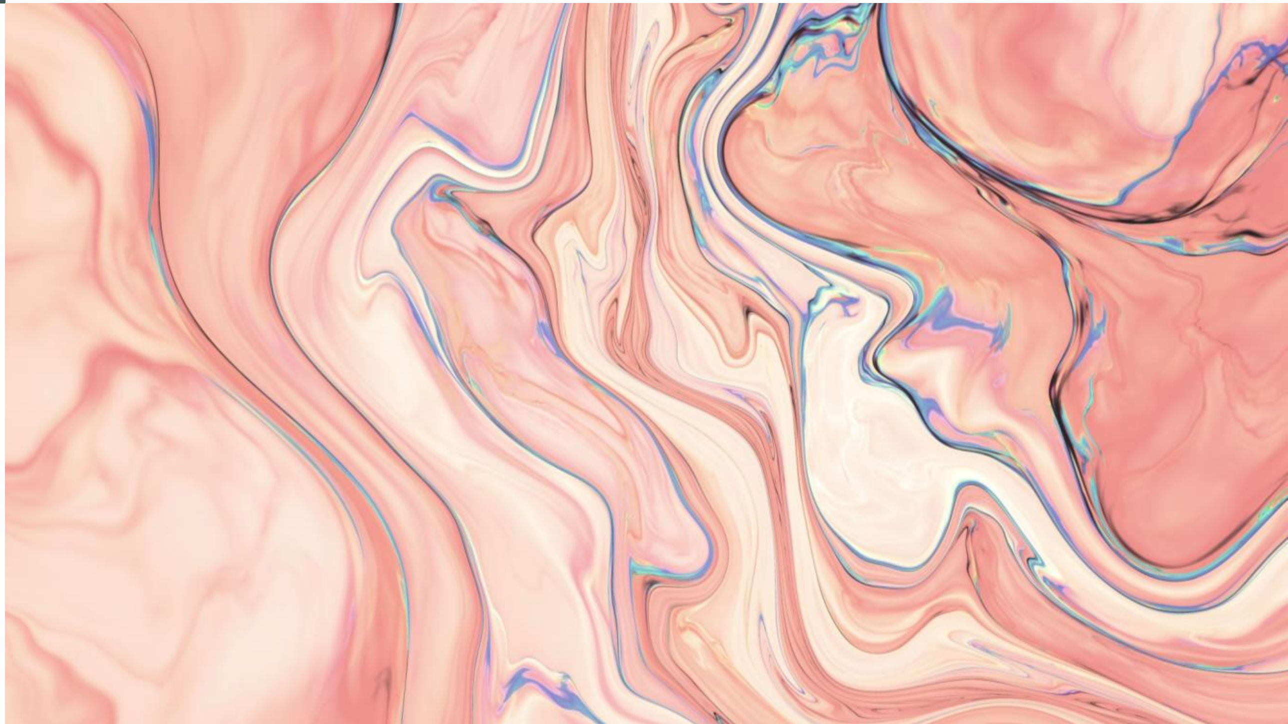
Dr. Lindi Liao Duoduo
AIT 614-005
23rd April 2024

Geospatial Analysis of Credit Card Fraudulent Transactions

GEORGE MASON UNIVERSITY

TEAM 7

Chanyoung Park, Nithiya Varsha Markandan Rajedren, Samhita Sarikonda, Vishal Orsu, Ravi Datta Rachuri



INTRODUCTION

Project Overview: Enhancing Credit Card Fraud Detection with Geospatial Analysis

Challenge: Current fraud detection systems struggle with evolving fraud tactics and their complexity.

Innovative Approach: Integrating geospatial analysis with traditional methods for improved fraud detection.

Key Differentiators:

Geographic Insights: Identifies high-risk areas through spatial transaction analysis.

Enhanced Detection: Combines transactional and geographic data to improve accuracy.

Targeted Prevention: Delivers region-specific anti-fraud strategies for financial institutions.

OBJECTIVES

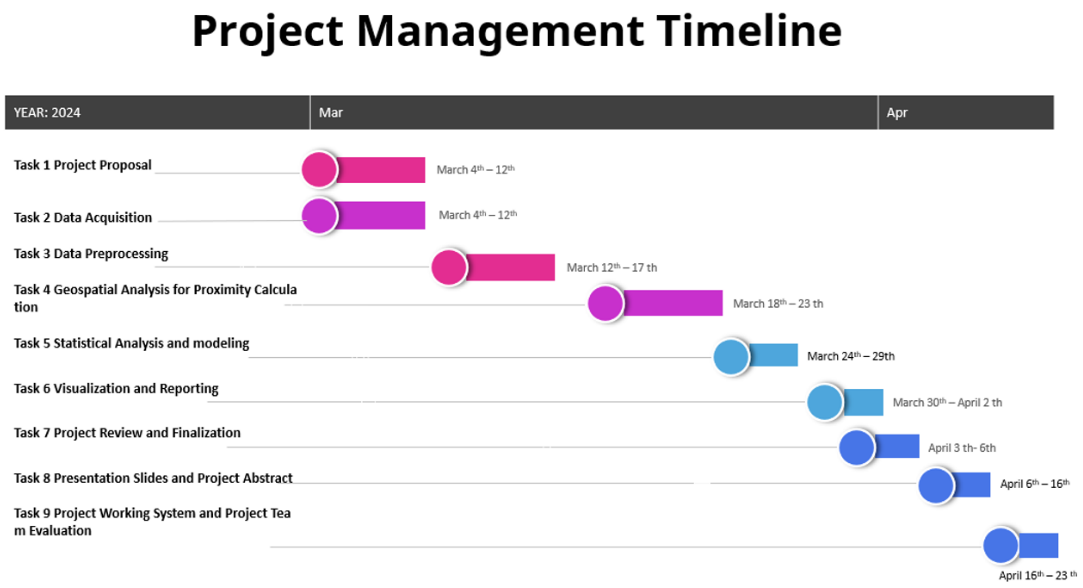
Primary Goal: Leverage geospatial analysis for more effective credit card fraud detection and prediction.

Objectives:

1. Identify high-risk geographic regions for credit card fraud.
2. Integrate geospatial data into existing fraud detection frameworks to enhance accuracy.
3. Develop a comprehensive model using transactional and geographic data to pinpoint potential fraud.

Outcome: Provide financial institutions with advanced tools for improved fraud prevention and customer security.

SUMMARIZED PROJECT TIMELINE



Project Management Timeline

March 4 – 12: Project Proposal & Data Acquisition

March 12 – 17: Data Preprocessing

March 18 – 23: Geospatial Analysis for Proximity Calculation

March 24 – 29: Statistical Analysis and Modeling

March 30 – April 2: Visualization and Reporting

April 3 – April 6: Project Review and Finalization

April 6 – April 16: Presentation Slides and Project Abstract

April 16 – April 23: Project Working System

SELECTED DATASET

Source: Available in Kaggle

Transactional Details: Time, anonymized card numbers, merchant info, categories, and amounts.

Geographic Coordinates: Latitude and longitude for transaction and merchant locations.

Demographic Information: Age, occupation, and city of residence of cardholders.

Fraud Flag: Transactions labeled as 'fraudulent' or 'non-fraudulent'.

Analysis Applications: Enables consumer behavior studies and geospatial fraud detection.

Project Relevance: Provides a rich dataset for enhancing fraud detection models using advanced analytics.

THE SYSTEM

databricks

Architecture/Framework:

Core Platform: Databricks - Unified analytics platform integrating data processing, machine learning, and collaborative workflows.

Data Processing & Analytics:

Tools: Spark with Python (PySpark) for distributed computing and efficient data transformations.

Machine Learning: Spark MLlib used for regression, Random Forest, and Gradient Boost tasks.

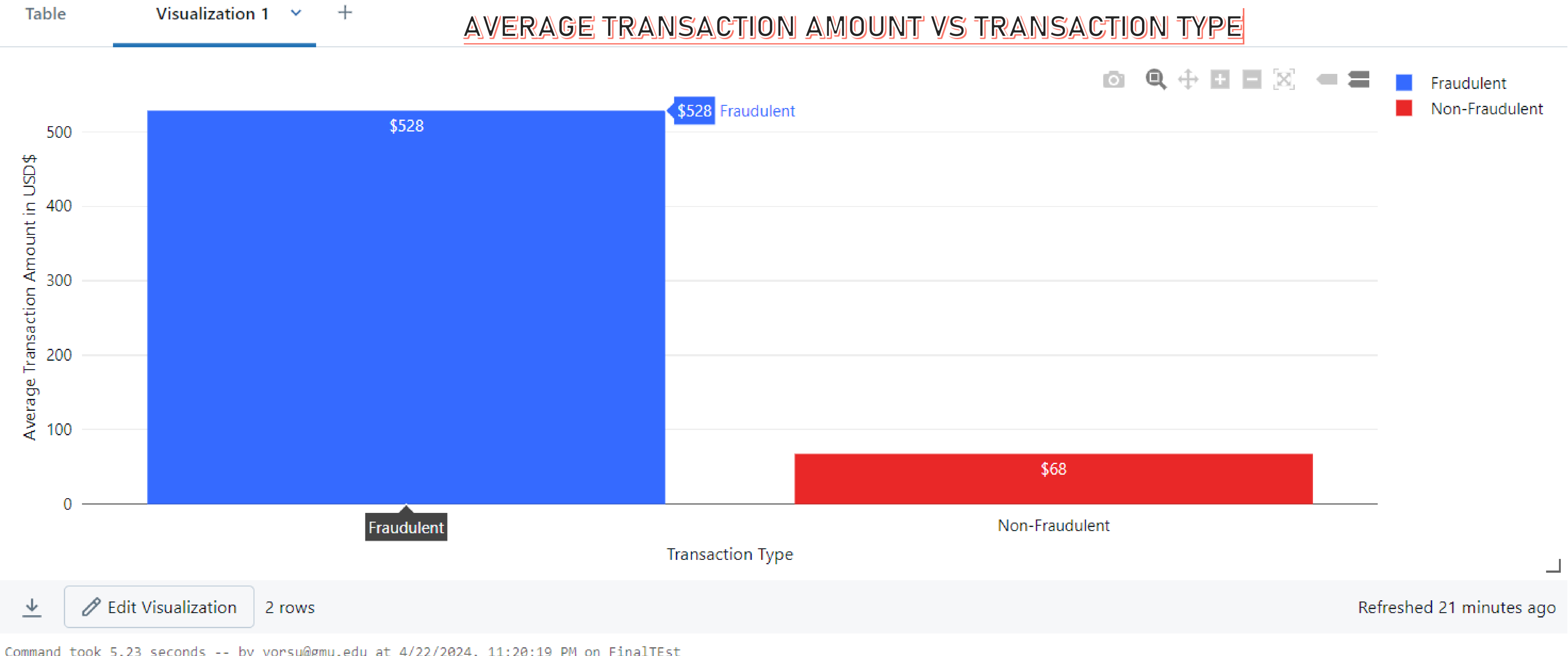
Software/Hardware Development:

Platform: Databricks as scalable cloud-based infrastructure for data analysis.

Services: Cloud-based tools within Databricks for advanced data management and processing.

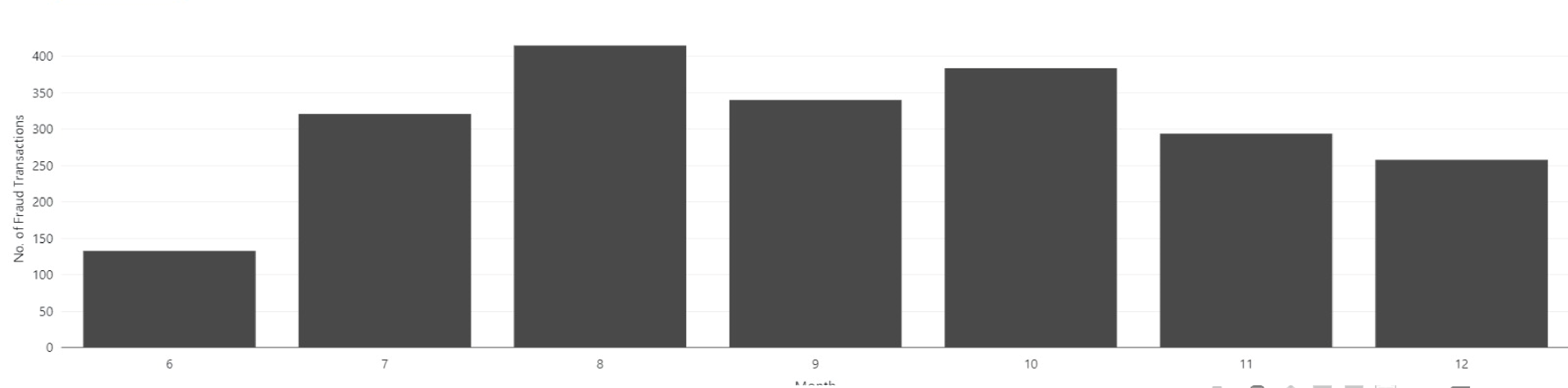
EXPERIMENTAL RESULTS AND ANALYSIS

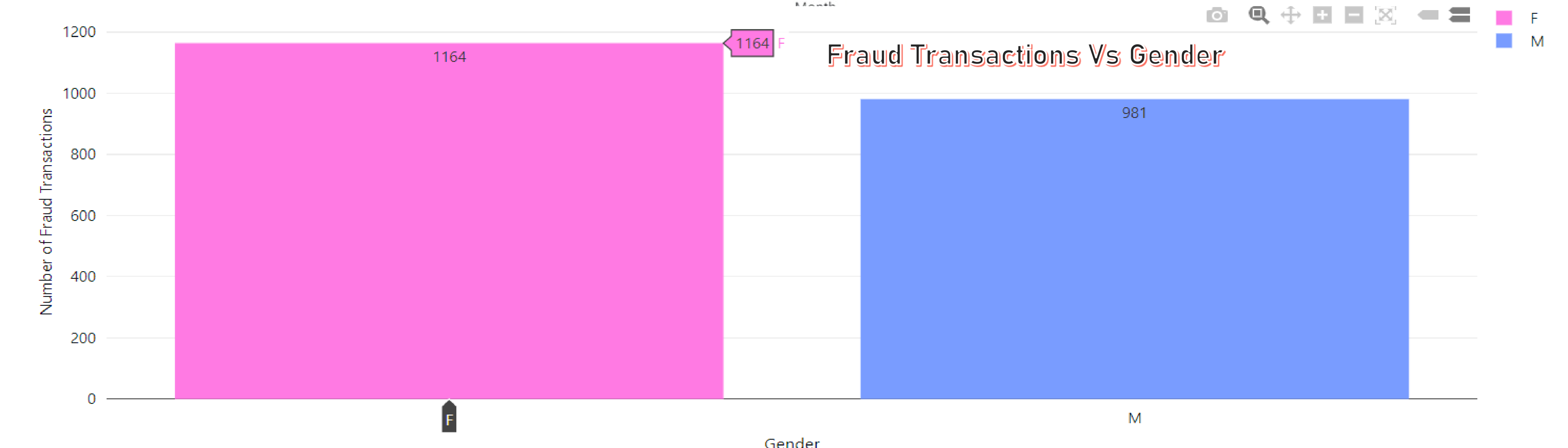
AVERAGE TRANSACTION AMOUNT VS TRANSACTION TYPE



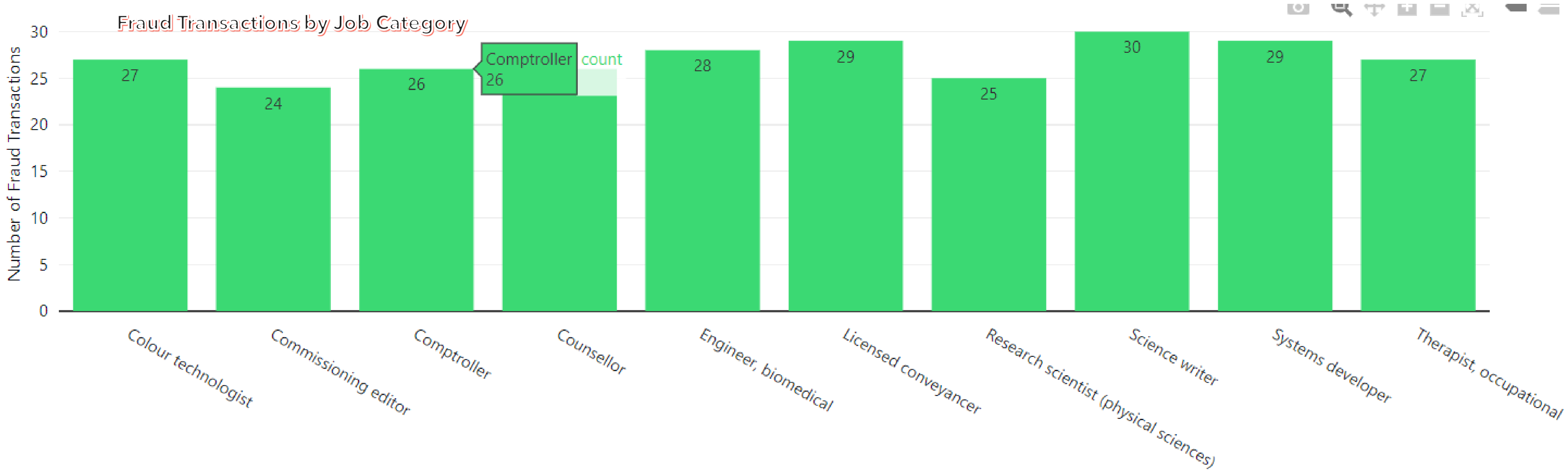
```
Command took 5.23 seconds -- by vorsu@gmu.edu at 4/22/2024, 11:20:19 PM on FinalTEst
```

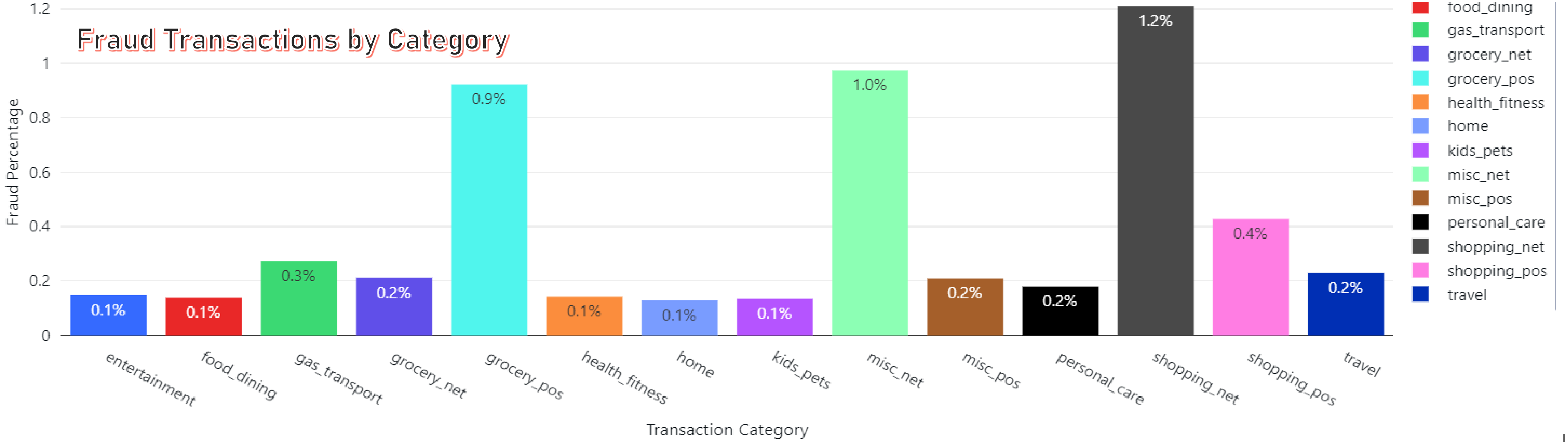
Fraud Transactions by Month



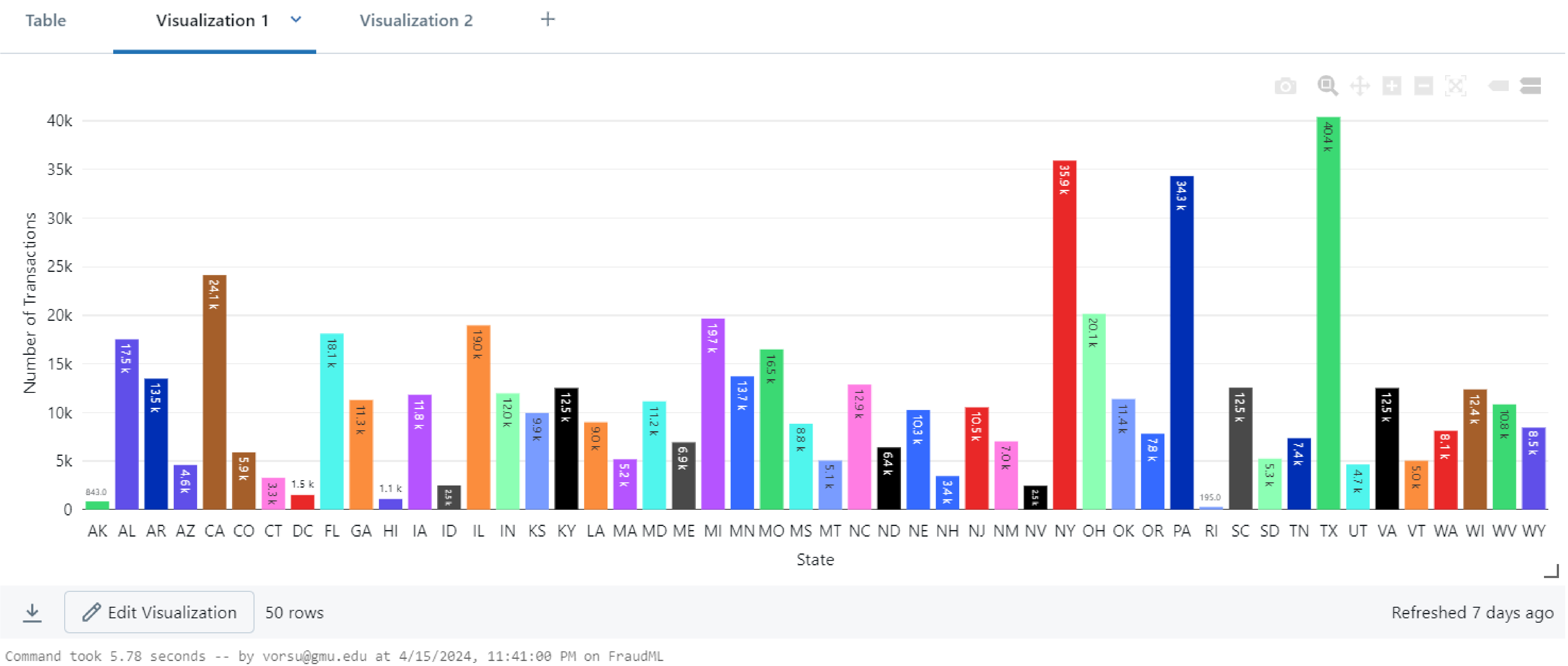


Fraud Transactions Vs Gender


Fraud Transactions by Job Category

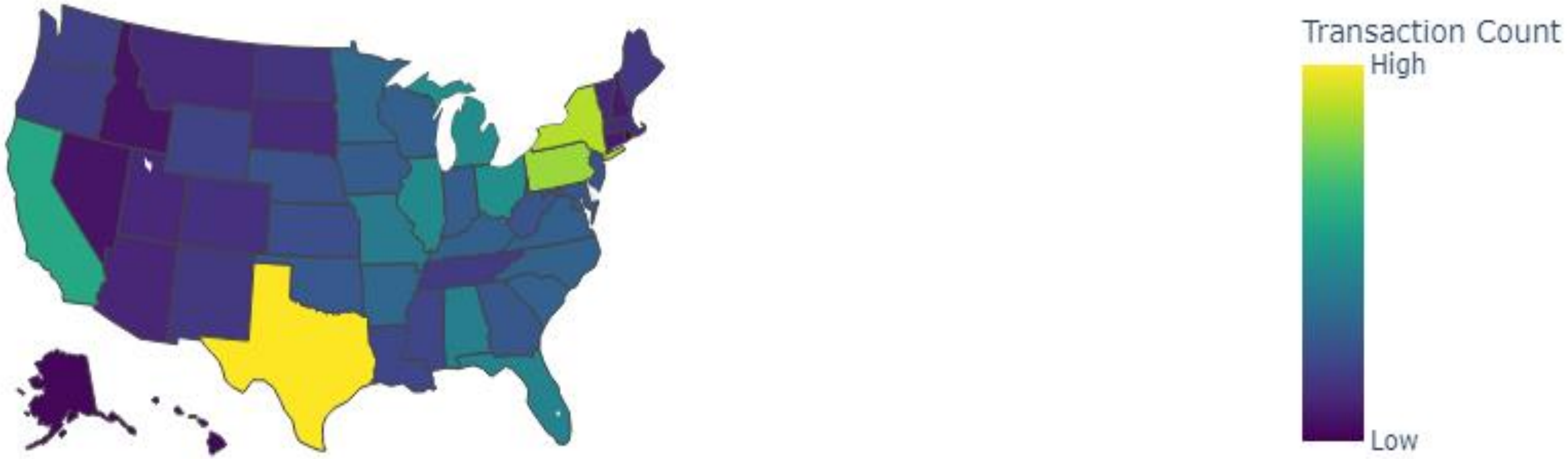

Fraud Transactions by Category

ALL TRANSACTIONS BY STATE



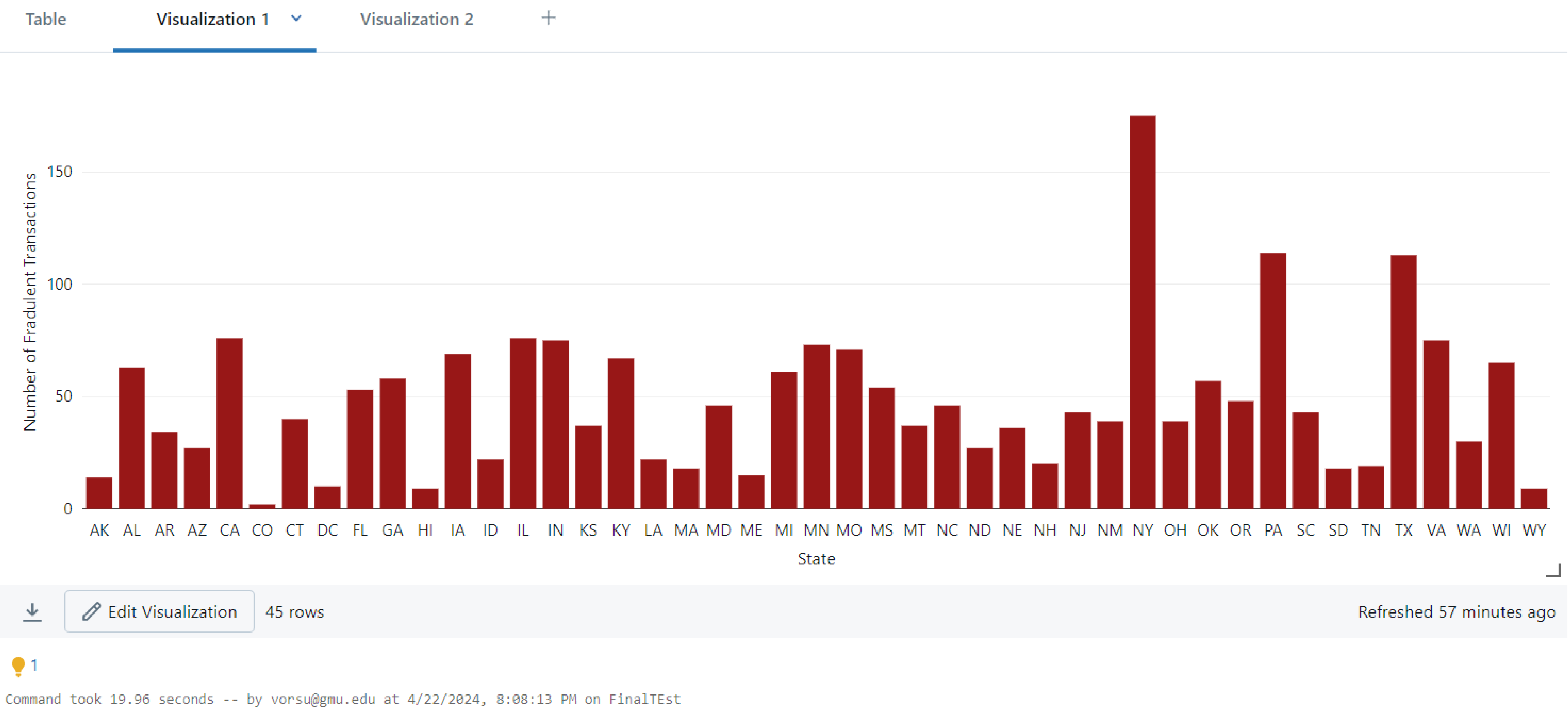
```
Command took 5.78 seconds -- by vorsu@gmu.edu at 4/15/2024, 11:41:00 PM on FraudML
```

GEOSPATIAL ANALYSIS



Heatmap of All Transactions by State

NUMBER OF FRAUD TRANSACTIONS BY STATE

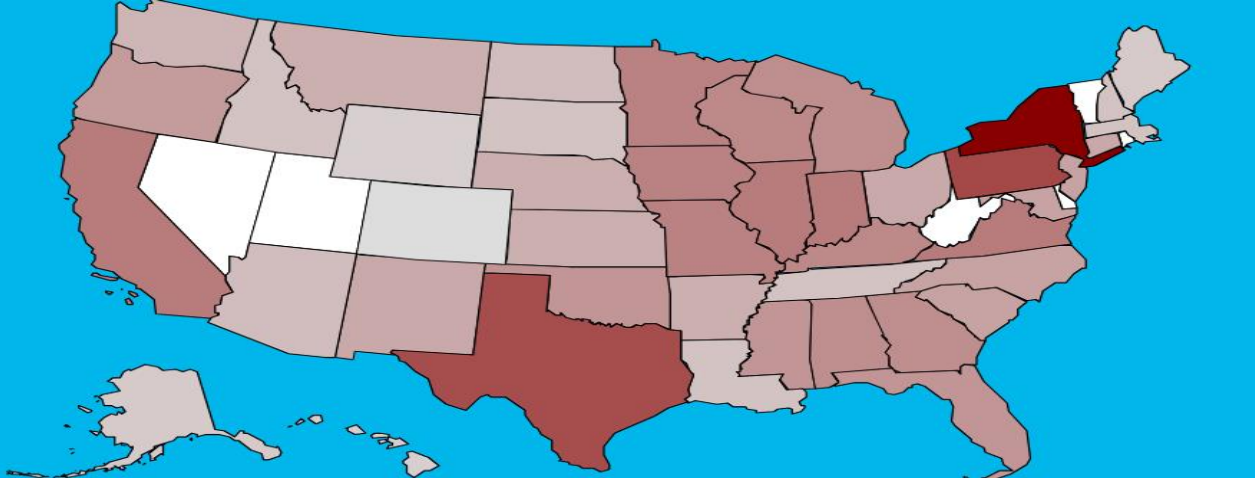


45 rows

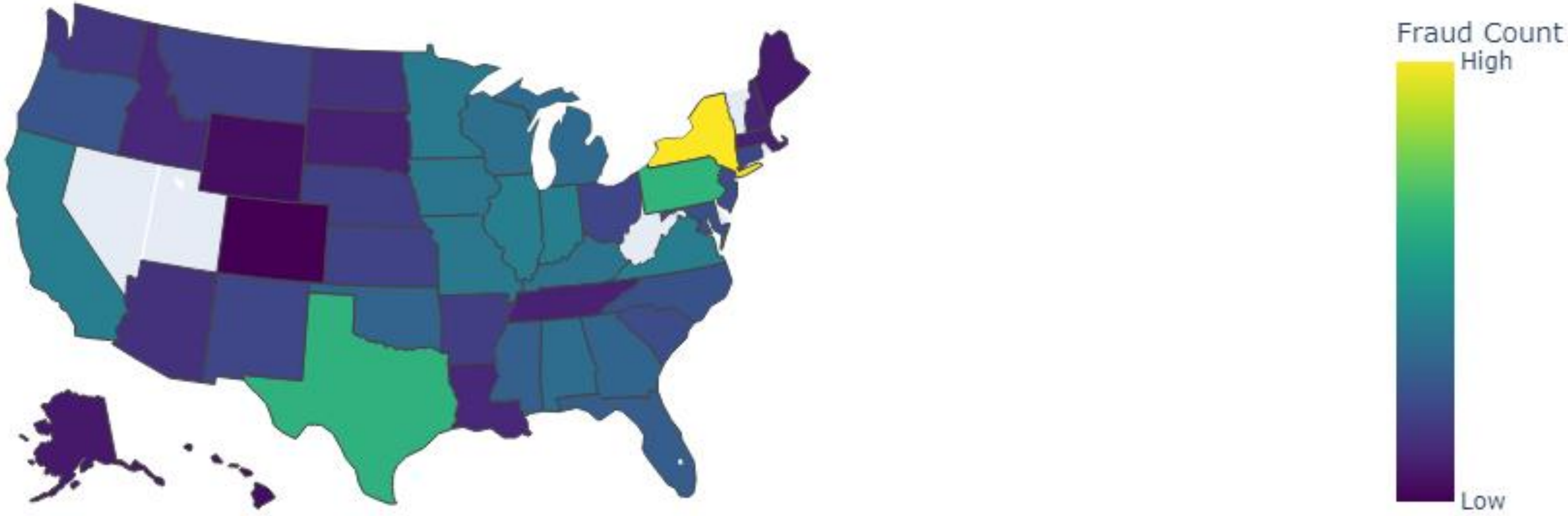
Refreshed 57 minutes ago

```
Command took 19.96 seconds -- by vorsu@gmu.edu at 4/22/2024, 8:08:13 PM on FinalTEst
```

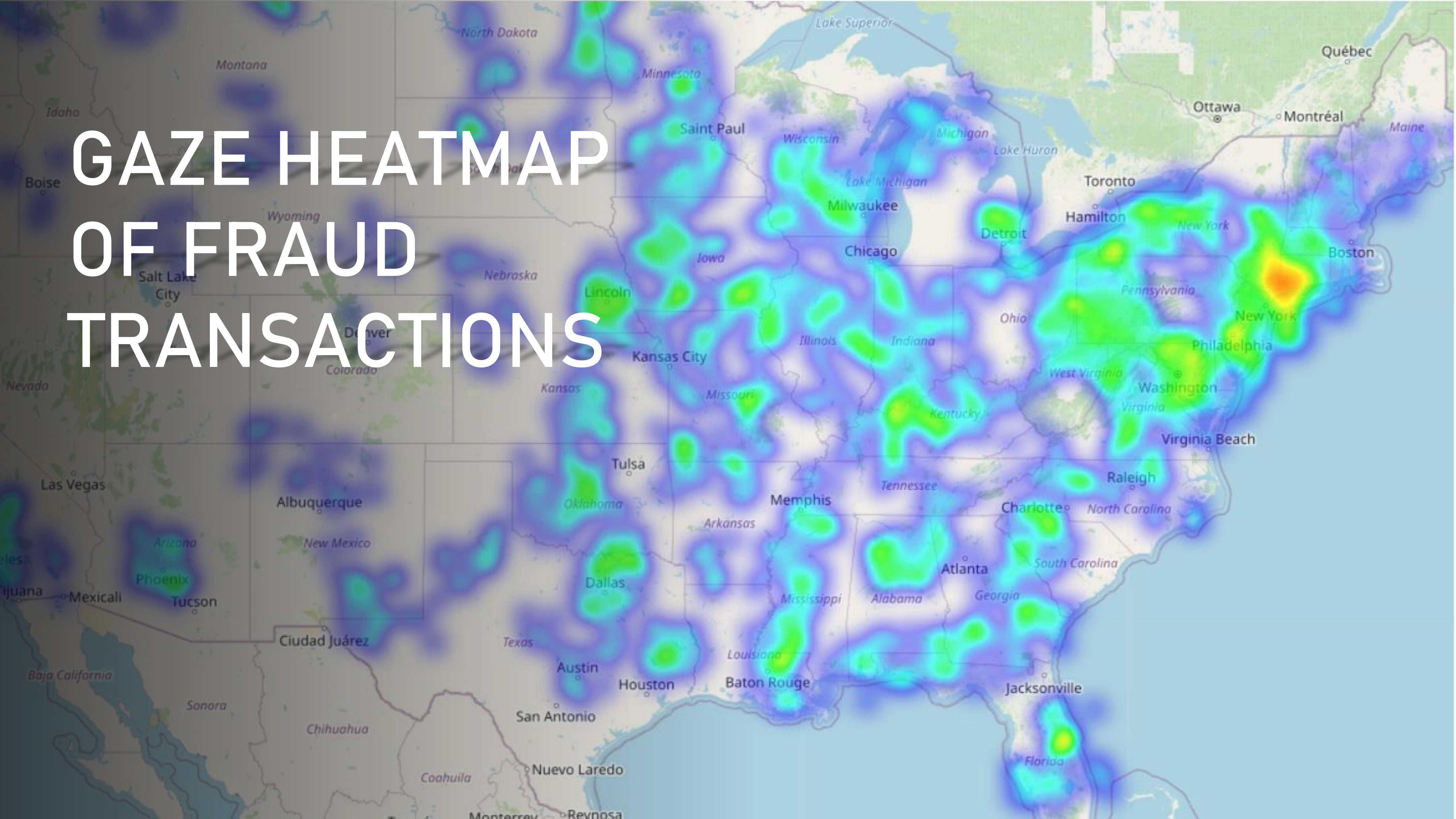
Heatmap of Fraudulent Transactions



Heatmap of Fraudulent Transactions by State

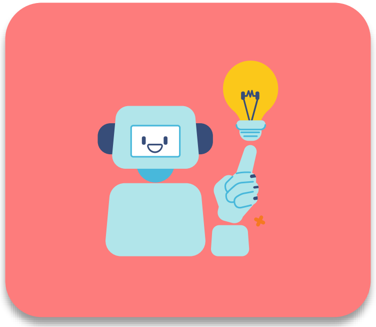


GAZE HEATMAP OF FRAUD TRANSACTIONS



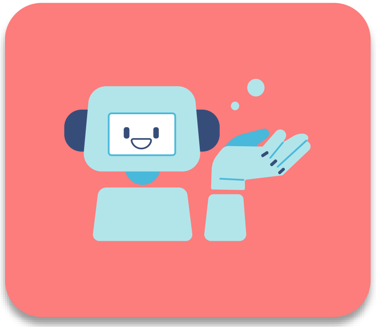
CLASSIFICATION MODEL

1. LOGISTIC REGRESSION



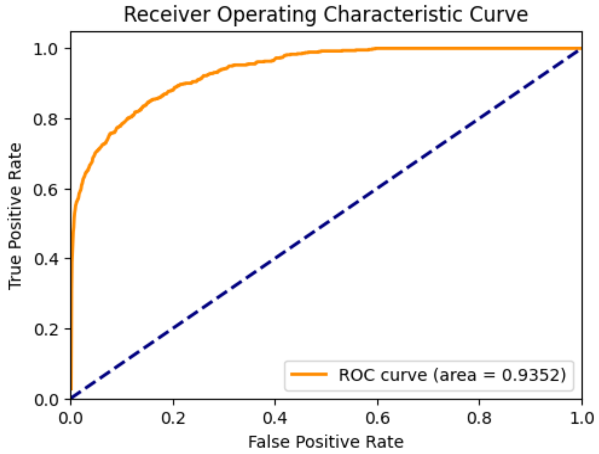
2. RANDOM FOREST

3. DECISION TREE

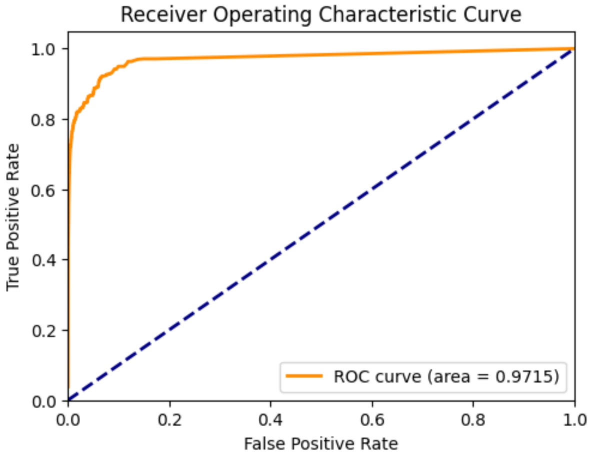


4. GRADIENT BOOSTING

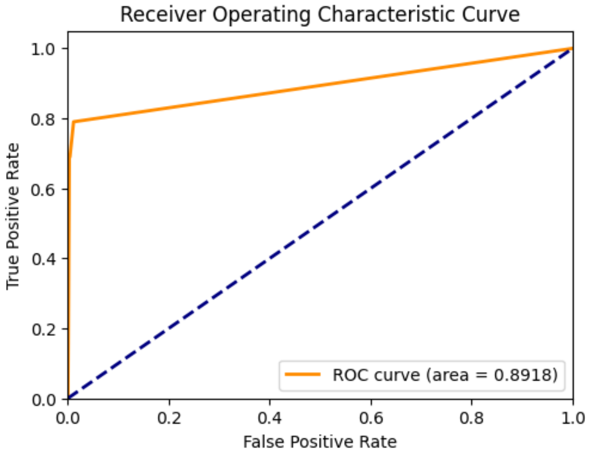
COMPARISON OF AREA UNDER ROC



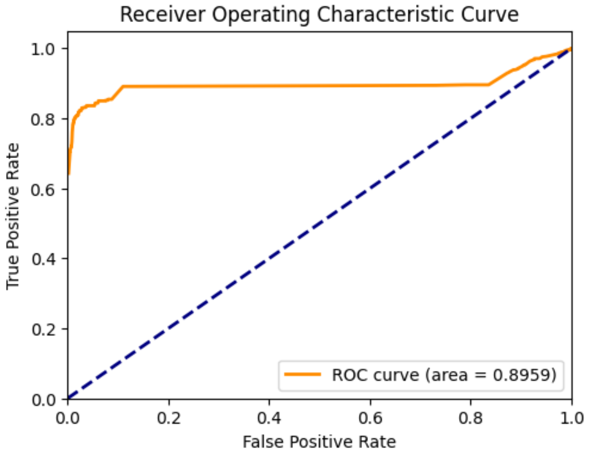
LOGISTIC REGRESSION



RANDOM FOREST

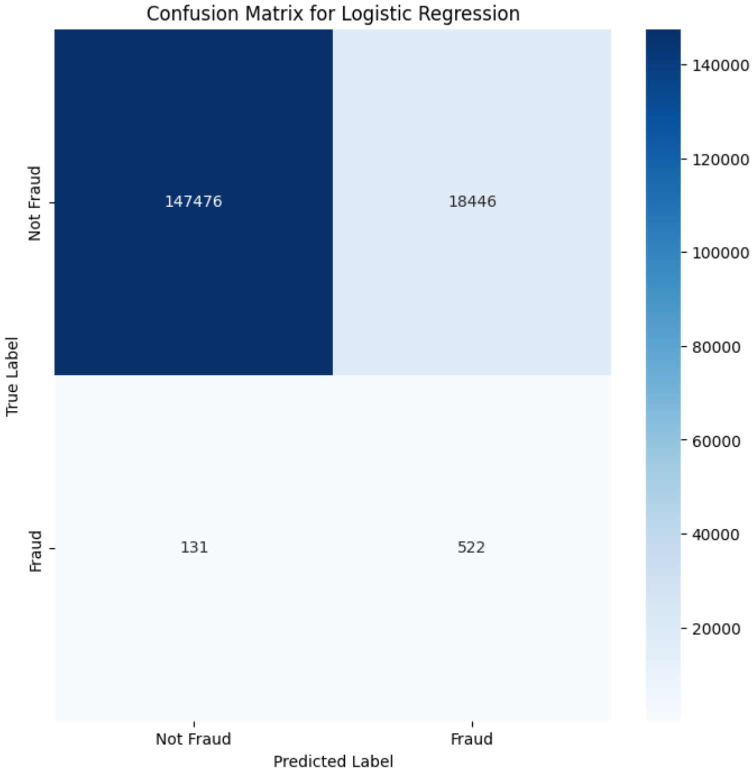


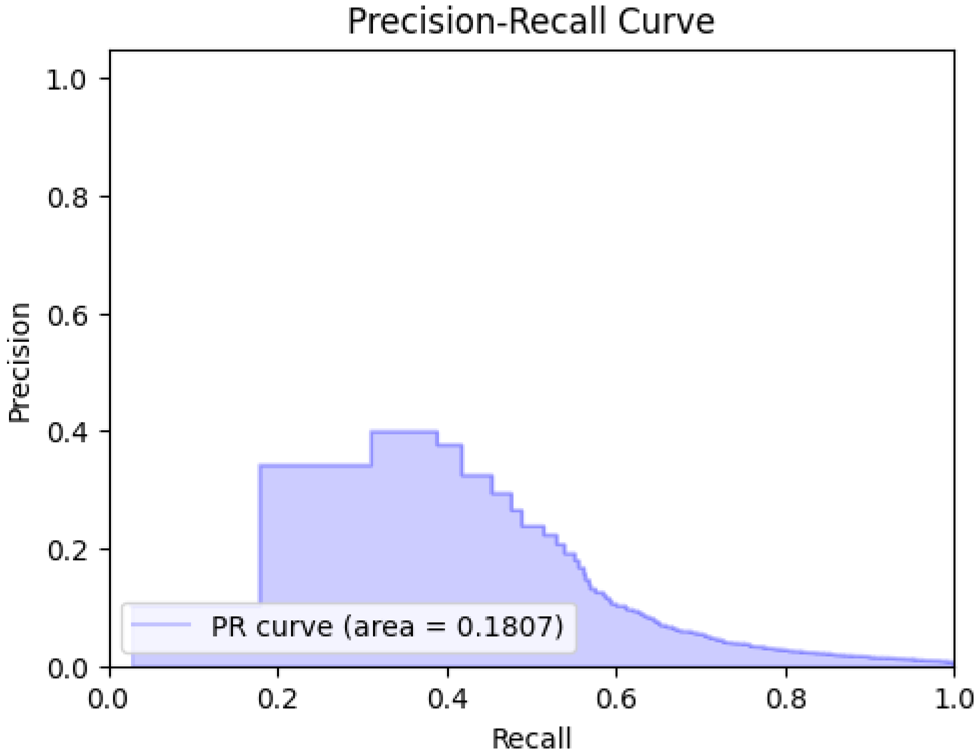
DECISION TREE



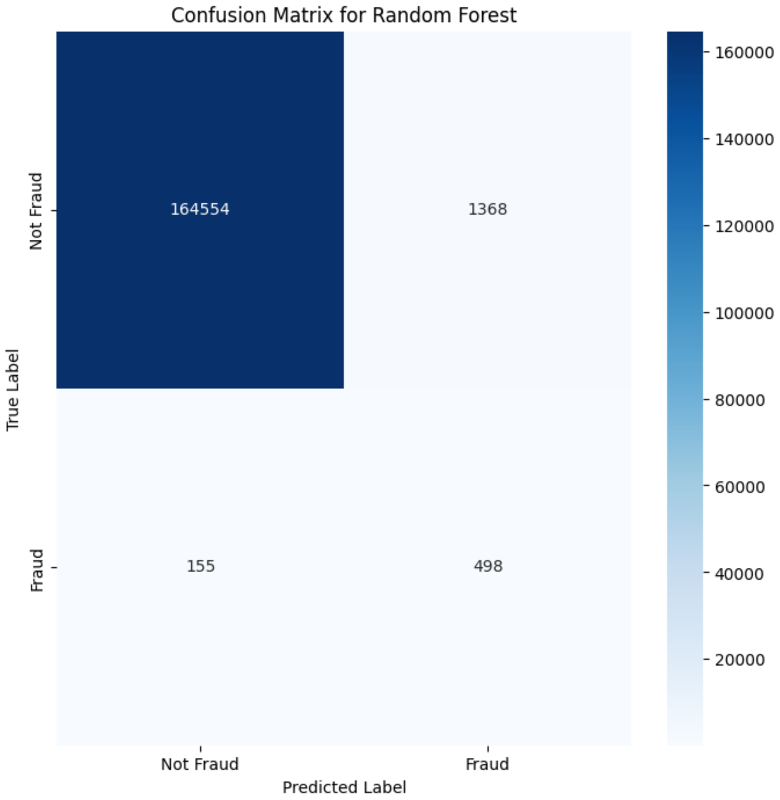
GRADIENT BOOSTING

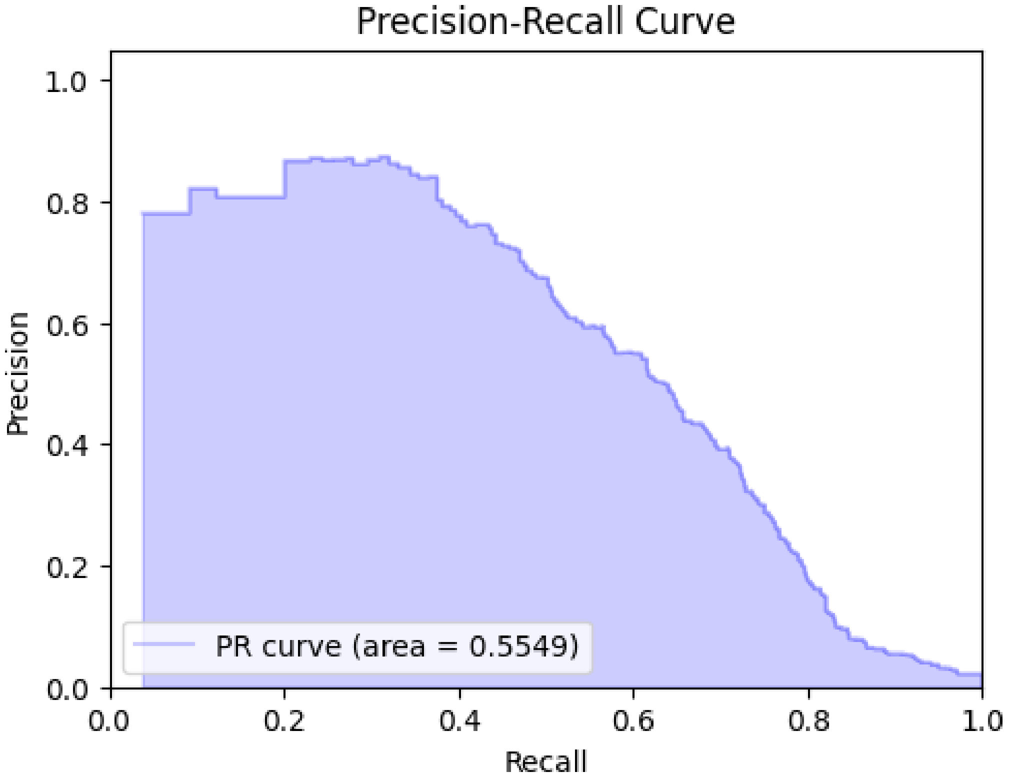
LOGISTIC REGRESSION(PR)



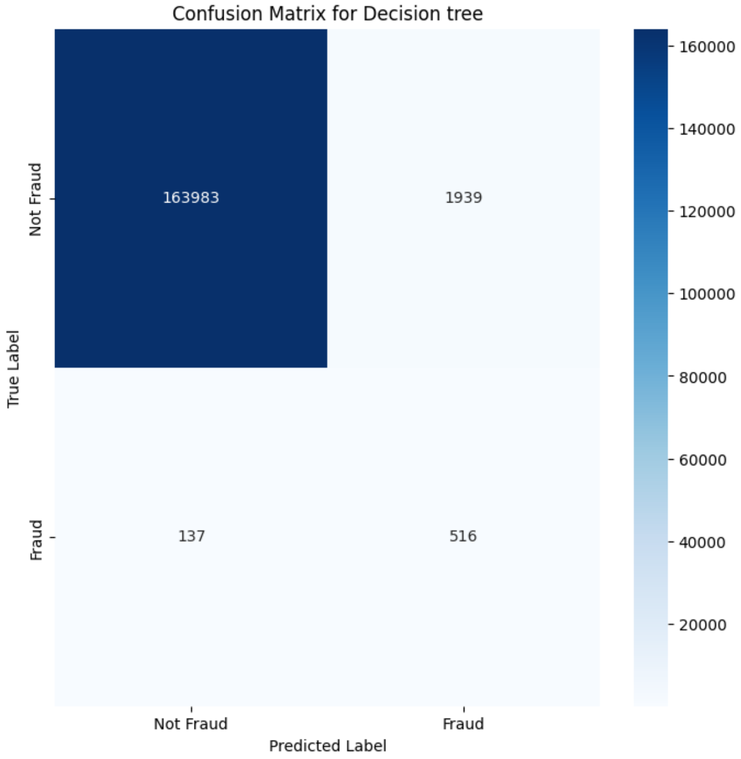


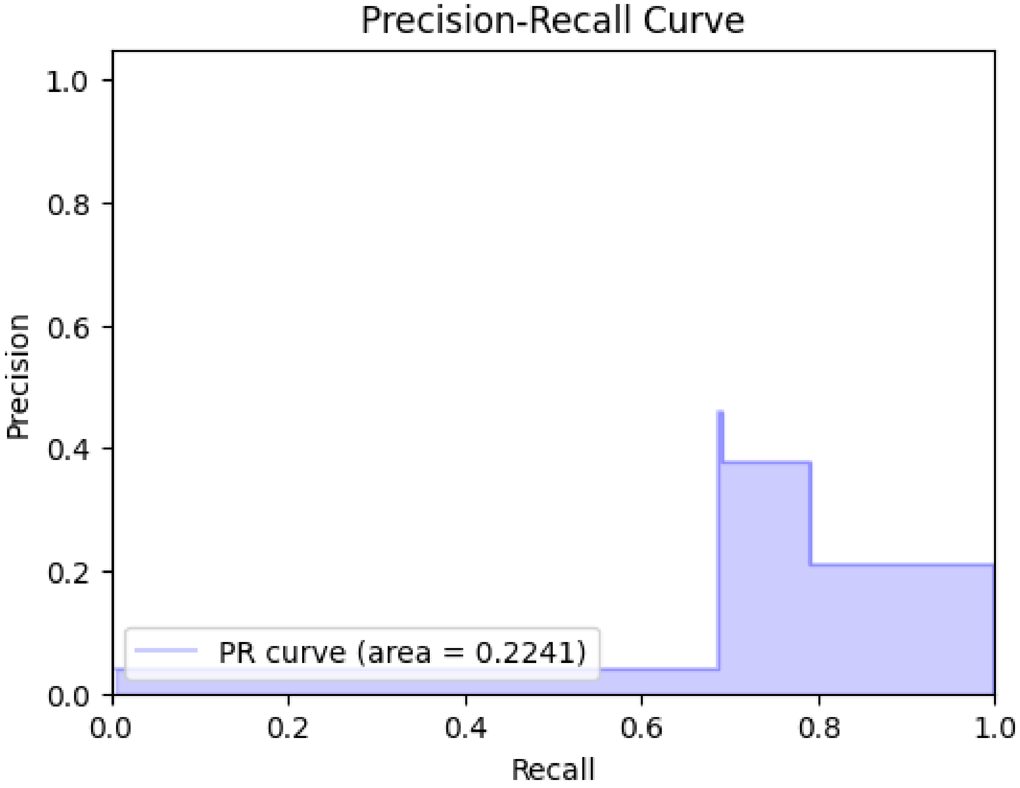
RANDOM FOREST (PR)



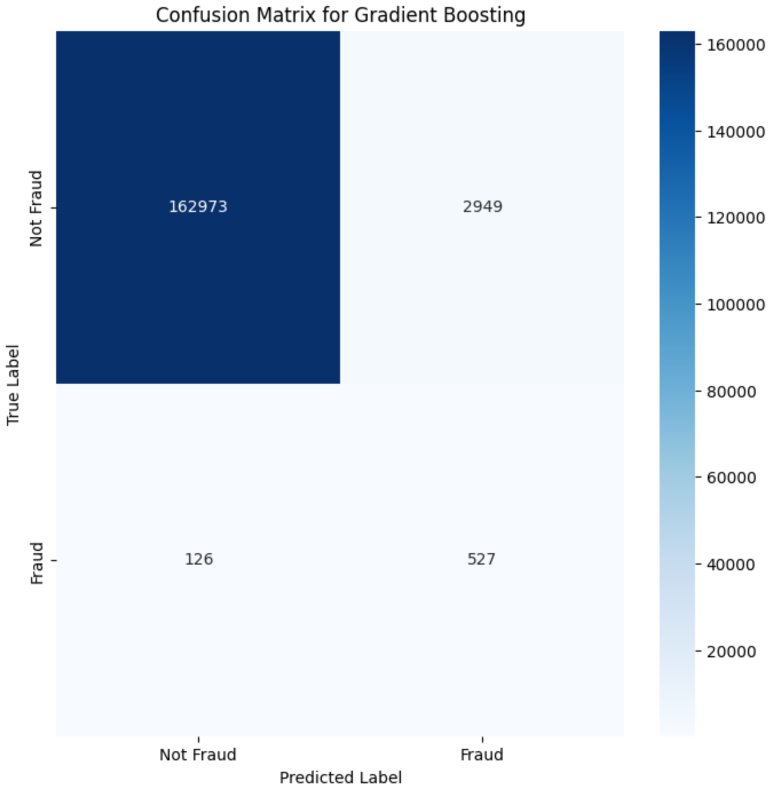


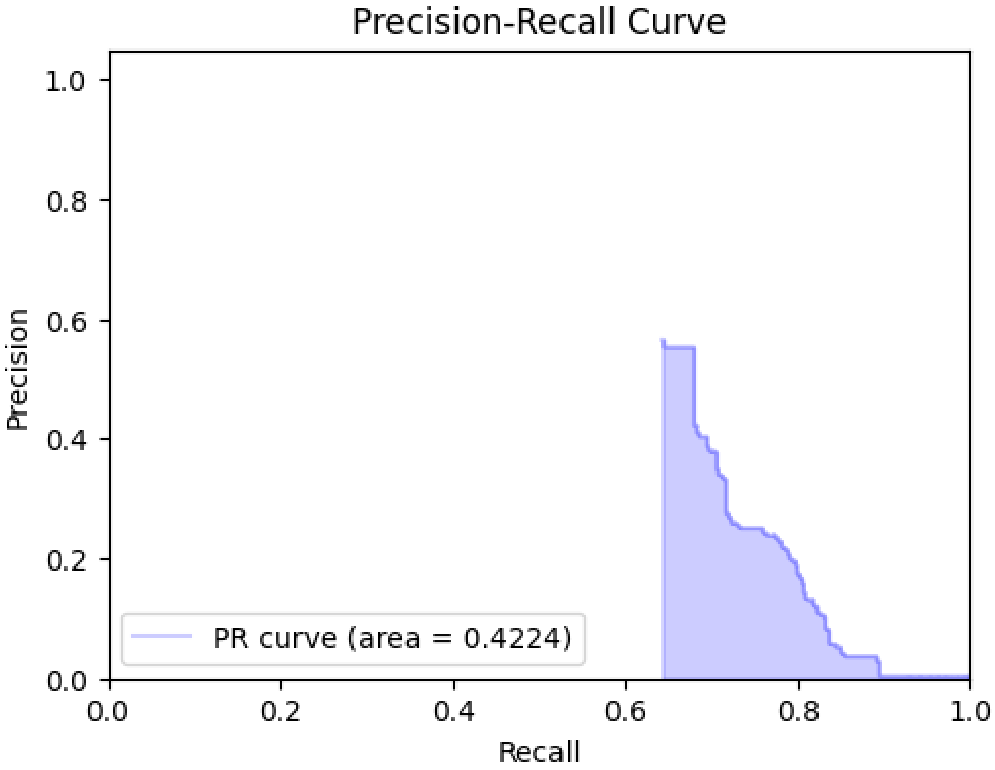
DECISION TREE (PR)



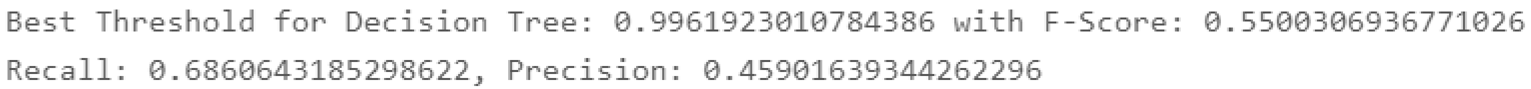


GRADIENT BOOST (PR)

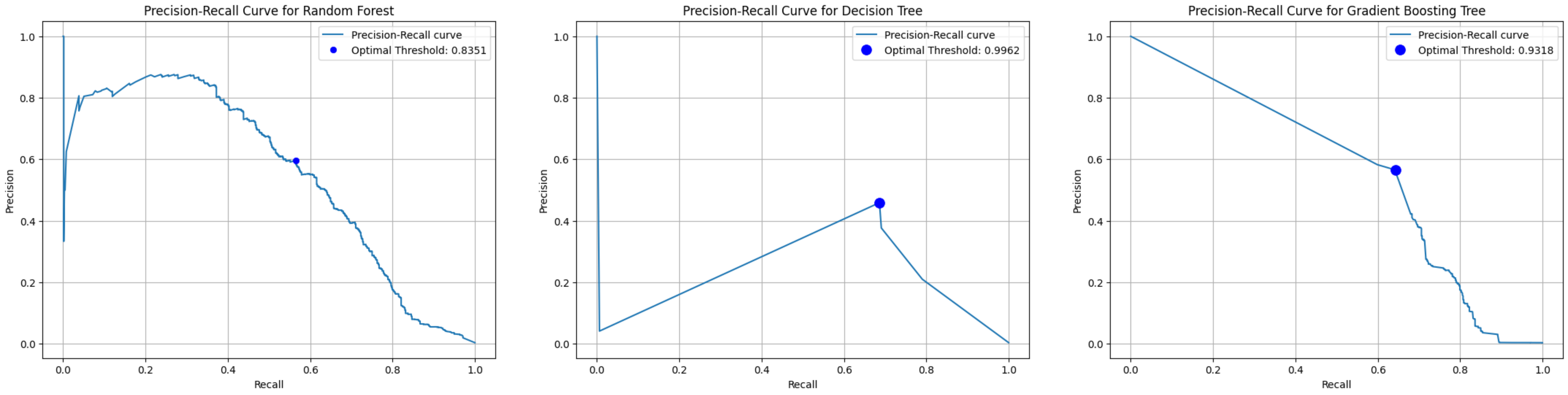


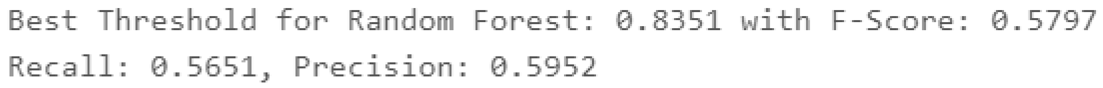


AREA UNDER PRECISION-RECALL

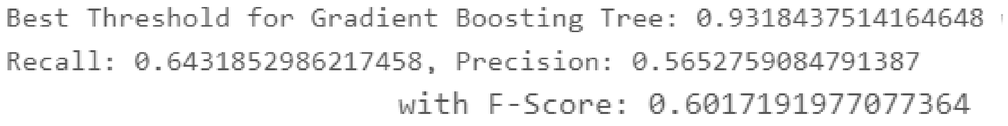


```
Best Threshold for Decision Tree: 0.9961923010784386 with F-Score: 0.5500306936771026
Recall: 0.6860643185298622, Precision: 0.45901639344262296
```



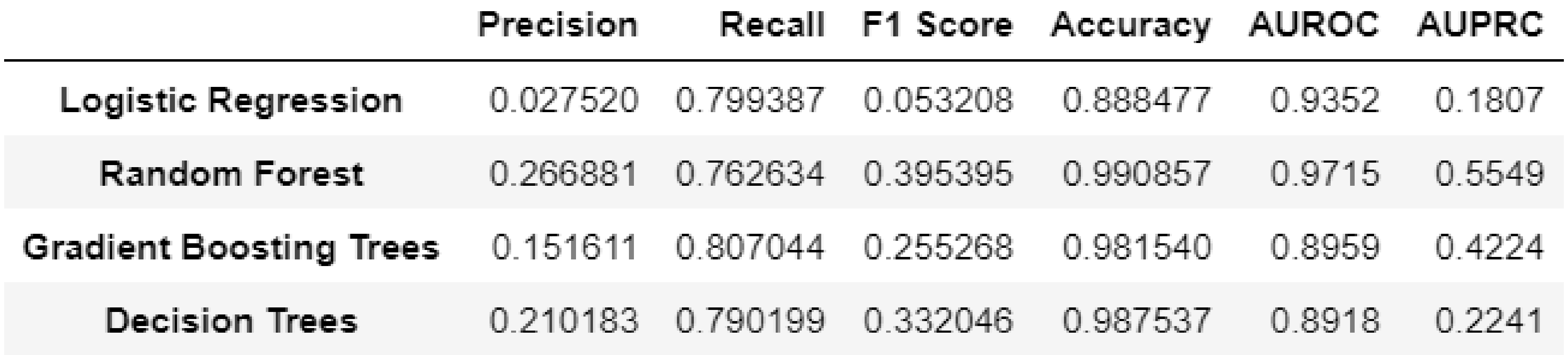


```
Best Threshold for Random Forest: 0.8351 with F-Score: 0.5797
Recall: 0.5651, Precision: 0.5952
```



```
Best Threshold for Gradient Boosting Tree: 0.9318437514164648
Recall: 0.6431852986217458, Precision: 0.5652759084791387
                         with F-Score: 0.6017191977077364
```

RESULT



	Precision	Recall	F1 Score	Accuracy	AUROC	AUPRC
Logistic Regression	0.027520	0.799387	0.053208	0.888477	0.9352	0.1807
Random Forest	0.266881	0.762634	0.395395	0.990857	0.9715	0.5549
Gradient Boosting Trees	0.151611	0.807044	0.255268	0.981540	0.8959	0.4224
Decision Trees	0.210183	0.790199	0.332046	0.987537	0.8918	0.2241

Why we got high accuracy:

Class Imbalance: Addressed through balancing techniques like oversampling.

Feature Relevance: Used feature engineering to boost model performance.

Overfitting/Underfitting: Applied regularization and cross-validation to avoid both.

Data Quality: Ensured data is clean and consistent.

Algorithm Limitations: Considered ensemble models for broader coverage.

Model Tuning: Optimized hyperparameters for better results.

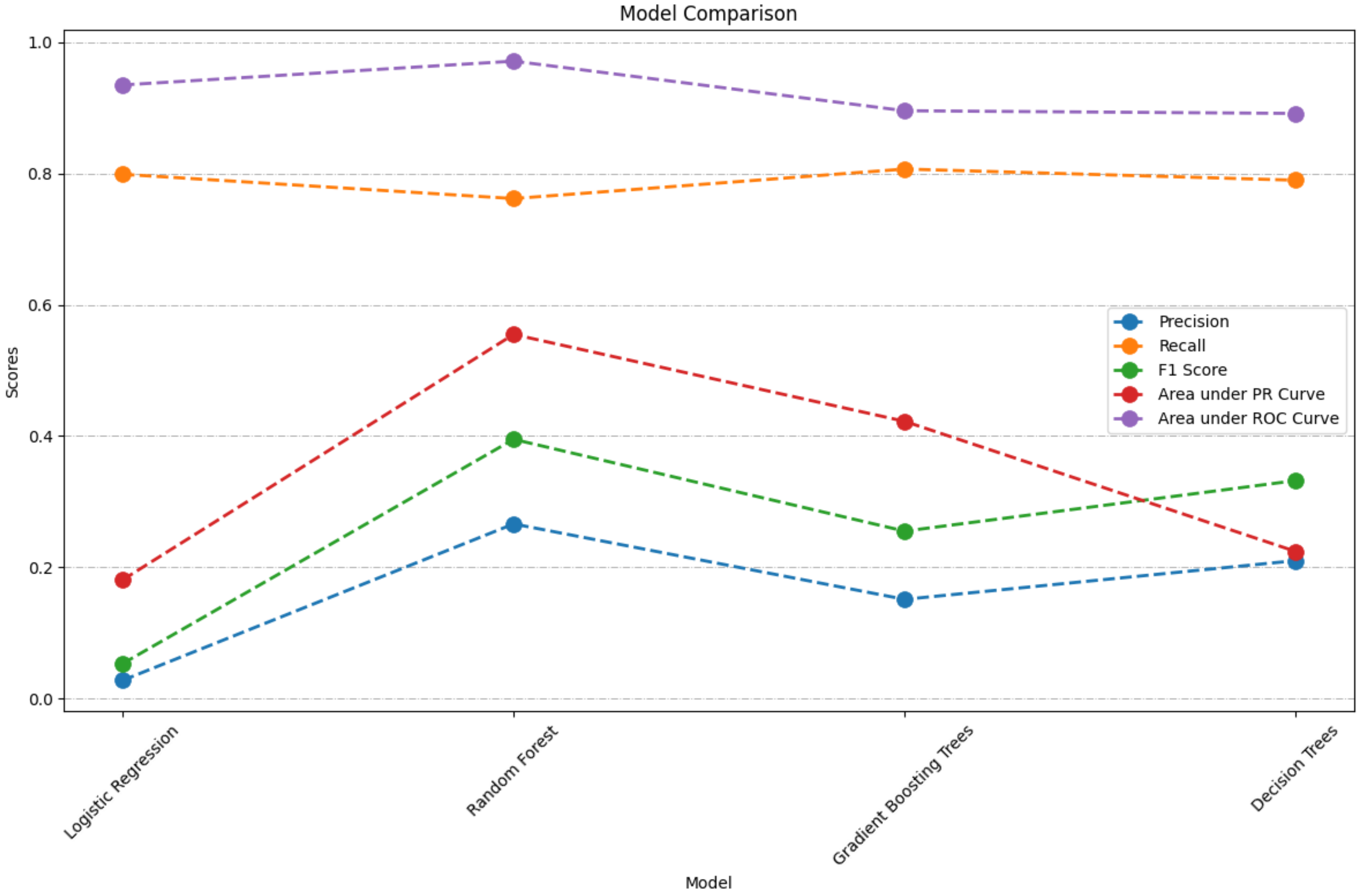
External Factors: Adapted to changes in trends and environment.

Way to Improve Low Precision:

Precision-Recall Balance: Adjusting the threshold can help to find a better balance between precision and recall. A higher threshold may decrease recall but increase precision, and vice versa. You can use precision-recall curves to find the optimal threshold for your specific use case.

Trade-off Tuning: By raising the threshold, you can make the model more conservative in predicting fraud; it will only predict fraud when it is more certain. This can reduce the number of false positives (i.e., transactions that are predicted as fraud but are actually legitimate), thereby increasing precision.

MODEL COMPARISON



Logistic Regression:

- High recall and AUROC, but low precision, F1 score, and AUPRC.

Random Forest:

- Good balance across metrics, with strong AUROC and accuracy.

Gradient Boosting Trees:

- High recall and accuracy, but lower precision and AUROC than Random Forest.

Decision Trees:

- Moderate scores across the board, with lower AUROC and AUPRC.

Each model has strengths and trade-offs, highlighting the importance of selecting a model based on key metrics.

CONCLUSIONS

- New York, Philadelphia & Texas Occupy top 3 positions where Fraud Transactions happened more in number.
- Fraudulent Transactions are recorded more with people working as Science Writer's.
- Fraudulent Transactions are mostly recorded in month of August
- Fraudulent Transactions are mostly recorded for online shopping when it comes to merchant category.
- The average Fraudulent transaction amount is above 500$.
- Females have faced more fraudulent transactions with their cards compared to males.

ACKNOWLEDGEMENT

We express our sincere gratitude to Dr. Liao Duoduo for her invaluable guidance and insightful feedback that significantly shaped our project. Her dedication to academic excellence greatly enhanced our experience. We are also grateful to our teaching assistant, Parth Manish Dalal, for his assistance and support throughout this project. Their combined efforts were crucial to our success.

REFERENCES

[1] A review of credit card fraud detection using Machine Learning techniques | IEEE conference publication | IEEE Xplore. (n.d.). https://ieeexplore.ieee.org/document/9365916

[2] real-time credit card fraud detection using streaming analytics. (n.d.-b). https://www.researchgate.net/publication/316530673_Realtime_credit_card_fraud_detection_using_Streaming_Analytics

DATASET SOURCE

[3] Credit Card Transactions Fraud Detection Dataset. (2020, August 5).

https://www.kaggle.com/datasets/kartik2112/fraud-detection/code