SmartRisk: Hybrid Prediction–Simulation Framework for Credit Default Analysis

Final Course Project Report

Juan Pablo Galindo Flórez - 20231020230
Stevens Camilo Llanos Acero - 20231020221
Universidad Distrital Francisco José de Caldas

2025

Contents

1 Introduction

Credit risk analysis is a fundamental component of modern financial systems. Institutions rely on statistical and machine-learning models to estimate the probability that a customer will default on a financial obligation. However, traditional prediction models offer only an individual-level perspective and often fail to capture systemic interactions, cascading failures, and chaotic behavior that can emerge under stress conditions. SmartRisk was developed as a hybrid analytical framework that combines the predictive power of machine learning with the emergent behavioral insights of a cellular automaton (CA).

This report integrates the complete development process from Workshops 1 through 4, including system analysis, high-level architecture, modeling, simulation, experiments, and conclusions. The early stages focus on understanding the American Express Default Prediction problem, identifying complexity sources, and exploring stakeholders and flows. Subsequent stages progressively build and validate the SmartRisk architecture, implement synthetic data generation, train a Random Forest classifier, execute systemic simulations, and analyze results.

The hybrid approach enables a comprehensive evaluation of credit-risk behavior at both micro and macro levels, addressing prediction accuracy, system resilience, and behavioral sensitivity. The results show that although the machine-learning model yields limited predictive performance due to class imbalance and weak signal structure, the CA simulation reveals clear systemic transitions, validating the importance of simulation components in credit-risk assessment.

2 System Analysis (Workshop 1)

2.1 Problem Context

The original problem, the American Express Default Prediction challenge, focuses on predicting whether a customer will default using a high-dimensional, time-dependent dataset. Stakeholders include financial institutions, customers, regulatory bodies, and algorithm designers. The system requires robust decision-making under uncertainty, balancing business goals with risk management responsibilities.

2.2 Key Elements

Key components identified during the analysis include:

- Customer financial history and behavior
- Data preprocessing and feature engineering
- Predictive model outputs
- Institutional decision-making policies
- Feedback loops between predictions and customer behavior

2.3 System Interactions

The interactions among these elements create a dynamic network of influence. Customer spending affects credit utilization, which influences risk scores produced by the model. These scores feed into decisions such as credit limits or interest rates, which influence future customer behavior. This feedback loop contributes to emergent nonlinear behavior.

2.4 Complexity and Sensitivity

The system exhibits complexity due to:

- large feature spaces and temporal patterns,
- imbalanced default outcomes,
- nonlinear interactions among variables,
- path-dependent and chaotic dynamics.

Sensitivity analysis during Workshop 1 showed that small variations in financial history or model parameters can produce disproportionately large shifts in predicted default probability. Such characteristics motivate the integration of simulation tools in later workshops.

2.5 Workshop 1 Conclusion

Workshop 1 concluded that the system requires a hybrid perspective. Prediction alone cannot fully address systemic risk emergence, thus the SmartRisk architecture must incorporate simulation and monitoring subsystems.

3 System Design (Workshop 2)

3.1 High-Level Architecture

Based on the analysis, SmartRisk was designed with four main subsystems:

- **Data Ingestion & Preprocessing** – Handles input data, cleaning, and structuring.
- **Feature Engineering & Model Training** – Extracts behavioral features and trains predictors.
- **Simulation & Stress Testing** – Implements the CA to study systemic propagation.
- **Monitoring & Drift Detection** – Oversees model performance and data stability.

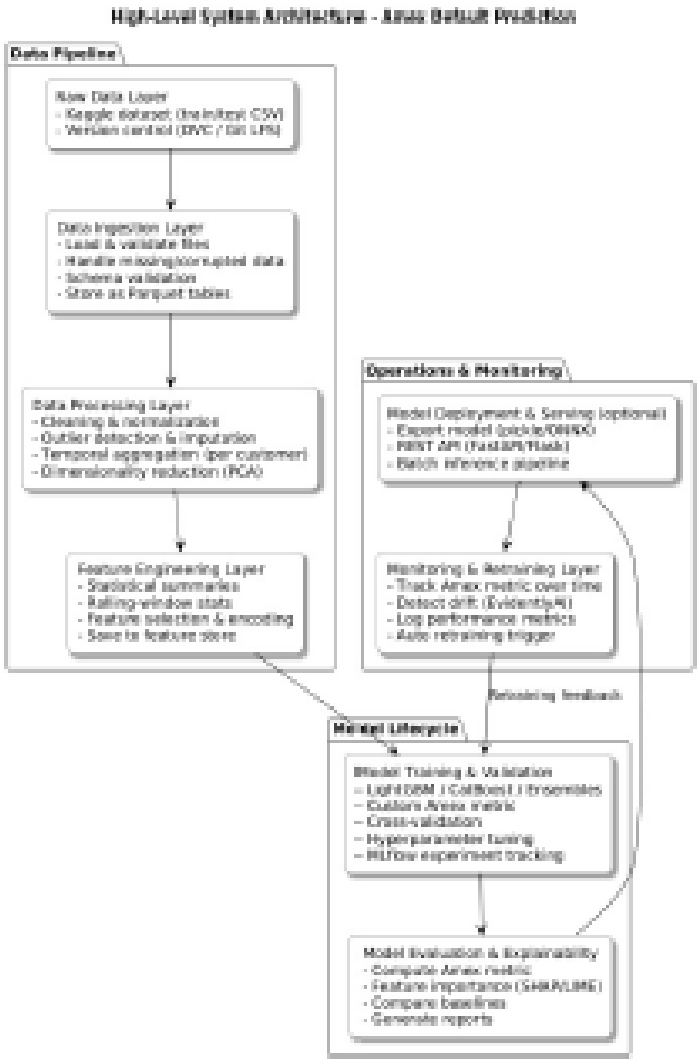

Figure 1: High-Level System Architecture for the Amex Default Prediction Project

Figure 1: SmartRisk High-Level Architecture (placeholder image)

3.2 Engineering Principles

The architecture leverages core engineering concepts:

- **Modularity** between modeling and simulation components.
- **Abstraction** for model interfaces.
- **Iterative refinement** to enable extension in later workshops.
- **Scalability** to accommodate larger datasets.
- **Lifecycle design** including drift detection and retraining.

3.3 Addressing Chaos and Sensitivity

Mechanisms to mitigate sensitivity include:

- monitoring data distributional shifts,
- using thresholds for retraining triggers,
- simulating high-stress propagation scenarios,
- implementing feedback loops to stabilize the system.

3.4 Workshop 2 Conclusion

The architecture is flexible and robust, prepared to incorporate machine-learning and simulation modules. Workshops 3 and 4 build on this foundation to produce the full implementation.

4 Modeling and Simulation (Workshop 3)

4.1 Synthetic Dataset Generation

A synthetic dataset was created to emulate customer financial behavior. For each of the 20,000 customers, multiple statistical features were generated, including:

- average balance and spending,
- volatility,
- temporal trends,
- credit utilization.

Default probability was computed using a logistic model combining financial behavior, noise, and interaction terms. The final dataset included engineered features and binary default labels.

4.2 Machine Learning Pipeline

The ML pipeline includes:

- Standardization of all numerical features,
- Train-test split with stratification,
- Random Forest classifier with 200 trees and max depth 10,
- Evaluation using ROC AUC and PR AUC.

4.3 Cellular Automaton Simulation

A 60x60 CA grid was used to model contagion-like behavior. Each cell represents a customer in a binary state: defaulted (1) or stable (0). CA dynamics are governed by:

- local neighborhood influence,
- global economic stress parameter,
- random noise.

Each time step updates the grid based on the probability of transitioning to default, allowing analysis of systemic transitions.

5 Workshop 3 Conclusion

Workshop 3 successfully implemented the core ML pipeline and CA simulation, setting the foundation for the experiments performed in Workshop 4.

6 Experiments (Workshop 4)

6.1 Machine Learning Experiments

The Random Forest model produced:

- ROC AUC: **0.500**
- PR AUC: **0.032**
- Accuracy: inflated due to class imbalance.

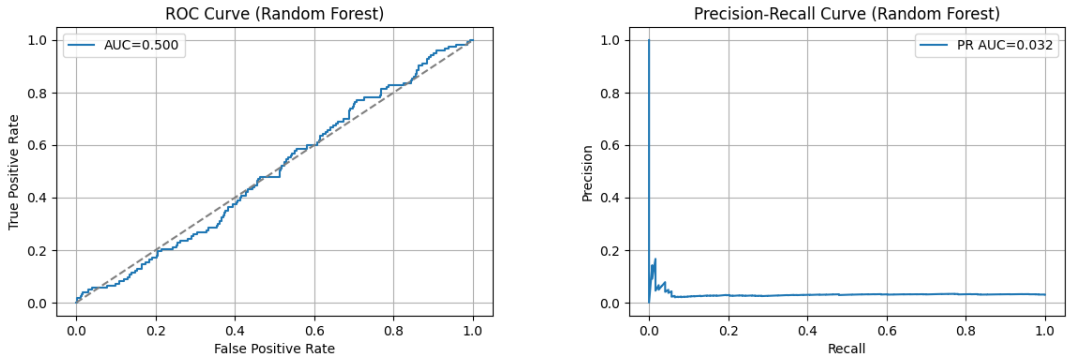

Figure 2: ROC and PR Curves of the Random Forest Model

Feature importance analysis revealed weak separability, confirming that the synthetic dataset does not embed strong default patterns.

6.2 CA Simulation Experiments

The CA simulation demonstrated systemic behavior absent from the ML predictions. As stress increased, small clusters of defaults expanded rapidly, producing full-grid contagion.

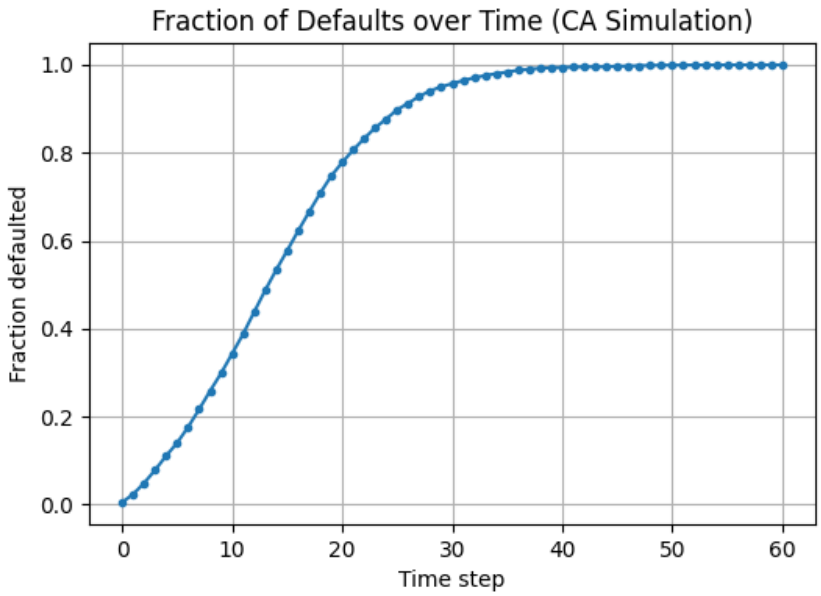

Figure 3: Fraction of Defaulted Customers Over Time in CA Simulation

6.3 Comparative Analysis

While the ML model fails to distinguish defaulters (ROC = 0.500), the CA reveals clear nonlinear transitions, supporting the inclusion of simulation for system-level robustness assessment.

7 Results and Discussion

Results show a strong contrast between predictive performance and systemic insights. The Random Forest classifier was unable to extract meaningful predictive signals due to weak feature separability and class imbalance. On the other hand, the CA displayed rapid and interpretable cascade effects, indicating sensitivity to small local disturbances under moderate stress.

The combination of these two tools validates the SmartRisk architecture: predictors estimate individual-level tendencies, while simulation captures systemic risk amplification.

8 Conclusion

SmartRisk successfully demonstrates a hybrid approach to credit-risk analysis, integrating machine-learning and simulation paradigms. Workshops 1 and 2 provided the conceptual and

architectural foundation, while Workshops 3 and 4 implemented full modeling, simulation, and experimentation pipelines.

Despite the low performance of the Random Forest classifier, the CA simulation showed valuable emergent behaviors and systemic transitions. Together, these components provide a more complete understanding of credit-risk dynamics and validate the importance of hybrid risk analysis frameworks.

9 References

References

[1] L. Breiman, "Random forests," Machine Learning, 2001.

[2] S. Wolfram, *Cellular Automata and Complexity*, Addison-Wesley, 1994.

[3] American Express Default Prediction, Kaggle, 2022.