Credit Card Fraud Analysis

Humphrey Hui, Nathan Arimilli, Minji Kim, Gabriel Sanders, Franco Salinas

August 2025

Introduction, Objectives, Data Overview

U.S consumers faced **$12.5 billion** in losses from credit card fraud in 2024, a **25% increase** from 2023[1]. This rise represents a significant financial and security issues for both credit card users and companies.

Research Questions

1. Is there a pattern in the feature values for those charges that were fraudulent to see **if specific customer groups were targeted**?
2. With these specific features, can we more **accurately and confidently predict fraudulent charges** to **keep our customer base safe and informed**?

Dataset Overview

Sourced from Kaggle

22 features

555,719 observations

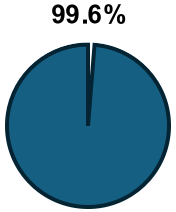
- Customer information
- Transaction time
- Transaction location

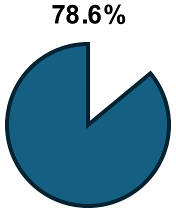
1. https://www.clearlypayments.com/blog/credit-card-fraud-statistics-in-2024-for-usa/

Exploratory Data Analysis (EDA)

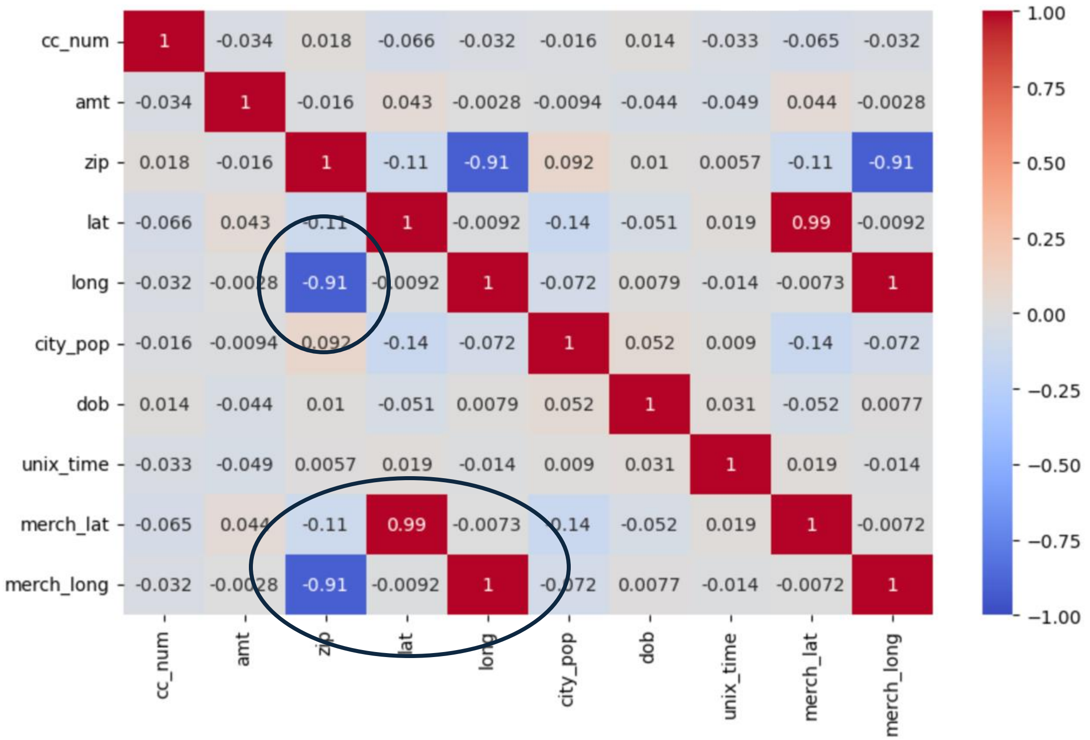
Our initial inspection of the data found one major issue: **99.6% of the observations were non-fraud**.

So, we **resampled the data** to truncate the non-fraud observations to create a more usable dataset.





Correlation Heatmap

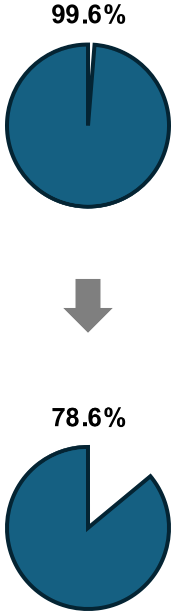


Action: Consider removing one of the correlated predictors (e.g. lat / long) for better model performance.

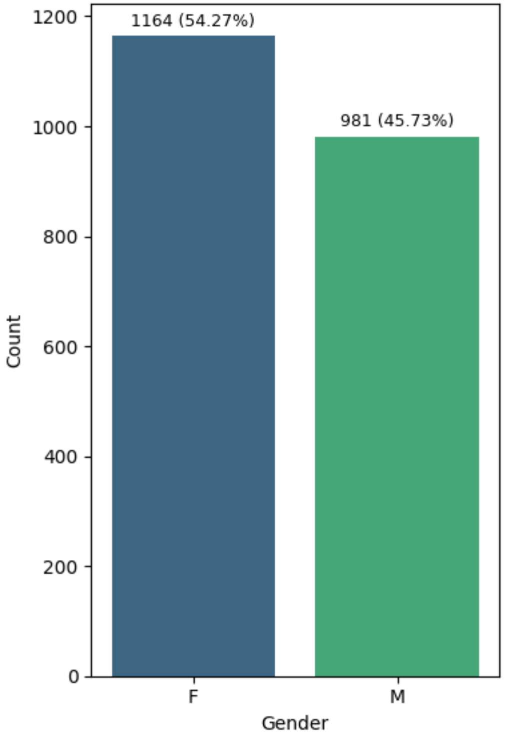
Exploratory Data Analysis (EDA)

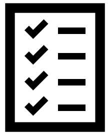
Our initial inspection of the data found one major issue: **99.6% of the observations were non-fraud**.

So, we **resampled the data** to truncate the non-fraud observations to create a more usable dataset.



Fraud by Gender



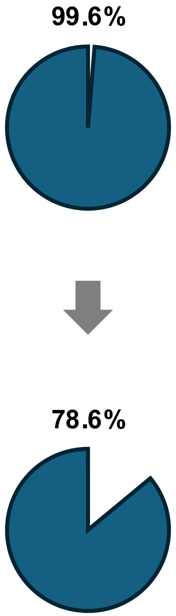


Consideration: As we resampled the original dataset, focus more on the relative positioning.

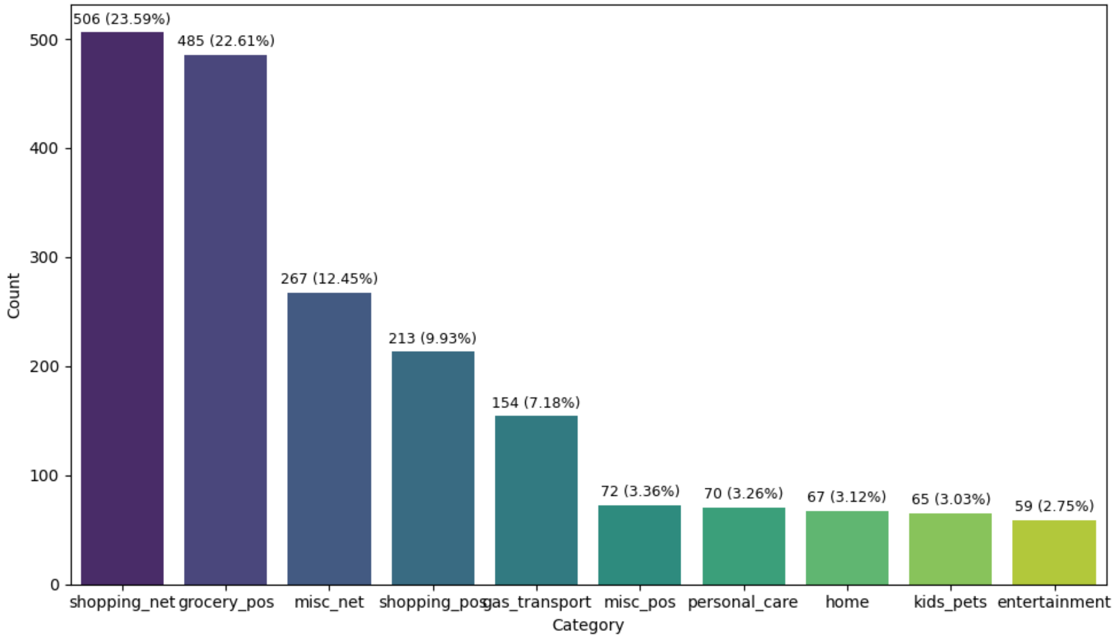
Exploratory Data Analysis (EDA)

Our initial inspection of the data found one major issue: **99.6% of the observations were non-fraud**.

So, we **resampled the data** to truncate the non-fraud observations to create a more usable dataset.



Fraud by Transaction Category



Consideration: As we resampled the original dataset, focus more on the relative positioning.

Model – Logistic Regression

Model Setup

Train/Test Split

80/20

Preprocessing

- Feature frequencies
- Time related factors
- Target encoding

Key Features

Night | Amount | Jobs - TE | Merchant - TE | Category - TE | Full Name Freq

Confusion Matrix

Actual \ Predicted	Not Fraud	Fraud
Not Fraud	**1502**	69
Fraud	138	**291**

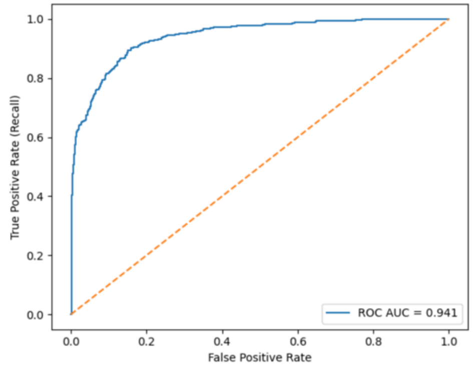
Key Metrics

Recall: 68%

Precision: 81%

Accuracy: 90%

ROC-AUC Graph



Feature Deep Dive

	Coefficient	*Odds Ratio*
Night	2.28	9.82
Amount	1.86	6.44
Jobs - TE	1.26	3.52
Merchant - TE	0.99	2.68
Category - TE	-0.70	0.50
Full Name Freq	0.56	1.76

Night transactions, higher amount, and higher frequency means more likely to be fraud

Model – Naïve Bayes

Model Setup

Train/Test Split

80/20

Preprocessing

- Jobs categories
- Independent features
- Binned numerical factors

Key Features

State – WV

State - UT

State - VT

Amount

Confusion Matrix

Predicted

Actual	Not Fraud	Fraud
Not Fraud	**1505**	66
Fraud	113	**316**

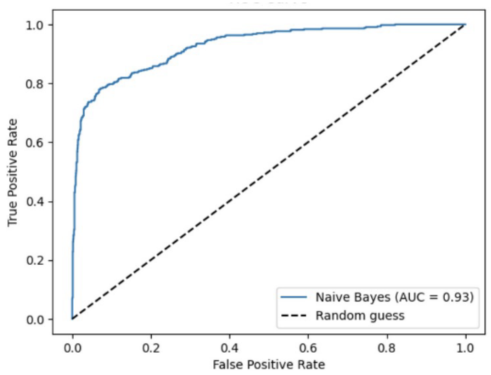
Key Metrics

Recall: 74%

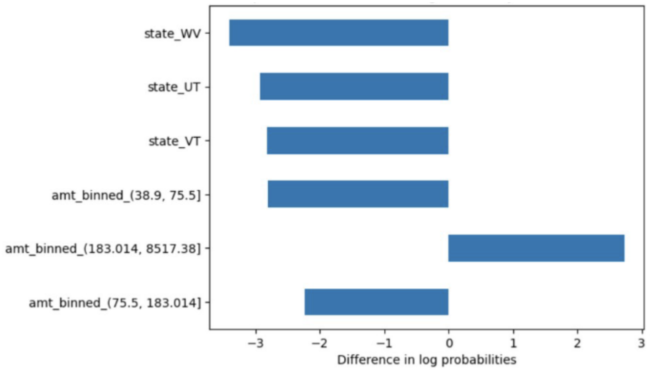
Precision: 83%

Accuracy: 91%

ROC-AUC Graph



Top 6 Influential Features



Lower recall, state location found to be more influential

Model – K Nearest Neighbors (KNN)

Model Setup

Train/Test Split

80/20

Preprocessing

- Log(amount), distance
- Time related factors
- Target encoding

Key Features

- Merchant - TE
- City - TE
- City Population
- Log(Amount)
- Night
- Distance

Confusion Matrix

Predicted

Actual	Not Fraud	Fraud
Not Fraud	**1524**	47
Fraud	55	**374**

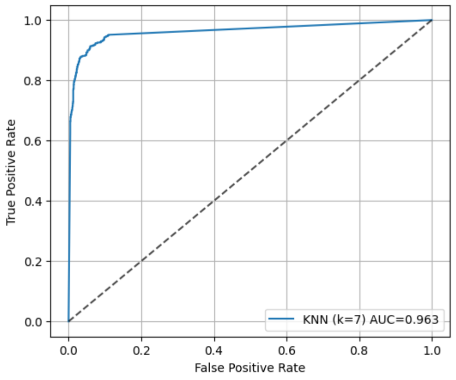
Key Metrics

Recall: 87%

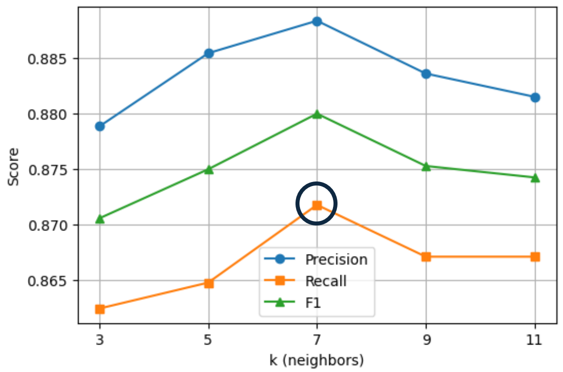
Precision: 89%

Accuracy: 95%

ROC-AUC Graph



KNN Metrics vs K



K=7 was used as it gave the highest recall, location and merchant features found to be most important

Model – Classification Tree

Model Setup

Train/Test Split

80/20

Preprocessing

- Drop features
- Encode categories

Key Features

Amount *72%*

Gas *6%*

Groceries *2%*

Confusion Matrix

		Predicted	
		Not Fraud	Fraud
Actual	Not Fraud	**1509**	44
	Fraud	32	**415**

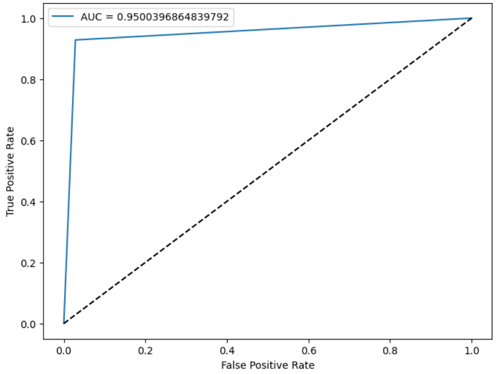
Key Metrics

Recall: 93%

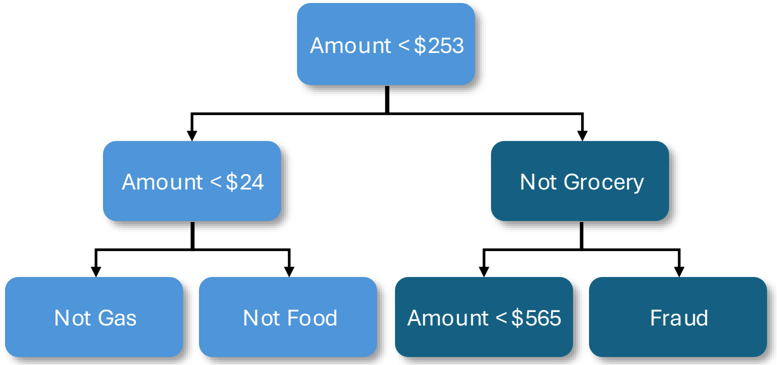
Precision: 90%

Accuracy: 96%

ROC-AUC Graph



Main Tree Splits



Best tree depth is 10, amount the overwhelmingly dominating predictor

Conclusion

Key Fraud Signals

Transaction Amount
The **higher the amount**, the more likely it is fraud

Time of Day
The **later the transaction in the day**, the more likely it is fraud

Merchant and Category
Historically high patterns of fraud for **certain merchants and categories**

Model Performance

Best Model: Classification Tree

- Recall: 93%
- Precision: 90%
- Accuracy: 96%

Reasoning and Caveat

- Strong ability to capture **non-linear patterns and interactions**
- High metrics due to **synthetic dataset**, but research has found classification trees to perform best on credit card fraud[2]

Recommended Actions

Real-time Controls

- More stringent checks on **high-amount late night transactions**
- **Additional verification steps** for higher risk merchants and categories

Model Deployment

- **Retrain regularly** with latest historical data
- While recall crucial for business, **precision is what the customer experience depends on**

2. https://www.sciencedirect.com/science/article/pii/S2772662223000036 and https://modelshop.dev/article/Top_5_machine_learning_models_for_fraud_detection_on_ModelShopdev.html#:~:text=Support%20Vect[...]is,Gradient%20Boosting