# A Systems Engineering Approach to IEEE-CIS Fraud Detection

Using Sensitivity and Chaos Theory

Leonardo Rodríguez Salas Santiago Marín Davidson Sánchez
Luis Mario Ramírez
Universidad Distrital Francisco José de Caldas – Faculty of Engineering, Bogotá, Colombia

**Abstract**

We present an exploratory review of the dataset from the IEEE-CIS Fraud Detection competition and the selection of the tool stack and validation criteria to build a reproducible fraud detection pipeline. The approach is iterative: exploratory analysis and feature engineering will be performed on key columns, applying compact representations for high-cardinality categoricals, imputation strategies and temporal validation; evaluation will prioritize metrics oriented to the minority class (AUC-PR, Precision@k and recall) and probability calibration. We expect to deliver data quality diagnostics, temporal visualizations and a baseline model accompanied by interpretability analyses that allow defining operational thresholds and human-verification policies.

## Contents

# 1 Introduction

Fraud in electronic transactions constitutes a financial and reputational risk for payment platforms, merchants and users. Real-time detection of fraudulent transactions reduces losses and improves security. This competition proposes classifying transactions using transactional and identity variables. The report materializes a technical development that follows the conventions of an academic/professional report and recommended practices in data science.

# 2 Literature Review

Transaction fraud detection is a mature applied research area that combines supervised, unsupervised and deep learning techniques. In industry contexts and competitions (such as IEEE-CIS), tree-based boosting approaches (GBDT) have been the dominant practical choice due to their performance, handling of sparse data and efficiency: LightGBM introduces optimizations (GOSS and EFB) to speed up training without sacrificing accuracy, and it is widely used in large-scale problems.

Recent reviews show the literature is evolving toward combining traditional techniques (trees, ensembles) with modern methods: advanced preprocessing for high-cardinality categorical variables, temporal aggregation windows to capture historical behavior, and increasingly deep learning approaches (especially for sequences and embeddings) where complex patterns abound. In addition, recent works stress the importance of solid experimental practices —temporal and grouped validation and drift analysis— to avoid overfitting and achieve robust production models.

# 3 Background

The goal is to design a system that, for each online transaction in the dataset provided by the IEEE-CIS challenge, estimates a probability of fraud (`isFraud = 1`). The dataset combines transaction records (amounts, product codes, cards, relative time) with identity information (device, email domain, `id_*` variables), which enables exploiting both transactional and identity signals for prediction.

### Expected data structure

- `train_transaction.csv` / `test_transaction.csv`
- `train_identity.csv` / `test_identity.csv`

### Example variables

`TransactionID`, `TransactionDT`, `TransactionAmt`, `ProductCD`, `card1..card6`, `addr1`, `addr2`, `P_emaildomain`, `DeviceType`, `DeviceInfo`, `isFraud` (train only), and identity variables `id_01..id_XX`.

# 4 Objectives

1. Build a reproducible pipeline that integrates EDA, preprocessing, feature engineering (temporal aggregations and encodings) and modeling.
2. Evaluate models with validation that simulates real conditions (temporal / group-based splits) to obtain honest generalization estimates.

3. Produce artifacts ready for deployment: serialized model, preprocessing mappings and documentation of model behavior under drift.

## 5 Scope

This project does not only seek to deliver a model that predicts `isFraud` on the Kaggle dataset but to lay the foundations for an operationally useful fraud detection system: reproducible, explainable, robust to distributional change, and capable of integrating with human review processes (alerts) and production systems (APIs, pipelines). In the short term the methodology will be defined and validated; in the medium and long term the work will progress toward deployment, monitoring and continuous improvement.

## 6 Assumptions

- `TransactionDT` allows building temporal splits.
- `isFraud` is correctly labeled in `train_transaction.csv`.
- PII will be handled in accordance with privacy policies (if anonymization is requested prior to sharing).

## 7 Limitations

- High class imbalance: fraud $\ll$ non-fraud.
- Potential feature leakage if not validated temporally or by groups.
- Variables with high cardinality and missingness that require specific encoding strategies.

## 8 Methodology

The proposed methodology adopts an iterative, goal-oriented approach: convert raw records (for example, transaction and identity tables) into representations that allow detecting behavioral patterns associated with fraud. Initially we will explore key columns such as `TransactionDT`, `TransactionAmt`, `ProductCD`, `card1..card6`, `addr1`, `P_emaildomain` and the `id_01..id_XX` identity variables to understand their distribution, seasonality and presence of missing values; this understanding will guide subsequent transformations. For high-cardinality categorical variables we will create compact representations that preserve discriminative information without inflating the feature space; likewise, numeric columns will be normalized or transformed where appropriate.

Validation will be analyzed with metrics that reflect both global separation between classes and performance on the minority class, and probability calibration will be studied so scores are useful for operational decision-making. Finally, interpretability reports and error analysis based on concrete examples will be produced to facilitate human review and to define thresholding and alarm management policies.

## 9 Results

This section presents the results of the fraud detection project on the IEEE-CIS dataset. The objective is to clearly and objectively show data quality diagnostics, applied transformations, pipeline performance and interpretative analysis that support operational decisions such

as threshold definition, alerts and review procedures. Baseline pipelines based on LightGBM/XGBoost will be compared using minority-centric metrics and calibration measures; additionally, model interpretation through global importance and local explanations and a representative error analysis will support operational recommendations.

During dataset analysis, relevant limitations were identified that condition both pipeline design and result interpretation. The marked class imbalance forces evaluation to focus on metrics reflecting the behavior on the minority class. Many identity variables show a high rate of missing values, so the imputation strategy and use of absence flags will be critical decisions to be evaluated for impact.

## 10 Conclusions

The system analysis identified critical variables and chaotic behaviors in fraud detection processes, giving rise to the need for continuous monitoring and adaptive retraining mechanisms. In addition, the modular design allows for scalability, maintenance, and adaptability to constantly evolving fraud patterns.

The systems engineering approach provides an integrated view that connects data, models, and behavioral dynamics to help implement a solution that is optimal, secure, and meets the system's objectives.

## References

[1] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*. Proc. 31st Conf. Neural Inf. Process. Syst. (NeurIPS), Long Beach, CA, USA.

[2] Hernandez Aros, L., Bustamante Molano, L. X., Gutierrez-Portela, F., Moreno Hernandez, J. J., Rodríguez Barrero, M. S., et al. (2024). *Financial fraud detection through the application of machine learning techniques: a literature review*. Humanities and Social Sciences Communications, 11, Article 1130. doi:10.1057/s41599-024-03606-0.

## A Glossary

**Class imbalance** Situation where the positive class is much rarer than the negative class.

**Data leakage** Inadvertent use of future or label-dependent information during training.

**Cardinality** Number of distinct levels in a categorical variable.

**Imputation** Technique to fill missing values.

**AUC-ROC** Area under the ROC curve; measures global separability between classes.

**AUC-PR** Area under the Precision-Recall curve; more informative with imbalanced classes.

**Precision@k** Precision among the top-k transactions by score.

**Brier score** Metric that measures the accuracy of predicted probabilities.

**Calibration** Degree to which predicted probabilities reflect true frequencies.

**SHAP** Method to explain feature contributions to predictions (local and global).

**LightGBM** Efficient implementation of gradient boosted trees, suitable for large tables.