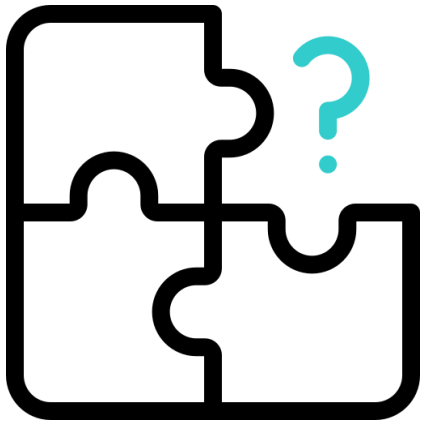
Experian Fraud Detection

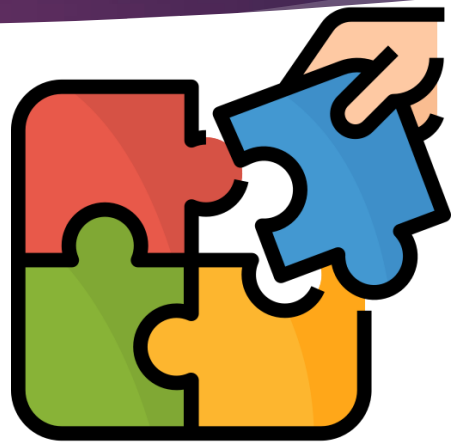
ABRAHAM OWODUNNI (ML ENG)

Introduction



Business Problem

- Importance/Impact
- Challenges



Solution

- Approach
- Key Component

Business Problem

Impact

Consumers

Financial Institutions

Businesses

Complexity

Evolving Techniques

Imbalance Data

Solution

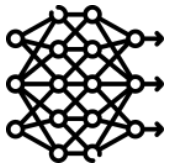
POV

Trust

Comfort

Technology

Seamless Integration

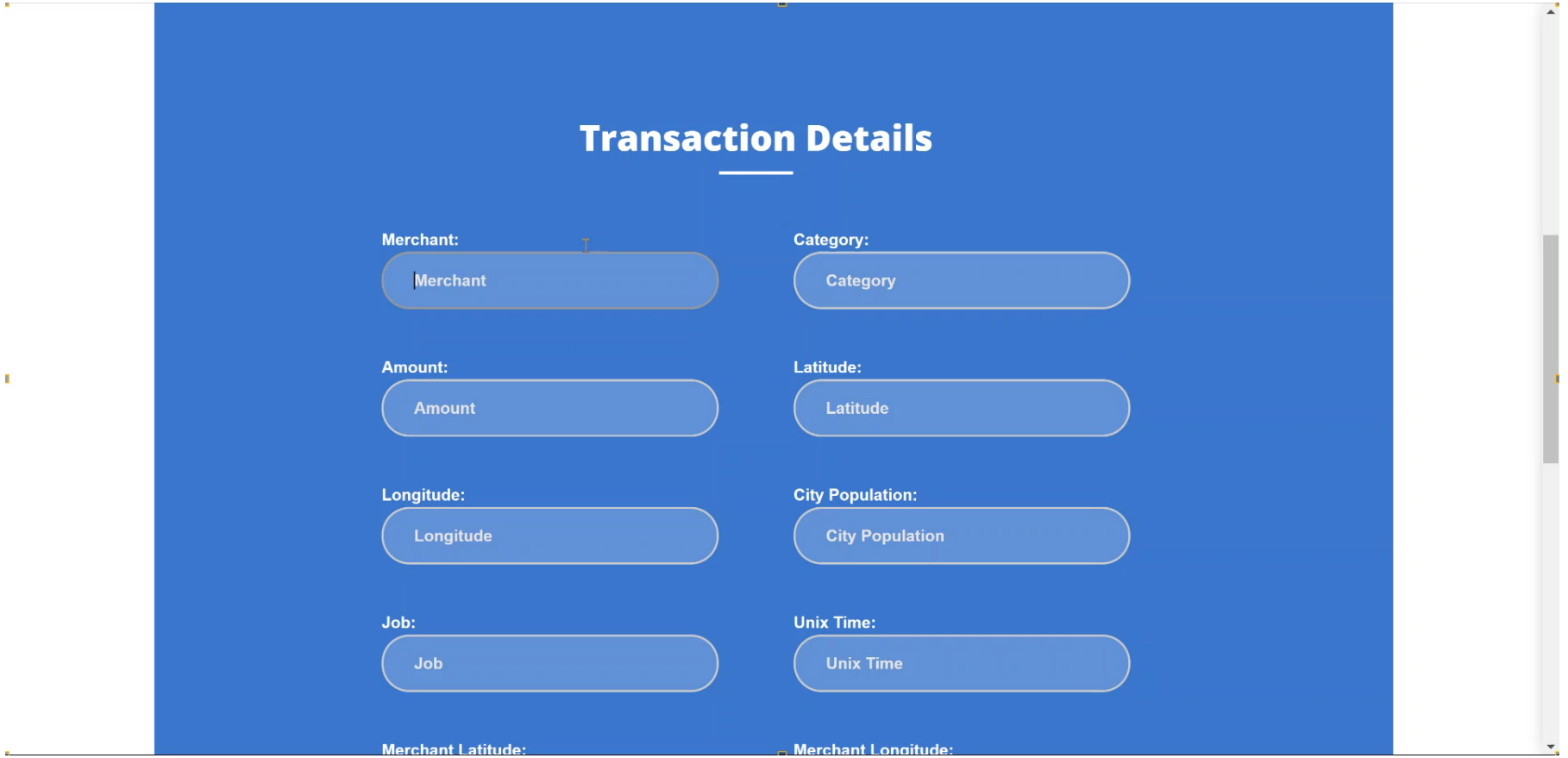


Machine Learning

Approach

Monitoring


DEMO


Transaction Details
Merchant:
Merchant
Category:
Category
Amount:
Amount
Latitude:
Latitude
Longitude:
Longitude
City Population:
City Population
Job:
Job
Unix Time:
Unix Time
Merchant Latitude:
Merchant Longitude:

Feature Engineering

Data

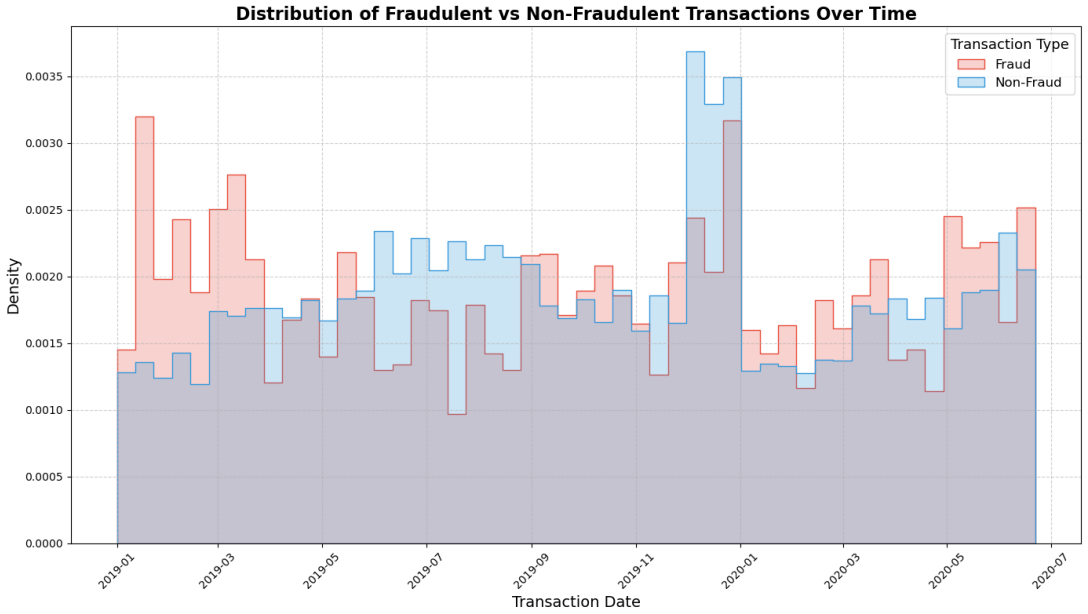
- Used simulated data
- 20 features (geo, address, personal)
- Training data (1 million plus rows)
- Test data (500k plus)

Feature Engineering

- Cleaning
- Feature extraction
- Handling categorical and numerical data
- Visualization
- Ethics

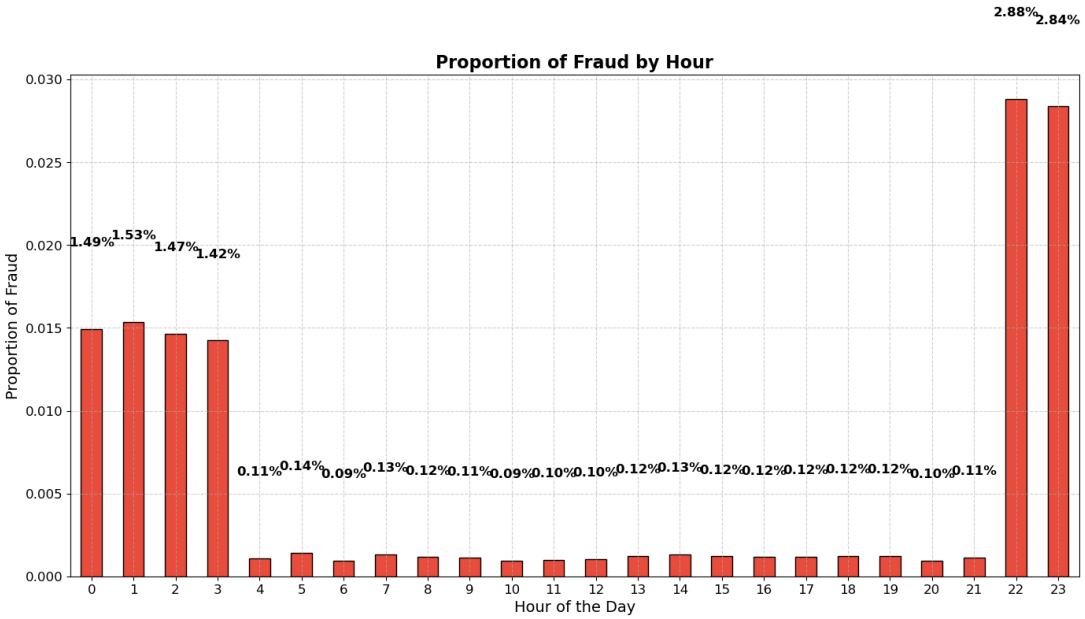
Distribution of Transaction (Fraud vs Non-Fraud)

The peaks and possible events



Fraud by Hour

Higher activities when users are typically in active.

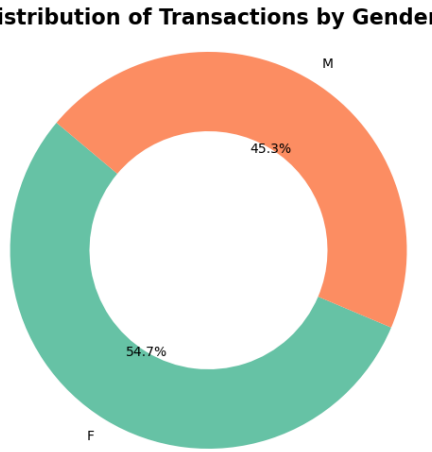


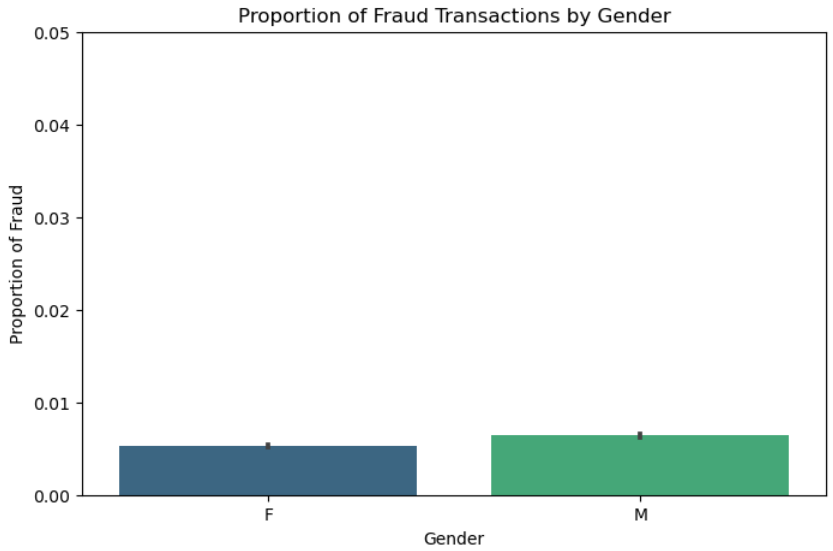
Top 10 Merchants

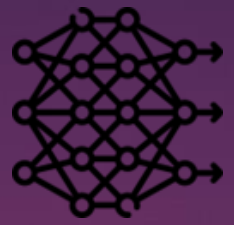
Do more investigation into this merchants.

Distribution of Transaction (Male vs Female)

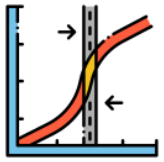
Even with females slightly edging but when it comes to fraudulent transactions males edge it.







Model Research

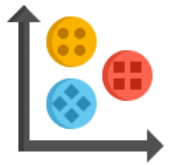


Linear

Gradient Boosting

Classifiers

Tree-Based



Clustering

Objective

Metrics

- Accuracy
- F1 (Recall & Precision)
- AUC-ROC

Considerations

- Speed
- Model interpretability
- Computational efficiency

Model Selection

Model	Precision	Recall	F1	Speed (seconds)
Logistic Regression	0.18	0.73	0.29	9.4
Decision Tree	0.72	0.74	0.73	99.5
Random Forest	0.94	0.69	0.80	1838.8
XGBClassifier	0.92	0.77	0.84	8.2
CatBoost	0.91	0.76	0.83	448.8

XGBClassifier

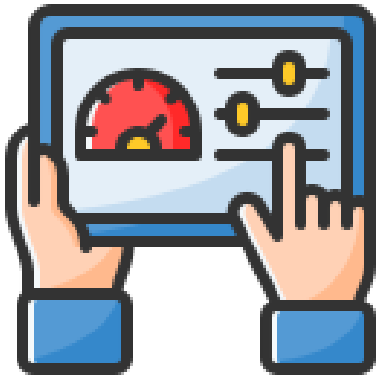
Model Tuning

Challenges

Computational Limitations
- CPU
- GPU

Approach
- Trial and Error

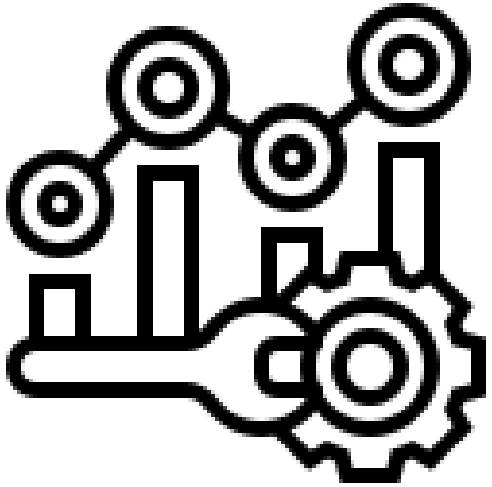
Parameters

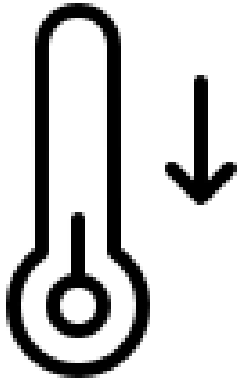


- Learning Rate: 0.6
- N_estimation: 400
- Max Depth: 8

Prediction Improvement

Threshold Optimization





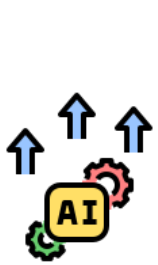
F1-Score improved from 0.86 to 0.88
Precision 0.92 to 0.90
Recall 0.84 to 0.86

Production and Deployment

Business Problem

Data Ingestion

Data Preprocessing & EDA



Model Research

Deployment & Monitoring

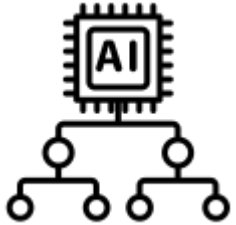
Key point:
This is an iterative process and not linear

Recommendation

More Data

- Category of fraud
- Payment method used
- Average monthly spend
- System information

Techniques



- Anomaly detection
- Behavioral analysis
- Clustering

So When Do I Start ;)


Q
A