Card Fraud Detection

for E-Commerce Transactions

Lillian Lakes
FinTech Advi$ors Inc.

August 24, 2023

LEAD INVESTIGATOR

LILLIAN LAKES

linkedin.com/in/lillianlakes

github.com/lillianlakes

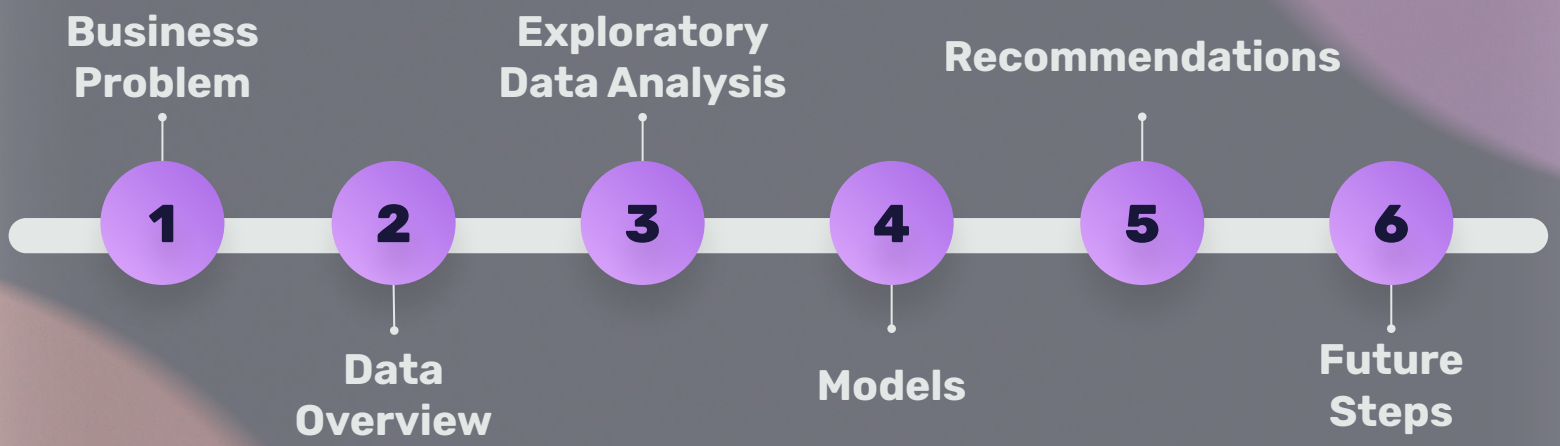
Experience

- **Software Engineering**, Treasure and AppFolio
- **Data Analysis**, MIT and US Census Bureau
- **Corporate Finance and Investment Banking**, Delta Airlines and Barclays Capital

Education

- **M.A. in Management Research**, MIT
- **M.B.A. in Finance & Strategy,** Emory University
- **B.A. in Economic & M.I.S.**, Chatham University

TABLE OF CONTENTS

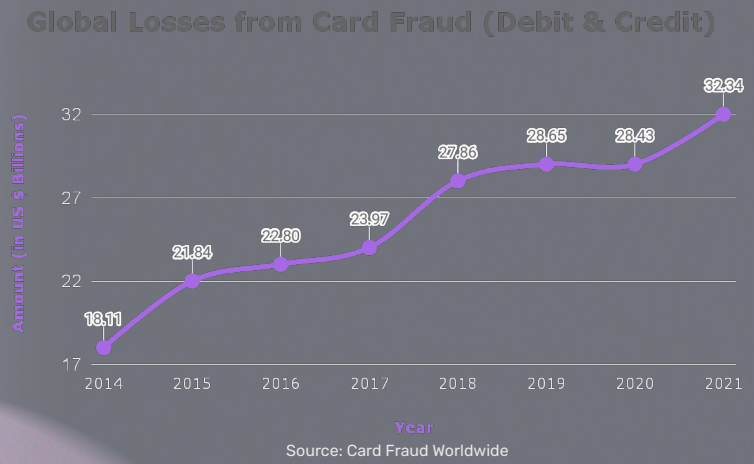


01. Business Problem

GOOD USERS INCORRECTLY FLAGGED

FRAUDULENT USERS NOT STOPPED

Facts and Figures



Source: Card Fraud Worldwide

$397.4 Billion

in global losses over next 10 years

Source: Nilson Report (2022)

In 2021, the FTC fielded

390,000 reports

of credit card fraud

2nd most common

form of identity theft

after government benefits & documents (FTC)

87% of US consumers use public WiFi for online shopping (Symantec)

BUSINESS PROBLEM

Accurately detect credit card fraud by minimizing the cost of:

- False positives: **Good users transactions interrupted**
  - Avoid upset customers and resulting churn

- False negatives: **Fraudulent transactions not flagged**
  - Avoid financial losses to clients

Identify variables that lead to fraud

STOP SCAMMERS

BOTTOM LINE

Quick and Same-Device E-Commerce Card Transactions are More Likely to be Fraudulent

- Made within two minutes of signup
- Multiple purchases from the same device

January, following the holiday shopping period, is higher risk

02. Data Overview

DATA OVERVIEW

- **Fraud Data**
  - 151,112 entries & 11 variables
  - Fraud Classification, Device ID, IP Address, Signup Time, Purchase Time & Purchase Value
- **IP Table: IP Address <-> Country**
  - 138,846 entries & 3 variables
  - Country
  - IP Address Upper & Lower Bounds

DATA LIMITATIONS

- **Real vs Synthetic Datasets and Data Transparency**
  - **Heavily anonymized features if real**
  - **Synthetic**
- **Additional Features**
  - **Card type (credit/debit)**
  - **Product SKU**
  - **Connection method (VPN, proxy, Tor usage)**
  - **Device type**
- **Class Imbalance**
  - **Only 9.36% fraudulent classification**

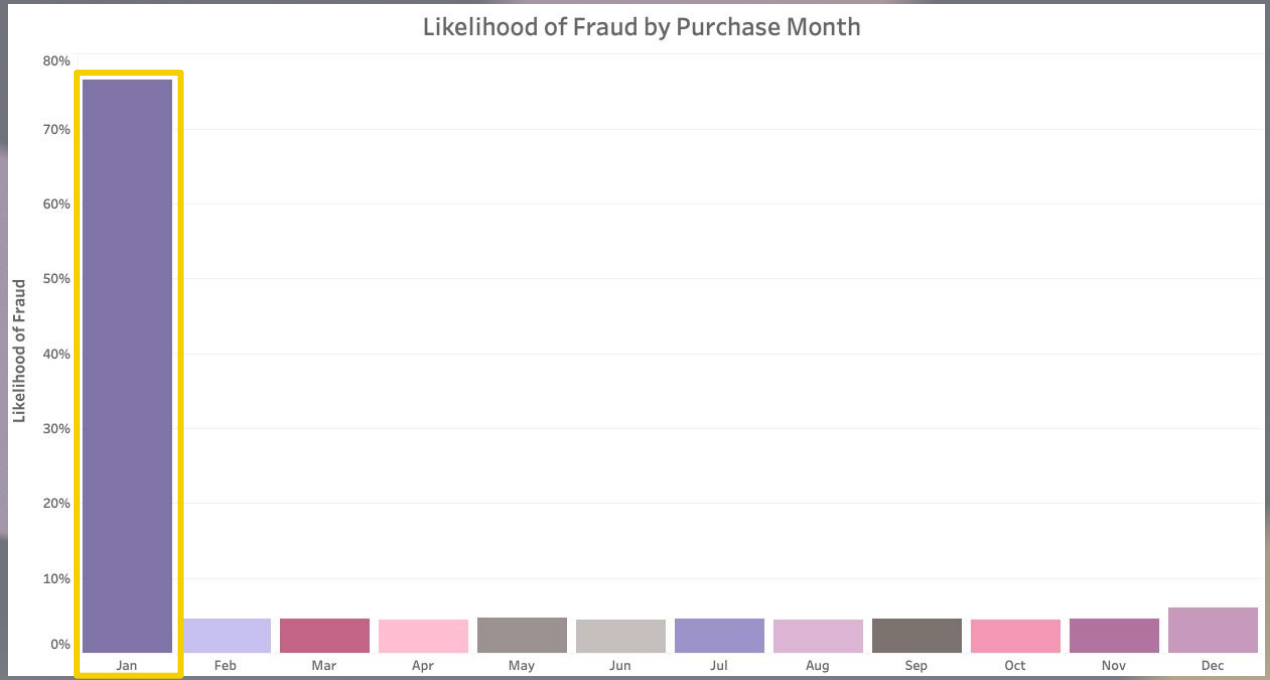
FEATURE ENGINEERING

- **Mapping IP Addresses to Country**
  - **Relevant Countries**
  - **Countries from Device**
- **Purchase Month, Weekday, Period of Day**
- **Device Frequency**
- **Quick Purchase**
  - **Less than 137 seconds since signup**

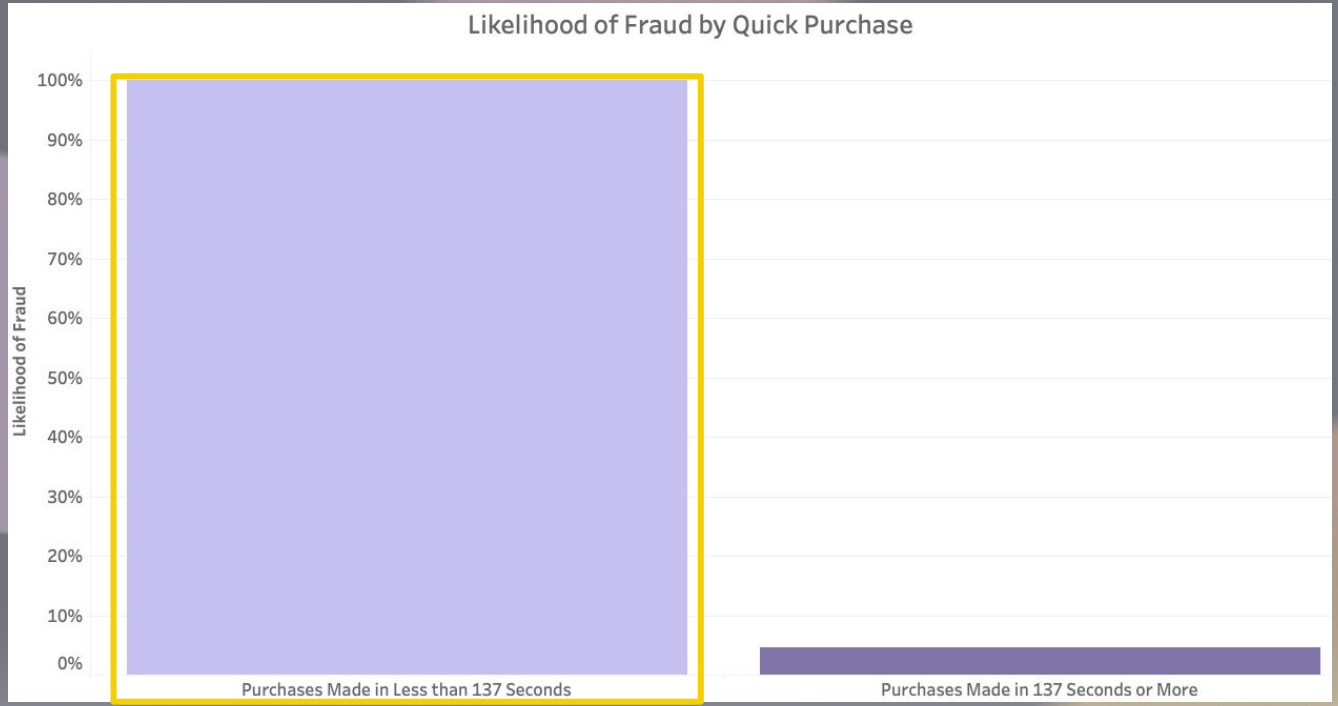
03.

Exploratory Data Analysis

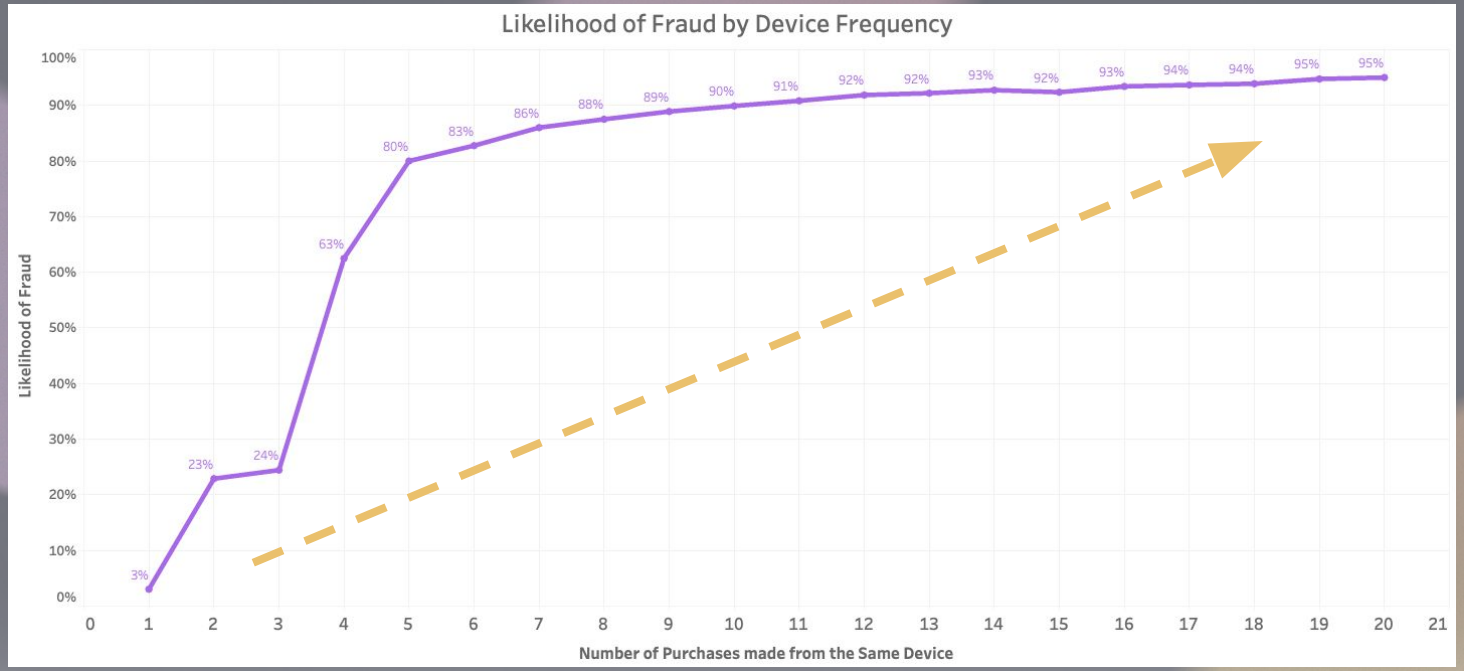
January, following the holiday shopping rush, has the highest likelihood of fraud



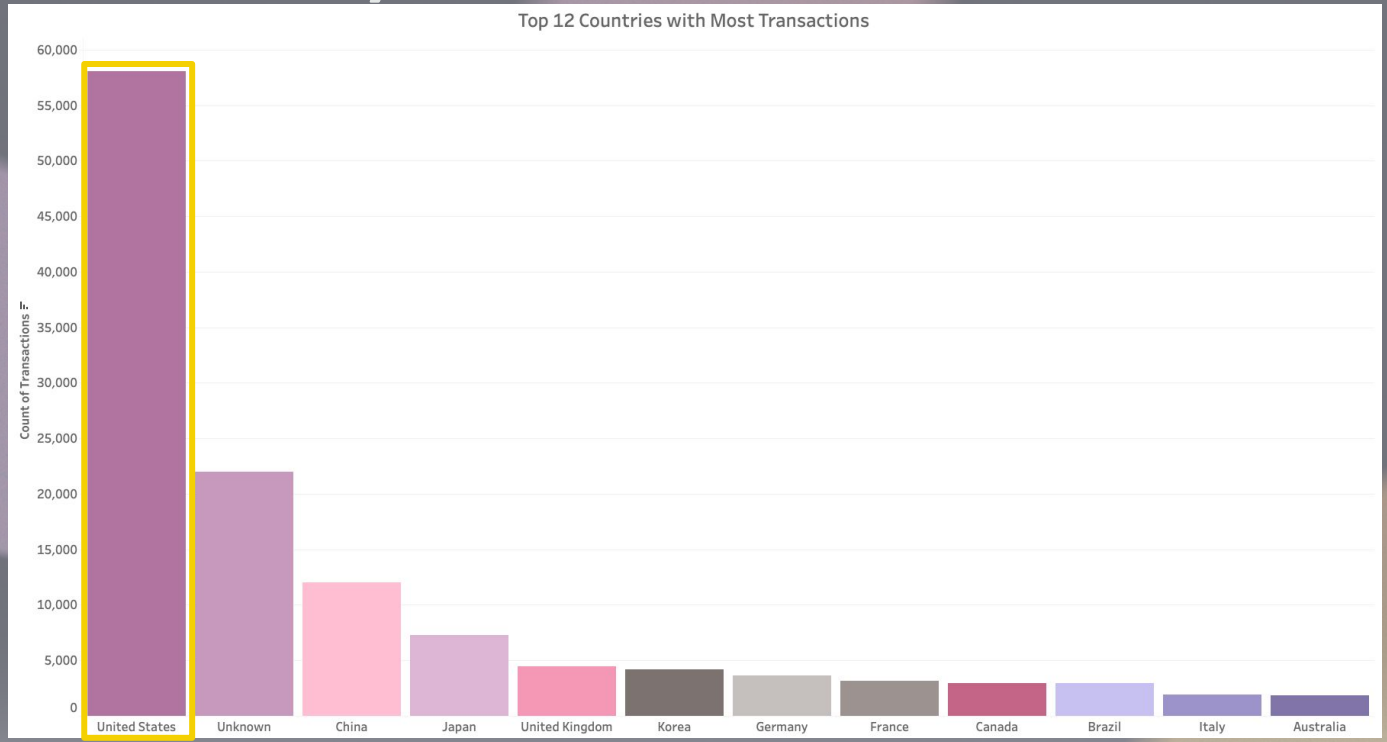
In the data, purchases made less than 137 seconds after signup are all fraudulent



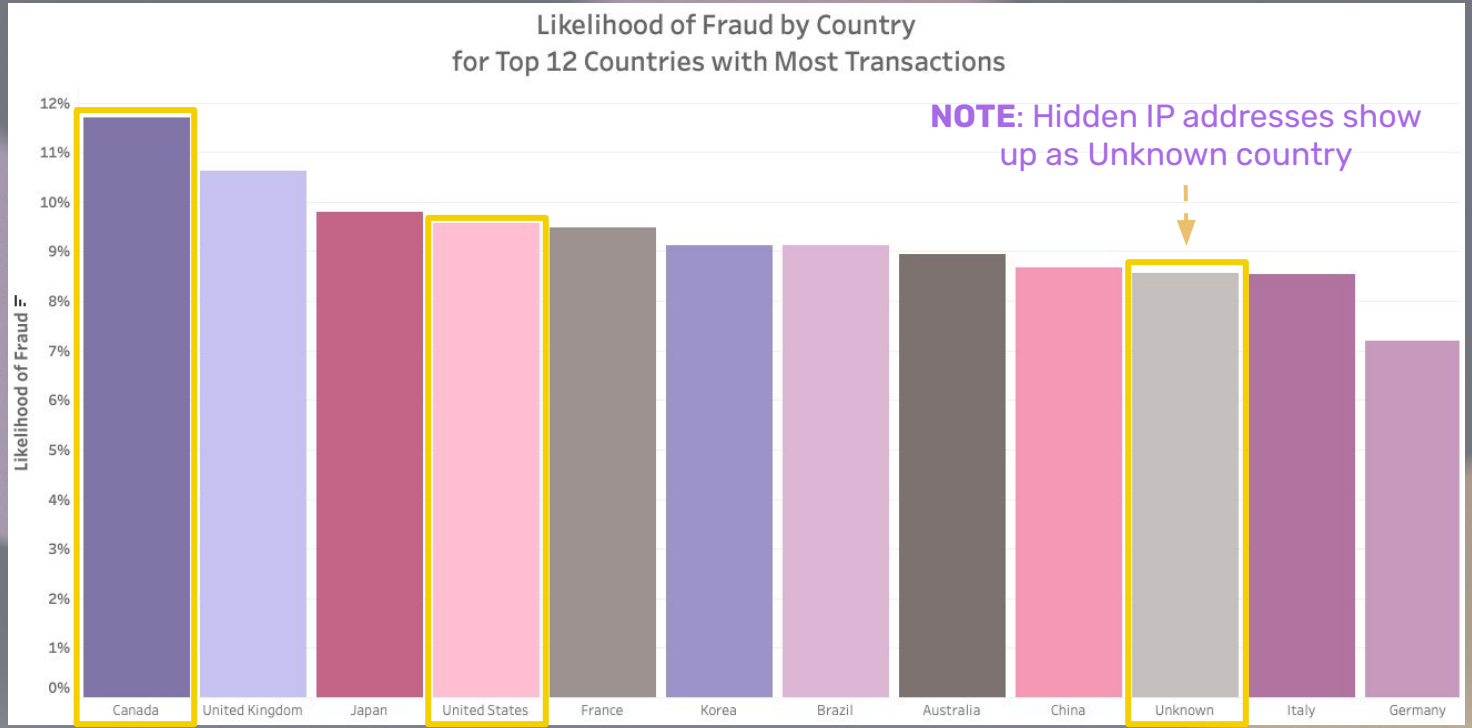
Fraud is more likely to occur when purchases are made from the same device



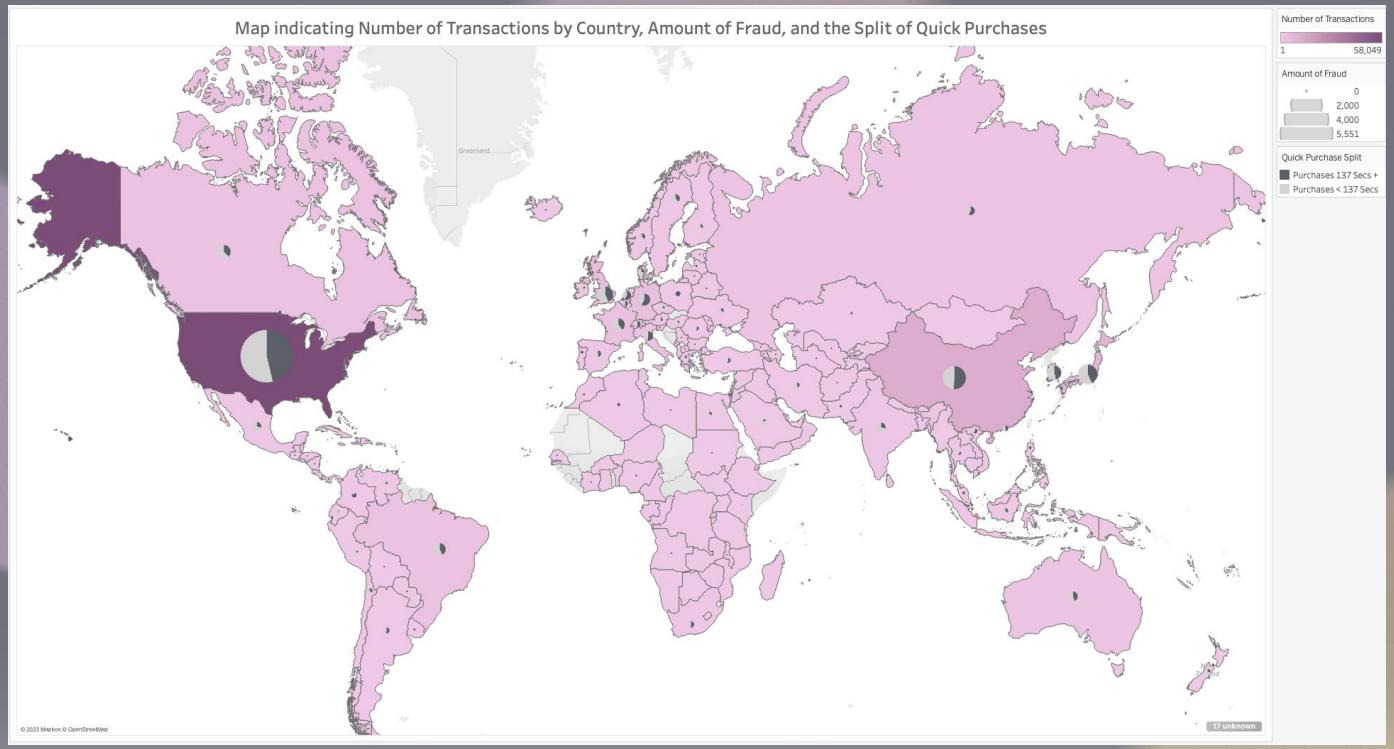
IP addresses from the United States are most represented, followed by IP addresses that are not trackable



Of the top 12, Canada has the most fraudulent transactions & the United States has high fraud levels



The United States has the most transactions, overall and fraudulent, with a high proportion of quick purchases



04.

Models

Logistic and Lasso models have the best Accuracy Score and Interpretability

Model	Accuracy Score	Interpretability
Dummy Classifier	90.6%	
Decision Tree Classifier	9.8%	
Logistic Regression	**92.0%**	**Best**
Logistic Regression w/ Lasso Regularization	**91.6%**	**Best**
Mixed Naive Bayes Classifier	92.2%	
Gaussian Naive Bayes Classifier	9.4%	
Random Forest Classifier	45.2%	
Gradient Boosting Classifier	15.9%	
Adaptive Boosting Classifier	10.1%	

LOGISTIC REGRESSION MODEL

- **Accurately predicts 92.0%**
  of e-commerce card transactions
- Purchases made within the
  1st 137 seconds of signup 110 X
  more likely to be fraudulent
- **Device used 1 additional time is 8 X**
  more likely to make a fraudulent purchase

LASSO REGRESSION MODEL

- **Accurately predicts 91.6%**
  of e-commerce card transactions
- Purchases made within the
  1st 137 seconds of signup 39 X
  more likely to be fraudulent
- **Device used 1 additional time is 29 X**
  more likely to make a fraudulent purchase

05.

Recommendations

RECOMMENDATIONS

- **Fraud Probability Scoring Model that heavily weighs Quick Purchases and Device Frequency**
  - Low Probability -> Proceed with transaction
  - Medium Probability -> 2-Factor Authentication
  - High Probability -> Call Fraud Prevention Line
- **Higher Fraud and Customer Service Staffing in January**

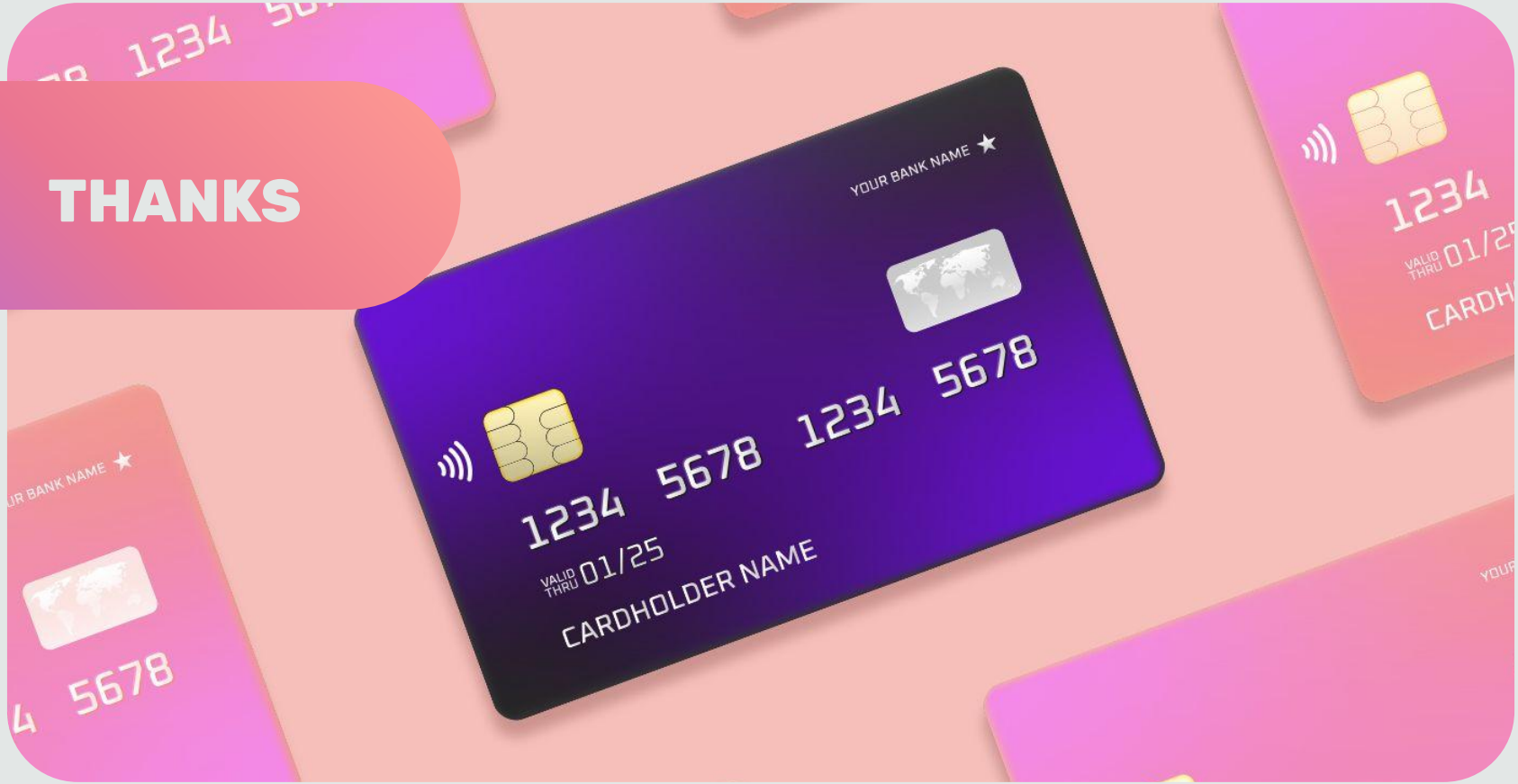
06.

Future Steps

FUTURE STEPS

- **Time-Series Analysis** to spot fraud trends over time
- **Find Publicly Available Real Datasets** with mostly transparent data
- **Detect other forms of Identity Theft e.g. Government Benefits and Documents**

THANKS


YOUR BANK NAME
1234 5678 1234 5678
VALID THRU 01/25
CARDHOLDER NAME
1234
5678

QUESTIONS?

LILLIAN LAKES

linkedin.com/in/lillianlakes

github.com/lillianlakes

07.

Appendix

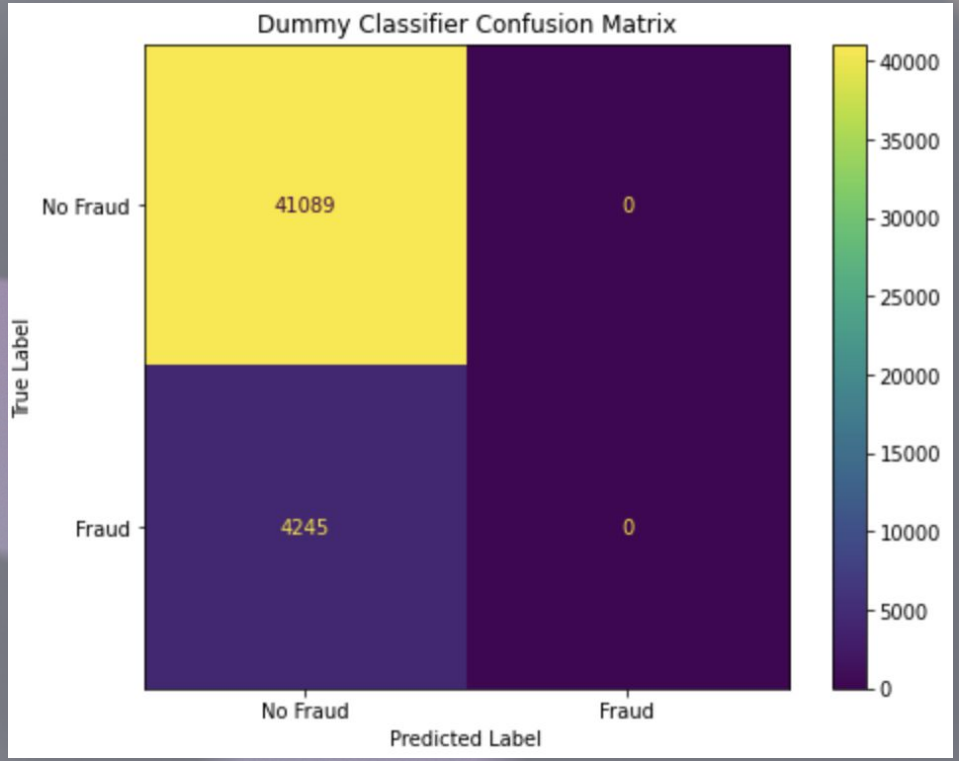
SOURCES

Fraud e-commerce transaction dataset, Kaggle, https://www.kaggle.com/datasets/vbinh002/fraud-ecommerce

Credit card fraud statistics, Bankrata.com, https://www.bankrate.com/finance/credit-cards/credit-card-fraud-statistics/

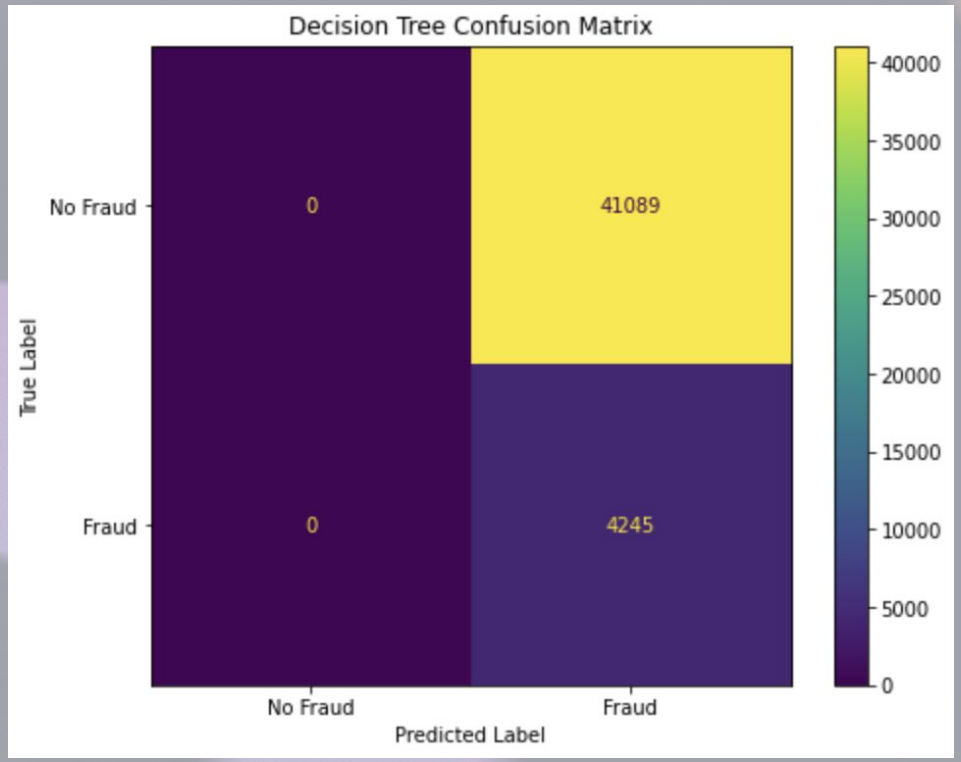
Consumer Sentinel Data Book, FTC, https://www.ftc.gov/system/files/ftc_gov/pdf/CSN%20Annual%20Data%20Book%202021%20Final%20PDF.pdf

Most people unaware of the risks of using public Wi-Fi, CNBC, https://www.cnbc.com/2016/06/28/most-people-unaware-of-the-risks-of-using-public-wi-fi.html



MODEL 1

Dummy Classifier Accuracy 90.6%

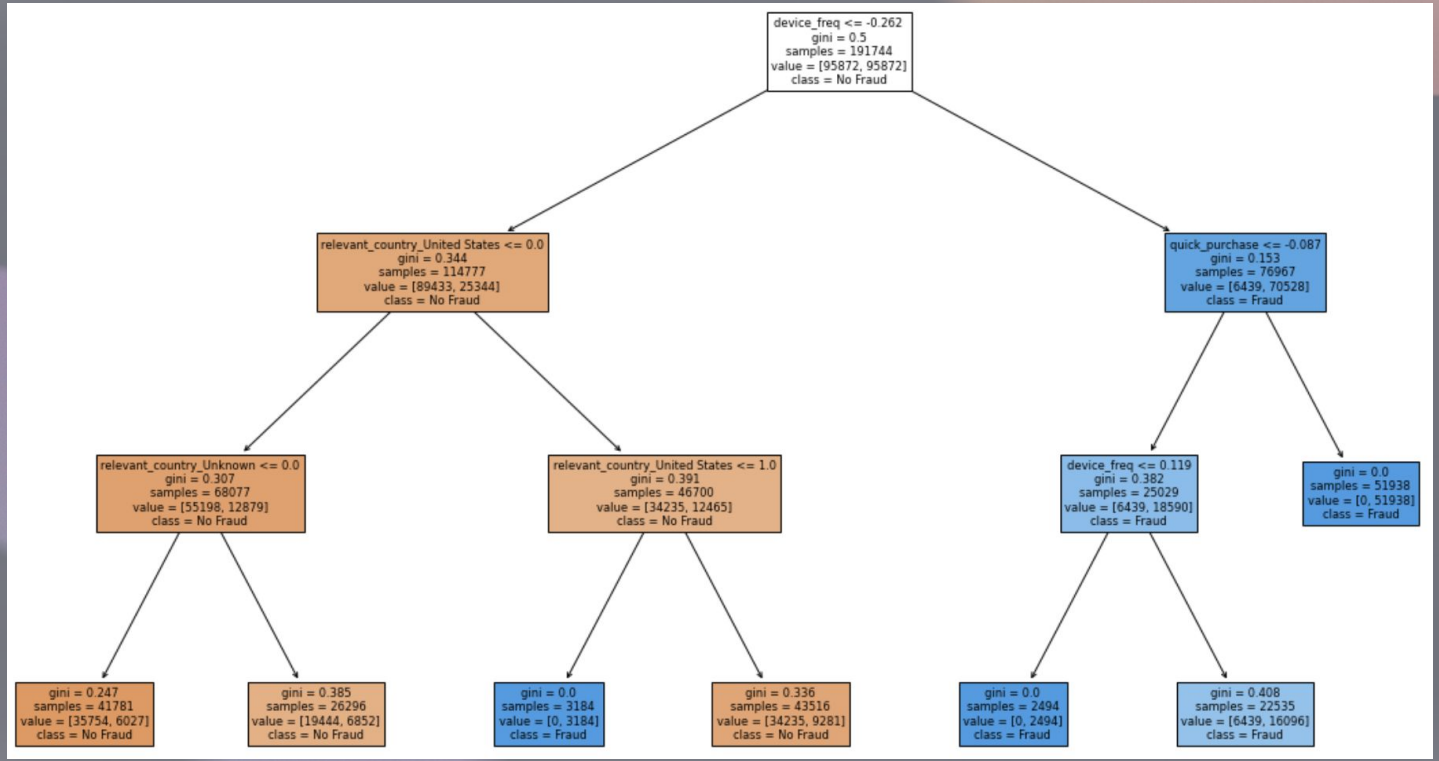


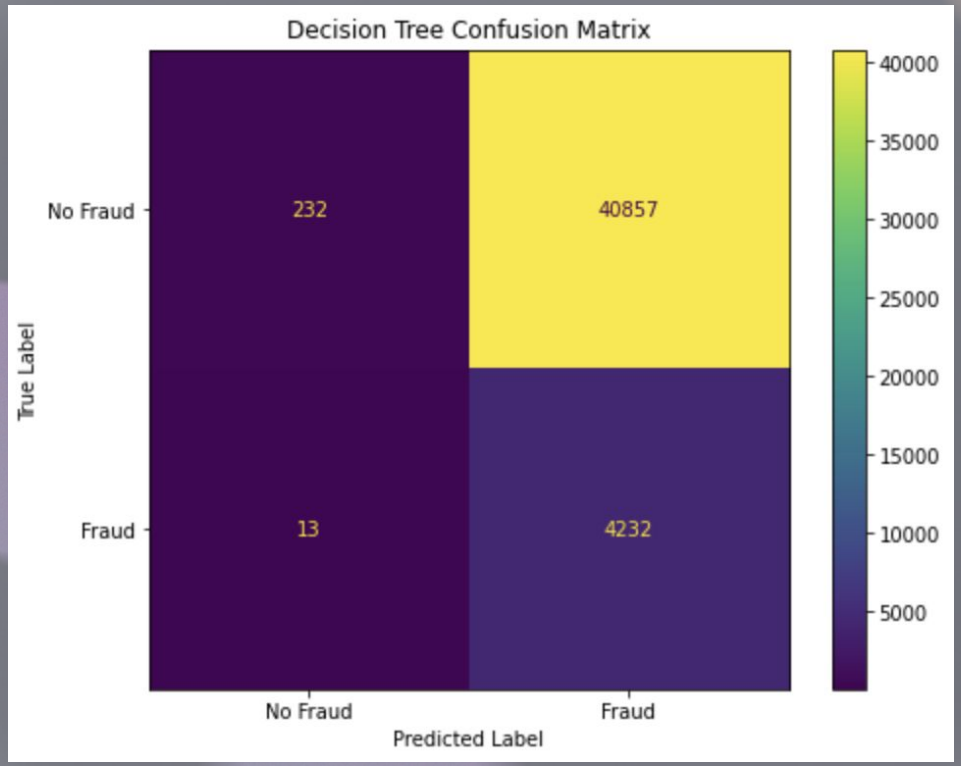
MODEL 2A

Decision Tree Classifier (Max Depth 3) Accuracy

9.4%

MODEL 2A : Decision Tree Classifier



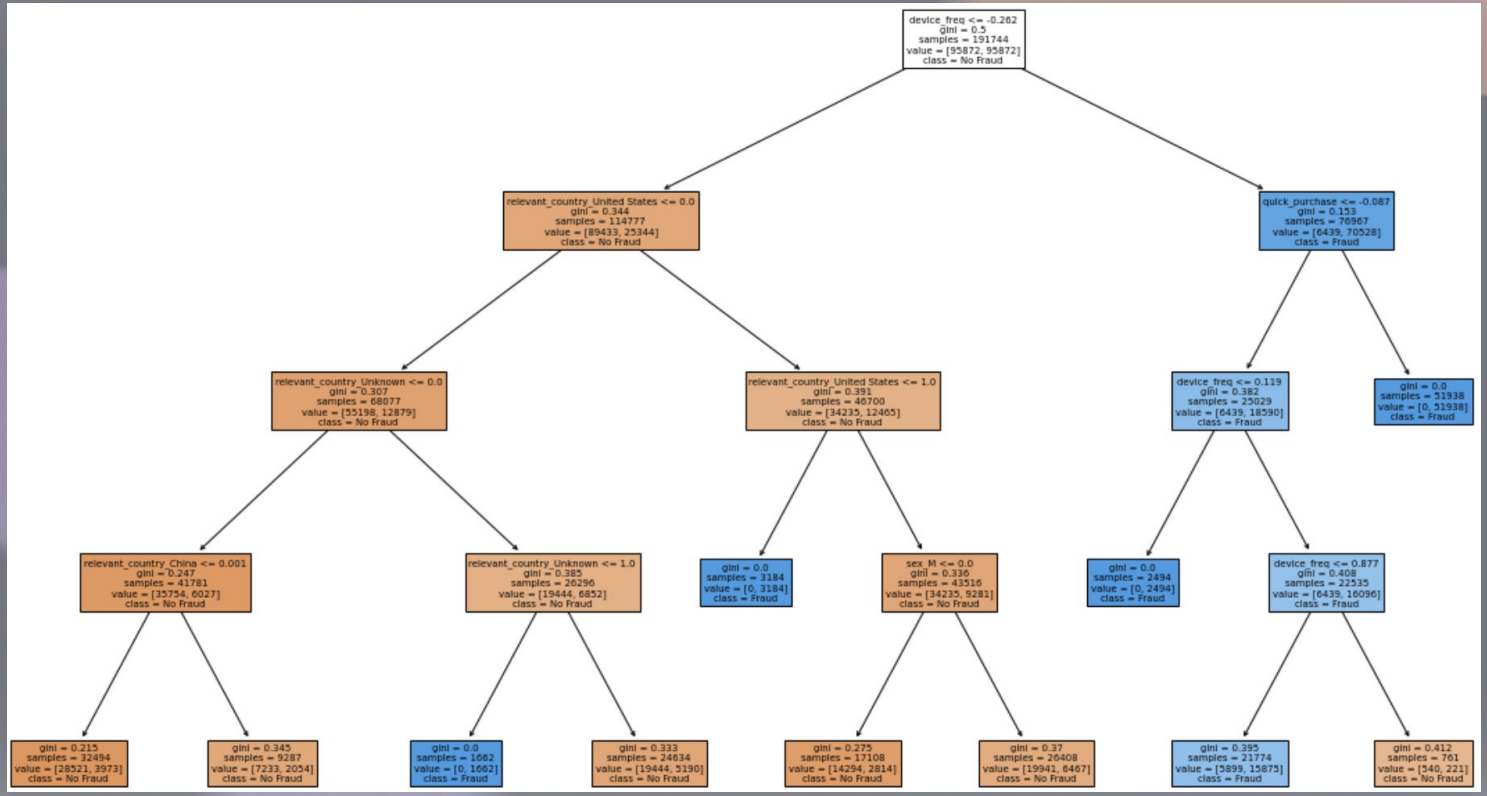


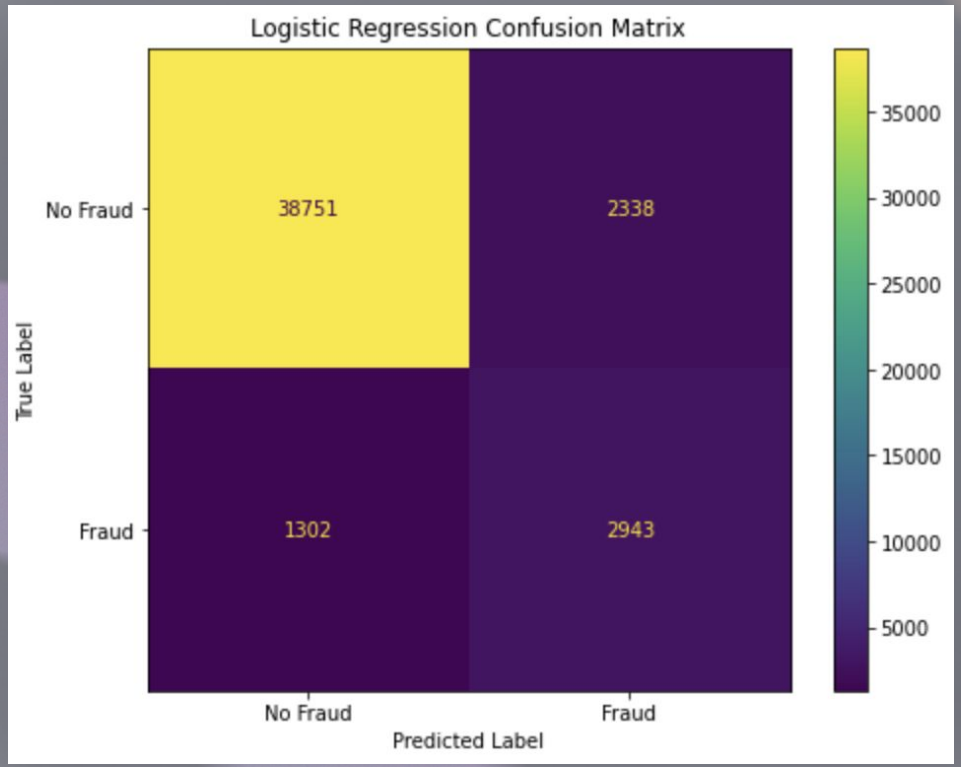
MODEL 2B

Decision Tree Classifier (Max Depth 4) Accuracy

9.8%

MODEL 2B : Decision Tree Classifier

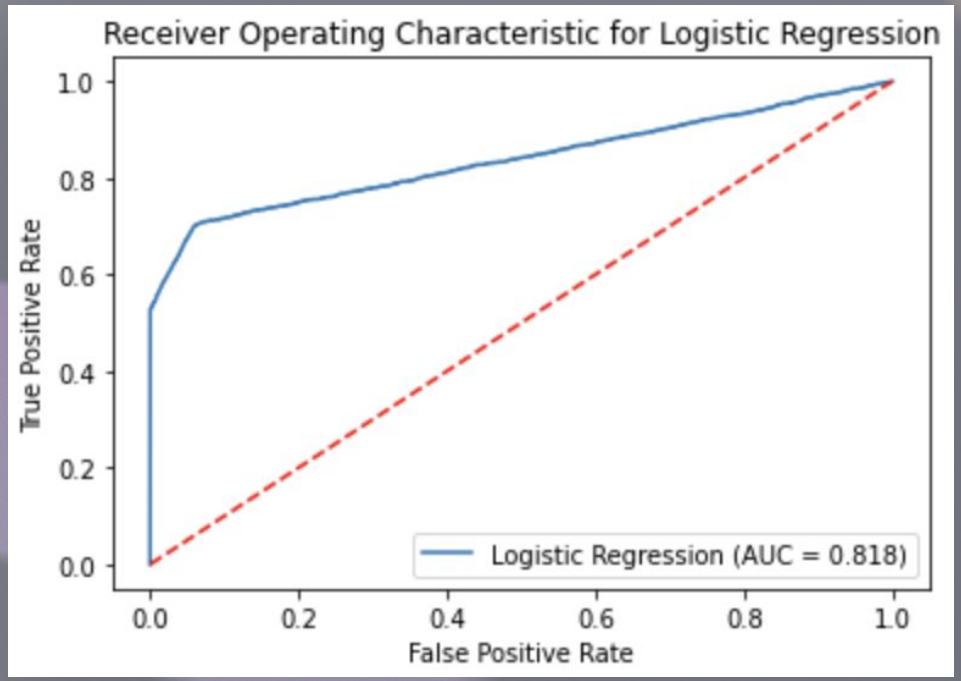




MODEL 3

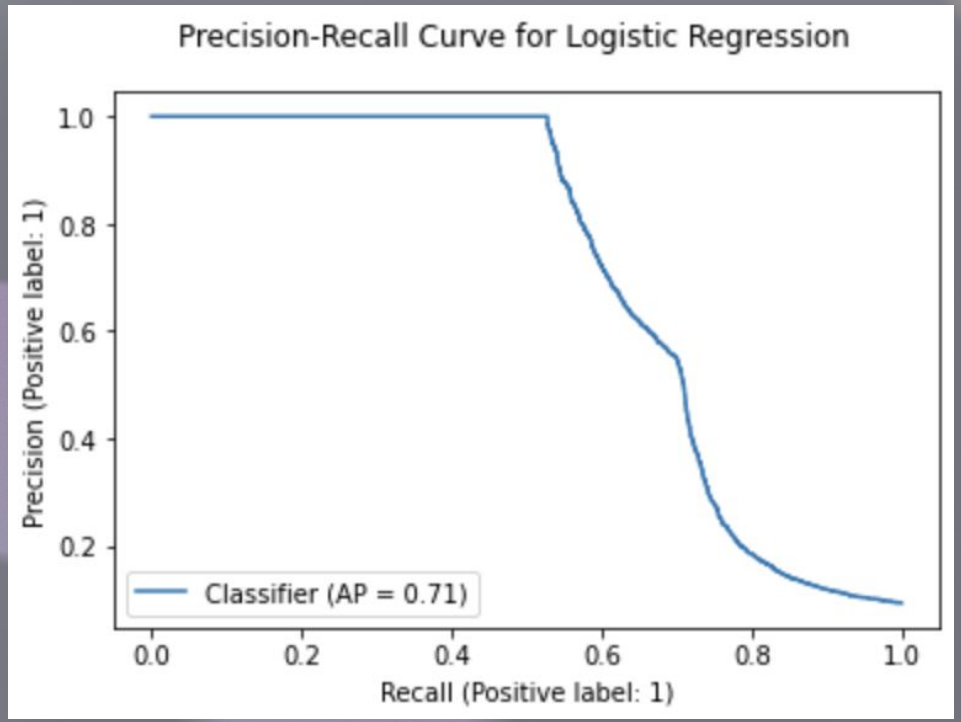
Logistic Regression Model Accuracy

92.0%



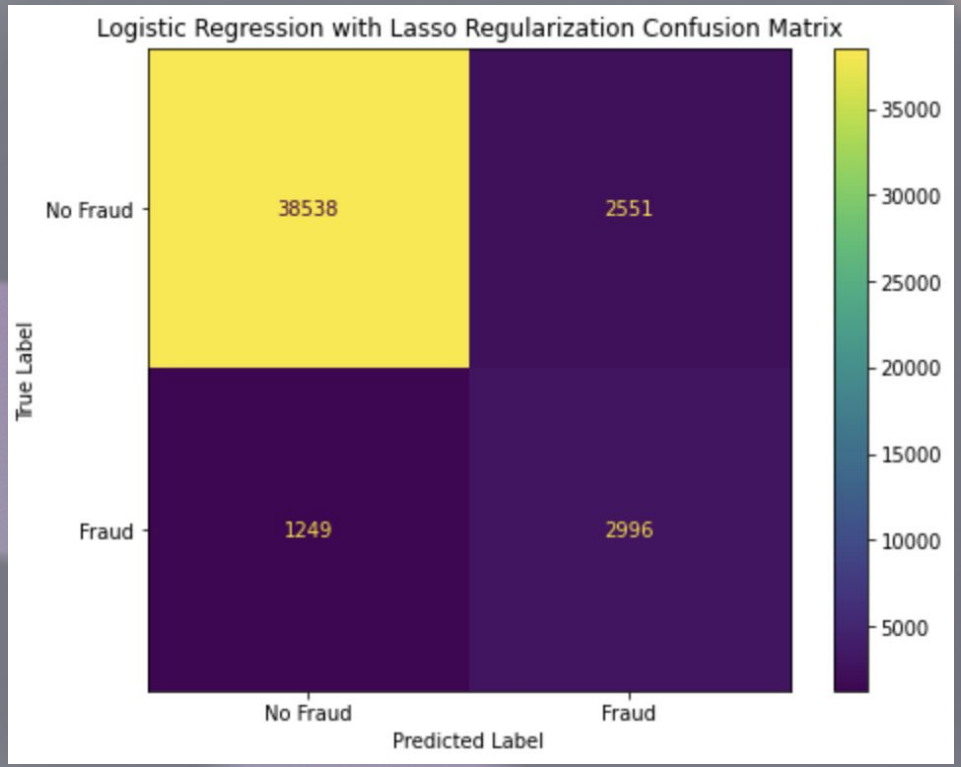
MODEL 3

Logistic Regression Model AUC 81.8%



MODEL 3

Logistic Regression Model AUC-PR 71%

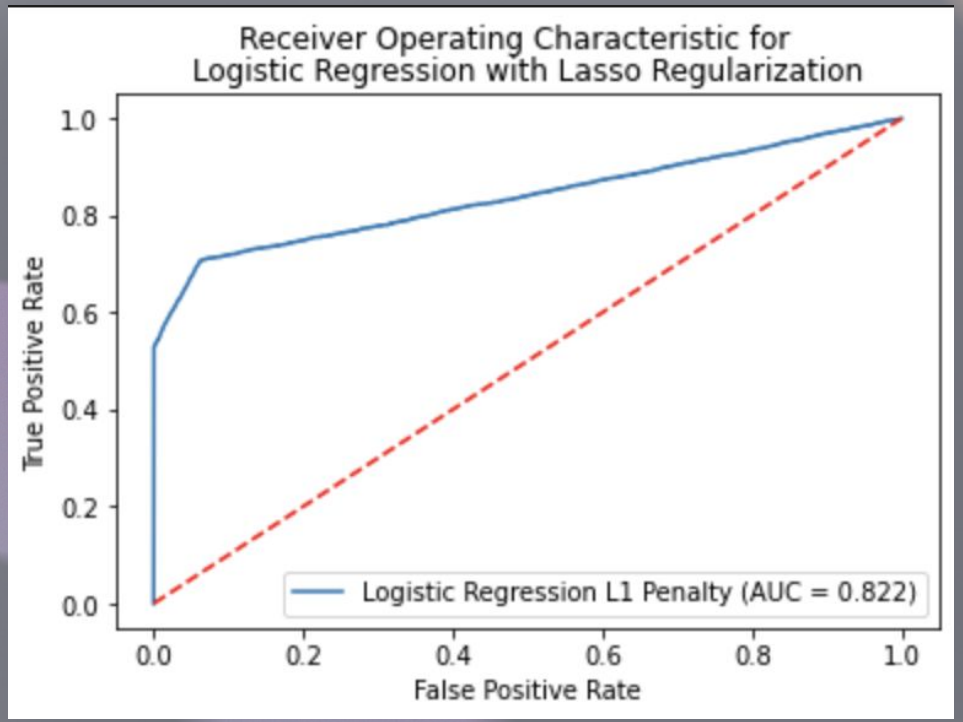


MODEL 4

Logistic Regression with Lasso Regularization Model Accuracy

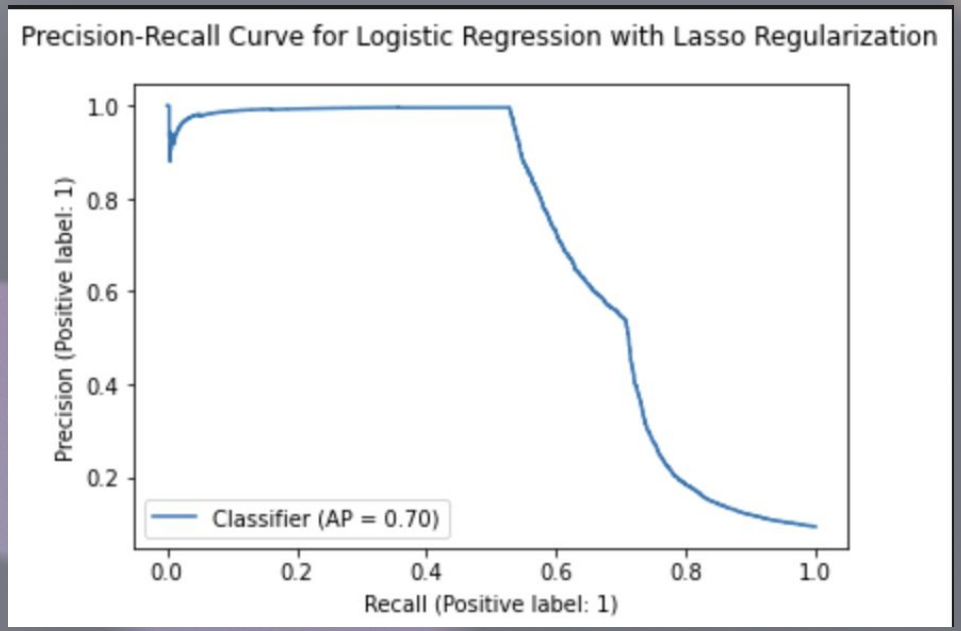
91.6%

MODEL 4



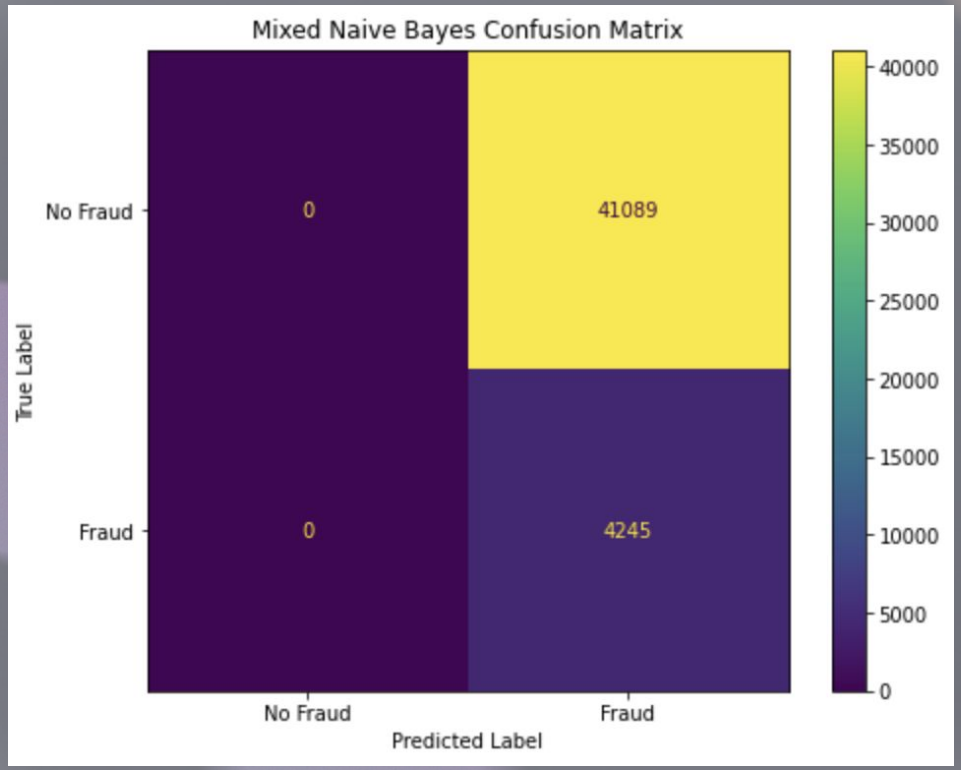
Logistic Regression with Lasso Regularization Model AUC

82.2%



MODEL 4

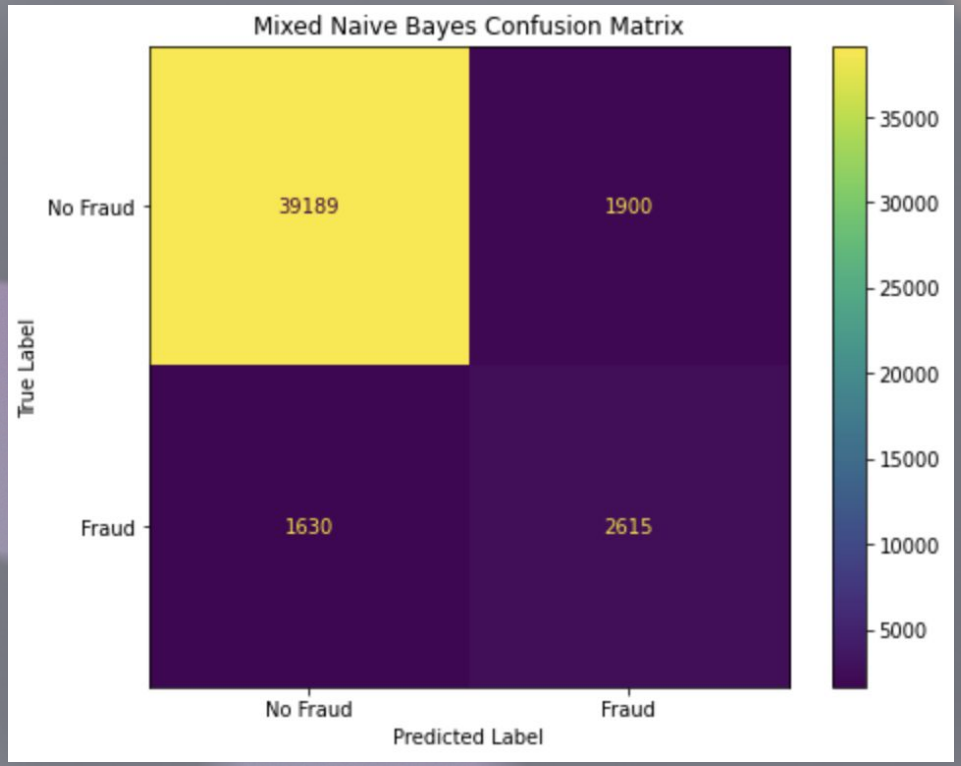
Logistic Regression with Lasso Regularization Model AUC-PR 70%



MODEL 5A

Mixed Naive Bayes Classifier (with pre-processed predictors) Accuracy

9.4%

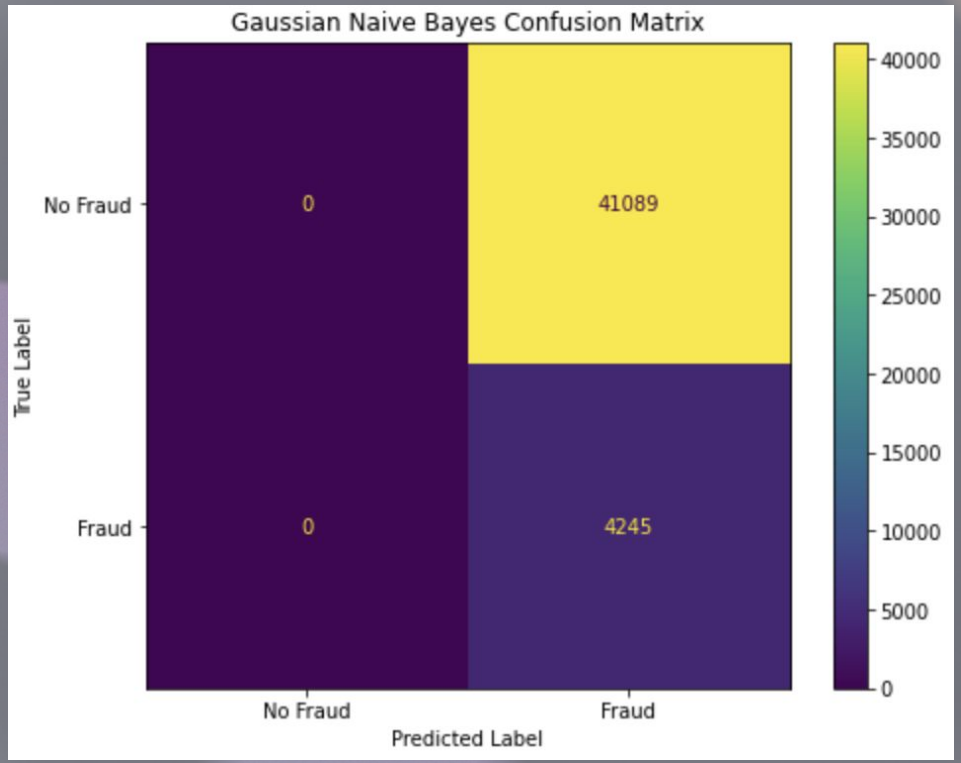


MODEL 5B

Mixed Naive Bayes Classifier (without pre-processing predictors)

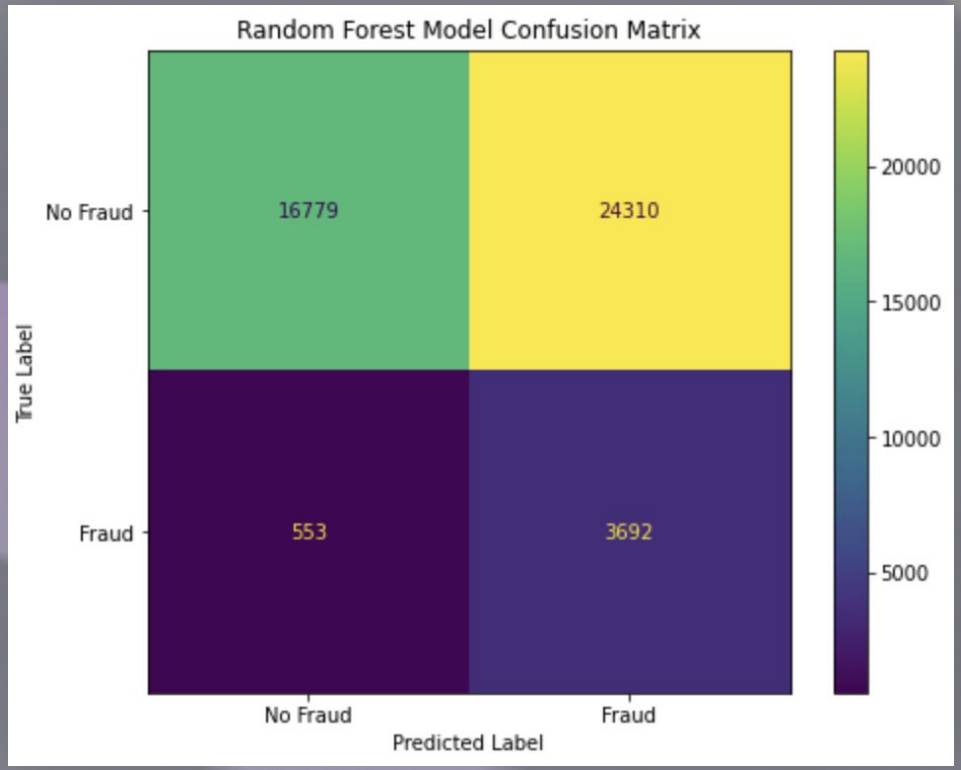
Accuracy

92.2%



MODEL 6

Gaussian Naive Bayes Classifier Accuracy 9.4%

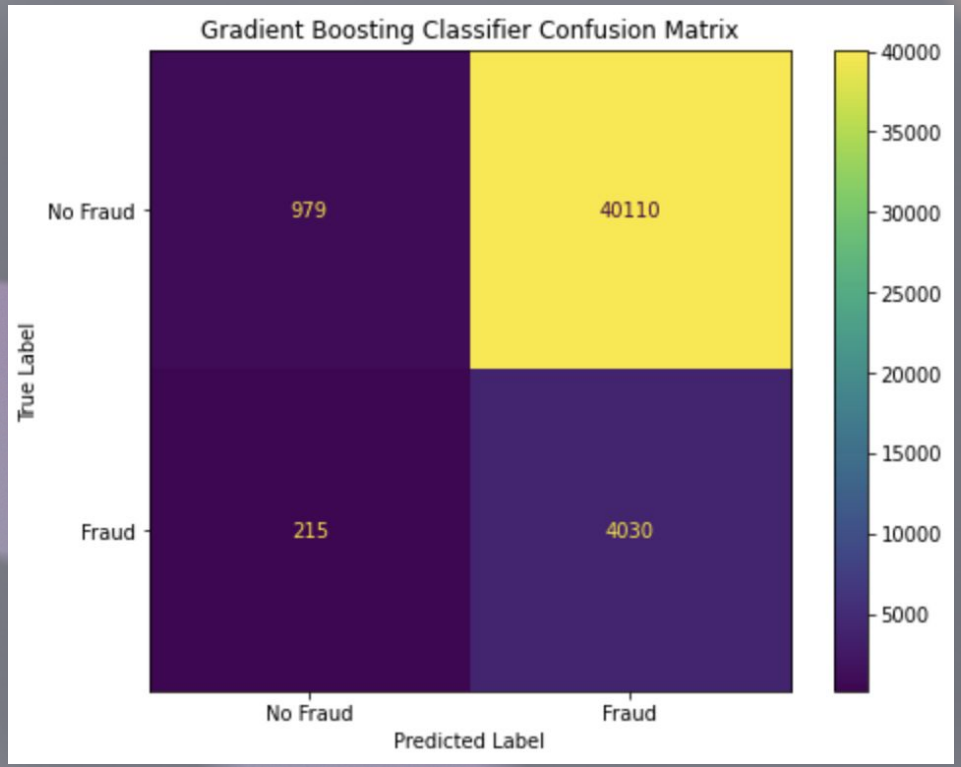


MODEL 7

Random Forest Classifier with Grid Search CV

Accuracy

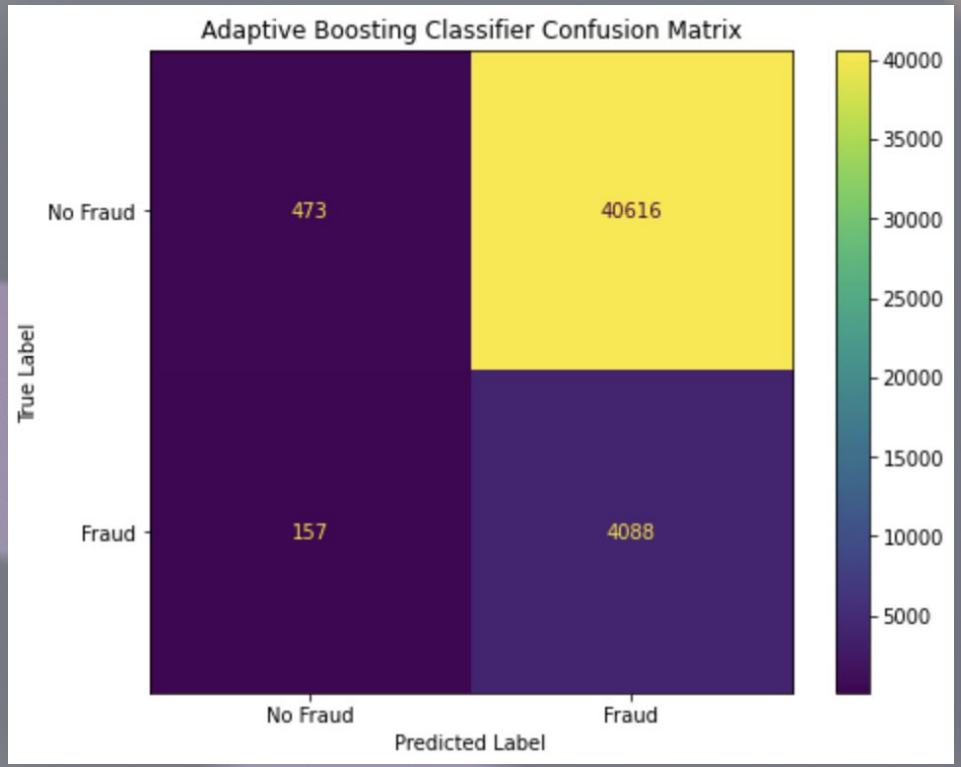
45.2%



MODEL 8

Gradient Boosting Classifier Accuracy

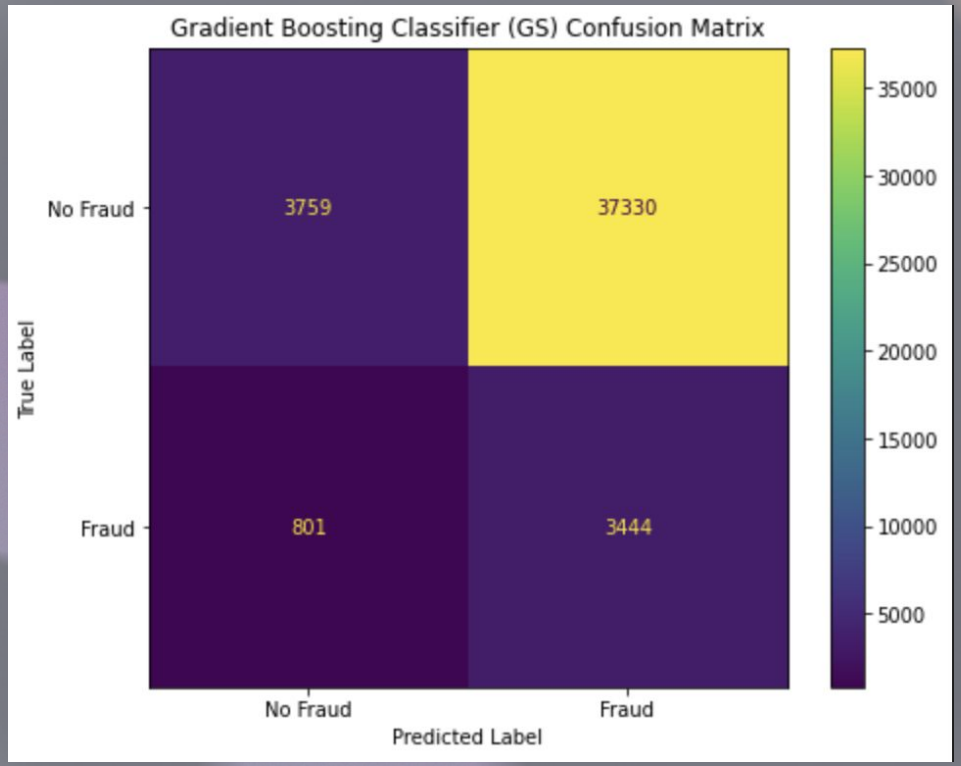
11.0%



MODEL 9

Adaptive Boosting Classifier Accuracy

10.1%



MODEL 10

Gradient Boosting Classifier with Grid Search CV Accuracy 15.9%