# Credit Card Fraud Detection: Classification Models

# Motivation


## Global and National Impact

- Total global losses from credit card fraud totaled around $34 billion in 2022
- On a national scale, roughly 426,000 cases of credit card fraud were reported to the FTC in 2023

## Business Considerations

- Proactive fraud detection using machine learning models can lead to significant savings by reducing the amount of money lost to fraud
- Businesses that use effective classification models to accurately detect fraud can enhance company reputation and solidify a loyal customer base

# DataSet

Rows: 555,179

Columns: 23

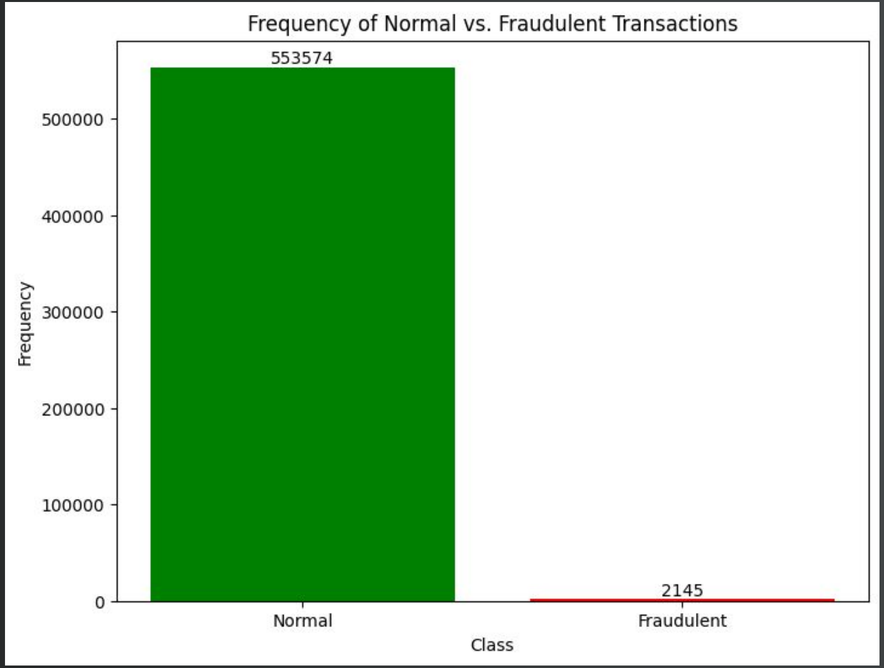
Significant Features: merchant, category, amount, city, state, dob, is_fraud (whether the transaction is fraud or not), trans_date_time(time and date of transaction)

# Objective

The objective of the project is to analyze a comprehensive dataset of credit card transactions and develop classification models that can accurately distinguish between legitimate and fraudulent transactions.

# EDA: Exploratory Data Analysis

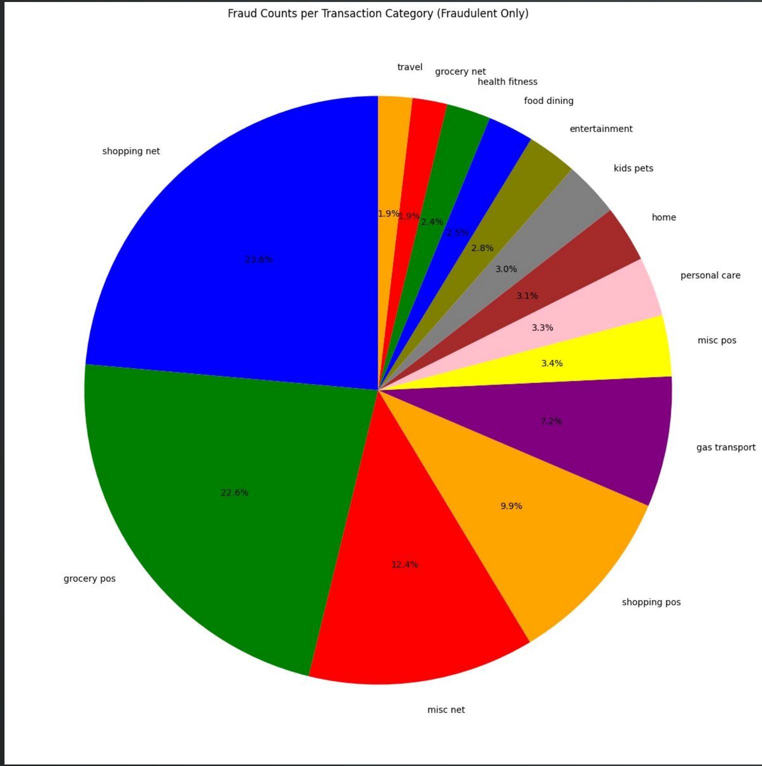
# Distribution of Transactions




Takeaway:

- Crucial to develop robust models to identify the small fraction of fraudulent transactions.
- Data balancing will be required to ensure robustness against false negatives.

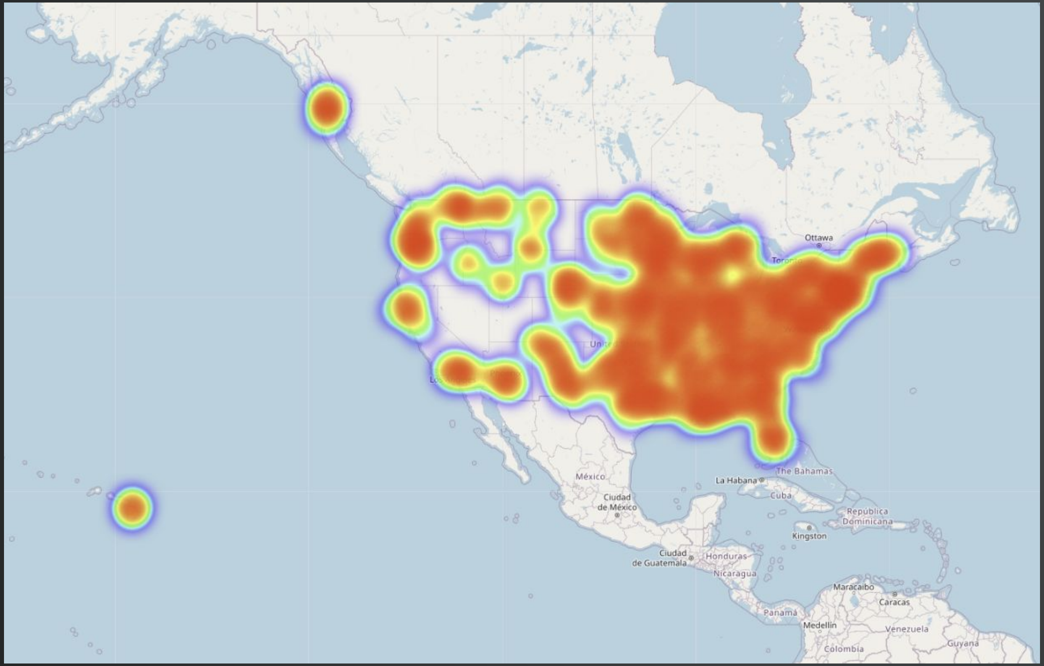
# Distribution of Fraudulent Transactions by Category




Takeaway:

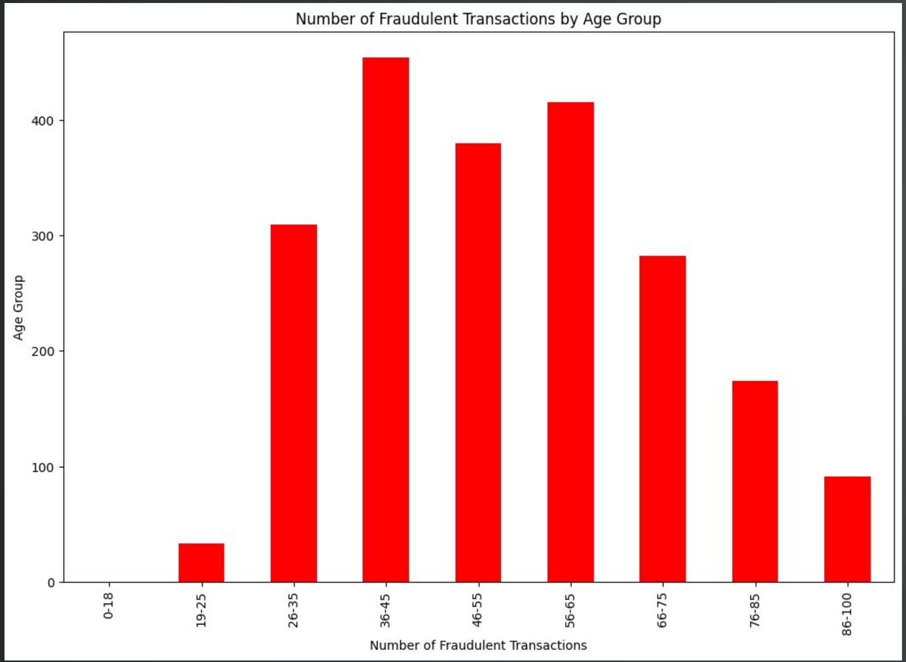
- Fraudulent transactions are predominantly concentrated in online shopping and grocery point-of-sale categories, which together account for nearly half of all fraud incidents.
- Underscores the complexity of credit card fraud activity, reinforcing the need for vigilance in all sectors.

# HeatMap of Fraudulent Transactions by State



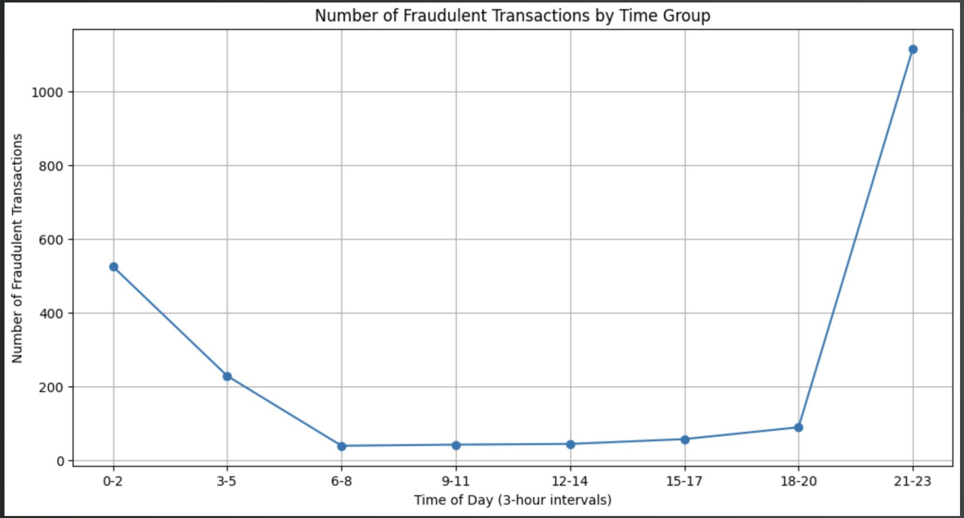

Takeaway: The central part of the United States shows less intensity compared to the coasts, possibly reflecting population density and urbanization levels to be factors correlated with fraudulent activities.

# Distributions of Fraudulent Transactions by Age Group




Takeaway: Middle-aged individuals, particularly those in the 35-45 and 45-55 age groups, experience the highest incidence of fraudulent transactions.

# Temporal Trends in Fraudulent Transactions




<u>Takeaway:</u>

- Fraudulent activities are most prevalent during later hours of the day, suggesting the need for extra vigilance.
- Indicates critical periods during the day which anti-fraud measures must be intensified

# Analyzing Fraudulent Transactions by Transaction Amount

| | amount_range | fraud_count |
|---|---|---|
| 0 | $500 - $999 | 760 |
| 1 | $100 - $499 | 629 |
| 2 | Below $100 | 480 |
| 3 | $1000 and above | 276 |

Takeaway:

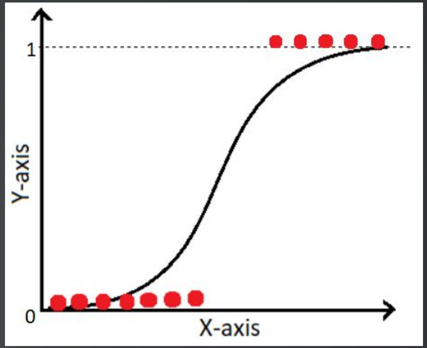
- Highest number of fraudulent transactions (760) falls within the "$500 - $999" range.
- Highest amount range has the least number of fraudulent transaction incidents.

# Modeling: Classification Models

# Model 1: Logistic Regression (SMOTE)

Model Explanation:

- Logistic Function
- Well-suited for binary classification
- SMOTE to artificially balance the dataset




Accuracy: 0.9299
Recall Score: 0.88

Actual Values

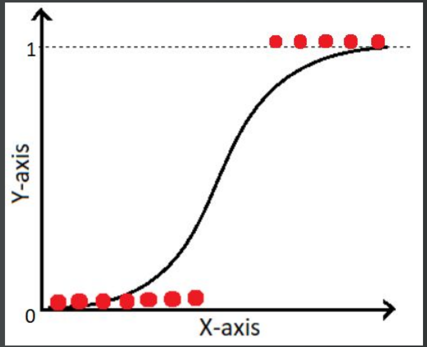
Predicted Values

| | Fraud (1) | Not Fraud (0) |
|---|---|---|
| Fraud (1) | 376 | 7732 |
| Not Fraud (0) | 50 | 102986 |

# Model 1: Logistic Regression (Class Weights)

Model Explanation:

- Class Weights
- More sensitive to the minority class

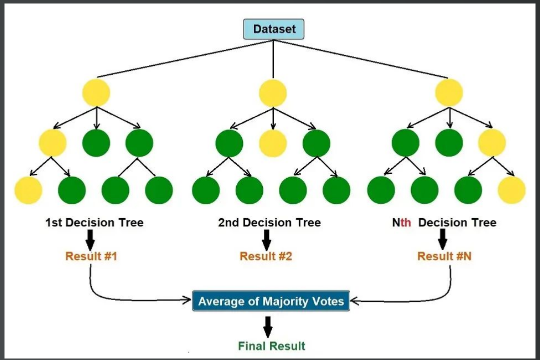



Accuracy: 0.926
Recall Score: 0.89

Actual Values

| Predicted Values | Fraud (1) | Not Fraud (0) |
|---|---|---|
| **Fraud (1)** | 381 | 8133 |
| **Not Fraud (0)** | 45 | 102585 |

# Model 2: Random Forest




Model Explanation:

- Ensemble learning method
- Constructs many decision trees and takes mode class

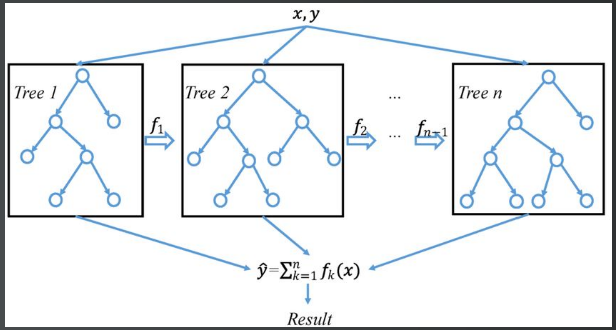
Parameters: n_estimators = 40,
max_depth = 6,
Random_state = 42

Accuracy: 0.976
Recall Score: 0.88

Actual Values

| Predicted Values | Fraud (1) | Not Fraud (0) |
|---|---|---|
| **Fraud (1)** | **374** | **2470** |
| **Not Fraud (0)** | **52** | **108248** |

# Model 3: XGBoost




Model Explanation:

- Build an ensemble of decision trees sequentially
- High performance and speed
- Flexible

Parameters: max_depth = 5

Accuracy: 0.99
Recall Score: 0.90

Actual Values

Predicted Values

| | Fraud (1) | Not Fraud (0) |
|---|---|---|
| **Fraud (1)** | **384** | **1051** |
| **Not Fraud (0)** | **42** | **109667** |

# Implications and Insights

- Enhanced customer trust and satisfaction
- More informed decisions on security measures to mitigate future losses
- Better risk management and operational efficiency
- Empowers institutions to stay ahead of evolving threats

# Challenges and Limitations

- Optimal number of components for PCA
- OneHotEncoder difficulties
- Limited RAM
- Fine tuning our models with the optimal parameters

# Potential Future Steps

- Further fine tune our models for better performance
- Improve recall score even more
- Implement more various models
- Incorporate additional datasets to extract relevant features to further improve our analysis