OFFLINE & ONLINE ANOMALY DETECTION IN CREDIT CARD DATA

AMINA TYNYBEKOVA,
MATTHEW BALOGH

The Problem behind the Project

- Financial losses for companies
  Many banks and financial companies face the common and serious problem of fraud.
- Unjust charges to customers
  Many customers meet scammers or thieves and are losing all their savings.
- Delayed fraud detection can lead to long investigation processes and legal complications.
- High volumes of daily transactions make it difficult to manually monitor suspicious activity.

Challenges

Imbalanced data

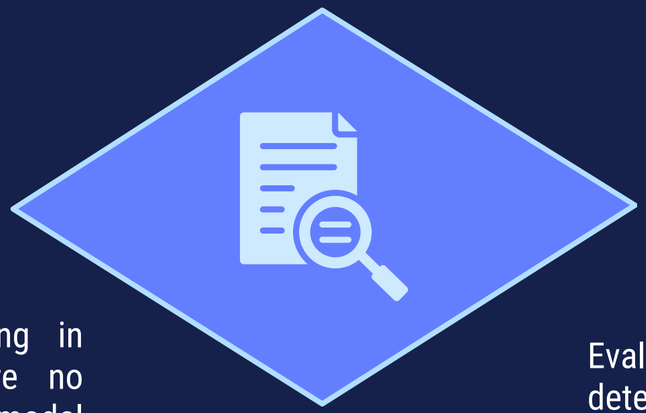
Imbalanced data makes fraud detection difficult because fraudulent transactions are very rare compared to normal ones.

Noise

Noise makes fraud detection harder because irrelevant or incorrect data can confuse the model and reduce its accuracy.

Unsupervised learning

Unsupervised learning is challenging in fraud detection because there are no labeled examples of fraud, so the model must identify suspicious patterns on its own.



Evaluation

Evaluation is challenging in fraud detection because accuracy alone can be misleading when the data is highly imbalanced.

Tools & Technologies

Python	Provides a flexible and powerful environment for data analysis, preprocessing, and implementing machine learning models efficiently.
Scikit-learn	Offers reliable and well-tested machine learning algorithms and evaluation tools for building and validating anomaly detection models.
RiverML	Enables real-time and online learning, allowing the model to detect fraud dynamically as new transaction data streams in.
Pandas	For efficient data cleaning, manipulation, and analysis of transaction datasets.
Numpy	For fast numerical computations and handling large arrays of transaction data.
Matplotlib	For visualizing data patterns, anomalies, and model performance.
Streamlit	For building an interactive web application to demonstrate and monitor the fraud detection model.

Plan and Expectations

Phase I: Offline Learning

Data Loading & Exploration (Python, Pandas)

- Load transaction dataset
- Analyze class distribution (fraud vs normal)
- Check missing values and data types

Data Preprocessing (Pandas, NumPy)

- Handle missing or noisy data
- Remove duplicates
- Address extreme imbalance (if supervised baseline is tested)

Model Implementation (Scikit-learn)

- Train unsupervised model (e.g., Isolation Forest)

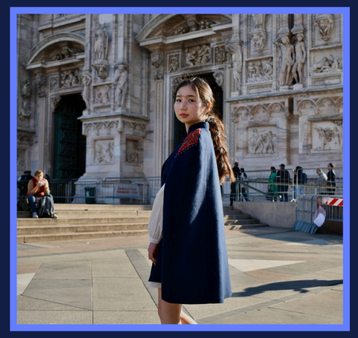
Model Interpretation

- Analyze anomaly scores
- Identify which features influence anomaly detection
- Investigate false positives and false negatives

Phase II: Online Learning

- Online counterparts of algorithms in RiverML
- Observe and visualize detection and performance over time
- Compare final performance with offline version

TEAM



Amina Tynybekova
Project Co-Lead

Matthew Balogh
Project Co-Lead

Open Positions (3):

- Offline modeling
- Online modeling
- Visualization

THANK YOU!