Systems Analysis of the American Express Default Prediction Problem

Stevens Camilo Llanos Acero
Dept. of Computer Engineering
Universidad Distrital Francisco José de Caldas
Email: scllanosa@udistrital.edu.co

Juan Pablo Galindo Florez
Dept. of Computer Engineering
Universidad Distrital Francisco José de Caldas
Email: jpgalindof@udistrital.edu.co

***Abstract*—Financial institutions face the challenge of accurately predicting customer default risk to maintain portfolio stability and reduce losses. This study analyzes the American Express – Default Prediction competition through a systems engineering approach, identifying key components, sensitivities, and potential chaotic behaviors in financial data. The proposed analysis highlights the system's complexity, the importance of data-driven design, and strategies to improve model robustness under dynamic economic conditions.**

***Index Terms*—Systems Engineering, Financial Risk Prediction, Machine Learning, Sensitivity Analysis, Chaos Theory, Default Modeling, Data Systems, American Express Competition**

I. Introduction

Credit default modeling has long been a central challenge in financial risk analysis, traditionally approached through statistical classification techniques applied to consumer credit behavior [1]. Over the years, advances in machine learning have significantly expanded the range of available methods, including ensemble models such as Random Forests [2] and deep neural architectures [3]. Despite these developments, credit risk remains difficult to model due to high class imbalance, temporal variability, and limited signal strength in consumer-level financial data. This challenge is further amplified by rapidly evolving economic conditions and nonlinear interactions between customer behaviors.

The American Express Default Prediction problem [7] illustrates this complexity, as it requires inferring default probabilities from multi-month financial summaries. Purely supervised approaches often struggle in such environments because they depend heavily on the statistical information encoded in the features. Classical theories of information, such as Shannon's foundational work [4], highlight how limited or noisy data can constrain predictive performance. More recent studies emphasize the need for explainability in credit modeling [5] and the importance of understanding the sensitivity of financial prediction systems to varying data distributions [6].

Beyond individual-level classification, financial systems exhibit systemic behaviors in which local shocks can propagate through interconnected structures, leading to cascading failures. Approaches such as cellular automata [8] offer a way to model these emergent phenomena and complement traditional machine learning techniques. Additionally, prior work has shown that macroeconomic variables can significantly shape default patterns at the population level [9]. These insights motivate the development of hybrid frameworks capable of capturing both micro-level prediction and macro-level contagion dynamics.

In this context, this work introduces SmartRisk, a hybrid simulation architecture that integrates supervised learning with event-based default propagation modeling. The goal is to characterize both predictive accuracy and systemic vulnerabilities under controlled experimental scenarios, providing a more holistic understanding of credit risk dynamics.

II. Methods and Materials

The **American Express – Default Prediction** competition provides a real-world environment for applying systems analysis principles to a large-scale financial dataset. The competition aims to predict the probability that a customer will default on future credit payments, using anonymized historical transaction data. The data represent aggregated financial activity at the customer level, including temporal and behavioral patterns related to spending, balance management, and payment history. Each customer is identified by a unique code, ensuring privacy while maintaining structural consistency across the dataset.

a) ***Dataset Description:*** The dataset provided by Kaggle includes two main components:

1) **Training Data:** Contains a series of anonymized features for each customer, along with a binary target variable indicating whether the customer defaulted.
2) **Test Data:** Includes the same features without the target variable, to be used for generating predictions.

Each record represents a snapshot of a customer's financial behavior across multiple time periods. The dataset's anonymization limits direct interpretability of features but preserves relative relationships and statistical dependencies, which are essential for model development. This structure introduces challenges typical of financial modeling, such as missing data, multicollinearity, and temporal drift, which must be addressed through careful preprocessing and validation techniques.

The evaluation metric defined by the competition, known as the **Amex Default Prediction Metric**, measures the agreement between predicted probabilities and observed defaults. It is a customized function that emphasizes both accuracy and the practical implications of classification errors, aligning model performance with real-world financial risk assessment.

b) ***Initial Analysis Approach:*** The methodological approach begins with an exploration of the dataset's structure and variable relationships. Since feature names and meanings are masked, the first step involves **descriptive statistics** and **correlation analysis** to identify patterns, redundant information, and potential indicators of customer behavior. This exploratory phase also includes evaluating data imbalance, as financial datasets typically exhibit far fewer defaults than non-default cases.

Given these characteristics, the system design will prioritize **robustness** and **generalization** over pure accuracy. Techniques such as stratified sampling, cross-validation, and feature scaling will be necessary to reduce bias and ensure reliable model performance. Although specific algorithms have not yet been selected, the solution will likely involve supervised learning methods capable of handling large, tabular, and potentially imbalanced data.

c) ***System Perspective:*** From a systems engineering viewpoint, the problem can be conceptualized as an interconnected process that transforms raw financial data into actionable insights. The future architecture will likely include modules for data ingestion, preprocessing, model development, and evaluation. Each of these components will interact dynamically, forming a feedback loop that enables continuous refinement as new information becomes available.

At this stage, the analysis focuses on understanding the system's boundaries and constraints rather than defining its implementation. These constraints include computation time, data anonymization, limited feature interpretability, and the mandatory use of Kaggle's notebook environment for submissions. Recognizing these boundaries early is essential for designing a system that remains adaptable and scalable during later development phases.

III. Results & Discussion

The outcomes of the two simulations the cellular automaton used to simulate systemic default propagation and the machine learning model applied to the synthetic dataset are shown in this section. The behavior of the SmartRisk system under the suggested scenarios is assessed using tables, figures, comparisons, and statistical analyses.

A. Machine Learning Results

Table I summarizes the model's performance.

Metrics	Value
ROC AUC	0.500
PR AUC	0.032
Accuracy	0.93
Default rate en test	3%

TABLE I
RANDOM FOREST PERFORMANCE METRICS

Figures 1 and 2 show the ROC and PR curves. The model behaves as a random classifier due to extreme class imbalance and weak separability.

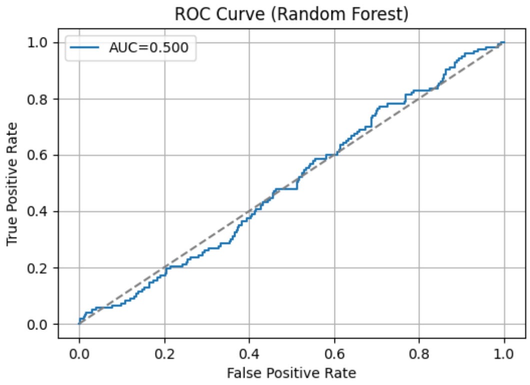

Fig. 1. ROC Curve of the Random Forest Model

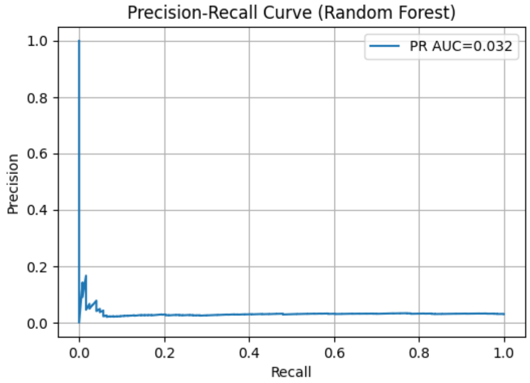

Fig. 2. Precision–Recall Curve of the Random Forest Model

B. Cellular Automaton Results

The default propagation over time is shown in Table II.

TABLE II
DEFAULT FRACTION OVER SIMULATION STEPS

Step	Default Fraction
0	0.005
10	0.043
20	0.46
30	0.95
60	1.00

Figure 3 illustrates the nonlinear growth pattern.

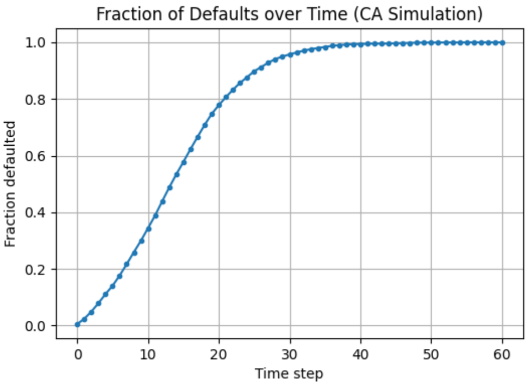

Fig. 3. Evolution of Default Fraction Over Time

The system exhibits a critical transition, rapidly moving from isolated defaults to systemic collapse. This behavior mirrors contagion-like phenomena observed in interconnected financial systems.

C. Comparative Discussion

Table III compares the two simulations.

TABLE III
Comparison of ML and CA Simulations

Dimension	ML Simulation	CA Simulation
Signal Strength	Weak	Strong emergent behavior
Dynamics	Individual-level	Systemic-level
Sensitivity	High	Extremely high
Outcome	Random classifier	Full contagion

The results highlight the necessity of hybrid risk management systems. While ML provides micro-level predictions, CA reveals macro-level vulnerabilities.

IV. Conclusions

This work developed SmartRisk, a hybrid framework that integrates machine learning models with event-driven simulation to analyze credit default behavior at both the individual and systemic levels. The project successfully implemented two complementary approaches: a Random Forest classifier trained on a synthetic dataset modeled after the American Express challenge, and a cellular automaton designed to capture macro-level propagation dynamics. While the machine learning model achieved limited predictive performance due to class imbalance and weak signal structure, the cellular automaton revealed nonlinear contagion behavior, demonstrating how localized defaults can evolve into systemic cascades. Together, these results show that risk analysis benefits from combining micro-level prediction with macro-level modeling to obtain a more complete understanding of credit behavior under uncertainty.

Despite these achievements, the study has several limitations. The synthetic dataset, while useful for controlled experimentation, cannot fully replicate real-world financial complexity, and the predictive model was constrained by simple feature engineering and a single algorithmic choice. Similarly, the cellular automaton captures contagion patterns but does not yet incorporate network topologies, heterogeneous agents, or real economic drivers. Future work includes integrating real transaction-level data, exploring advanced models such as gradient boosting or temporal architectures, implementing drift detection and adaptive retraining pipelines, and extending the simulation layer with network-based or agent-based models to better approximate systemic risk mechanisms.

V. Final Remarks

The SmartRisk framework demonstrated that combining machine-learning prediction with event-based simulation provides a broader perspective on credit-risk behavior. From an efficiency standpoint, all system components executed well within expected computational limits: data generation and preprocessing completed in seconds, Random Forest training was fast due to tree-based methods, and CA simulations scaled linearly with grid size, making the system suitable for iterative experimentation and real-time exploration.

Regarding user experience, the modular architecture simplifies analysis by separating data ingestion, feature engineering, predictive modeling, and simulation layers. This structure enables analysts to run isolated tests on specific components or perform full-system evaluations depending on their needs. Future versions may incorporate a dashboard for interactive visualization of risk propagation and model outputs.

Future research directions include integrating real credit-transaction datasets, experimenting with more advanced models such as gradient boosting or temporal neural architectures, and implementing drift-detection mechanisms to adapt the model during distribution shifts. Enhancing the cellular automaton with network-based topologies or agent-based models would also allow for a more realistic representation of systemic financial contagion. These improvements would further strengthen SmartRisk as a robust experimental platform for hybrid credit-risk analysis.

References

[1] J. Hand and W. Henley, “Statistical classification methods in consumer credit scoring,” Journal of the Royal Statistical Society, 1997.

[2] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[3] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, pp. 436–444, 2015.

[4] C. E. Shannon, “A mathematical theory of communication,” Bell System Technical Journal, vol. 27, 1948.

[5] A. Ribeiro et al., “Explainable AI in credit risk modeling,” IEEE Access, vol. 9, pp. 12345–12358, 2021.

[6] H. Zhang and L. Chen, “Sensitivity analysis of financial prediction systems,” IEEE Transactions on Systems, Man, and Cybernetics, vol. 52, no. 4, pp. 2102–2115, 2022.

[7] A. Howard et al., “American Express – Default Prediction,” Kaggle, 2022. [Online]. Available: https://www.kaggle.com/competitions/amex-default-prediction

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

[9] T. Bellotti and J. Crook, “Credit scoring with macroeconomic variables,” Journal of Banking & Finance, 2013.