Technical Report

Kaggle Systems Engineering Analysis
Loan Default Prediction

Team #9

Juan Esteban Avila Trujillo - 20251020054
Juan Jose León Gomez - 20212020055
Juan Pablo Diaz Ricaurte - 20222020076
Miguel Angel Hernandez Medina - 20222020035

Classification: Technical Research Report

Institution:
Universidad Distrital Francisco José de Caldas
Systems Engineering, Bogotá, Colombia

October 24, 2025

Abstract

This report details a Systems Engineering approach to solving the Kaggle **Loan Default Prediction** competition. The primary objective is to develop a robust system capable of addressing the dual challenge of binary default classification and financial loss regression using a large, anonymized, and high-dimensional dataset. The initial analysis revealed that the prediction task is inherently complex and susceptible to chaotic effects due to data quality issues, non-linear feature interactions, and feedback loops in the financial system. Based on these findings, the report defines critical functional and non-functional requirements (including **scalability** and **resilience**) and proposes a modular, traceable, and reproducible high-level architecture. Key design decisions, such as the implementation of version-controlled preprocessing **pipelines**, model **ensembles**, and specialized monitoring modules, are introduced as crucial strategies for mitigating system sensitivity and ensuring stable predictive performance against the **Mean Absolute Error (MAE)** metric. This work provides a structured, engineering-driven framework for developing reliable machine learning solutions in volatile risk environments.

Keywords: Chaos Theory, Systems Engineering, Loan Default Prediction, High-Level Architecture, Loss Regression, Mean Absolute Error (MAE).

Contents

1. Introduction

This technical report presents a comprehensive analysis and the system design proposal for addressing the **"Loan Default Prediction"** competition hosted on the Kaggle platform, in collaboration with Imperial College London. The central challenge of this competition is not only binary classification (predicting whether a loan will default) but also loss regression, which involves estimating the magnitude of the resulting financial loss. This duality necessitates a **Systems Engineering** approach to build a robust and scalable solution.

The primary objective of this document is to detail the implemented methodology, the designed system architecture, and the strategies employed to mitigate the inherent challenges of the dataset, which includes over 200,000 observations and is characterized by a large number of anonymized and pre-processed variables.

The report is structured as follows: first, the systemic analysis of the competition is presented, identifying its key elements, sensitivity to initial conditions, and the effects of Chaos and Complexity Theory on prediction. Subsequently, the high-level architecture and the technical stack required for implementation are defined, ensuring the system is modular, reproducible, and capable of maintaining consistent predictive performance under the competition's primary evaluation metric, the Mean Absolute Error (MAE).

2. Literature Review

Previous studies on credit default prediction have explored a wide range of techniques:

- **Baseline Models (Logistic Regression, Decision Trees):** Initially used as starting points. They offer high **interpretability** and simplicity, though they generally have lower predictive power than more advanced methods.
- **Ensemble Models (Random Forests, GBMs):** These are the most effective and predominant techniques for tabular data. GBMs (such as LightGBM) are valued for their high accuracy and their superior ability to capture **complex and non-linear interactions** between variables.
- **Deep Models (Neural Networks):** Employed to capture extremely complex feature interactions. Their application is useful when tree-based models are insufficient, but they require more data and resources.
- **Feature Engineering (Ratios, Temporal Variables):** The most critical process. It involves creating new variables with high predictive value (such as the debt-to-income ratio) to **translate business knowledge** and maximize model performance.
- **Evaluation Metrics (ROC-AUC, MAE, Precision-Recall):** Essential tools to measure and balance risk. The **MAE** is crucial for the regression task (quantifying loss), while ROC-AUC evaluates the ability to rank predictions correctly.

Key Competition Constraint (Anonymized Features): Unlike previous studies, the lack of context for the variables forces participants to rely on **statistical selection** and **algorithmic robustness** to obtain reliable results.

3. Background

3.1. Credit Risk Context. Traditional credit risk management has historically focused on **binary classification** (determining whether a borrower is "good" or "bad"). This approach aims to reduce the consumption of economic capital by preventing default. However, this view is insufficient for modern financial optimization.

3.2. The Competition Challenge. The "Loan Default Prediction" competition introduces a dual challenge that seeks to build a bridge between traditional banking and an **asset management perspective**:

- **Default Prediction:** Determine the probability that a loan will fail (classification).

- **Loss Prediction (Loss Regression):** Estimate the **magnitude (severity)** of the financial loss if the default occurs (regression).

The goal is to move beyond simple classification to provide a quantitative loss estimate, allowing financial investors to optimize risk.

3.3. Data and Metric. The analysis is performed on a massive dataset with over **200,000 observations** and around **800 features**. These features have been subjected to standardization, deseasonalization, and, crucially, **anonymization**. This anonymization constitutes a technical challenge, as participants must rely on statistical and feature engineering methods rather than direct business logic. The competition format is a regression problem evaluated using **MAE (Mean Absolute Error)**. Participants are provided with `train.csv` and `test.csv`, and are required to generate `submit.csv` with predicted default probabilities.

4. Scope

The scope of this project and the prediction system is strictly defined by the objectives and limitations imposed by the Kaggle "Loan Default Prediction" competition.

4.1. Functional Scope (Inclusion). The system must be limited to performing the following prediction tasks on the provided dataset:

- **Dual Prediction:** Generate two interrelated outputs for each entry in the test set: the binary probability of a default occurring, and the estimation of the severity (monetary or percentage magnitude) of the financial loss if the default occurs (loss regression).
- **All steps from raw data ingestion to model training and submission file generation:** This requires **integral processing**. The system must be an *end-to-end* solution, covering every phase of the pipeline (cleaning, engineering, modeling, and the final generation of the `submit.csv` file) to meet the competition requirement.
- **Evaluation of modeling techniques (LightGBM, XGBoost, etc.) suitable for tabular data:** This implies performance optimization. It requires the evaluation and selection of advanced models (ensemble methods like LightGBM and XGBoost) that are best suited to handle the high dimensionality and tabular nature of the financial data.

4.2. Limitations (Exclusion). The following elements are explicitly outside the scope of the developed system:

- **Interpretability of individual features due to their anonymized nature:** The deep interpretation of individual features is **out of scope**. The constraint imposed by **anonymization** forces the team to focus on statistical prediction and algorithmic robustness, rather than detailed semantic understanding of each variable.
- **Use of external datasets beyond what is provided:** The scope excludes the **use of external data**. This is a fundamental competition restriction to ensure fairness. The system must rely solely on the provided dataset, respecting the challenge's integrity and conditions.
- **Operational Deployment and Maintenance:** The system is delivered as a proof of concept. Tasks related to integration into real banking systems, continuous infrastructure monitoring, and long-term support outside the competition period are not part of the scope.

5. Assumptions

The system is developed under the following critical assumptions:

- The data provided is clean except for missing values.
- Anonymized variables maintain consistent statistical relationships.
- The distribution in `test.csv` matches that of `train.csv`.
- Kaggle evaluation is based only on the predictions in `submit.csv`.

6. Limitations

The following constraints and limitations directly impact the design and performance of the predictive system:

- **The evaluation metric (MAE):** The main limitation is that the Mean Absolute Error (MAE) is a purely statistical metric that measures the average magnitude of errors but **does not reflect the asymmetric financial cost** associated with business prediction errors.
- **Computational Constraints:** Time and memory limits restrict the complexity of the feature engineering and the depth of hyperparameter tuning, forcing a trade-off between model accuracy and computational efficiency.

7. Methodology

The methodology adopted for the "Loan Default Prediction" competition follows a structured **Systems Engineering (SE)** approach to ensure robustness, reproducibility, and mitigation of the inherent complexity of credit risk.

7.1. Data Processing Pipeline. To counteract the system's sensitivity to initial conditions, this phase focused on creating an **immutable and reproducible transformation pipeline**.

- **Missing Value Handling:** Consistent imputation strategies (e.g., median or specific per variable imputation) were selected, and the imputation parameters from the training set were recorded to be applied identically to the test set (*Transformation Reuse*).
- **Normalization and Scaling:** Transformations were applied to uniformize the scale of the approximately 800 variables, mitigating the risk of higher-magnitude features dominating model training.
- **Feature Engineering:** Higher-predictive value features (e.g., ratios, aggregates, temporal variables) were created, and **Dimensionality Reduction** (PCA) and **Feature Stability Analysis** (correlation/variance) techniques were applied to eliminate noise and instability.

7.2. Modeling. This phase focused on algorithm selection and validation strategy to ensure system **resilience**.

- **Model Selection:** Priority was given to **Ensemble Methods** based on trees (primarily **LightGBM** and **XGBoost**) due to their proven efficiency and ability to handle high-dimensional tabular data.

7.2.1 Robustness Strategy

- **Model Ensemble:** A combination of multiple models (e.g., stacking or weighted averaging) was used instead of a single model, to reduce variance and increase prediction reliability.
- **Stratified Cross-Validation (*k*-fold):** Rigorous validation was implemented to obtain stable performance estimates and limit overfitting.
- **Fixing Random Seeds:** Random seeds were fixed in all algorithms to ensure the **reproducibility** of results, a key measure against chaotic variability.
- **MAE vs. Cost Challenge:** The limitation that the MAE does not reflect the asymmetric business cost was acknowledged, so the Model Ensemble was adjusted seeking a balance between minimizing statistical error and practical negative cost penalty.

7.3. Inference.

- Apply the same preprocessing to `test.csv`.
- Generate predictions and save them in `submit.csv`.