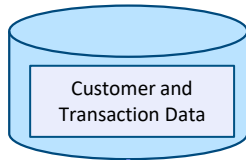




Group 5: Credit Card Fraud Prediction

Elizaveta Kotikova, Martyna Janina Kopyta, Péter Ferenc Gyarmati

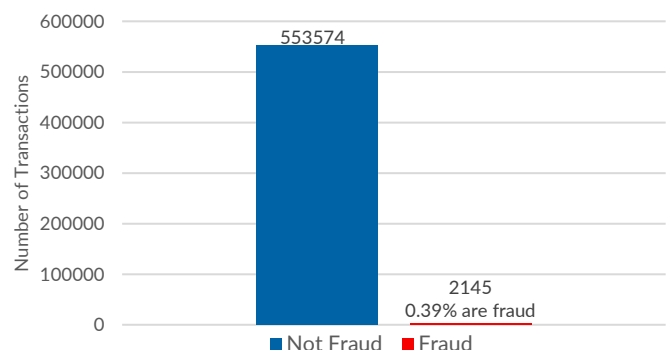
Dataset



Customer and Transaction Data

- 555,719 instances
- 22 attributes
- Mix of categorical and numerical data types
- No null values
- **Imbalanced** dataset

Fraud vs not Fraud count



Number of Transactions

553574 2145
0.39% are fraud

■ Not Fraud ■ Fraud

- Timestamp of the transaction
- Customer id number
- Merchant
- Category Transaction type
- Transaction amount
- Cardholder's first and last name
- Cardholder's address
- Latitude and Longitude of cardholder's location
- Population of the cardholder's city
- Cardholder's date of birth
- Unique transaction id
- Transaction timestamp...

You can find the documentation of the entire process here:



Challenge

Why is it challenging?

- High cost of errors
- Fraudsters evolution
- Real-Time detection required
- Personal data → privacy regulations

What simplifies the analysis?

- No domain knowledge required

Expected benefits :

- Reduced financial losses
- Increased trust and satisfaction
- Efficient fraud investigation

Data Understanding & Preparation

Data Preparation

- Data Cleaning
- Data Validation

Feature Engineering

Time-based features

- Hour of Day
- Day of Week
- Is Weekend

Transaction

- Amount
- Log Amount
- Distance from Merchant

Customer Features

- Age
- City Population
- Gender

Time Series Analysis

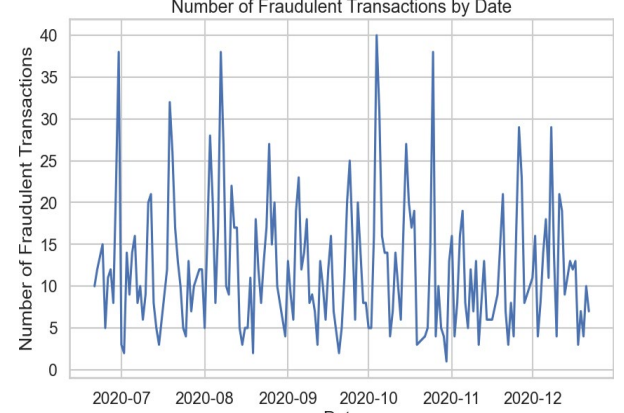
Components

- Stationary Tests
- Trend Analysis
- Seasonality Analysis

Models

- SARIMA MODEL

Number of Fraudulent Transactions by Date



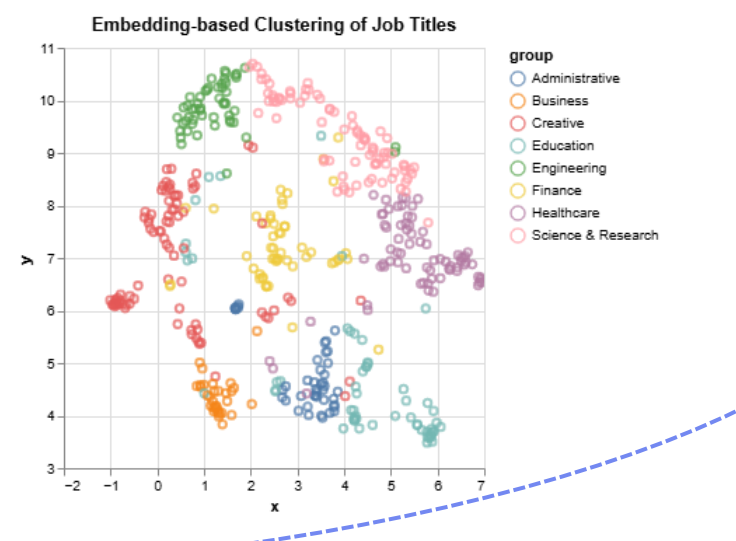
Job Title Processing via LLM

- Job Title Processing
- Nomic Embed + Qwen2.5-7B
- KMeans Clustering
- Job Groups

Correlation between Transaction Frequency and Fraud

	Tr.* per Day	Tr. per Week	Tr. per 2 Weeks	Tr. per Month	Is Fraud
Tr.* per Day	1.00	0.60	0.61	0.50	-0.00
Tr. per Week	0.60	1.00	0.90	0.76	-0.02
Tr. per 2 Weeks	0.61	0.90	1.00	0.83	-0.03
Tr. per Month	0.50	0.76	0.83	1.00	-0.03
Is Fraud	-0.00	-0.02	-0.03	-0.03	1.00

Embedding-based Clustering of Job Titles



Modeling

Classification

- 3 datasets: original, SMOTE-oversampled, and NearMiss-undersampled.
- Training of decision tree and random forest classifiers.
- → 6 models with varying Class-wise Performance.

Clustering

- Visualization of prepared features in two-dimensional space using UMAP.
- Exploration with KMeans, DBSCAN, and agglomerative hierarchical clustering.
- None were effective in distinguishing fraudulent transactions from legitimate ones.

Training

Sampling

- Original Data
- SMOTE Oversampling
- NearMiss Undersampling

Models

- Random Forest
- Decision Tree

Model Evaluation

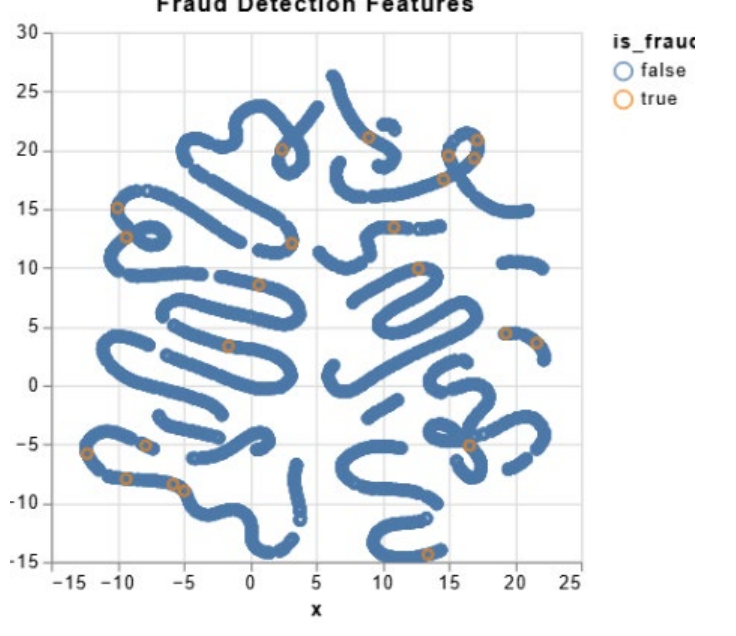
- Precision
- Recall
- F1 Score
- Confusion Matrix
- Feature Importance Evaluation

Final Model Selection

- Random Forest with NearMiss
- Optimized for Fraud Detection

Model	Overall Metrics	Feature Importance
Decision Tree, Original Samples	High accuracy, imbalance issues evident	Transaction amount, category, time
Random Forest, Original Samples	High macro precision, lower macro recall	Transaction amount, time
Random Forest, SMOTE Oversampling	Good macro precision, weaker recall, trade-off	Transaction amount, gender, merchant frequency
Random Forest, NearMiss Undersampling	Better recall and accuracy, limited precision	Transaction amount, time-related factors

Fraud Detection Features



is_fraud

- false
- true

Results

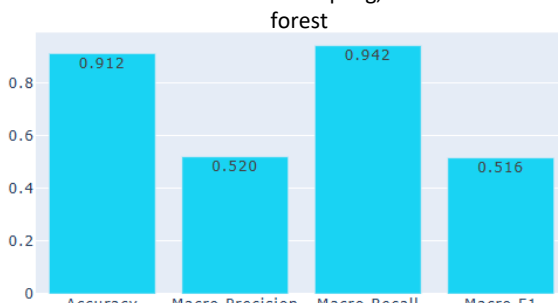
Random Forests: higher precision.

Decision Trees: simpler, faster, task-dependent.

SMOTE: improves recall by reducing false negatives.

NearMiss: balances errors, higher fraud detection with more false positives.

NearMiss-undersampling, random forest



Accuracy Macro Precision Macro Recall Macro F1

0.912 0.520 0.942 0.516

Conclusion

Clustering

Explored KMeans, DBSCAN, hierarchical clustering—none effectively distinguished fraud.

→ Shifted to supervised learning.

Classification

Random Forest or Decision Tree trained on the undersampled dataset with NearMiss might be the best choice.