Competitive analysis of IEEE-CIS Fraud Detection

Leonardo Rodríguez Salas - 20231020150
Santiago Marín - 20231020159
Davidson Sanchez - 20231020183
Luis Mario Ramirez - 20231020166

September 2025

1 Introduction

This report presents a systems analysis of the IEEE-CIS Fraud Detection Challenge on Kaggle. The objective of the challenge is to predict the probability that an online transaction is fraudulent, using real e-commerce data provided by Vesta. By applying systems engineering principles, we aim to understand the structure, interactions, and dynamics of the challenge, incorporating elements of sensitivity analysis and chaos theory to explore the system's complexity.

2 Objectives

- Identify and describe the key elements of the competition, including datasets, features, and target variables.
- Map inter-element relationships and system boundaries.
- Apply systems engineering principles to frame the problem.
- Conduct a sensitivity analysis to evaluate the impact of input variations.
- Discuss chaotic behaviors within the system.

3 Competition overview

The IEEE-CIS Fraud Detection competition focuses on improving the accuracy of fraud detection in online transactions. The dataset is split into two main files:

- transaction.csv: Contains transactional features such as ProductCD, card1–card6, addr1–addr2, $P_emaildomain, R_emaildomain, andtransactionalmetadata.$

- identity.csv: Includes identity-related features like DeviceType, DeviceInfo, and ID.

The target variable is isFraud, a binary indicator of fraudulent activity. Not all transactions have corresponding identity records, adding complexity to the dataset.

4 System Analytics

The competition can be viewed as a system with the following elements:

- Inputs: Transactional and identity features, user behavior data, device information.
- Processes: Data processing, model training, validation, and prediction.
- Outputs: Probability of fraud (isFraud).
- Constraints: Data imbalance, missing identity records, privacy restrictions, and computational limits.

The relationships between the different elements include:

- The data flow from the input data to the processed features.
- The interdependence between transactional data and identity data, facilitated by the transaction ID attribute.
- The feedback loops involved in model iteration and validation.

A system architecture diagram would illustrate these components and their interactions, highlighting the data flow and decision points.

5 Sensitivity Analysis

Sensitivity analysis shows how changes in input data or implementation parameters influence model results and predictions. For the IEEE-CIS Fraud Detection competition, this analysis helps determine which features or causes have the most impact upon the likelihood of a transaction being classified as fraud (isFraud).
The main variables considered were:

- TransactionAmt: small changes in transaction amount can lead to large differences in predicted probability of fraud.

- ProductCD (card1-card6, addr1-addr2): these are categorical variables; categroical encoding can change the way models behaves.
- Identity features (DeviceType, DeviceInfo) - patterns also applied here since each missing value is important to map correctly (fake email). Gain: sensitivity around how missing values were treated (impute, remove or encode), hint about booking.

After conducting an analysis of sensitivity, the following was discovered that the model performance is sensitive to the handling of missing identity data (MI) and scaling of numerical features.

Finally, based on these results, we can see that even minor differences in input parameters can have an impact on the ability of a model to accurately detect fraud. Thus, care must be taken to keep data preparation pipelines constant between datasets and closely track the distribution of data features to ensure model stability.

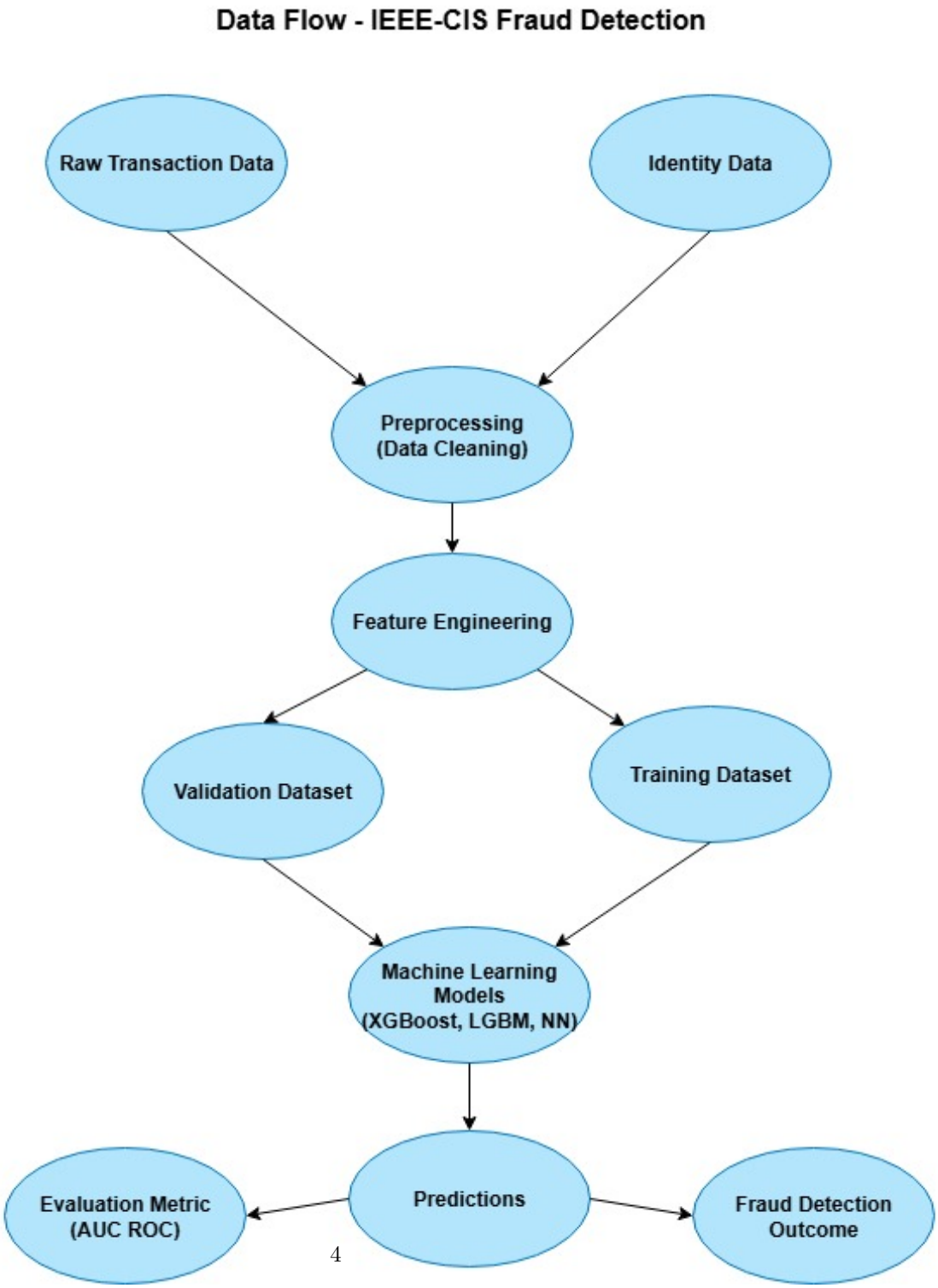
6 Chaos Theory

The fraud detection system exhibits complex and potentially chaotic behavior due to the following factors:

- Non-linearity: Fraud patterns can evolve unpredictably over time.
- Feedback loop: Fraudsters adapt to detection mechanisms, making it increasingly difficult to accurately identify fraudulent activity.
- Small changes, big effects: Minor variations in feature engineering or model selection can lead to significant changes in classification accuracy.
- Randomness: Human behavior and fraudulent tactics introduce a high degree of uncertainty.

These elements align with chaos theory, which posits that initial conditions can lead to vastly different outcomes.

7 Data Flow


Data Flow - IEEE-CIS Fraud Detection
Raw Transaction Data
Identity Data
Preprocessing (Data Cleaning)
Feature Engineering
Validation Dataset
Training Dataset
Machine Learning Models (XGBoost, LGBM, NN)
Predictions
Evaluation Metric (AUC ROC)
Fraud Detection Outcome

8 Conclusion

The analysis carried out on the IEEE-CIS Fraud Detection competition allowed us to understand the problem from a systems engineering perspective, identifying the key elements that make up the process of detecting fraud in online transactions.
It was determined that the quality and treatment of input data are critical factors for model performance. Sensitivity to parameter variations showed that small changes in the handling of missing values or normalization can significantly alter predictive power.
The systemic approach allows us to visualize the interrelationships between technical, human, and contextual components, facilitating a better understanding of the problem and the search for more robust solutions.