PROJECT ABSTRACT

Credit Card Fraud Detection using Isolation Forest

An unsupervised Isolation Forest that scores card transactions for fraud on the ULB dataset (284,807 transactions, 0.172% fraud), with a live monitoring dashboard — delivered as a built-to-order student project with code, notebook, report, PPT and viva kit.

01. Title

Credit Card Fraud Detection using Isolation Forest

02. Introduction / Background

Card fraud is a needle-in-a-haystack problem: in the public ULB dataset only 492 of 284,807 transactions are fraudulent. Rule systems drown in false alarms, and supervised classifiers struggle because labeled fraud is rare while tactics mutate. Anomaly detection flips the problem — learn what normal looks like and flag what does not fit. Isolation Forest implements this elegantly: random recursive partitioning isolates anomalies in far fewer splits than normal points, requiring no fraud labels at all. This project builds that idea into a complete student system with a monitoring dashboard.

03. Problem Statement

Labeled fraud is too rare and too mutable for comfortable supervised learning, and static rules generate unmanageable false alarms. The project addresses this by building an unsupervised Isolation Forest trained only on legitimate ULB transactions, scoring every transaction by isolation path length, tuning the decision threshold by contamination fraction, and serving it all through a dashboard with an anomaly map, alert feed and precision/recall tradeoff explorer.

04. Objectives

- Build the ULB data pipeline: loading, scaling of Time/Amount, stratified splits preserving the 0.172% fraud rate.
- Train scikit-learn IsolationForest (200 trees, subsample 256) unsupervised on legitimate transactions and implement contamination-based threshold tuning.
- Build a Flask monitoring dashboard: anomaly map, KPI cards, streaming alert feed, scored transaction ledger, threshold tuner.

- Deliver the complete student kit: code, evaluation notebook, report, PPT and viva Q&A.

05. Proposed Methodology

The ULB creditcard.csv is loaded; Time and Amount are standardized while V1–V28 are used as-is (already PCA-scaled by the provider). Stratified splitting preserves the fraud fraction in validation and test sets. IsolationForest fits 200 random trees on legitimate transactions — each tree recursively splits random features at random values until points isolate. Mean path length per transaction converts to an anomaly score (near 1 = anomaly, near 0.5 = normal). A contamination fraction sets the flagging threshold; precision, recall, F1, ROC-AUC and PR curves are computed on the held-out split.

06. System Architecture / Working

The system has three stages. (1) Data pipeline: ULB schema handling, scaling and stratified splitting. (2) Model: unsupervised Isolation Forest training on legitimate traffic with contamination-tuned threshold selection on the validation split. (3) Demo application: a Flask dashboard streams scored transactions onto a V1–V2 anomaly map, routes high scores to a live alert feed, lists every transaction in a sortable ledger with verdicts, and lets the user drag the contamination slider to watch precision/recall trade off in real time.

07. Hardware / Software Requirements

- Any modern laptop/desktop (CPU only; Isolation Forest trains in minutes)
- Python 3.10, scikit-learn (IsolationForest, metrics), pandas, NumPy
- Matplotlib, Seaborn (score distributions, ROC/PR curves, anomaly maps)
- Flask for the monitoring dashboard
- ULB Credit Card Fraud Detection dataset from Kaggle (mlg-ulb/creditcardfraud; free account required)
- Jupyter notebook for the buyer-run training, tuning and evaluation

08. Expected Outcomes

A trained unsupervised fraud scorer with precision, recall, F1, ROC-AUC and PR curves computed on the buyer's own held-out split during their build; a dashboard that makes the anomaly structure, the alert stream and the threshold economics tangible; and a report with an honest treatment of imbalance, uninterpretable PCA features and concept drift. No metric is pre-claimed.

09. Applications

- Teaching material for unsupervised learning: isolation principle, contamination, precision/recall tradeoffs

- Prototype transaction-monitoring interface for fraud-ops demos
- Reference build for anomaly detection on other rare-event data (intrusion, faults)
- Base for hybrid supervised+unsupervised ensembles as future scope

10. Future Scope

- Online/continual retraining windows to counter concept drift in fraud tactics
- Hybrid scoring combining the forest with a supervised head on confirmed fraud labels
- Feature-attribution methods adapted to PCA-anonymized inputs for analyst review
- Latency benchmarking and streaming deployment (e.g. Kafka + FastAPI) as engineering future scope

11. References

- Dal Pozzolo, A. et al. — Credit Card Fraud Detection dataset, Machine Learning Group, ULB (Kaggle mlg-ulb/creditcardfraud).
- Liu, F. T., Ting, K. M. and Zhou, Z.-H. — Isolation Forest, ICDM 2008.
- Dal Pozzolo, A. et al. — Calibrating Probability with Undersampling for Unbalanced Classification, IEEE SSCI 2015.
- Pedregosa, F. et al. — Scikit-learn: Machine Learning in Python, JMLR 2011.