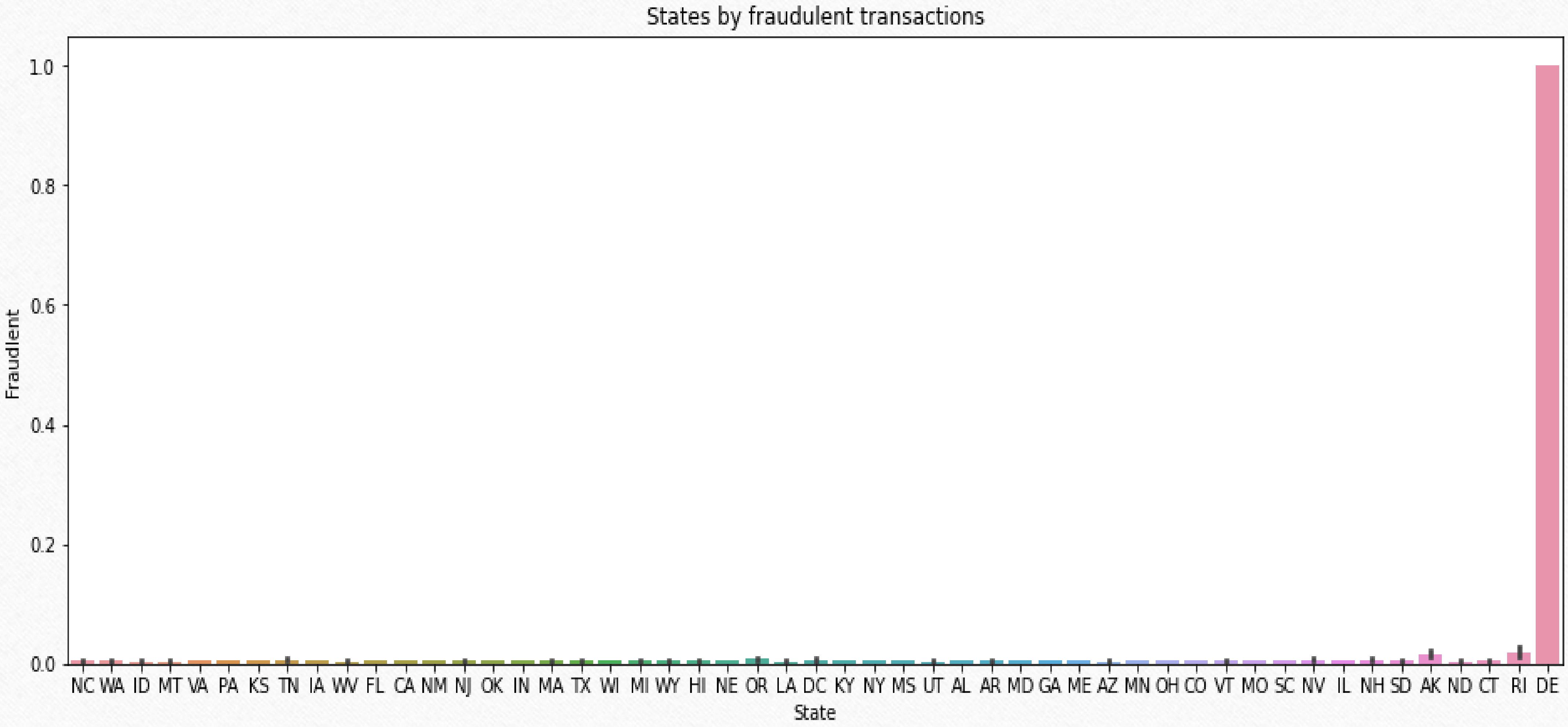


TOP 5 PEOPLE HAVE HAD FRAUD TRANSACTIONS

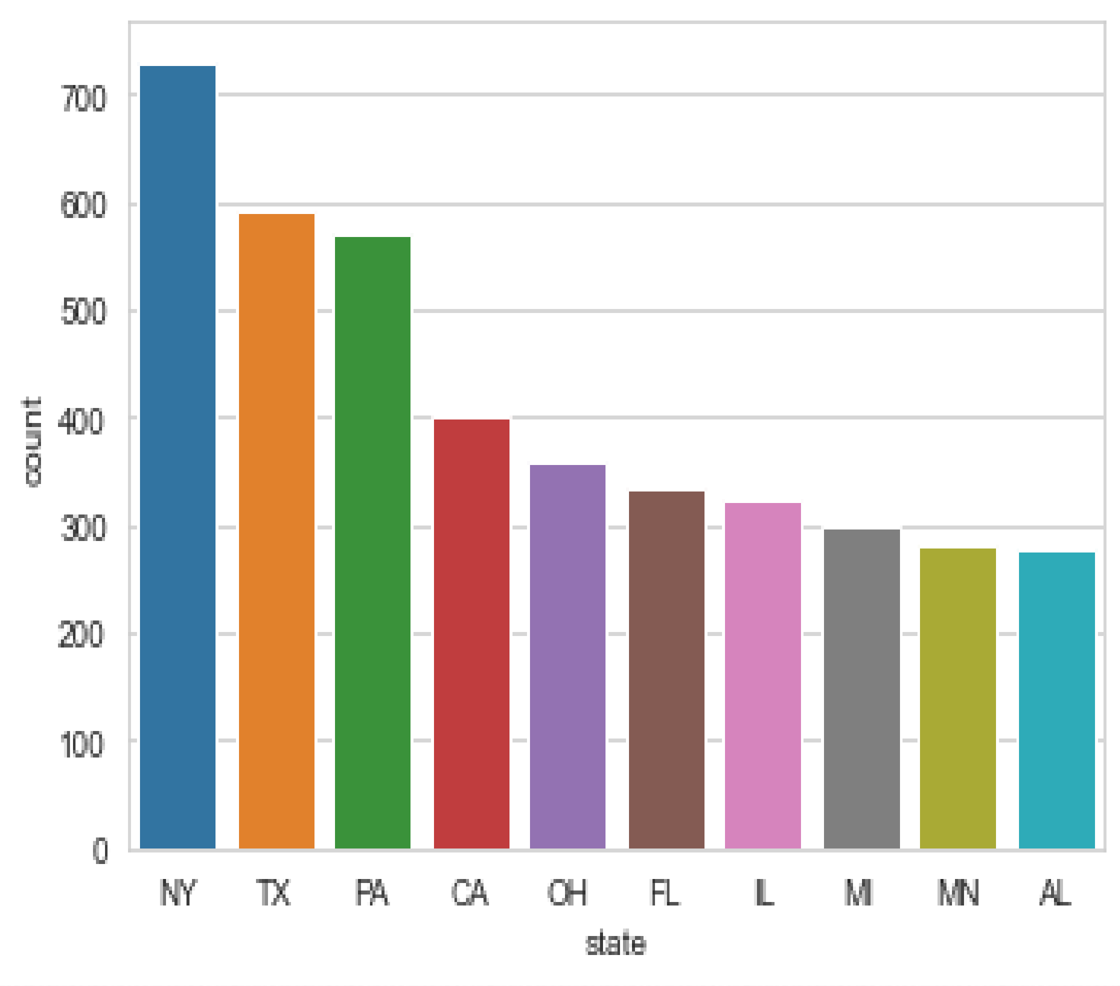


TOP 5 PEOPLE MAKING MOST OF ALL TRANSACTIONS

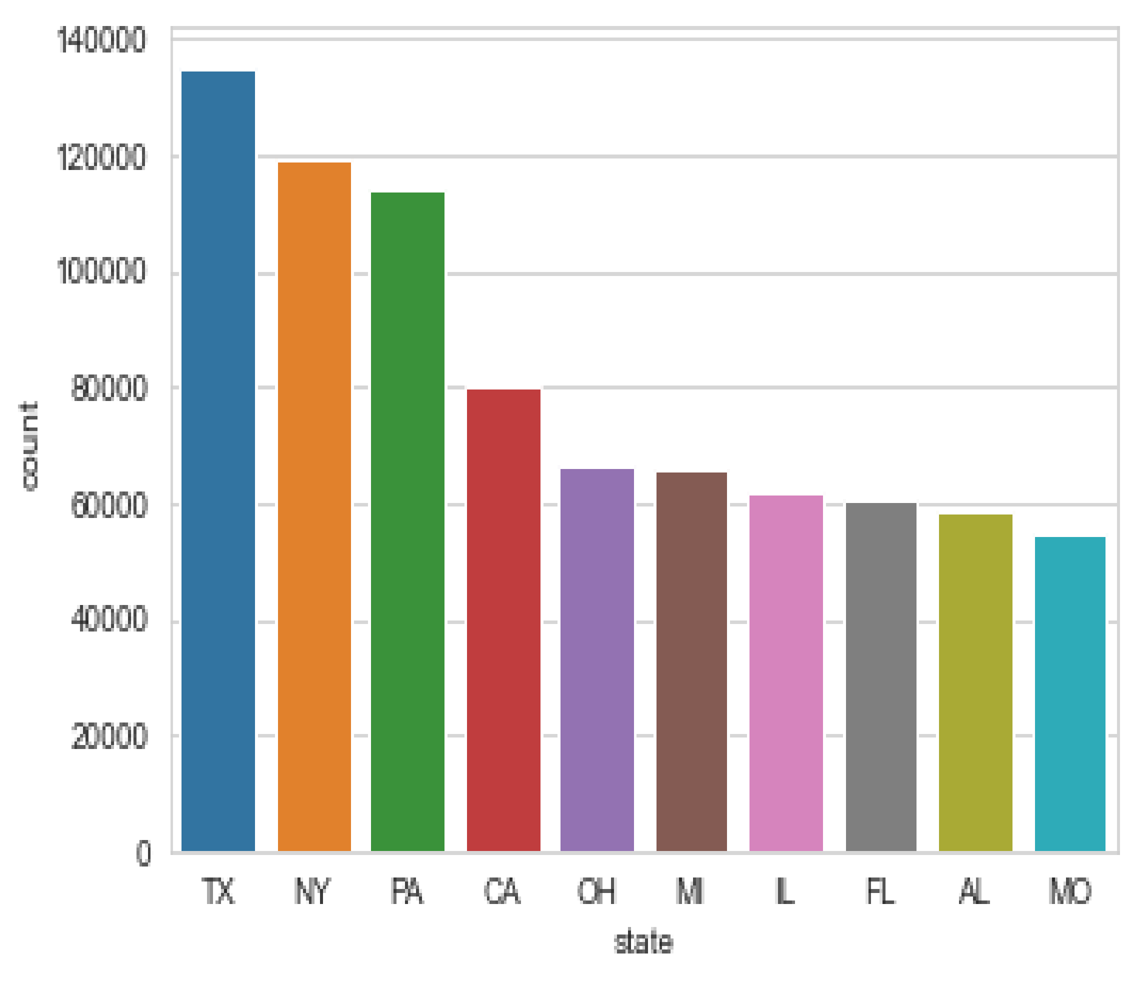
State Delaware is on top position by Fraud transactions.



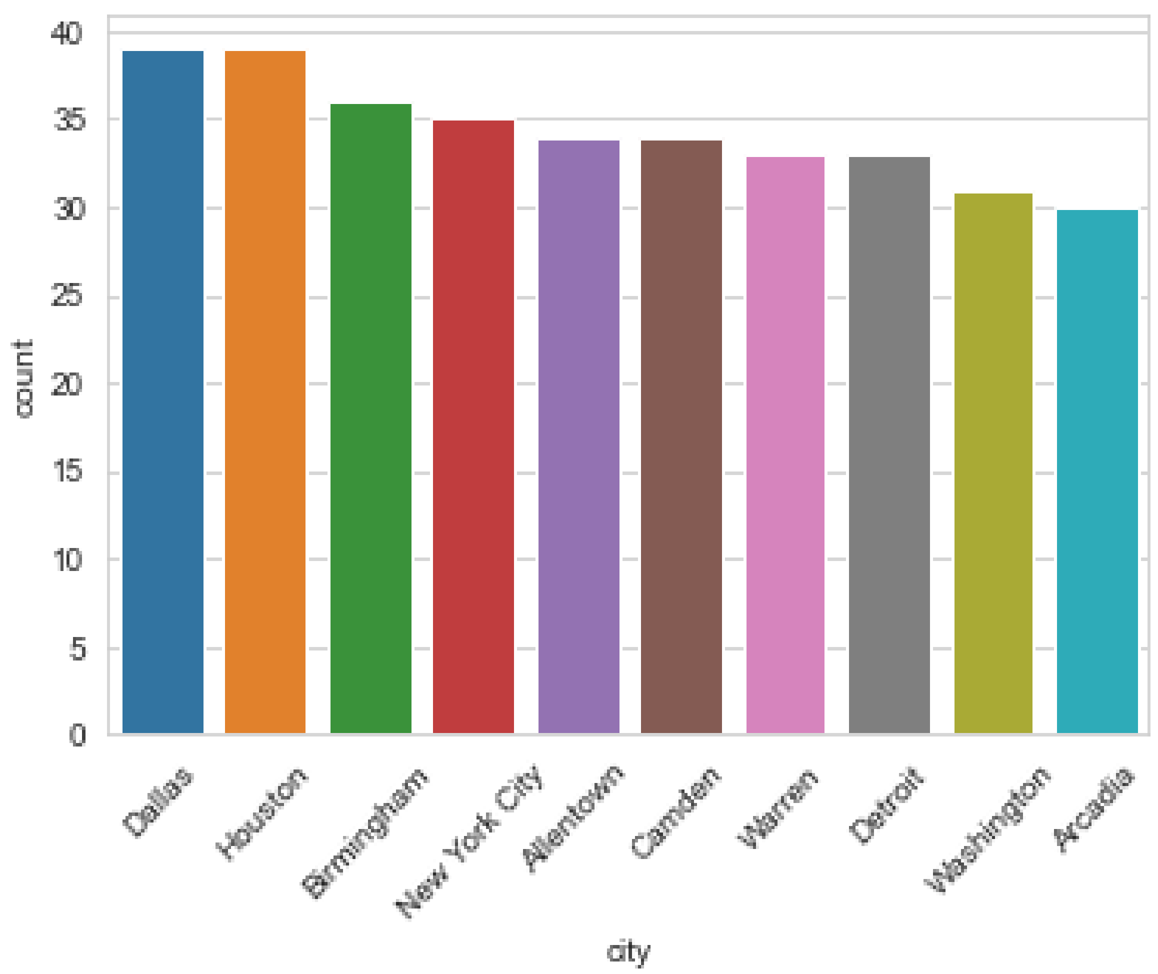
Top 10 States of Fraudulent Transactions



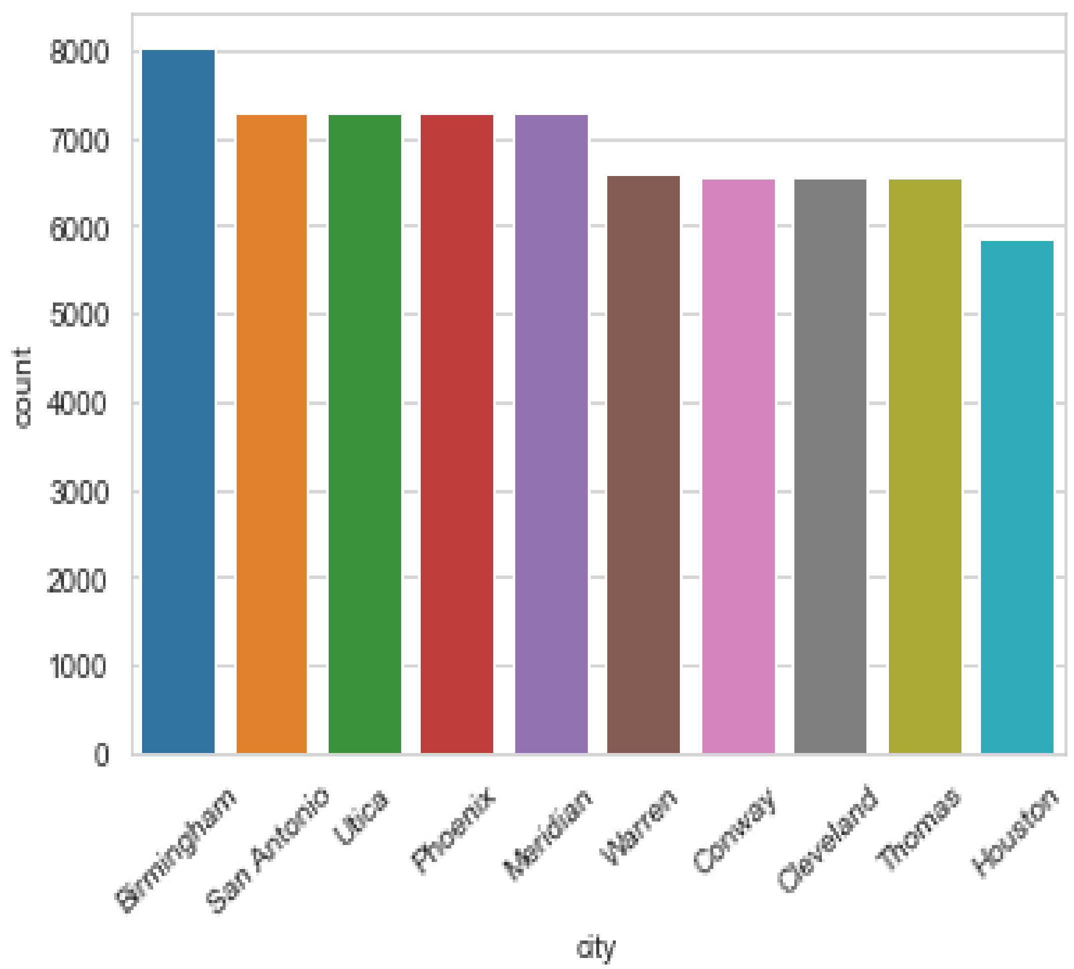
Top 10 States of All Transactions



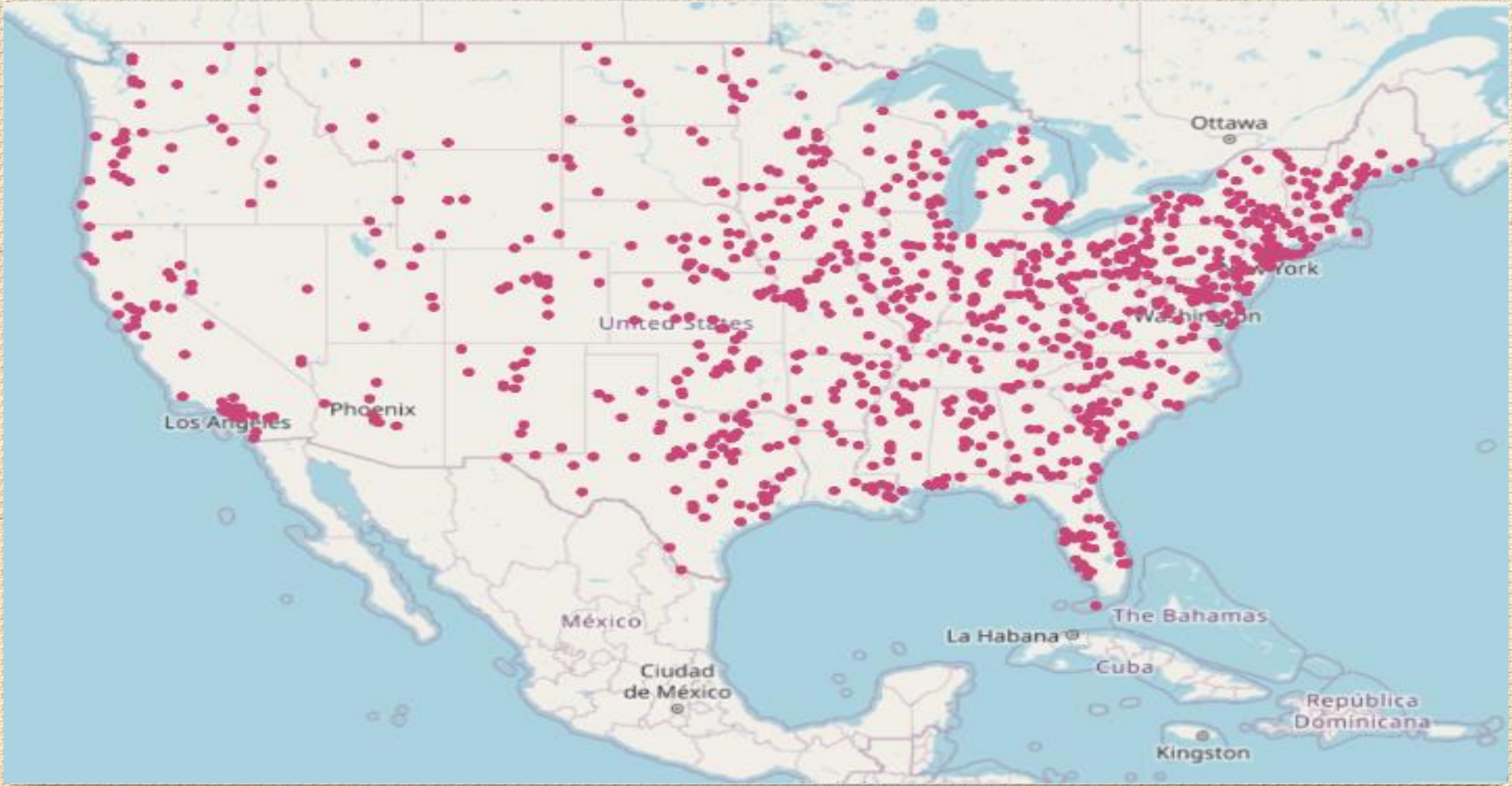
Top 10 Cities of Fraudulent Transactions

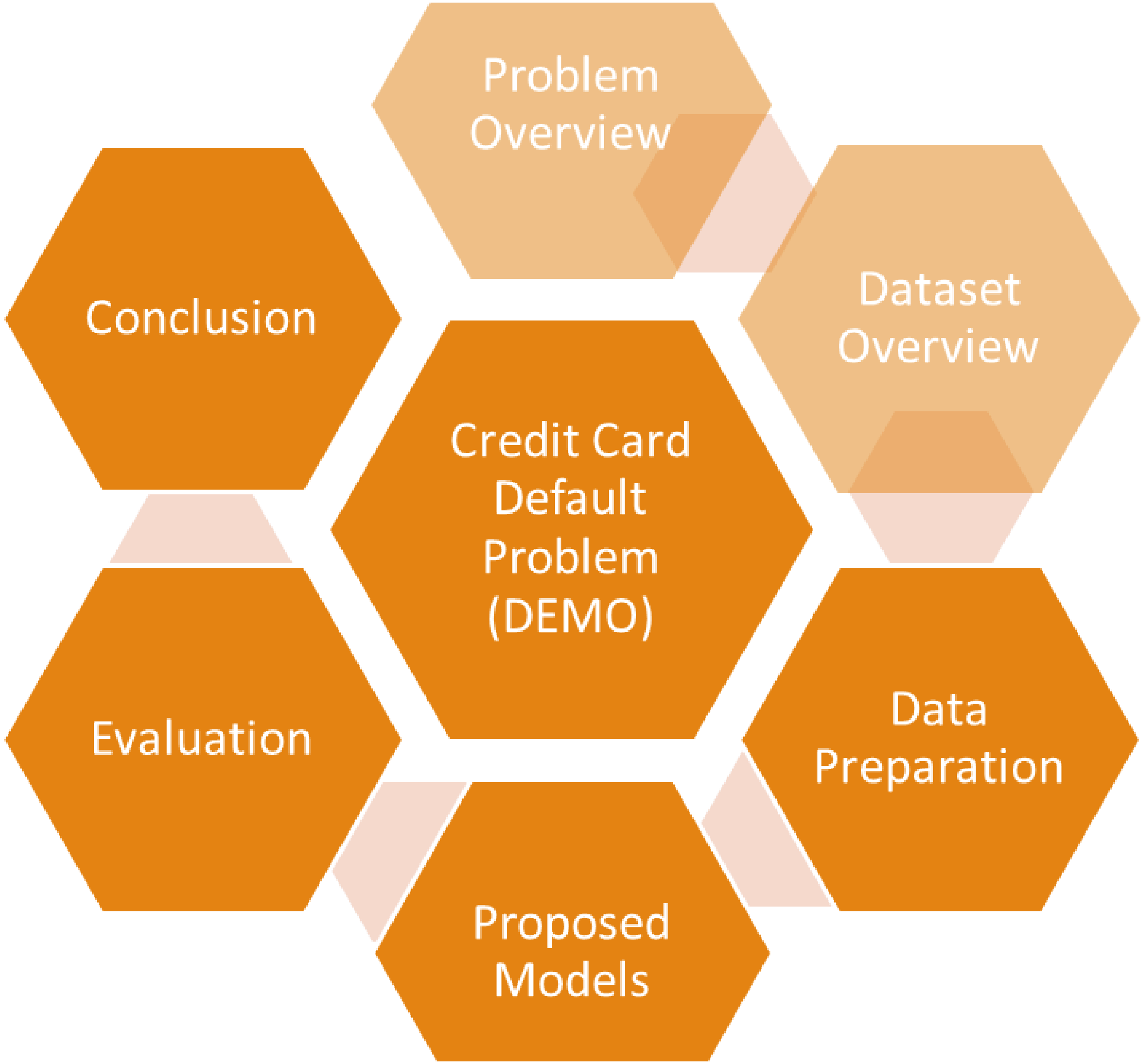


Top 10 Cities of All Transactions



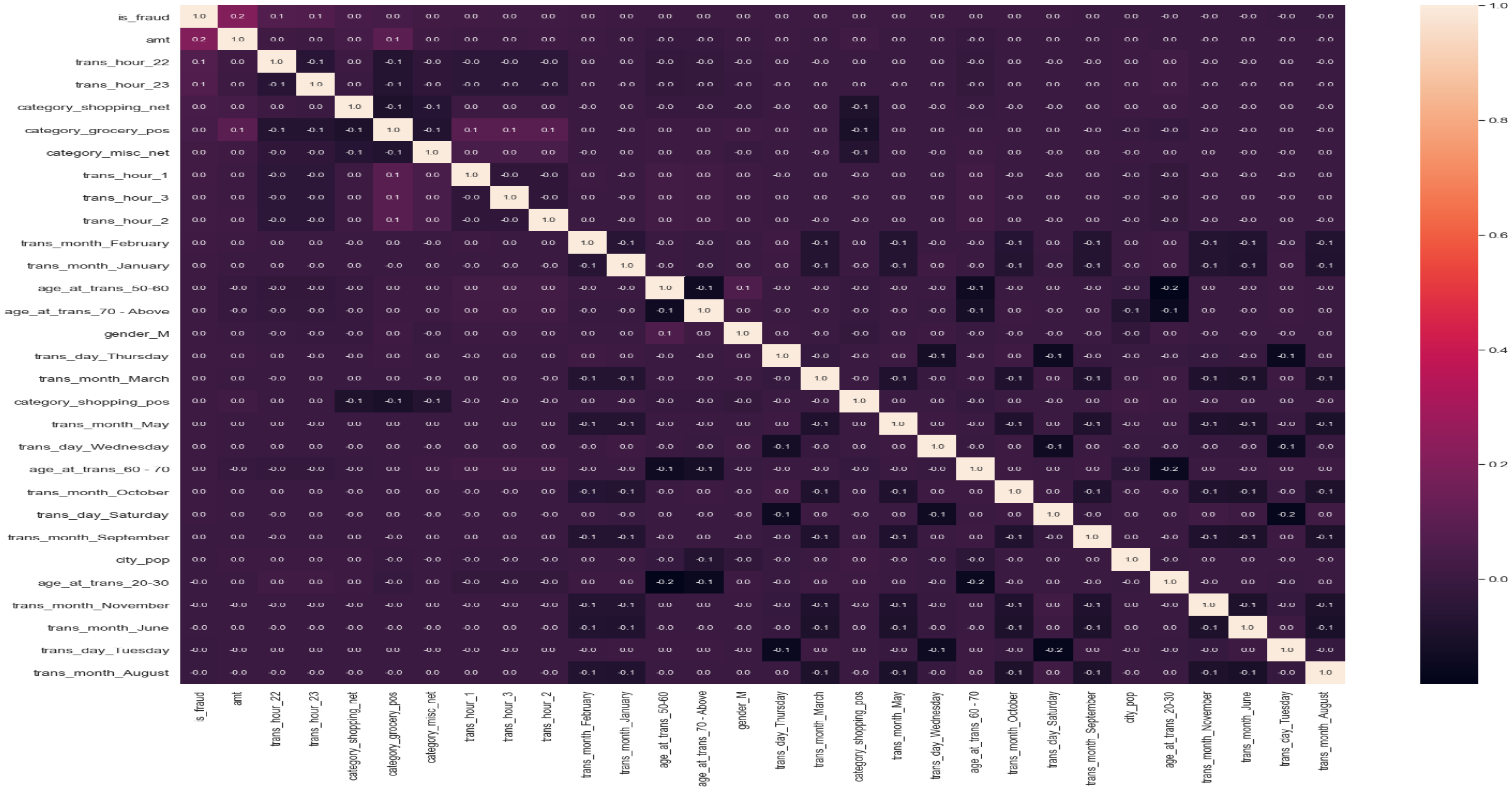
Fraud Transactions by location



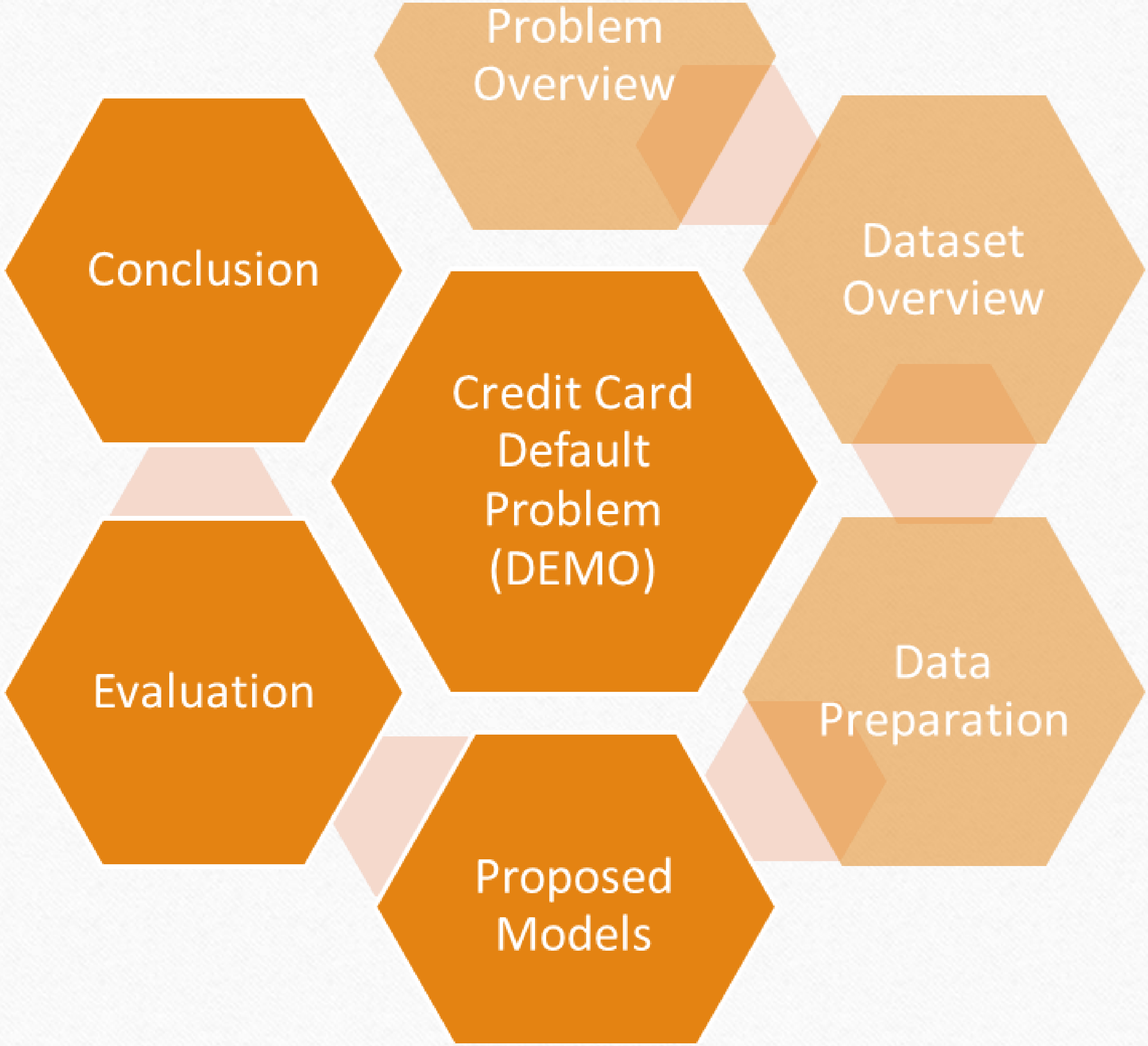

Problem Overview
Dataset Overview
Conclusion
Credit Card Default Problem (DEMO)
Data Preparation
Evaluation
Proposed Models

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
is_fraud	1.0	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.0
amt	0.2	1.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	-0.0	-0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.0	0.0	0.0	0.0	0.0	0.0	-0.0	0.0	-0.0	-0.0
trans_hour_22	0.1	0.0	1.0	-0.1	0.0	-0.1	-0.0	-0.0	-0.0	-0.0	0.0	0.0	-0.0	-0.0	-0.0	0.0	-0.0	0.0	0.0	0.0	-0.0	0.0	-0.0	-0.0	0.0	0.0	-0.0	-0.0	0.0	-0.0
trans_hour_23	0.1	0.0	-0.1	1.0	0.0	-0.1	-0.0	-0.0	-0.0	-0.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0	0.0	-0.0	0.0	-0.0	0.0	0.0	0.0	0.0	0.0	-0.0	-0.0	0.0	-0.0
category_shopping_net	0.0	0.0	0.0	0.0	1.0	-0.1	-0.1	0.0	0.0	0.0	-0.0	0.0	-0.0	-0.0	-0.0	-0.0	0.0	-0.1	0.0	-0.0	-0.0	-0.0	-0.0	0.0	0.0	0.0	-0.0	-0.0	0.0	-0.0
category_grocery_pos	0.0	0.1	-0.1	-0.1	-0.1	1.0	-0.1	0.1	0.1	0.1	0.0	-0.0	0.0	0.0	0.0	0.0	0.0	-0.1	0.0	0.0	0.0	0.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0	-0.0
category_misc_net	0.0	0.0	-0.0	-0.0	-0.1	-0.1	1.0	0.0	0.0	0.0	-0.0	0.0	0.0	0.0	-0.0	0.0	-0.0	-0.1	0.0	-0.0	0.0	0.0	-0.0	0.0	-0.0	0.0	0.0	-0.0	-0.0	0.0
trans_hour_1	0.0	0.0	-0.0	-0.0	0.0	0.1	0.0	1.0	-0.0	-0.0	0.0	-0.0	0.0	0.0	0.0	-0.0	0.0	-0.0	0.0	0.0	0.0	-0.0	-0.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0
trans_hour_3	0.0	0.0	-0.0	-0.0	0.0	0.1	0.0	-0.0	1.0	-0.0	0.0	-0.0	0.0	0.0	0.0	0.0	0.0	-0.0	0.0	0.0	0.0	-0.0	0.0	0.0	-0.0	-0.0	-0.0	0.0	-0.0	0.0
trans_hour_2	0.0	0.0	-0.0	-0.0	0.0	0.1	0.0	-0.0	-0.0	1.0	0.0	-0.0	0.0	0.0	0.0	0.0	-0.0	-0.0	0.0	0.0	0.0	0.0	-0.0	0.0	-0.0	-0.0	-0.0	-0.0	0.0	-0.0
trans_month_February	0.0	0.0	0.0	0.0	-0.0	0.0	-0.0	0.0	0.0	0.0	1.0	-0.1	-0.0	-0.0	0.0	0.0	-0.1	0.0	-0.1	-0.0	-0.0	-0.1	0.0	-0.1	0.0	0.0	-0.1	-0.1	-0.0	-0.1
trans_month_January	0.0	0.0	0.0	-0.0	0.0	-0.0	0.0	-0.0	-0.0	-0.0	-0.1	1.0	-0.0	-0.0	0.0	0.0	-0.1	-0.0	-0.1	0.0	-0.0	-0.1	-0.0	-0.1	-0.0	0.0	-0.1	-0.1	0.0	-0.1
age_at_trans_50-60	0.0	-0.0	-0.0	-0.0	-0.0	0.0	0.0	0.0	0.0	0.0	-0.0	-0.0	1.0	-0.1	0.1	-0.0	-0.0	-0.0	-0.0	-0.0	-0.1	-0.0	0.0	-0.0	-0.0	-0.2	0.0	-0.0	0.0	-0.0
age_at_trans_70 - Above	0.0	-0.0	-0.0	-0.0	-0.0	0.0	0.0	0.0	0.0	0.0	-0.0	-0.0	-0.1	1.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.1	0.0	-0.0	0.0	-0.1	-0.1	0.0	-0.0	0.0	0.0
gender_M	0.0	0.0	-0.0	-0.0	-0.0	0.0	-0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	1.0	0.0	-0.0	-0.0	0.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0	-0.0	-0.0	-0.0	0.0
trans_day_Thursday	0.0	0.0	0.0	-0.0	-0.0	0.0	0.0	-0.0	0.0	0.0	0.0	0.0	-0.0	-0.0	0.0	1.0	-0.0	-0.0	0.0	-0.1	-0.0	0.0	-0.1	-0.0	-0.0	-0.0	-0.0	-0.0	-0.1	0.0
trans_month_March	0.0	0.0	-0.0	0.0	0.0	0.0	-0.0	0.0	0.0	-0.0	-0.1	-0.1	-0.0	-0.0	-0.0	-0.0	1.0	-0.0	-0.1	-0.0	-0.0	-0.1	0.0	-0.1	0.0	0.0	-0.1	-0.1	0.0	-0.1
category_shopping_pos	0.0	0.0	0.0	0.0	-0.1	-0.1	-0.1	-0.0	-0.0	-0.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	1.0	0.0	0.0	-0.0	0.0	-0.0	0.0	0.0	0.0	-0.0	0.0	0.0	-0.0
trans_month_May	0.0	0.0	0.0	-0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	-0.1	-0.0	-0.0	0.0	0.0	-0.1	0.0	1.0	0.0	-0.0	-0.1	0.0	-0.1	-0.0	0.0	-0.1	-0.1	-0.0	-0.1
trans_day_Wednesday	0.0	0.0	0.0	0.0	-0.0	0.0	-0.0	0.0	0.0	0.0	-0.0	0.0	-0.0	-0.0	0.0	-0.1	-0.0	0.0	0.0	1.0	-0.0	0.0	-0.1	0.0	-0.0	-0.0	-0.0	-0.0	-0.1	-0.0
age_at_trans_60 - 70	0.0	-0.0	-0.0	-0.0	-0.0	0.0	0.0	0.0	0.0	0.0	-0.0	-0.0	-0.1	-0.1	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	1.0	0.0	0.0	0.0	-0.0	-0.2	0.0	-0.0	0.0	0.0
trans_month_October	0.0	0.0	0.0	0.0	-0.0	0.0	0.0	-0.0	-0.0	0.0	-0.1	-0.1	-0.0	0.0	-0.0	0.0	-0.1	0.0	-0.1	0.0	0.0	1.0	0.0	-0.1	-0.0	-0.0	-0.1	-0.1	-0.0	-0.1
trans_day_Saturday	0.0	0.0	-0.0	0.0	-0.0	0.0	-0.0	-0.0	0.0	-0.0	0.0	-0.0	0.0	-0.0	-0.0	-0.1	0.0	-0.0	0.0	-0.1	0.0	0.0	1.0	-0.0	0.0	0.0	0.0	0.0	-0.2	0.0
trans_month_September	0.0	0.0	-0.0	0.0	0.0	-0.0	0.0	0.0	0.0	0.0	-0.1	-0.1	-0.0	0.0	-0.0	-0.0	-0.1	0.0	-0.1	0.0	0.0	-0.1	-0.0	1.0	0.0	-0.0	-0.1	-0.1	0.0	-0.1
city_pop	0.0	0.0	0.0	0.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0	-0.0	-0.0	-0.1	-0.0	-0.0	0.0	0.0	-0.0	-0.0	-0.0	-0.0	0.0	0.0	1.0	-0.0	0.0	0.0	-0.0	-0.0
age_at_trans_20-30	-0.0	0.0	0.0	0.0	0.0	-0.0	0.0	-0.0	-0.0	-0.0	0.0	0.0	-0.2	-0.1	0.0	-0.0	0.0	0.0	0.0	-0.0	-0.2	-0.0	0.0	-0.0	-0.0	1.0	-0.0	0.0	-0.0	0.0
trans_month_November	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0	-0.0	-0.0	-0.0	-0.1	-0.1	0.0	0.0	-0.0	-0.0	-0.1	-0.0	-0.1	-0.0	0.0	-0.1	0.0	-0.1	0.0	-0.0	1.0	-0.1	-0.0	-0.1
trans_month_June	-0.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0	-0.0	-0.1	-0.1	-0.0	-0.0	-0.0	-0.0	-0.1	0.0	-0.1	-0.0	-0.0	-0.1	0.0	-0.1	0.0	0.0	-0.1	1.0	0.0	-0.1
trans_day_Tuesday	-0.0	-0.0	0.0	0.0	0.0	0.0	-0.0	-0.0	-0.0	0.0	-0.0	0.0	0.0	0.0	-0.0	-0.1	0.0	0.0	-0.0	-0.1	0.0	-0.0	-0.2	0.0	-0.0	-0.0	-0.0	0.0	1.0	-0.0
trans_month_August	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0	0.0	0.0	-0.0	-0.1	-0.1	-0.0	0.0	0.0	0.0	-0.1	-0.0	-0.1	-0.0	0.0	-0.1	0.0	-0.1	-0.0	0.0	-0.1	-0.1	-0.0	1.0



Credit Card Default Problem (DEMO)

Problem Overview

Dataset Overview

Data Preparation

Proposed Models

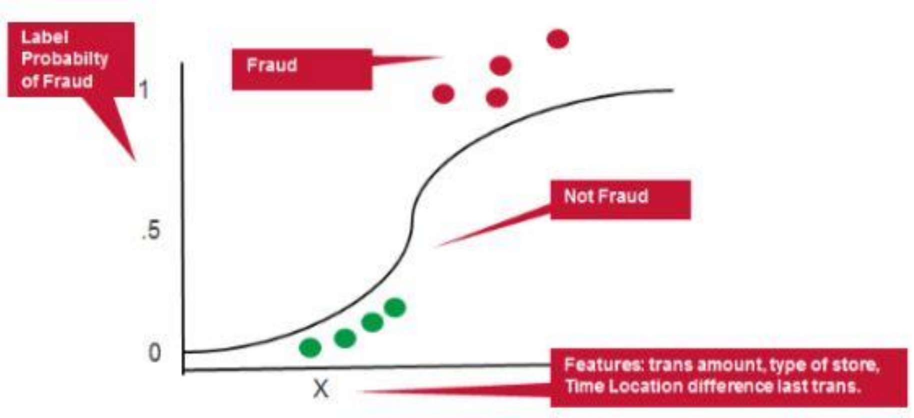
Evaluation

Conclusion

Proposed Models

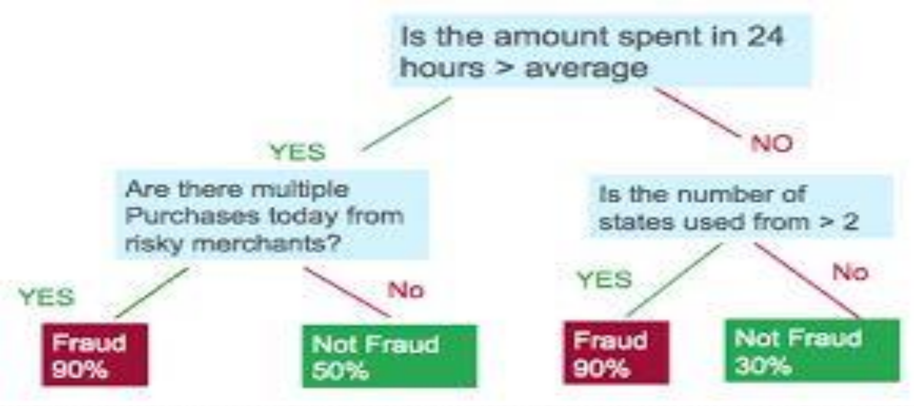
Logistic Regression:

- 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.

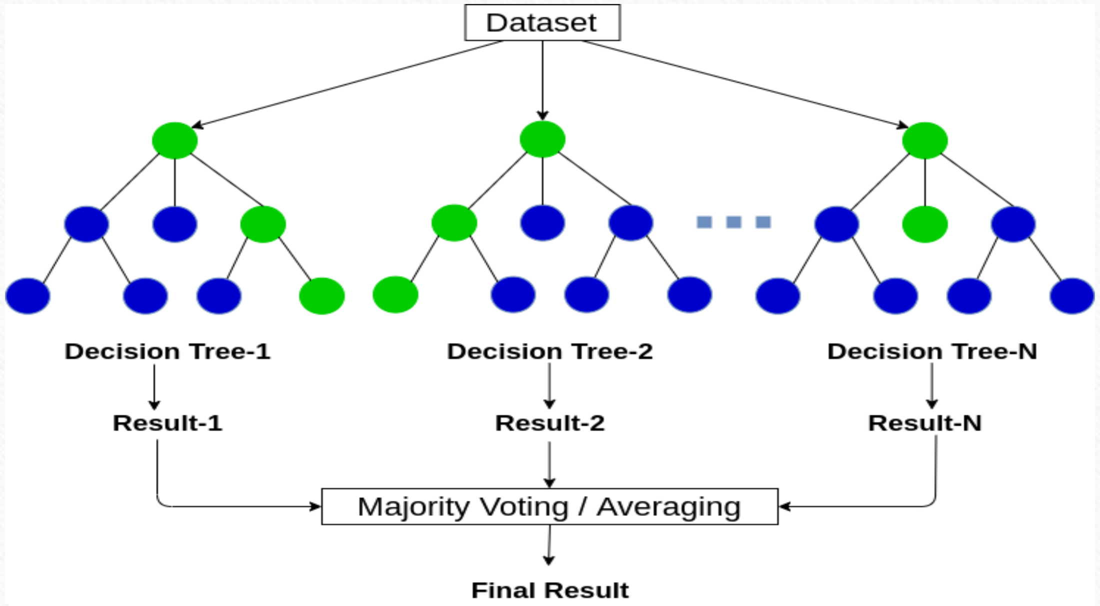

Is the amount spent in 24 hours > average
YES
NO
Are there multiple Purchases today from risky merchants?
Is the number of states used from > 2
YES
No
YES
No
Fraud 90%
Not Fraud 50%
Fraud 90%
Not Fraud 30%

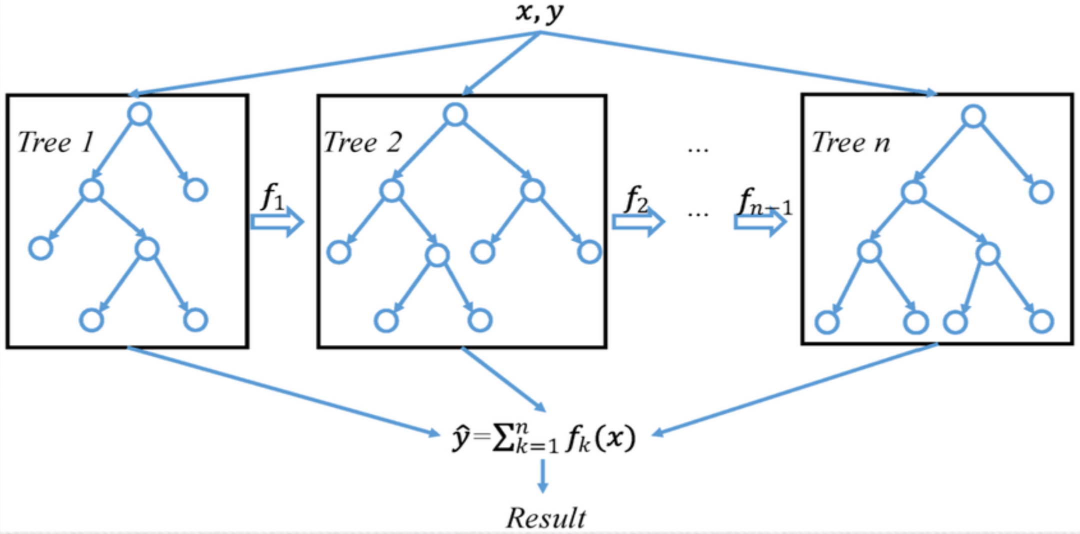
Random Forests:

- 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:**

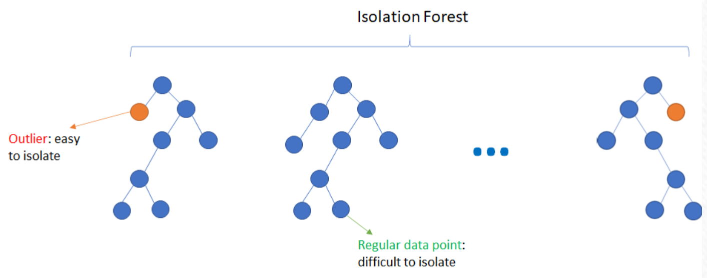
Is an unsupervised algorithm for anomaly detection that works on principle of isolating anomalies. Instead of trying to build a model of normal instances, it explicitly isolates anomalous points in the dataset. It is a very fast algorithm with a low memory demand.

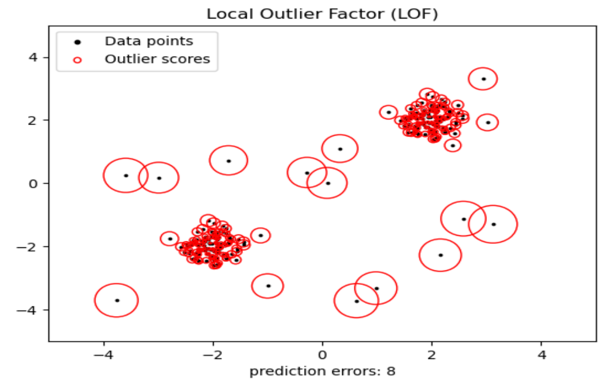
- **Local Outlier Factor (LOF):**

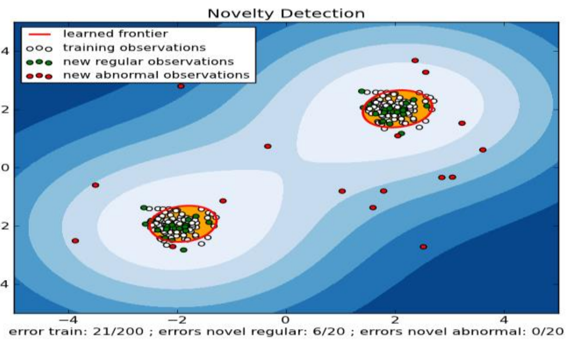
Is an unsupervised anomaly detection method which computes the local density deviation of a given data point with respect to its neighbors. It considers as outliers the samples that have a substantially lower density than their neighbors.

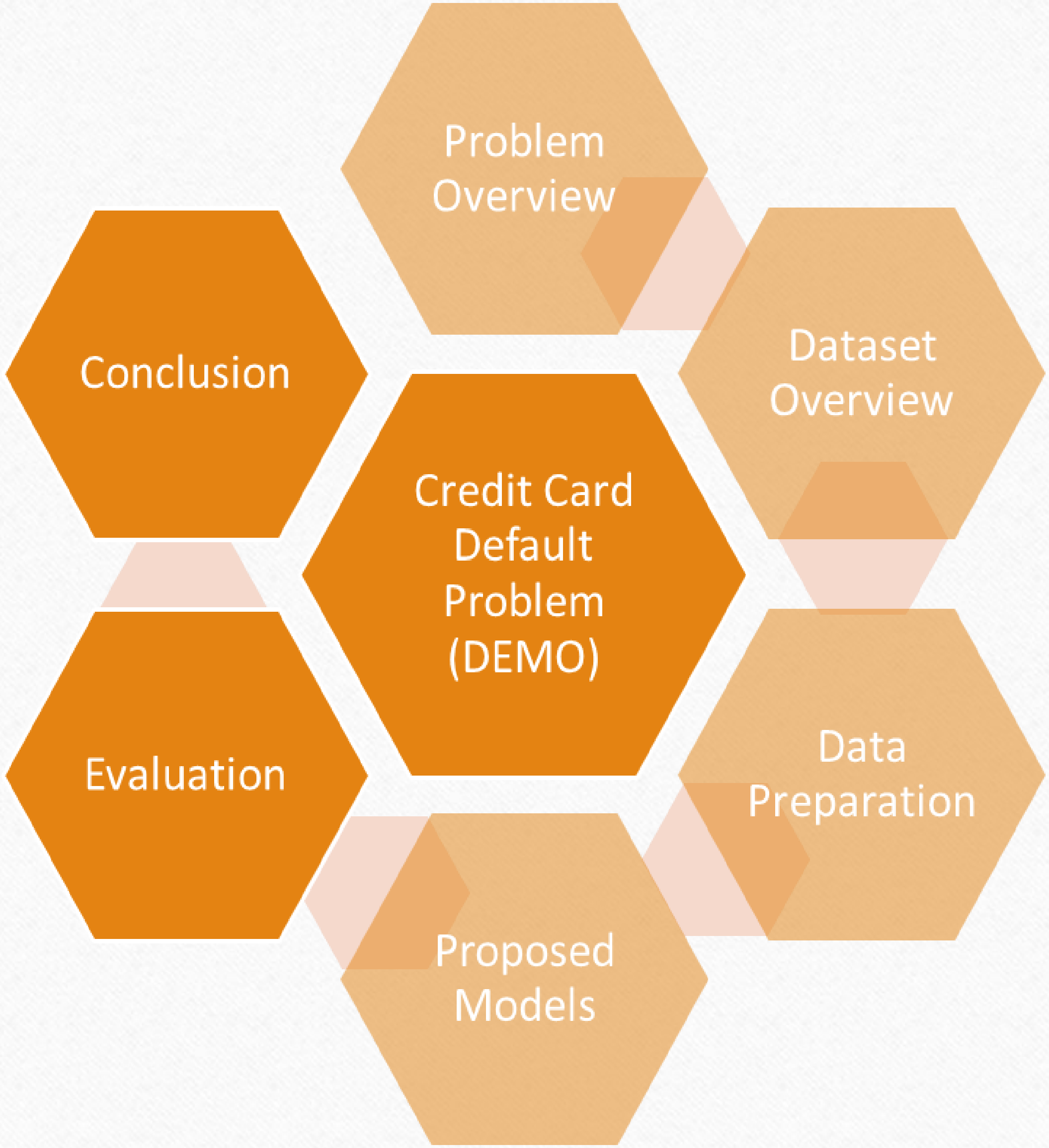
- **One-Class SVM:**

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.


Isolation Forest


Local Outlier Factor (LOF)


Novelty Detection



Credit Card Default Problem (DEMO)

Problem Overview

Dataset Overview

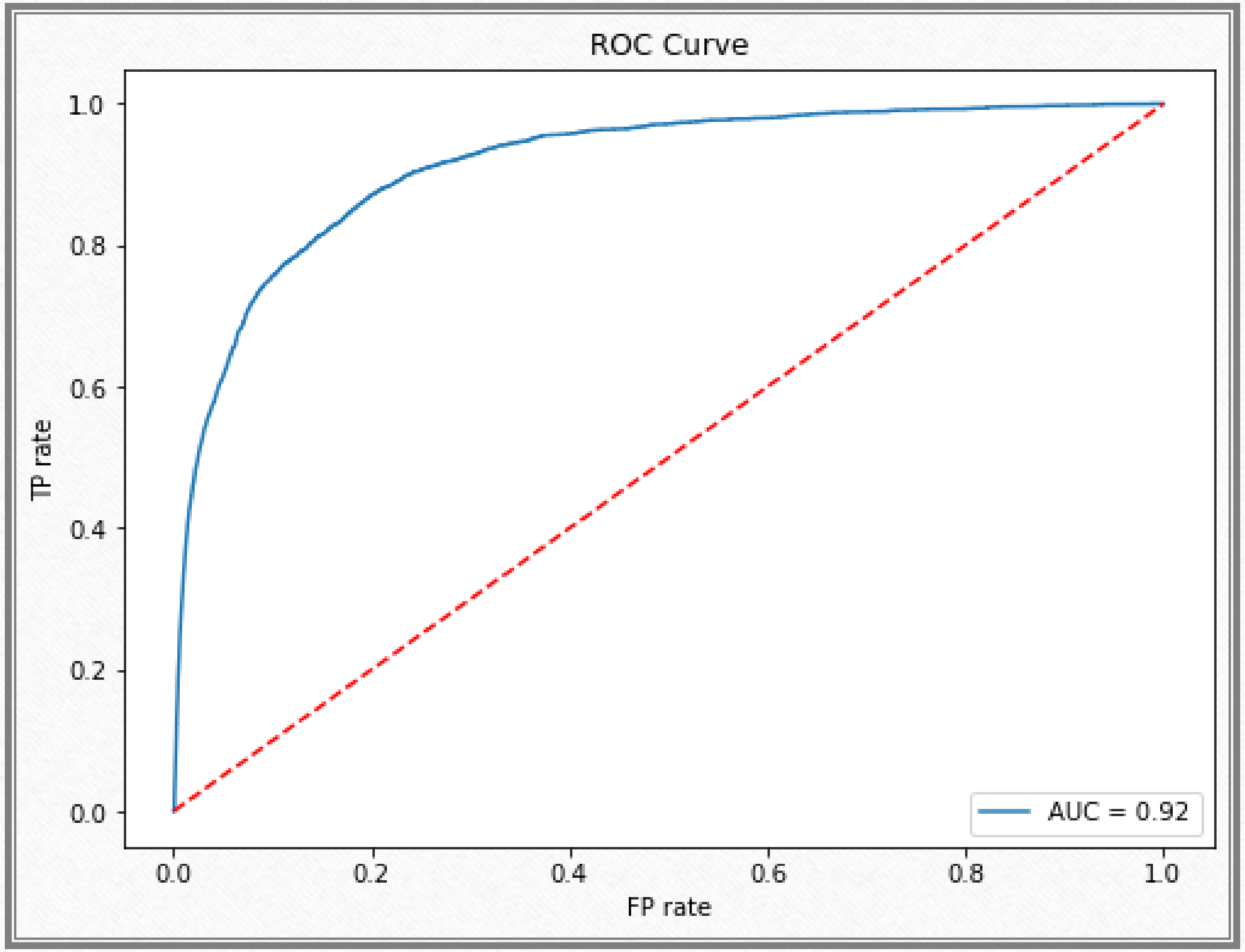
Data Preparation

Proposed Models

Evaluation

Conclusion

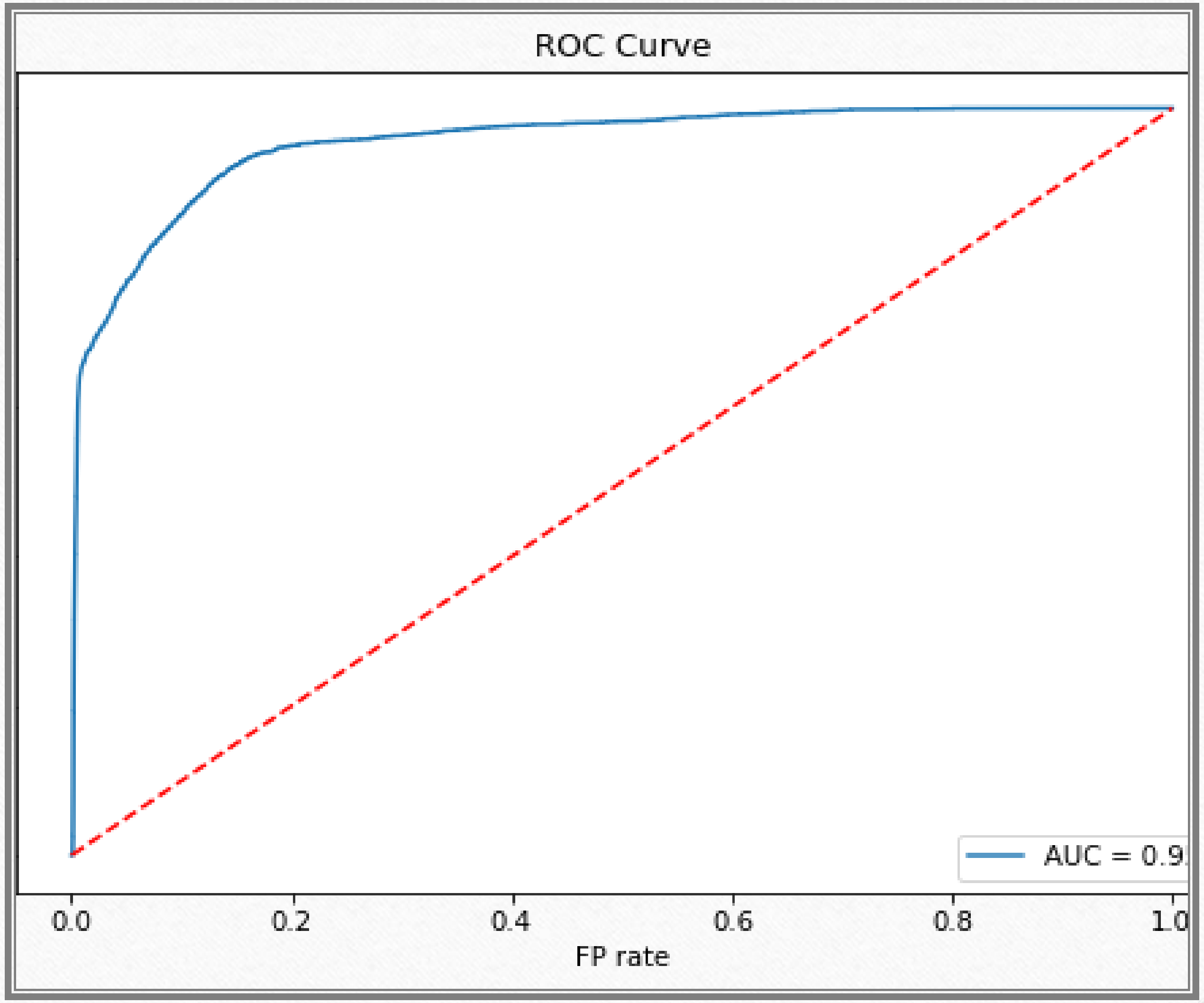
Logistic Regression using SMOTE



Logistic Regression using SMOTE technique.

- **Accuracy train score 0.98**
- **Accuracy test score 0.98**
- **Average Cross-Validation score 0.98**
- **Confusion Matrix :[547138 5686]**
  [1920 975]
- **Precision 0.15**
- **Recall 0.34**
- **F1 score 0.20**

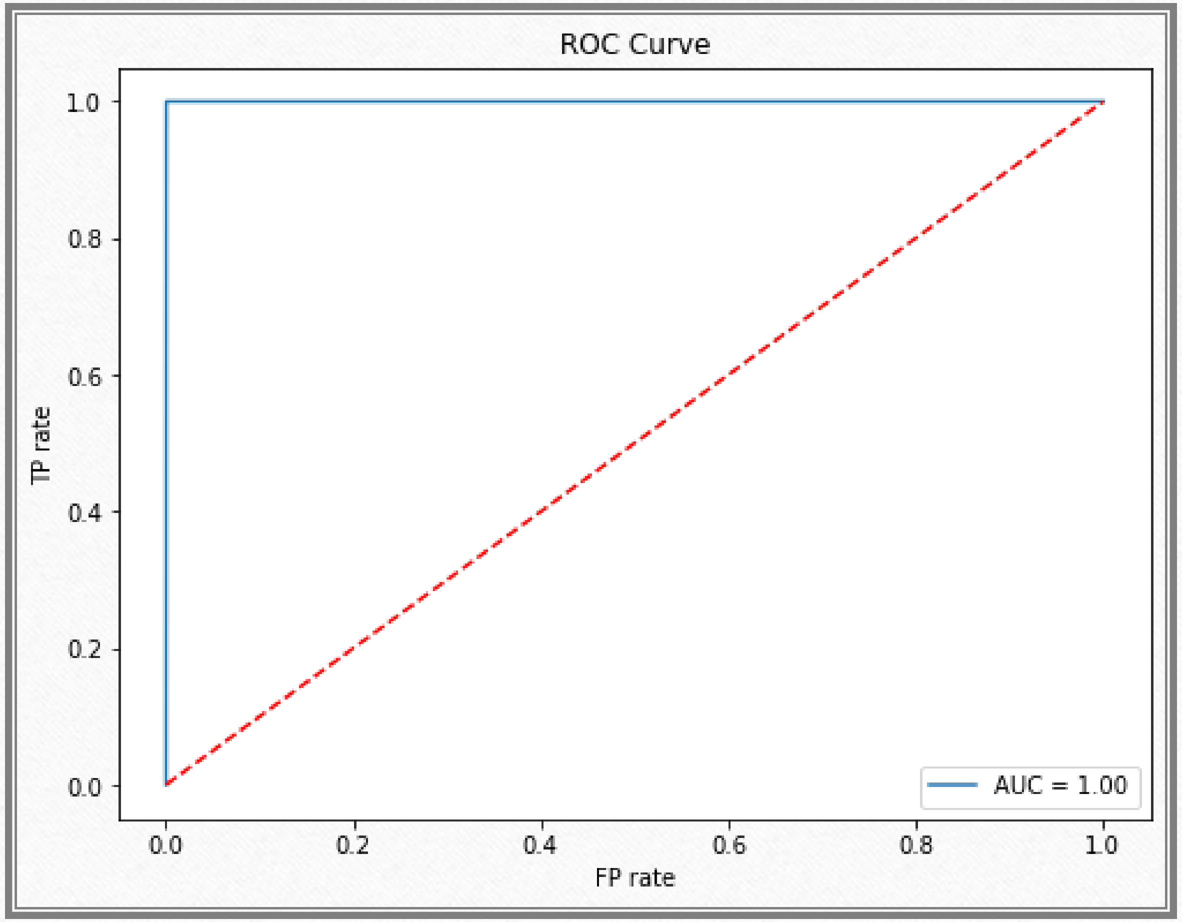
Logistic Regression using Undersample technique.



Logistic Regression using Undersample technique.

- Accuracy train score 0.88
- Accuracy test score 0.87
- Average Cross-Validation score 0.83
- Confusion Matrix : [482758   70066]
  [293   2602]
- Precision 0.04
- Recall 0.90
- F1 score 0.07

Decision Tree using SMOTE

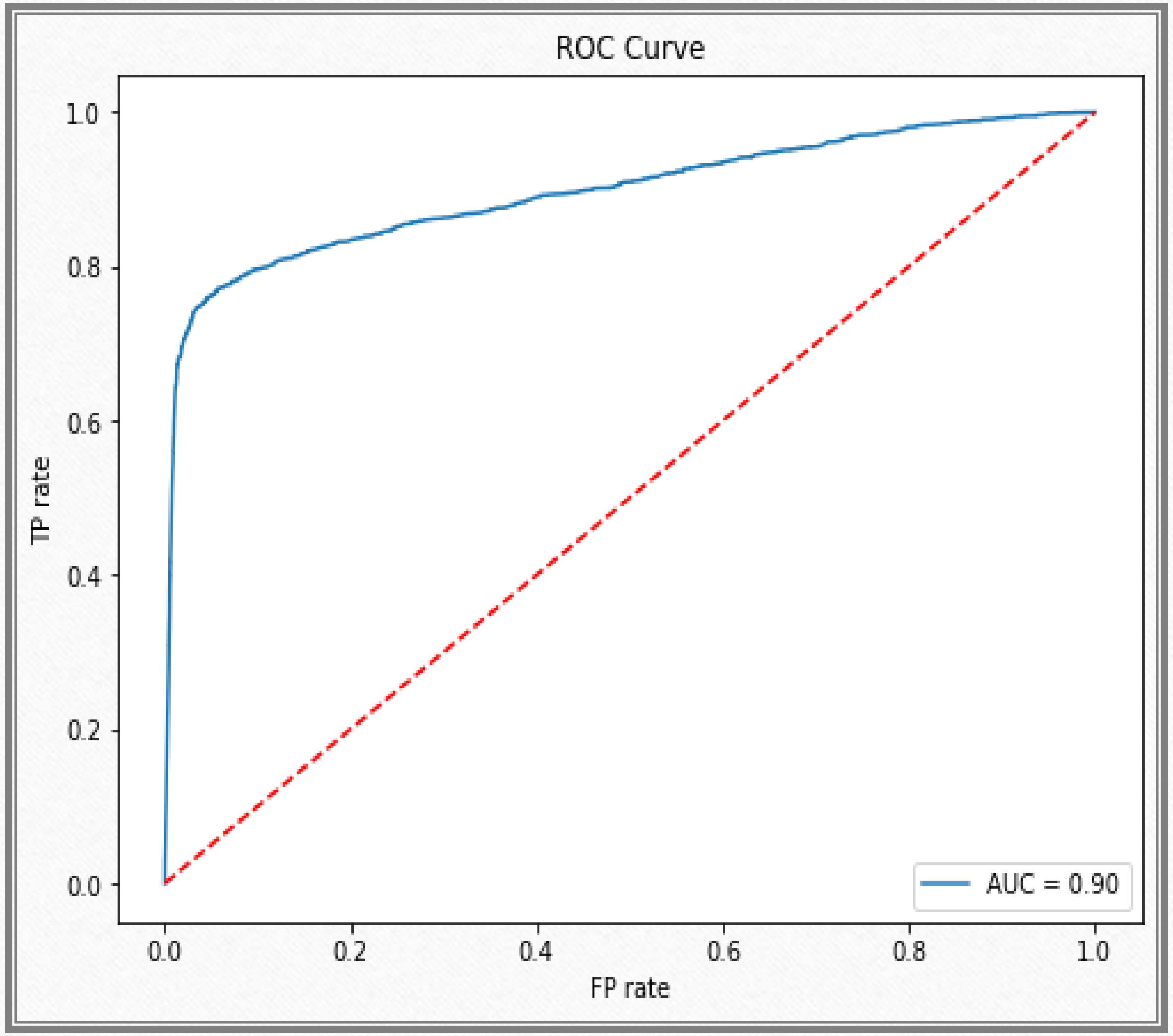


Decision Tree with SMOTE

- Accuracy train score 0.98
- Accuracy test score 0.98
- Average Cross Validation score 0.85
- Confusion Matrix : [542071 10753]
  [191 2704]
- Precision 0.20
- Recall 0.93
- F1 score 0.33

Random Forest Classification using SMOTE technique.

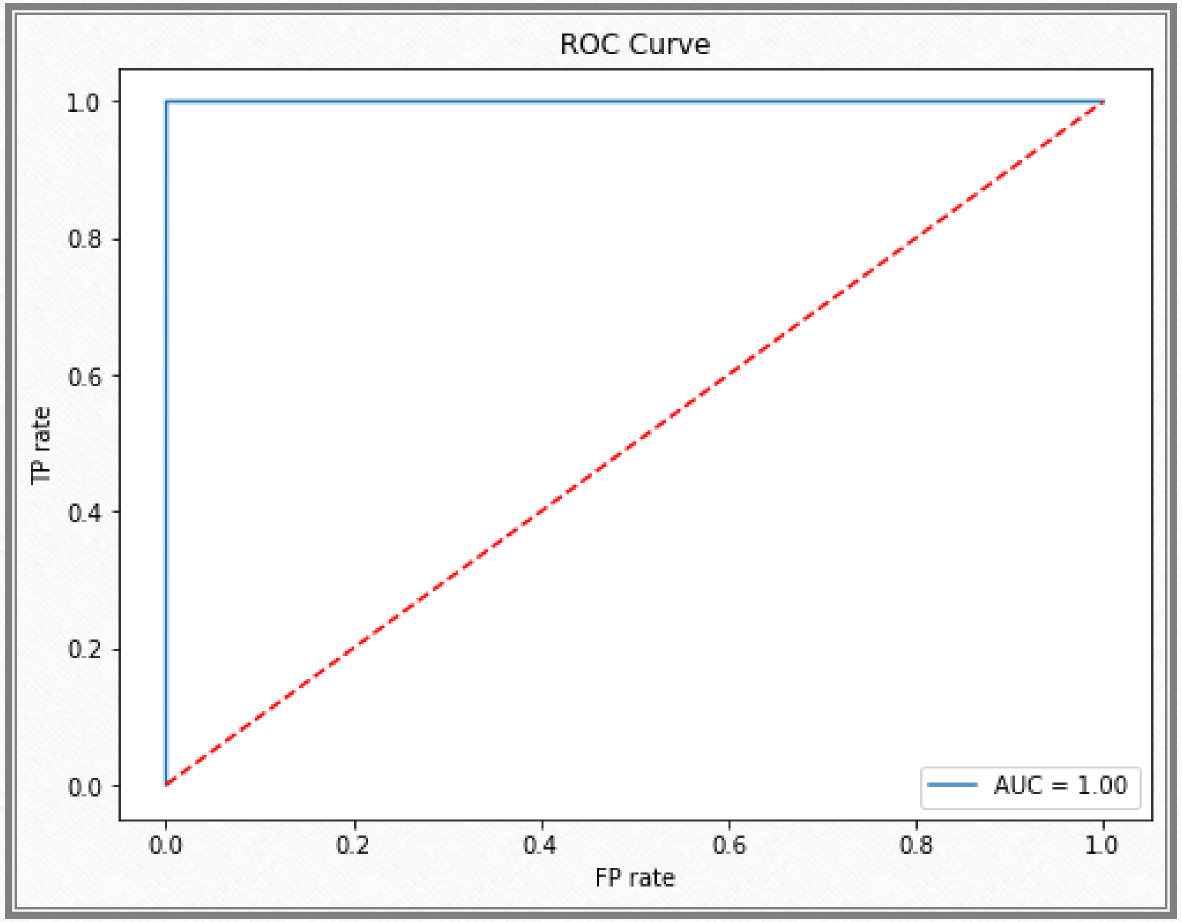
- Accuracy train score 0.94
- Accuracy test score 0.96
- Average Cross-Validation score 0.89
- Confusion Matrix : [534162 18662]
  [740 2155]
- Precision 0.10
- Recall 0.74
- F1 score 0.18



Random Forest using Random Undersample technique

Random Forest
using Random Undersample technique and
GridSearch for best parameters and best score.

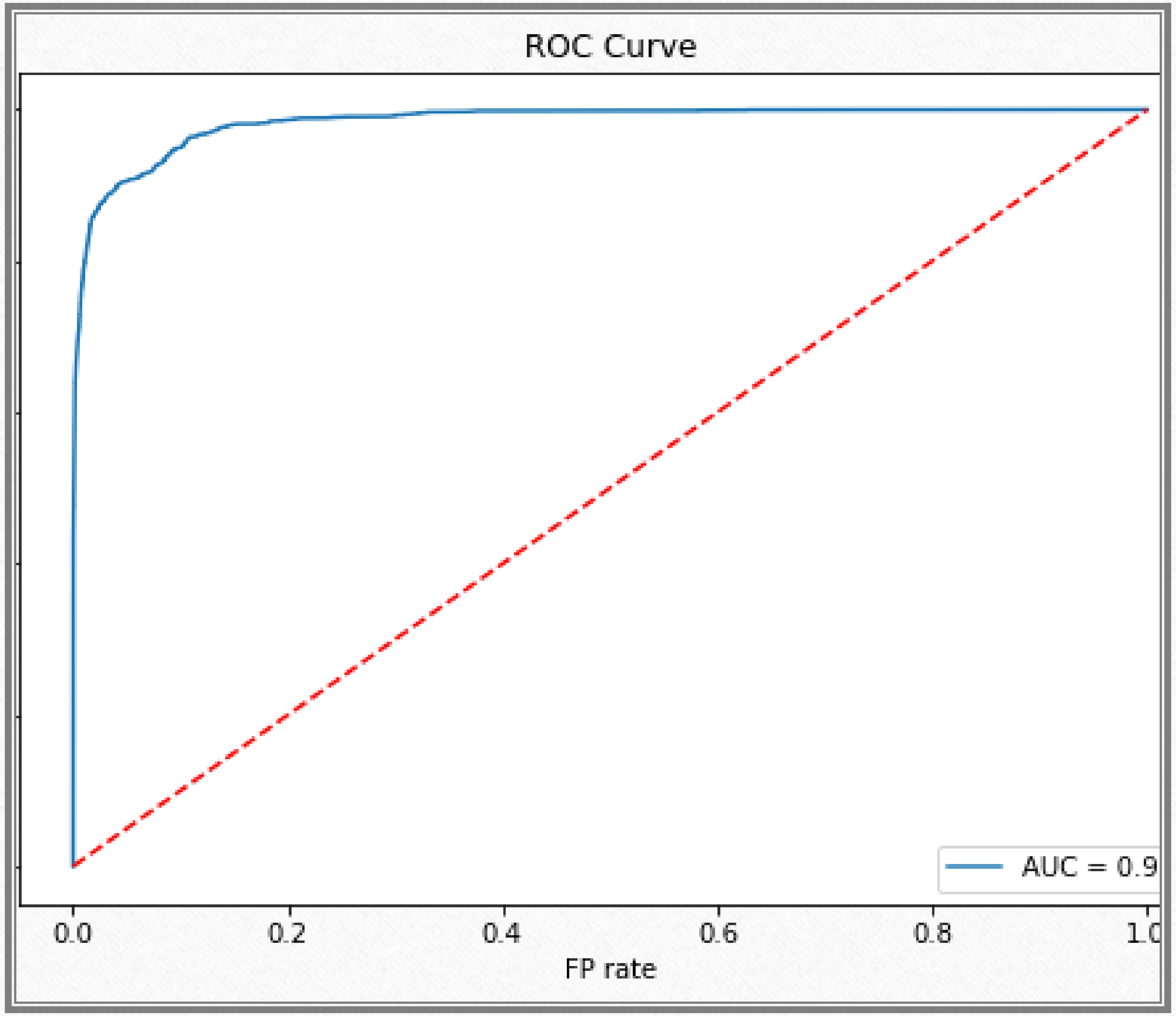
- Accuracy train score 1
- Accuracy test score 0.97
- Average Cross Validation score 0.94
- Confusion Matrix : [538956 13868]
  [0 2895]
- Precision 0.20
- Recall 0.93
- F1 score 0.33



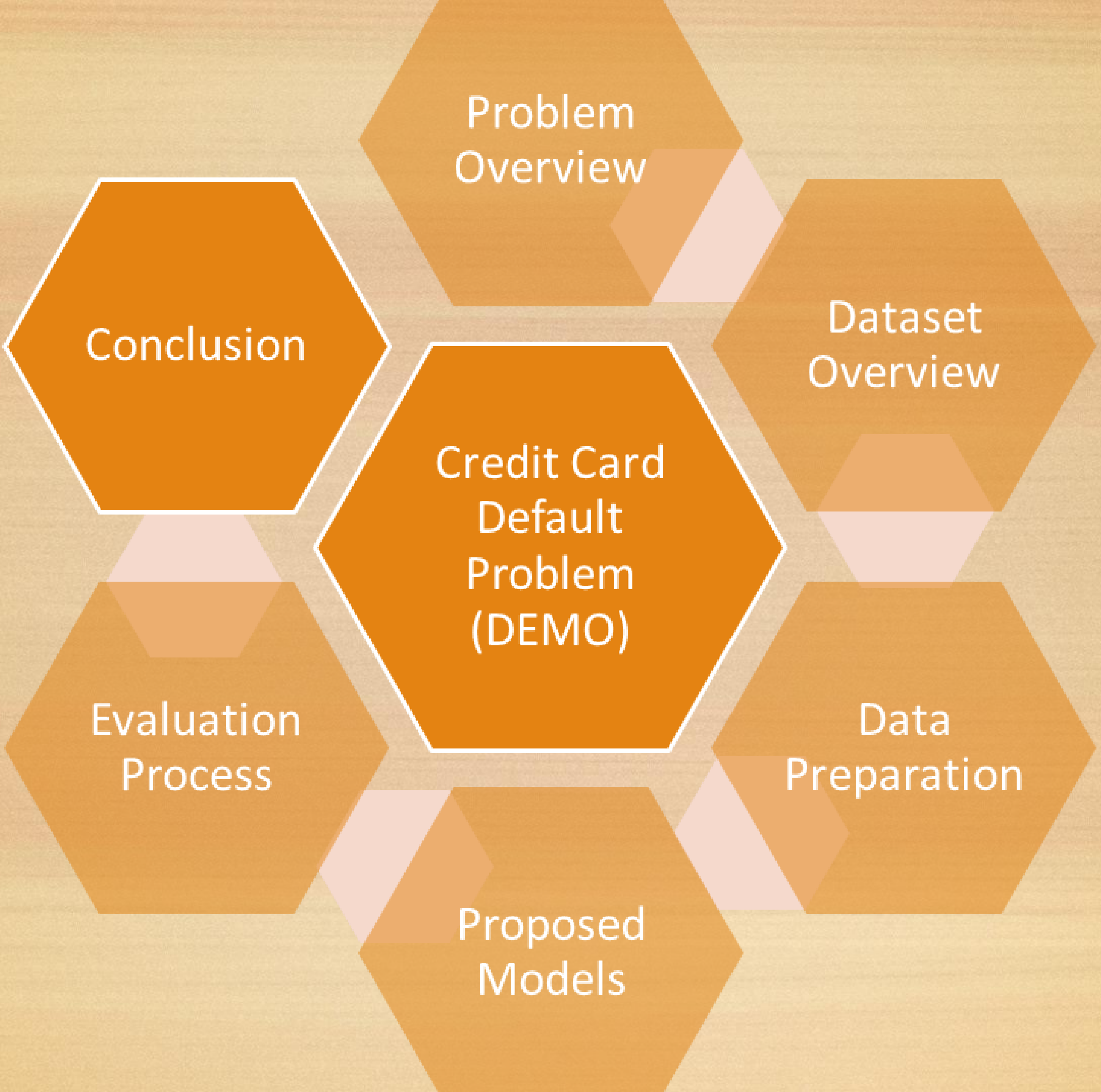
XGBoost Classifier using Random UnderSampled technique

XGBoost Classifier using Random UnderSampled technique

- Accuracy train score 0.92
- Accuracy test score 0.94
- Average Cross Validation score 0.91
- Confusion Matrix : [523147 29677]
  [266 2629]
- Precision 0.08
- Recall 0.91
- F1 score 0.15



Model	Accuracy	Precision	Recall	F1	Confusion Matrix	AUC Score
Logistic Regression With Smote	0.986	0.15	0.34	0.20	[547138 5686 1920 975]	0.92
Logistic Regression using Unsampled technique	0.873	0.04	0.90	0.07	[523147 29677] [266 2629]	0.95
Decision Trees With Smote	0.980	0.20	0.93	0.33	[542071 10753] [191 2704]	1.00
Random Forest With Smote	0.965	0.10	0.74	0.18	[534162 18662] [740 2155]	0.90
Random Forest Under Sampled	0.975	0.17	1.00	0.29	[538956 13868] [0 2895]	1.00
XGBoost Classifier Under Sampled	0.943	0.08	0.91	0.14	[523147 29677] [266 2629]	0.98



Credit Card Default Problem (DEMO)

Problem Overview

Dataset Overview

Data Preparation

Proposed Models

Evaluation Process

Conclusion

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 TheRandom Undersampled technique and SMOTE technique. The GridSearch was used to find optimal hyper parameters of Random Forest and XGBoost models . As a result of modeling , best score performed Random Forest with Optimized Hyperparameters with UnderSample technique.

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 !