

Credit Card Fraud Detection

Machine Learning Project Summary

March 29, 2025

Chervin Daniel

BIU DS-18

Project Overview



Credit card fraud causes major financial losses and reduces user trust.



Goal: Build a machine learning model to detect fraud with minimal false positives.



Challenge: Only ~0.5% of transactions are fraudulent.



Approach: Use advanced techniques like feature engineering and SMOTE for better detection.

Data Handling

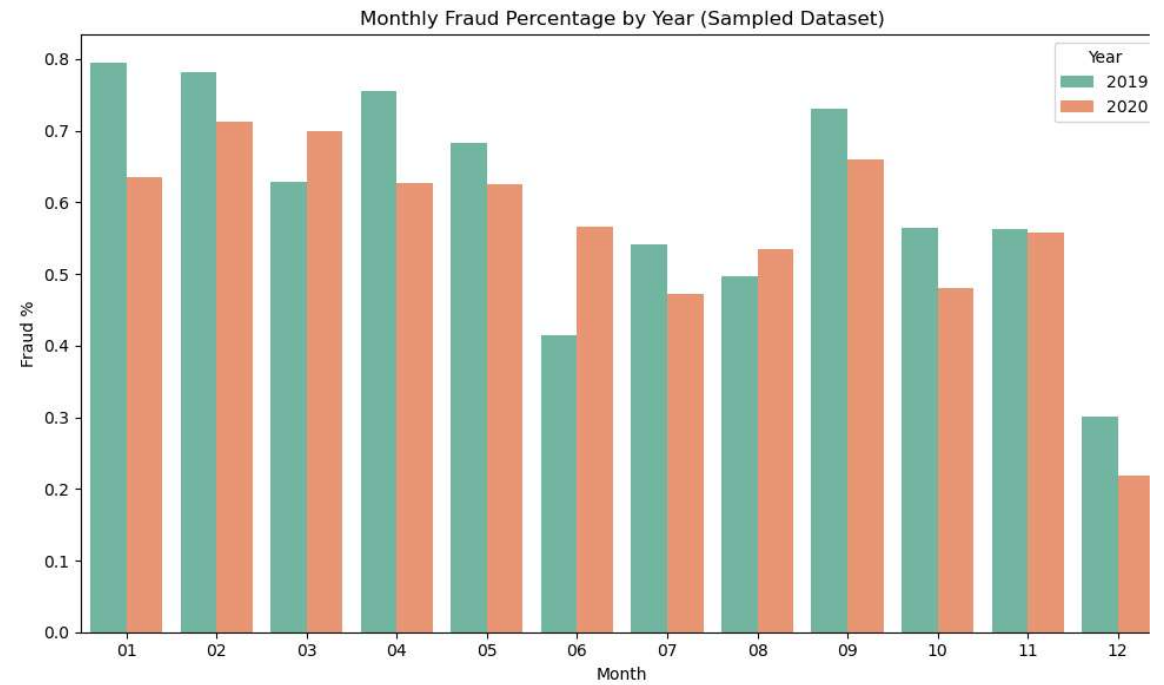
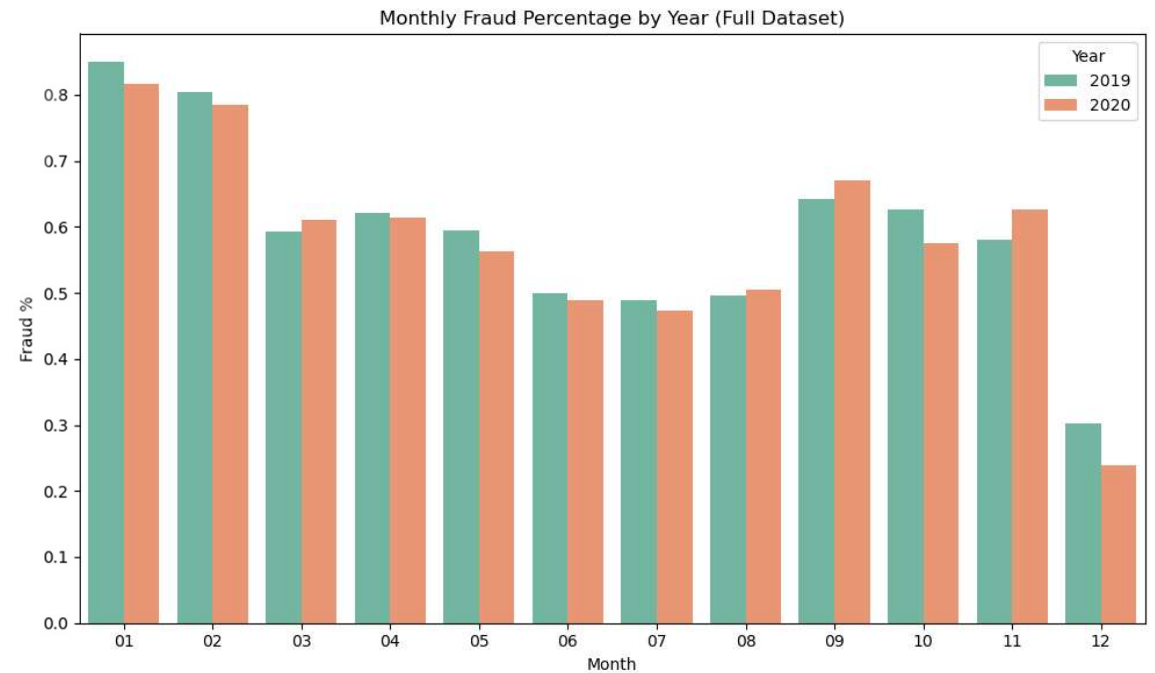
Initial dataset: 34.6 million rows; reduced to 300,000 for analysis.

Tools employed: Utilized Dask for effective data processing.

Feature engineering: Included Temporal, Geographical, Demographic, and Transactional features.

Data quality assessment: Ensured no missing values, preserved outliers, and implemented log transformations.

Fraud Visualization



Exploratory Data Analysis (EDA)



Fraud rate:
approximately 0.5%.



Characteristics of
fraudulent transactions
often include:



Higher transaction
amounts



Occurrence in
categories such as
entertainment and
groceries



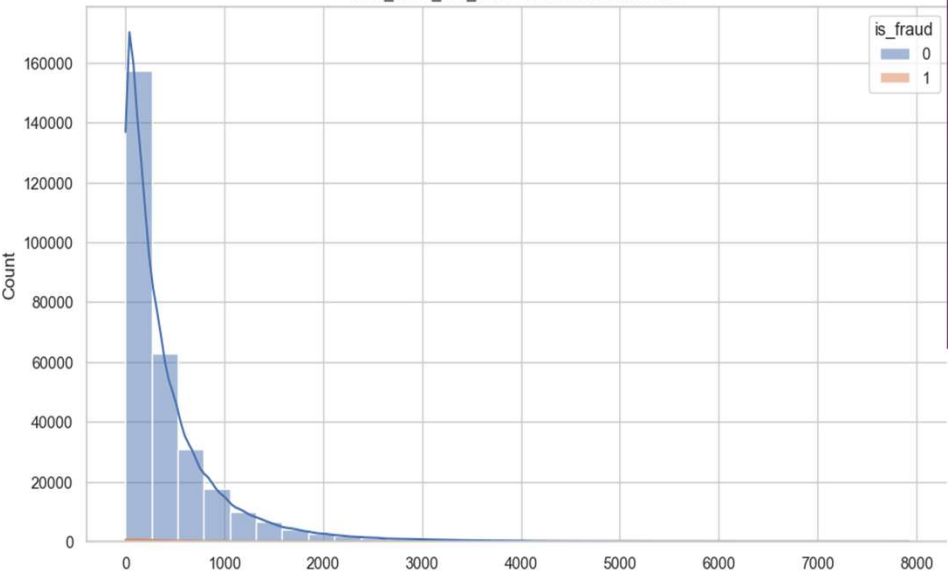
Happening during
atypical hours



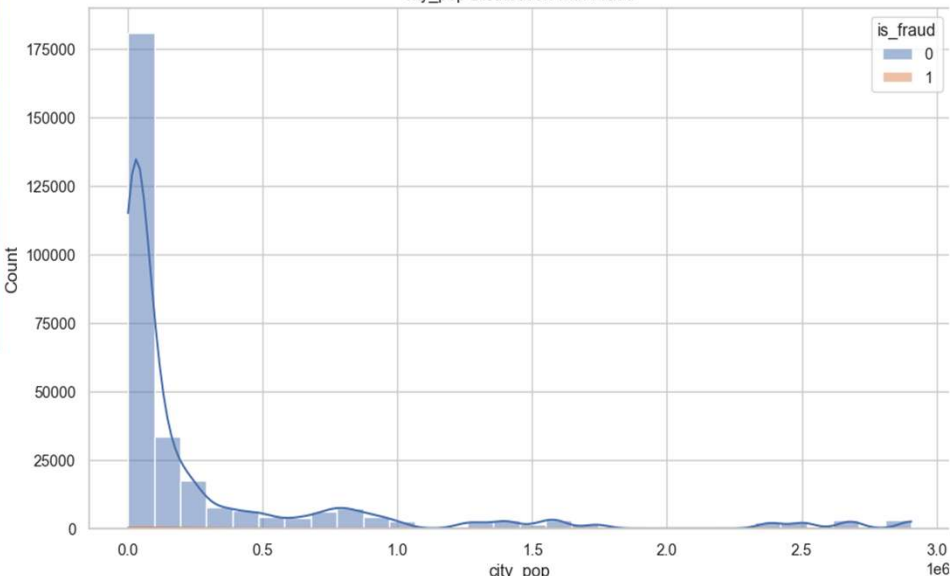
Major indicators:
transaction category,
amount (log_amt), and
hour of occurrence.



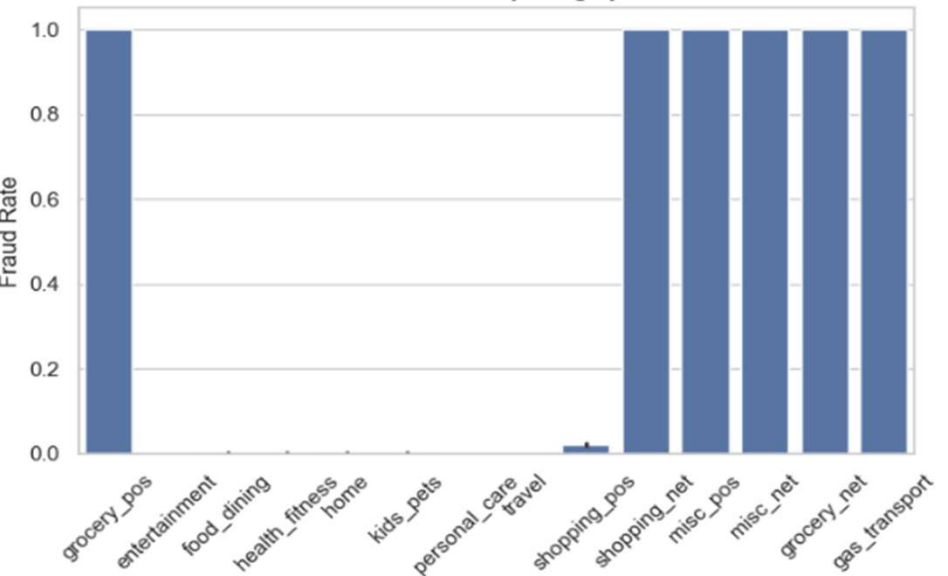
time_since_last_trans Distribution with Fraud



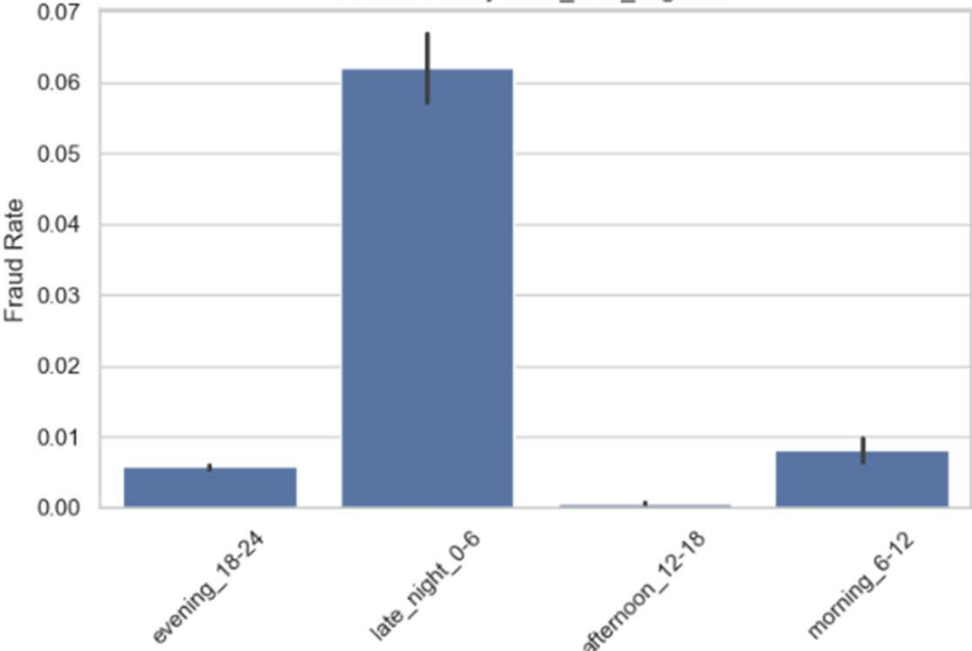
city_pop Distribution with Fraud



Fraud Rate by category



Fraud Rate by trans_time_segment





Feature Selection & Class Imbalance

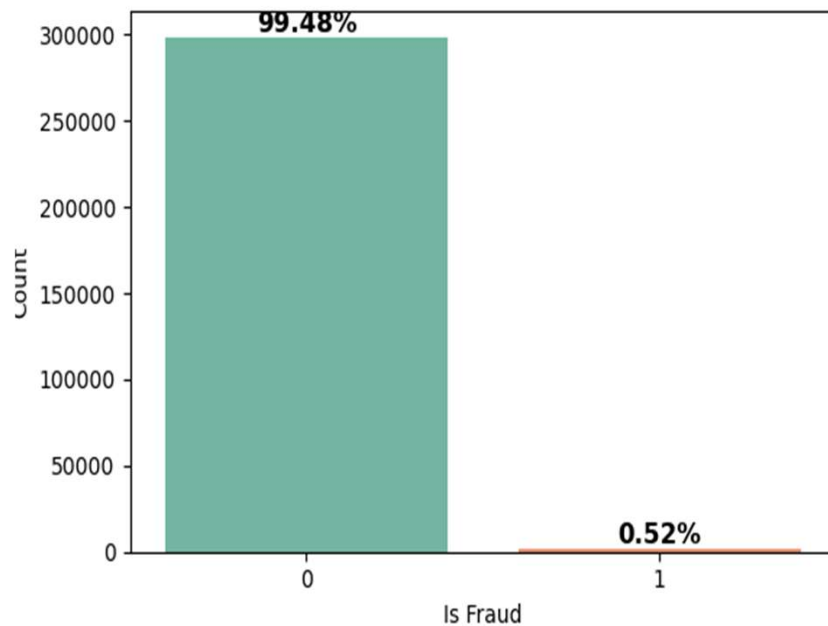
- Feature selection employed techniques including Lasso, SVC, Random Forest, and Ridge.
- The resampling methods evaluated included:
 - No Balancing, Random Over Sampling (ROS), Random Under Sampling (RUS), SMOTE, and SMOTETomek.
 - Of these, SMOTE emerged as the most effective approach for achieving a balance between precision and recall.

Feature Selection

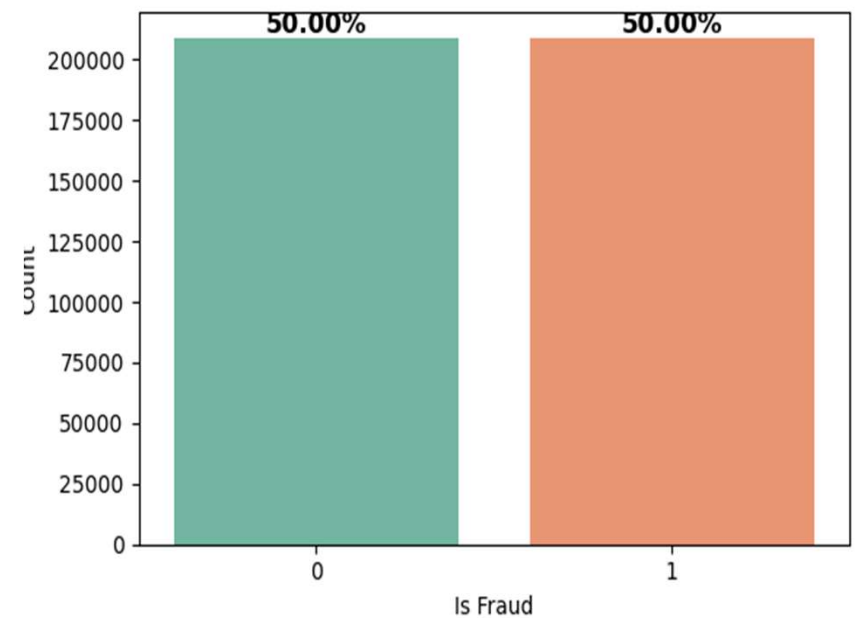
Feature	Lasso	Ridge	LinearSVC	GradientBoosting	RandomForest
category	✓	✓	✓	✓	✓
log_amt	✓	✓	✓	✓	✓
unix_time	✓	✗	✗	✓	✓
trans_day_of_week	✓	✗	✓	✗	✗
is_weekend	✓	✗	✗	✗	✗
trans_hour	✓	✗	✓	✗	✓
age_group	✓	✗	✗	✗	✗
area_cat	✓	✗	✗	✗	✗
log_time_since_last_trans	✓	✗	✓	✗	✗
log_city_pop	✓	✗	✗	✗	✗

Class Imbalance

Distribution of Fraud vs Non-Fraud Transactions



Distribution of Fraud vs Non-Fraud Transactions



Technique	Accuracy	Precision	Recall	F1-Score
ROS	0.9992	1.0000	0.8416	0.9140
RUS	0.9793	0.1867	0.9593	0.3126
SMOTE	0.9991	0.9222	0.8846	0.9030
SMOTETomek	0.9990	0.9093	0.8846	0.8968

Model Development

Models evaluated: Linear Regression, Decision Tree, Random Forest, AdaBoost, Gradient Boosting Machine, Support Vector Machine, XGBoost.

Optimization conducted using RandomizedSearchCV.

Leading models: Random Forest and XGBoost.

Model Performance

- Accuracy: 99.97% for both Random Forest (RF) and XGBoost.
- XGBoost: Exhibits greater precision (False Positives = 3).
- RF: Demonstrates superior recall (False Negatives = 7).
- Recommendation: Opt for XGBoost to reduce false positives.




Model Selection

Model	Accuracy	Precision	Recall	F1-Score
Linear Regression	0.8801	0.8969	0.8601	0.8781
Decision Tree	0.9972	0.9964	0.9979	0.9972
Random Forest	0.9997	0.9996	0.9998	0.9997
Adaptive Boosting (ADABOOST)	0.9672	0.9856	0.9485	0.9667
Gradient Boosting Machine (GBM)	0.9868	0.9948	0.9788	0.9868
XGBoost Classifier	0.9997	0.9998	0.9997	0.9997

Model Optimization




RF Version	TN	FP	FN	TP	Notes
Default	41,595	18	10	41,923	Solid performance
Optimized	41,601	12	7	41,926	Fewer errors (both FP and FN)

Interpretation

-  The **optimized model** shows a **reduction in both false positives and false negatives**, improving both **precision** and **recall**.
-  **False Negatives (FN)** dropped from **10** → **7**, which is important in fraud detection (fewer missed frauds).
-  **False Positives (FP)** also dropped from **18** → **12**, helping avoid incorrectly flagging legitimate transactions.

XGB Version	TN	FP	FN	TP	Notes
Default	41,604	9	14	41,919	Strong baseline performance
Optimized	41,610	3	12	41,921	Improved precision and reduced false positives

Interpretation

-  The **optimized XGBoost model** **reduced false positives** from **9** → **3**, improving **precision** and reduce interruptions.
-  **False negatives** also dropped slightly (**14** → **12**), improving **recall** and catching more frauds.
-  **True positives (TP)** and **true negatives (TN)** slightly increased, indicating overall better classification accuracy.

Deployment and Key Insights

Deployed as a real-time API for fraud detection, it can accommodate two strategies:

- Tailored for in-person transactions, focusing on reducing false positives (FP), minimizing interruptions, and prioritizing accuracy.
- Tailored for online transactions, aiming to lower false negatives (FN) while emphasizing recall.

Key Takeaways:

- Importance of class balancing
- Impact of spatial and temporal factors
- The application of ensemble models and resampling techniques enhances performance.