Transaction Fraud Detection System

Heuristic Rules and Simple Machine Learning (Logistic Regression)

Anouk Moreno

January 2026

Abstract

This project presents the design and implementation of a prototype transaction fraud detection system inspired by real-world financial anti-fraud architectures. The system follows a two-layer architecture: first, an interpretable rule-based baseline using transaction analysis heuristics (amount anomaly, velocity, impossible travel, and unusual hours); second, a simple machine learning model compares detection performance and assesses the trade-off between false positives and false negatives through threshold tuning.

Repository: https://github.com/anouk-moreno/fraud-detection-system

Contents

Chapter 1

Introduction

1.1 Background and Motivation

From paying for a morning coffee to ordering something online at night, we constantly move money throughout the day. Such operations are referred to as **financial transactions**, and in most cases, they are perfectly normal. Yet, things become interesting when a transaction does not match a user's regular behavior. This can occur for various reasons, such as actual fraud (e.g., a compromised card), a technical error, or simply an exceptional situation. For this reason, real-world systems refer to **suspicious transactions** rather than confirmed fraud. The goal is not to immediately "condemn" a transaction, but to detect signs of risk and determine whether additional verification is needed.

Financial institutions and payment platforms process a high volume of transactions every day, making manual review slow, expensive, and often ineffective. Automated detection systems therefore act as a first filter, assigning risk scores or generating alerts so that only a small subset requires closer inspection.

In recent years, digital fraud and automated abuse have increased in complexity and speed. For example, when a card is compromised, it is common for multiple transactions to be attempted within a few minutes, or for small amounts to be tested before larger charges are executed. In this context, early detection and prioritization of alerts become critical, as they reduce losses and protect the user without unnecessarily interrupting legitimate transactions.

This personal project aims to build a simple, explainable prototype inspired by early-stage anti-fraud systems, using behavioral signals derived from basic transaction data. Beyond addressing a realistic technical problem, it reflects a personal interest in how financial systems operate and how data-driven methods support risk detection in fintech environments.

1.2 Problem Statement

The problem this project tackles is how to flag **suspicious financial transactions** using only basic transaction features (such as timestamps, amounts, locations, or merchants). Rather than assuming that every anomaly corresponds to fraud, the system focuses on identifying deviations from a user's typical behavior that warrant further inspection or verification.

The system is not intended to make definitive decisions or automatically block transactions, but to act as a lightweight **risk prioritization mechanism**. Given a set of transactions, the system highlights those that appear most anomalous, allowing subsequent processes (manual or automatic) to determine the appropriate response. In this sense, the system provides an initial filtering layer on top of which more advanced mechanisms can later be developed.

The problem is addressed under the following constraints:

- The system only has access to basic transaction data.
- Decisions must be interpretable and explainable.
- A high-volume processing environment is assumed.

Under these conditions, the main challenge is to balance effective detection of suspicious behavior with the minimization of false alarms. Excessive alerts increase friction and operational cost, while insufficient detection allows risky activity to remain unnoticed.

1.3 Project Objectives

The overall objective of this project is to design and implement a simple and explainable prototype for detecting suspicious transactions. The system does not aim to definitively identify fraud, but rather to prioritize risk based on behavioral signals, making it easier to review transactions that deviate from a user's common pattern.

To achieve this objective, the following specific goals are defined:

1. Design a reduced set of behavioral signals derived from basic transaction information, capable of capturing relevant anomalous patterns.
2. Implement a detection system based on heuristic rules that acts as a filtering layer, generating understandable risk indicators and an aggregate risk score for each transaction.
3. Evaluate the performance of the rule-based approach using standard metrics (precision, recall, and confusion matrix), analyzing the balance between effective detection and false alarm generation.
4. Explore the use of a simple machine learning model (Logistic Regression) as a complement to the rule-based system, in order to assess whether combining multiple signals improves detection while preserving interpretability.

5. Systematically compare the purely rule-based approach with the hybrid approach that incorporates ML, highlighting the advantages, limitations, and trade-offs of each in terms of performance and complexity.

Together, these objectives aim not only to build a functional system, but also to clarify the design decisions involved in detecting suspicious transactions. They also reflect on how simple approaches can gradually scale toward more advanced solutions in real-world settings.

1.4 Scope and System Boundaries

The scope of this project is deliberately limited to detecting suspicious financial transactions based on simple behavioral signals and basic data. The system is designed as an educational and exploratory prototype that reflects the early detection layers found in real-world anti-fraud systems, while prioritizing clarity and reproducibility.

Specifically, the project focuses on behavioral signals derived from the following attributes:

- the transaction amount,
- the frequency of transactions per user,
- the geographical location of the transaction,
- and the time of day when the transaction takes place.

These signals form a common basis in real-world early risk detection systems, as they can be computed efficiently and easily explained to analysts or end users.
In contrast, more advanced techniques and data sources, such as device fingerprinting, network or IP address reputation, behavioral biometrics, graph-based models, or sequential deep learning approaches, are intentionally excluded. Although widely used in production environments, these techniques require larger data volumes, complex infrastructures, and demanding training processes, which are beyond the design boundaries.

Several key simplifications are also made. First, a simulated dataset is used, allowing the types of anomalies to be controlled and the system's performance to be evaluated in a reproducible and objective manner. Second, user behavior patterns are assumed to remain relatively stable over time, without explicitly modeling phenomena such as concept drift. Finally, the system is evaluated as an isolated component, without integration into a complete payment authorization or transaction blocking pipeline.

These decisions are not intended to reflect the full complexity of a production-grade anti-fraud system, but rather to establish a clear and understandable basis for analyzing the fundamental principles of suspicious transaction detection. From this foundation, the approach can be extended to more realistic and complex scenarios, as discussed in later sections.

1.4.1 Data and Modeling Assumptions

Furthermore, access to real financial transaction data is often restricted due to privacy, security, and regulatory constraints. As a result, many academic studies and experimental prototypes rely on simulated datasets designed to reproduce realistic behavioral patterns without compromising user confidentiality.

This project explicitly operates under two conditions: an imbalanced data environment and the controlled data setting described above. These assumptions directly influence system design decisions, the selection of evaluation metrics, and the interpretation of the resulting performance.

1.5 General Approach

The approach adopted in this project follows a layered and incremental design, inspired by how many real-world fraud detection systems are built in practice. Rather than relying immediately on complex models, the overall pipeline starts with simple, transparent mechanisms and progressively incorporates more flexible techniques where appropriate.

Initially, the project implements a rule-based detection layer grounded in domain intuition and behavioral heuristics. Each rule captures a specific type of anomalous behavior, such as unusually high transaction amounts, bursts of transactions in short time windows, rapid changes in geographical location, or activity occurring at atypical hours for a given user. These rules operate independently and produce understandable signals that can be easily inspected. The outputs of the individual rules are then combined into an aggregated risk score. This score does not represent a probability of fraud, but rather a relative measure of risk that allows transactions to be ranked and prioritized. By introducing a scoring mechanism instead of binary decisions, the system reflects real operational settings in which alerts are reviewed according to urgency and available resources.

Building on this rule-based baseline, a simple machine learning model is subsequently introduced. The role of the machine learning component is not to replace the rules, but to learn how multiple signals interact and to assess whether a data-driven combination of features can improve detection performance. A Logistic Regression model is chosen due to its simplicity, transparency, and suitability for imbalanced classification problems.

Throughout the project, special emphasis is placed on interpretability, reproducibility, and realistic evaluation. The architecture is assessed using time-based train–test splits and standard performance metrics, and model behavior is analyzed through confusion matrices and precision–recall trade-offs. This structured approach ensures that each design decision can be clearly motivated and that the resulting system remains understandable at every stage.

1.6 Data Strategy

The design of the dataset plays a central role in this project, as it directly influences both the detection logic and the evaluation of the proposed system. In real-world financial settings, access to transaction-level data is typically restricted due to privacy, security, and regulatory constraints. As a consequence, developing and testing experimental fraud detection systems often requires the use of simulated or synthetically generated data.

For this case, a simulated dataset is deliberately chosen. This decision allows full control over the data generation process, including the definition of user profiles, transaction patterns, and anomalous behaviors. By simulating data, it becomes possible to explicitly model realistic scenarios, such as unusually high transaction amounts, bursts of activity, rapid geographical changes, or atypical transaction times, while maintaining complete transparency regarding which transactions are considered suspicious.

Another important advantage of using simulated data is reproducibility. Because the data generation process is fully specified and parameterized, experiments can be repeated under identical conditions, and the impact of individual design choices can be analyzed systematically. This is particularly relevant when comparing different detection approaches, such as rule-based systems and machine learning models, where consistent evaluation conditions are essential.

The simulated dataset is designed to reflect several characteristics commonly observed in real transaction data. First, it exhibits class imbalance, with suspicious transactions representing only a small fraction of the total volume. Second, user behavior is heterogeneous: different users show distinct spending habits, preferred time windows, and geographical patterns. Third, anomalous behaviors are injected in a controlled manner, enabling a clear assessment of whether the system successfully captures the intended deviations.

At the same time, the limitations of simulated data are explicitly acknowledged. The dataset cannot fully capture the noise, unpredictability, and evolving nature of real-world financial behavior. Phenomena such as concept drift, coordinated fraud across multiple users, or subtle long-term behavioral changes are not modeled. As a result, the dataset is not intended to replace production data, but rather to serve as a controlled environment for studying the fundamental principles of suspicious transaction detection.

In summary, the data strategy adopted in this project prioritizes clarity, control, and reproducibility over realism at all costs. This choice aligns with the educational and exploratory nature of the project and provides a solid foundation for analyzing detection mechanisms before considering more complex data sources and modeling strategies.

Bibliography