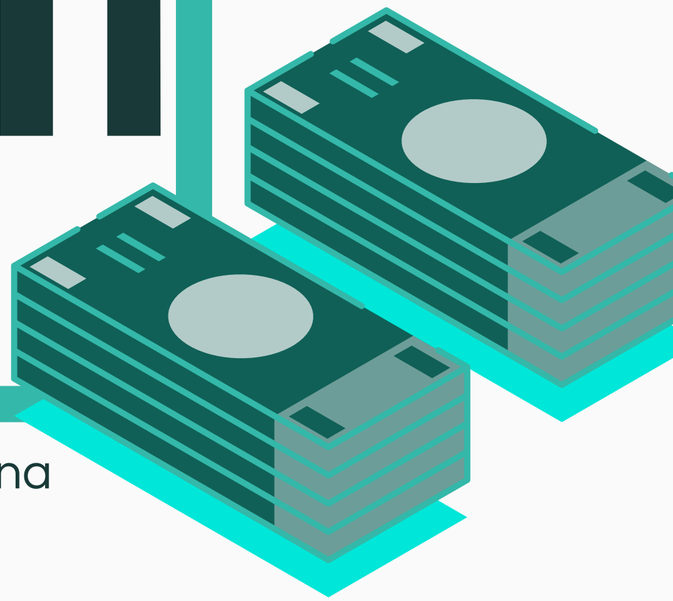
Loan Default Prediction

A Systems Engineering Approach



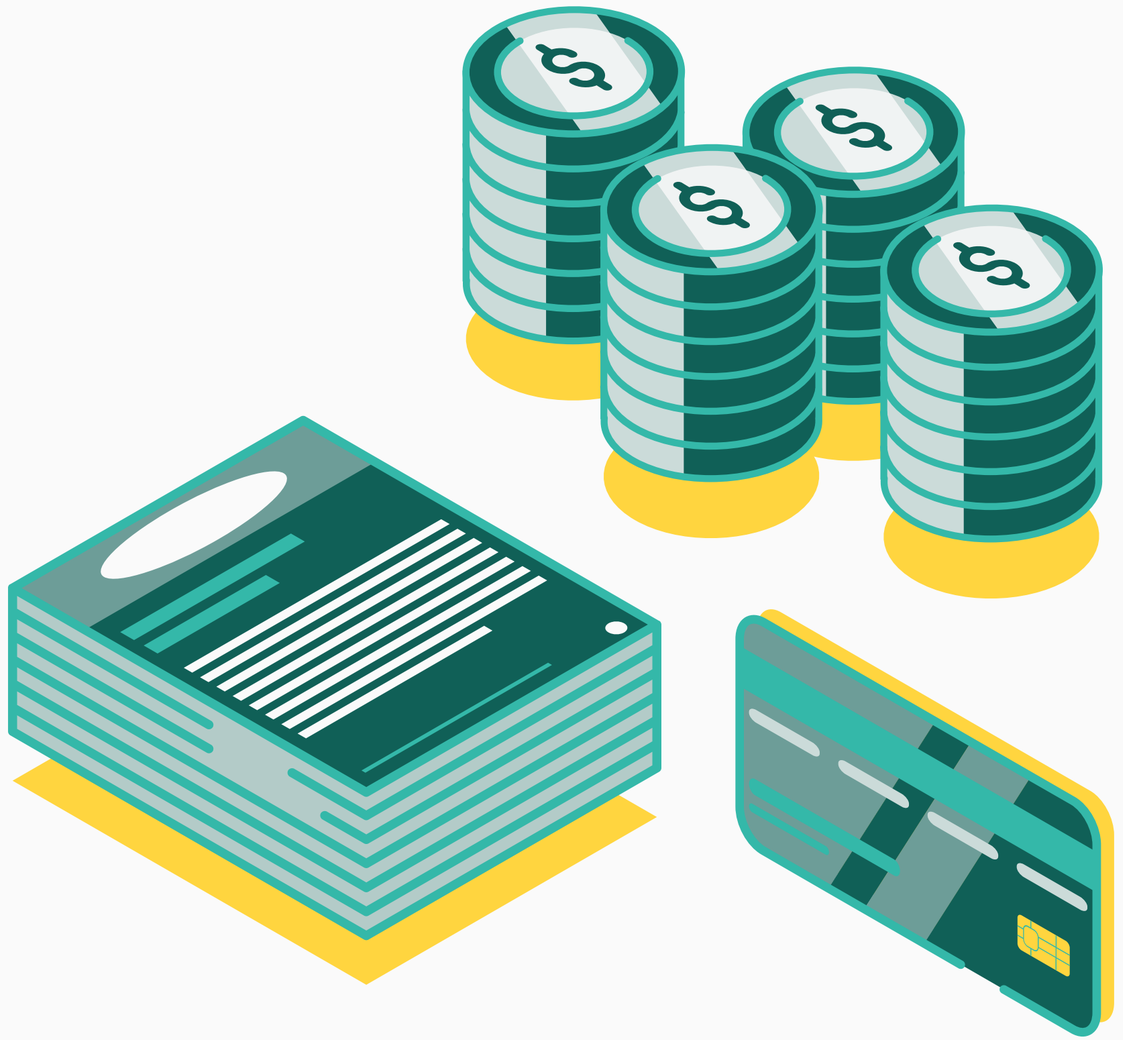
Juan Jose León Gomez

Miguel Angel Hernandez Medina

Juan Pablo Diaz Ricaurte

Juan Esteban Avila Trujillo

INDEX



Introduction

Overview

Traditional risk management is often insufficient

Challenges

Mitigating the inherent system instability

Motivation

Design a system capable of performing a dual prediction

Approach

Systems Engineering principles

SYSTEMATIC ANALYSIS OF THE PROBLEM

1. System Elements

- Financial clients, transactions, and credit histories form a dynamic system where each component influences others.

2. Interdependencies

Variables like income, debt, and payment behavior are deeply connected — a small change in one alters the whole structure.

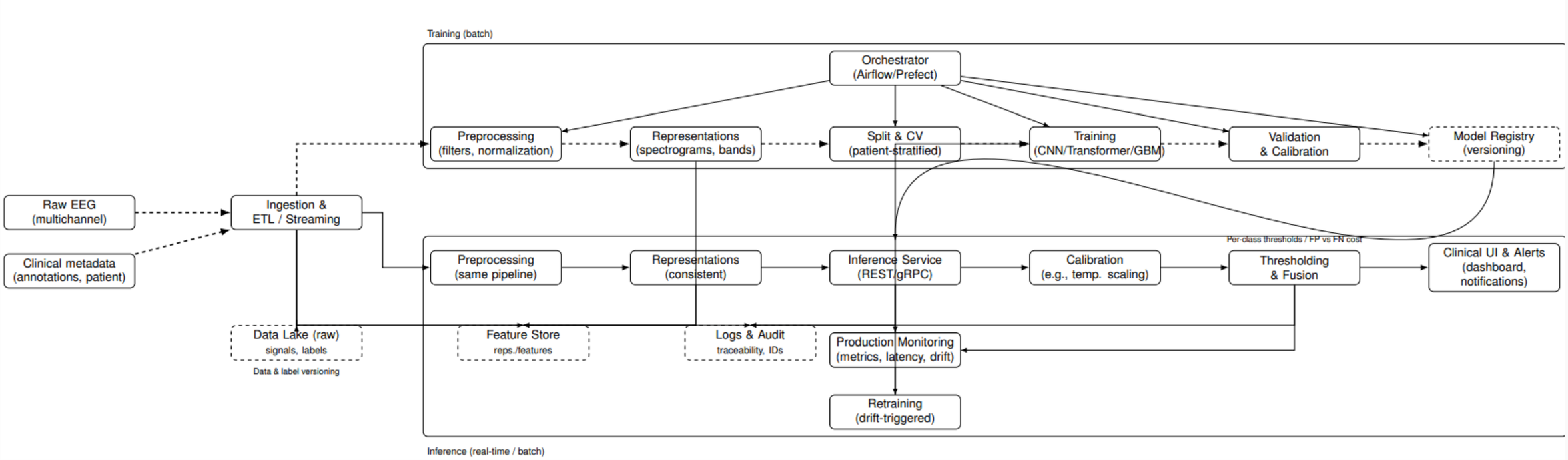
3. Sensitivity and Chaos

- The system shows nonlinear behavior: minor data variations can lead to large prediction shifts, reflecting chaotic dynamics.

4. Systemic Risk

- When unstable patterns emerge in data, errors amplify through the model, producing unreliable or biased outcomes.

System Design



Quality and risk analysis

Focus on system stability, reliability, and security

- Identified key risks: data drift, bias, security breaches, pipeline failures
- Mitigation strategies: monitoring, validation, retraining, encryption
- Quality assurance: unit tests, integration tests, model versioning

Risk	Impact	Mitigation
Data drift	High	Monitoring + Retraining
Bias	Medium	Fairness metrics
Security	Critical	Encryption + RBAC
Pipeline failures	Medium	Fault tolerance

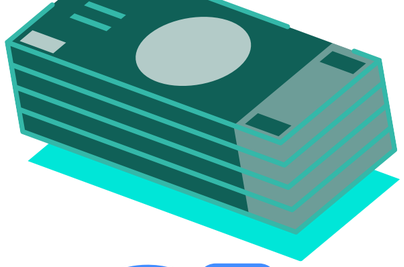
Dataset & Preprocessing

OLD

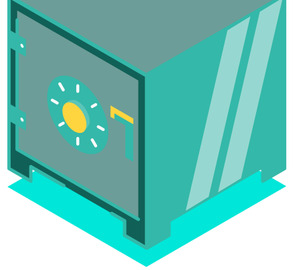
NEW

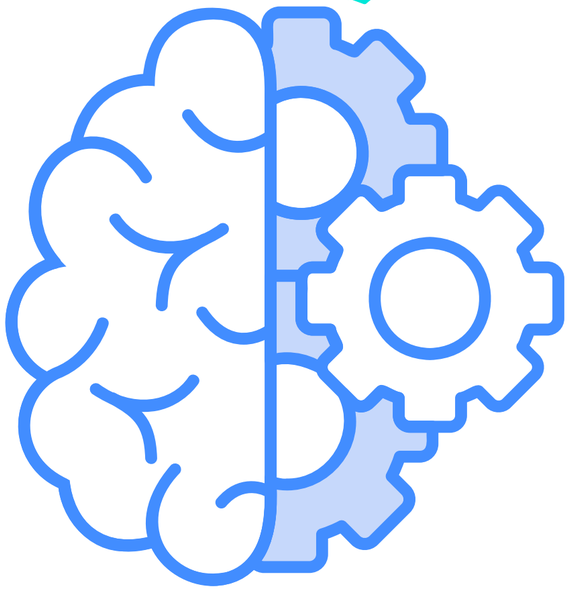
- MISSING DATA
- RAW DATA
- HIGH UNCERTAINTY AND ENTROPY
- MASSIVE VOLUME OF DATA

- ANOMALIES HANDLED
- DIMENSIONALITY REDUCTION
- FEATURE ENGINEERING
- CLEANING AND NORMALIZATION



Built Models



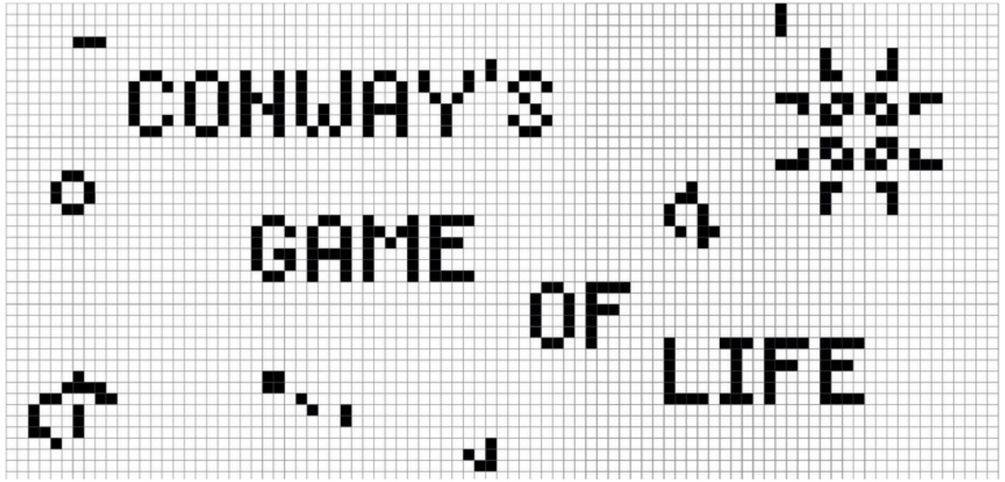


MACHINE LEARNING MODEL

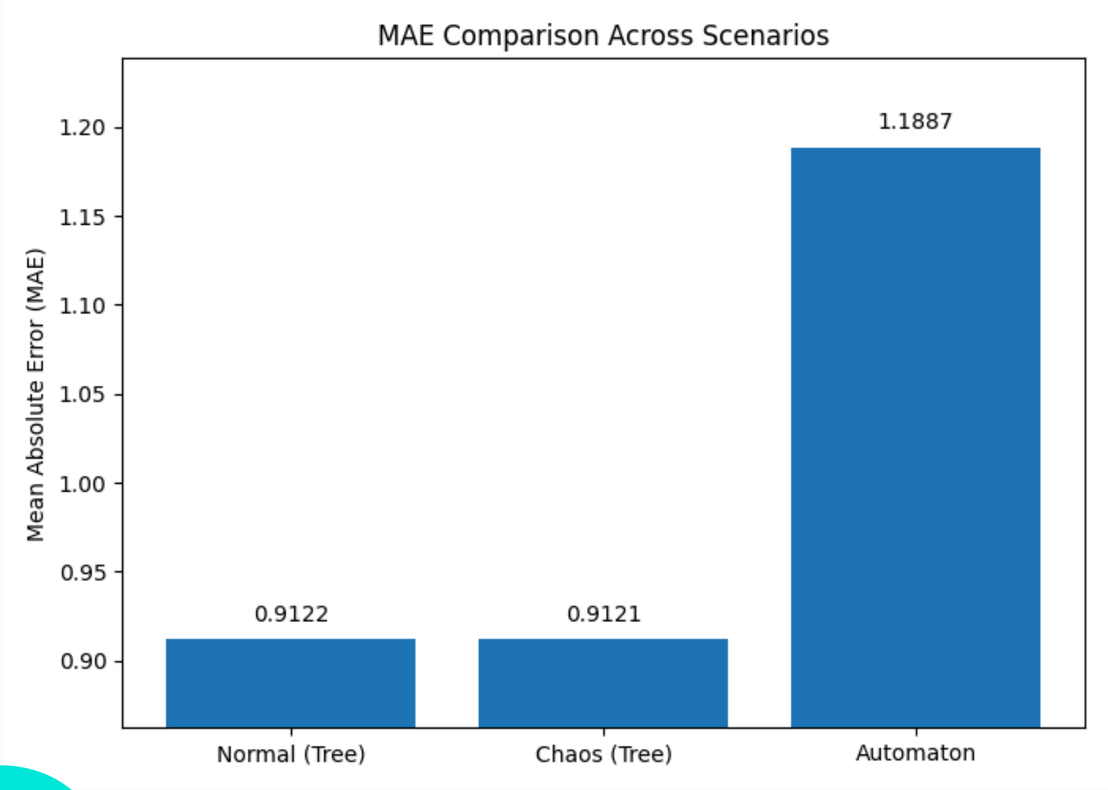
- TREE ENSEMBLE
- PREDICTS: DEFAULT + LOSS
- MAE ≈ 0.91
- HIGH STABILITY UNDER DISTURBANCES

RANDOM FORESTS

- MODELS RISK PROPAGATION
- CAPTURES NONLINEAR AND EMERGENT BEHAVIORS
- MAE ≈ 1.18
- EIGENVALUE PERSISTENCE=0.8
- PROBABILITY OF CONTAGION=0.6



Results



- MAE in different scenarios:

  Normal (Tree): 0.9122
  Chaos (Tree): 0.9121
  Automaton: 1.1887

- Key findings:

  Tree model = extremely stable
  Chaos barely affects performance
  Automaton captures nonlinear patterns

CONCLUSIONS

- A robust data processing pipeline is essential when working with anonymized and high-dimensional datasets. Ensuring consistency between training and test sets significantly improves model reliability.
- Gradient boosting models like LightGBM and CatBoost demonstrated strong performance and adaptability, even in the absence of domain-specific feature names or labels.
- Despite the challenges, the final solution achieved competitive accuracy by focusing on systematic preprocessing, model tuning, and error minimization under the MAE metric.

THANK You!