VINUNIVERSITY

Enhancing Financial Security: Credit Card Fraud Detection Using XGBoost

Tran Tue Nhi

Agenda

1. Project Background
2. Model Background
3. Dataset Description
4. Project Pipeline
5. Challenges & Recommendation

Project Background

Project Background | Problem Definition

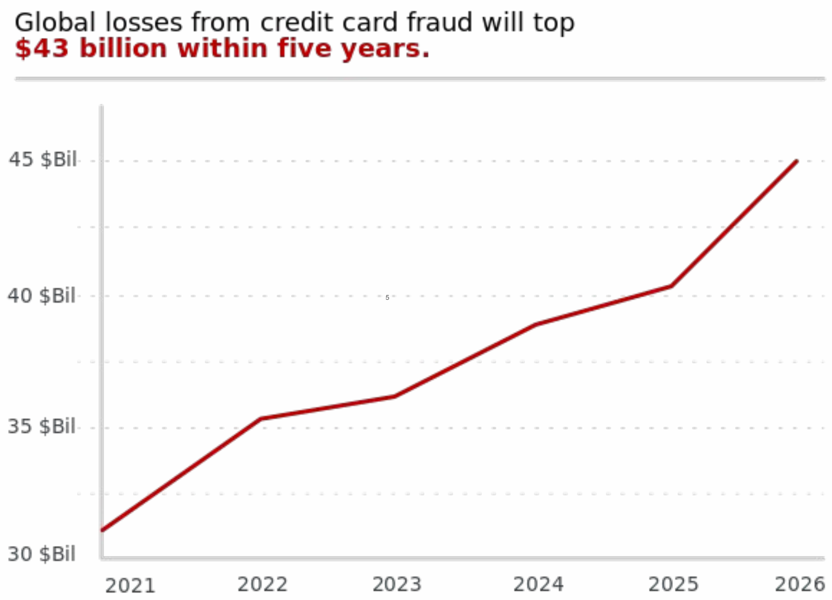
	2021 Rank	2022 Rank	2023 Rank	Global % Experiencing (2023)
Phishing / pharming / whaling	3	1	1	**43%** ↑
First-Party Misuse (i.e., friendly / chargeback fraud)	1	4	2 ●	34%
Card testing	2	2	3 ●	33%
Identity theft	4	3	4 ●	33%
Coupon / discount / refund abuse	5	7	5 ●	30%
Account takeover	7	5	6 ●	27%
Loyalty fraud	6	6	7 ●	22%
Affiliate fraud	8	8	8	22%
Re-shipping	12	11	9 ●	**20%** ↑
Botnets	10	9	10 ●	19%
Triangulation schemes	9	10	11 ●	17%
Money laundering	11	12	12	15%
AVG. # of attacks experienced	**3**	**3**	**3**	**3**

● Increased Ranking ● Decreased Ranking ↑ = Sig. Higher vs. 2022

Figure: Types Of Fraud Experienced By Merchants – Past 3 Year Rankings & Global Incidence (2023)

Reference: Global Fraud and Payments Survey Report 2023

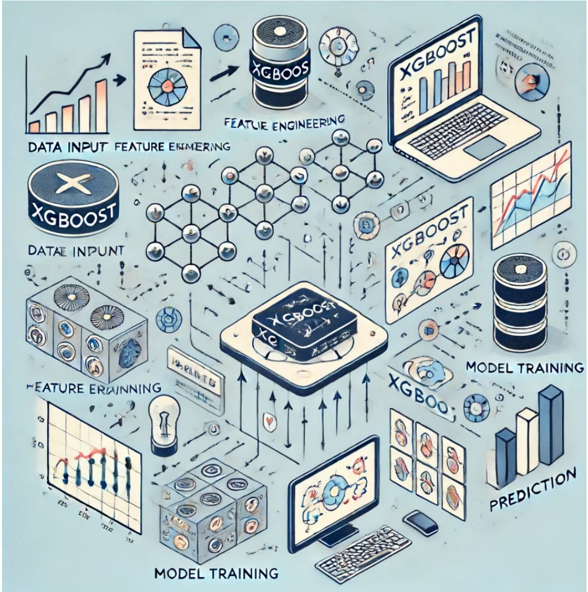
Project Background | Problem Definition



How to identify credit card fraud fast and efficiently?

Reference: Key Credit Card Fraud Statistics to Know for 2024, 2024

Project Background | Overview



XGBoosting detects credit card fraud

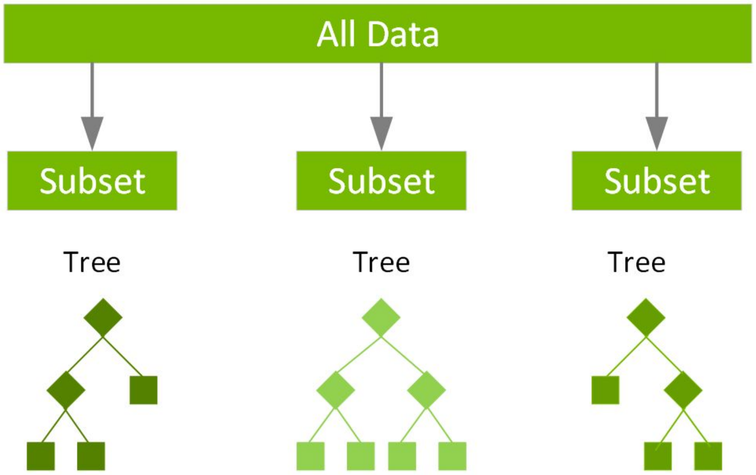
Project Background | Novelty of Solution

01	**Assumes a linear relationship between predictors and the outcome**	• XGBoosting has high accuracy and performance • Build and combine multiple decision trees in an optimal way
02	**Needs large amounts of labeled data for training**	• XGBoost has built-in mechanisms like scale_pos_weight to handle imbalanced datasets
03	**Requires comparing each new instance to all training instances**	• XGBoost's algorithm includes regularization terms (L1 and L2), which prevent overfitting

Reference: Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System.
Akila, V. P., & Sivasankari, S. (2018). Imbalanced Data Handling for Credit Card Fraud Detection using Hybrid Machine Learning Techniques

Model Background

Model Background



XGBoost is a scalable and highly accurate implementation of gradient boosting that pushes the limits of computing power for boosted tree algorithms, being built largely for energizing machine learning model performance and computational speed

Reference: Nvidia. What is XGBoost?

VINUNIVERSITY

Dataset Description

10

Dataset Description

01 Data Overview and Sources

- Sourced from **Kaggle**
- from January 1, 2019, to December 31, 2020
- 1,000 customers and 800 merchants
- Highly **imbalanced dataset**

02 Types of Data

- **numerical** data
- **categorical** data
- **time series** data

03 Data Relevance

- Diverse features
- Identify **patterns** and **anomalies**
- **Historical data** for model learning
- Effective fraud detection with Random Forest

04 Compatibility with XGBoosting

- Data **Cleaning**
- Normalize
- Encode
- Feature **Engineering**

VINUNIVERSITY

Project Pipeline

12

Project Pipeline | Data Preprocessing

Clean

- Handle missing values
- Separate data

Normalize

Normalize numerical columns using StandardScaler

Encode

Target encode on categorical features using TargetEncoder

SMOTE

Apply SMOTE to handle imbalanced data

Feature Engineering

Creates the “hour of the day” from the transaction time

Project Pipeline | Exploratory Data Analysis (EDA)

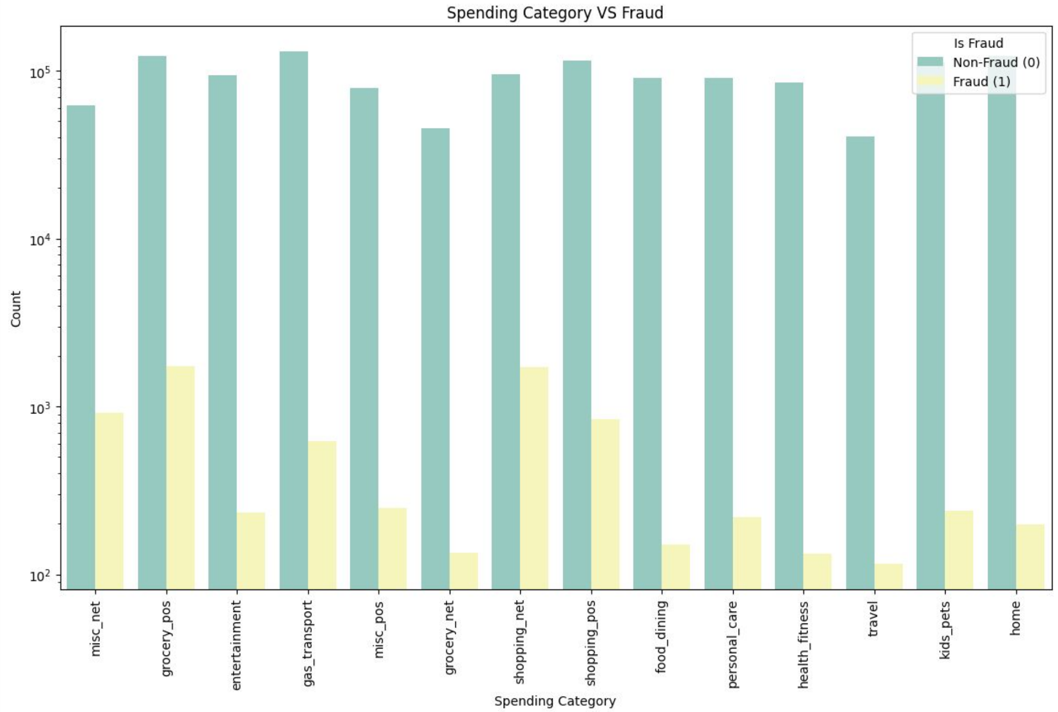


Figure: Which spending categories are more susceptible to fraud

Project Pipeline | Exploratory Data Analysis (EDA)

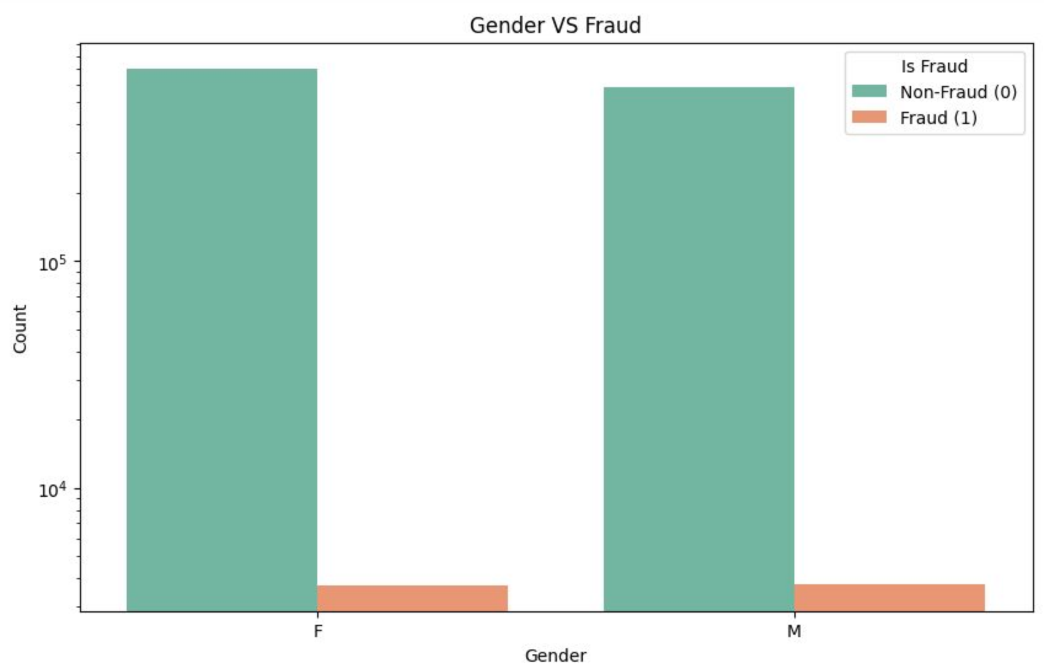


Figure: Is there gender bias in fraudulent transactions?

Project Pipeline | Exploratory Data Analysis (EDA)

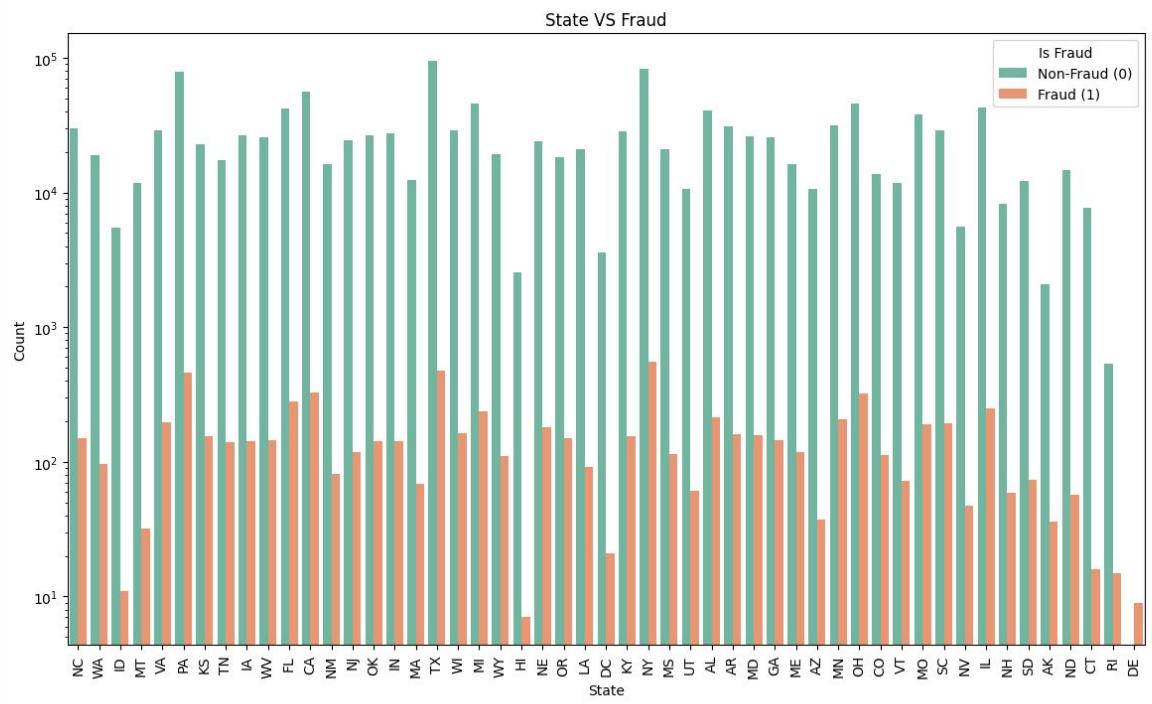


Figure: States with its fraud rates for geographically targeted interventions

Project Pipeline | Exploratory Data Analysis (EDA)

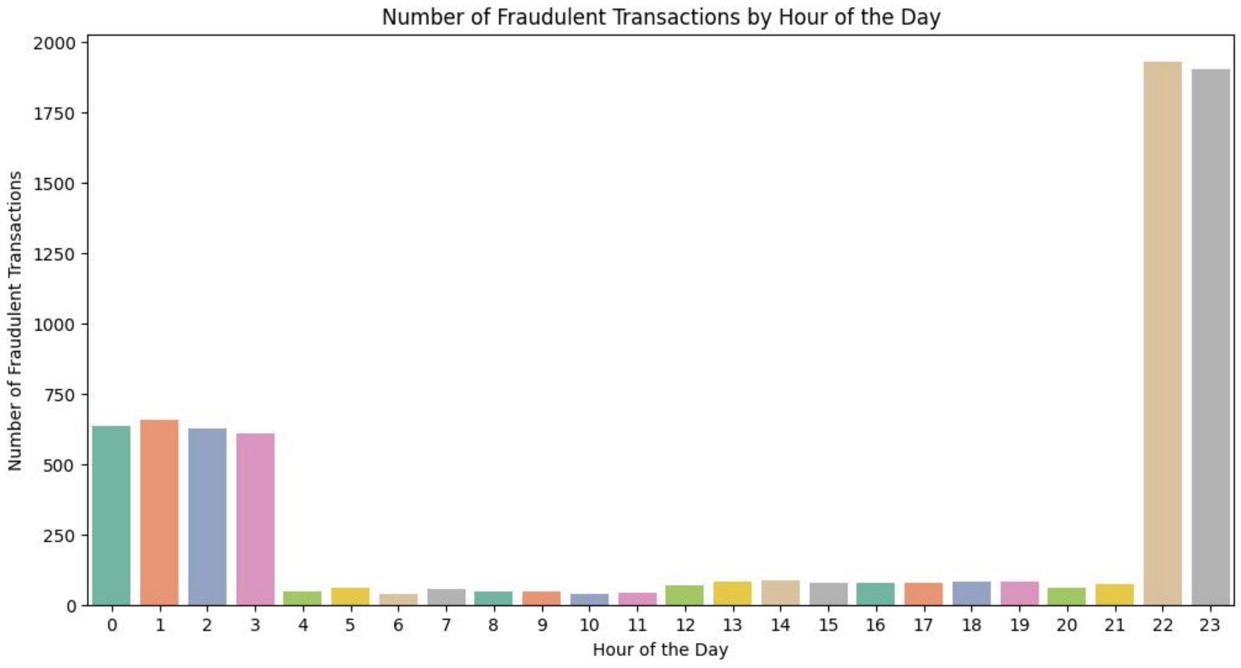


Figure: Temporal patterns in fraud

Project Pipeline | Initial Model Training

- **Incorporating EDA Insights into Model Training:**
    + Feature Engineering Based on Time of Day
    + Prioritize Spending Category in Feature Engineering
    + Feature Engineering Geographical Patterns

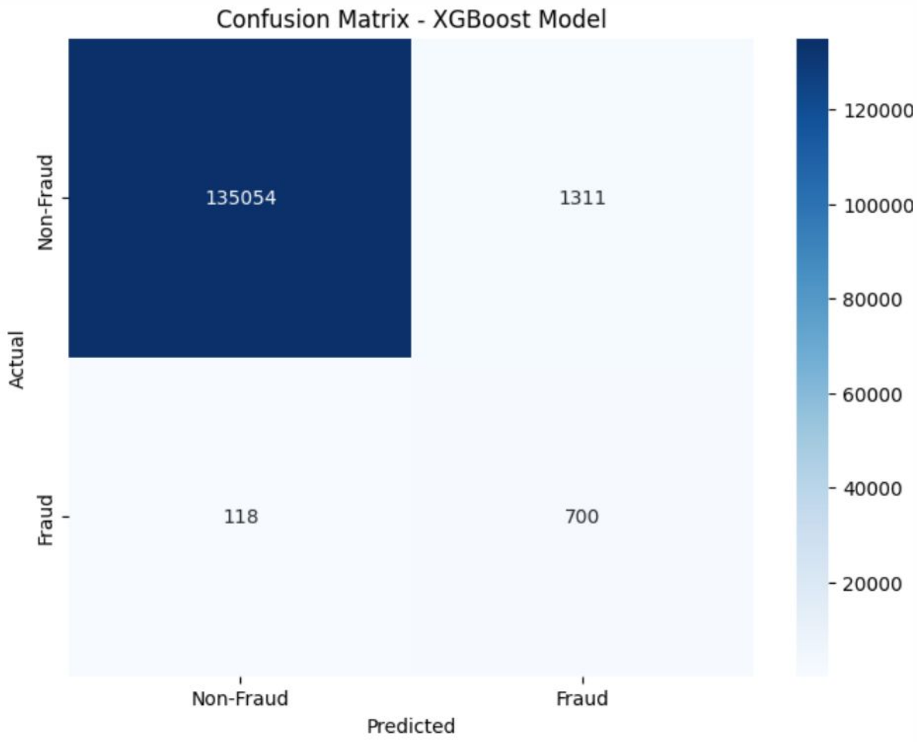
- **Train-Test Split:** Divide the data into 80% training and 20% testing sets.

- **Handling Imbalance:** Utilize SMOTE to address class imbalance.

Project Pipeline | Model Evaluation

Initial Model Classification Report:

	precision	recall	f1-score	support
0.0	1.00	0.99	0.99	136365
1.0	0.35	0.86	0.49	818
accuracy			0.99	137183
macro avg	0.67	0.92	0.74	137183
weighted avg	1.00	0.99	0.99	137183



Project Pipeline | Model Tuning

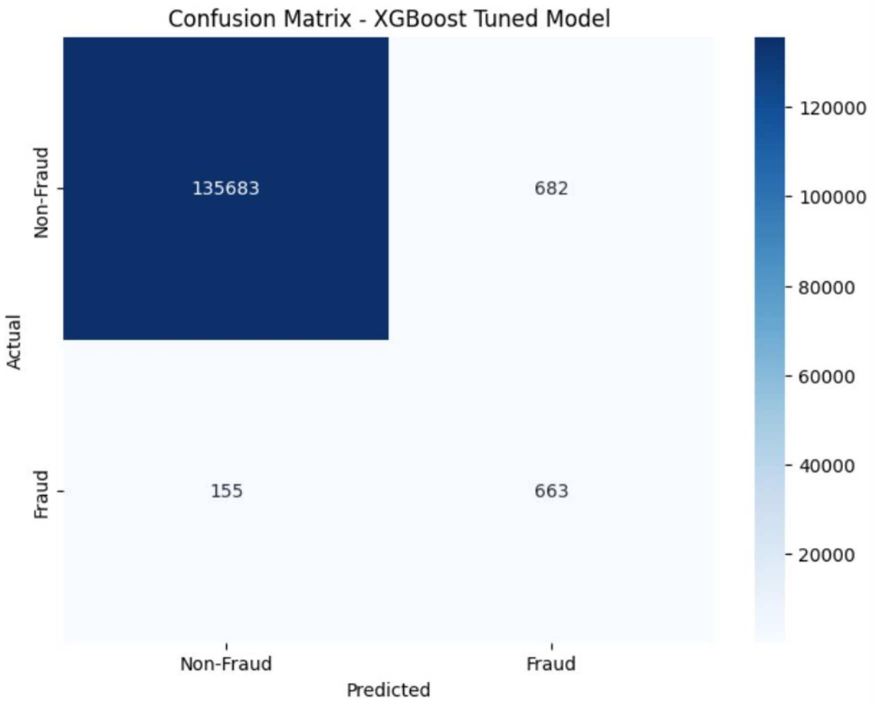
Random grid search for hyperparameter tuning

- Number of boosting rounds (trees): 100, 200, 300
- Maximum depth of each tree: 3, 5, 7
- Learning rate: 0.01, 0.1, and 0.2
- Fraction of samples to be used for fitting each tree: 0.6, 0.8, 1.0
- Fraction of features to be used for fitting each tree: 0.6, 0.8, 1.0

Project Pipeline | Tuned Model Evaluation

Tuned Model Classification Report:

	precision	recall	f1-score	support
0.0	1.00	0.99	1.00	136365
1.0	0.49	0.81	0.61	818
accuracy			0.99	137183
macro avg	0.75	0.90	0.80	137183
weighted avg	1.00	0.99	0.99	137183



VINUNIVERSITY

Challenges & Recommendation

Challenge & Recommendation

Data Accuracy

Differences between current dataset and real-world transaction data => Utilize more real-world datasets

Limited Data

Limited data in terms of time and quantity
=> Constantly expand and update the dataset

Privacy

Data Privacy and Ethical Use => Conduct regular audits and protection layers for model

Thank you for listening!