German International University of Applied Sciences
Informatics and Computer Science
Dr. Caroline Sabty
TA Nouran Khaled
TA Sandra Samuel
TA Sarah Hatem

Machine Learning, Winter 2025
Project 2
Deadline 2nd of December at 11:59 pm

1 DataOrbit - HealthCare Provider Fraud Detection Project

Welcome to your capstone machine learning project with Data Orbit! Your team has been hired to assist Medicare in detecting fraudulent healthcare providers. Healthcare fraud costs the U.S. healthcare system over $68 billion annually, and your work could make a measurable impact.

This project emphasizes not only building an accurate model but also understanding the data structure, justifying modeling decisions, and communicating your reasoning effectively to both technical and non-technical audiences.

1.1 Overview

Data Orbit has been contracted by the Centers for Medicare & Medicaid Services (CMS) to develop an intelligent fraud detection system. Currently, CMS can only investigate a small fraction of suspicious cases, allowing many fraudulent activities to go undetected. Existing systems rely on basic rule-based methods that capture obvious patterns but fail to identify more sophisticated fraud schemes.

Your goal is to design a data-driven model that identifies high-risk providers while maintaining interpretability and minimizing false positives, which can lead to unnecessary investigations and reputational damage.

Types of Healthcare Fraud:

- Billing for services never rendered.
- Upcoding - billing for higher-cost procedures than those performed.
- Unbundling - billing separately for procedures that should be combined.
- Submitting claims for deceased patients.
- Prescribing unnecessary treatments for financial gain.
- Engaging in kickback or referral schemes.

1.2 Team Registration

Teams of four members should register before starting the project. Each team is responsible for collaborative execution and documentation of the full pipeline - from exploration to reporting and presentation. You can register the teams maximum by 13th of November: https://forms.gle/wBPLnf4MAR5FwAnc8. **For any team that are missing a team member, or can not find a team, you can specify that in the form.**

1.3 Dataset Description

The project uses the **Healthcare Provider Fraud Detection dataset**, which contains anonymized Medicare data labeled for fraudulent and non-fraudulent providers.

Download from: https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis

Files Included:

a) **Train_Beneficiarydata.csv** - Demographics, coverage, and chronic conditions for each patient (BeneID).

b) **Train_Inpatientdata.csv** - Hospital admission claims with financial, procedural, and physician details.

c) **Train_Outpatientdata.csv** - Outpatient claim data (visits, tests, minor procedures).

d) **Train_labels.csv** - Provider-level fraud labels (Yes or No).

Key Identifiers: `BeneID` links patients to claims; `Provider` links claims to the fraud label.

1.4 Objective

This project aims to build an end-to-end fraud detection pipeline capable of identifying potentially fraudulent providers. Your model should:

a) Detect fraudulent providers from multi-table claims data.

b) Handle severe class imbalance (approximately 10% of providers are labeled fraudulent).

c) Provide explainable predictions for investigators and regulators.

d) Demonstrate business value by prioritizing high-risk providers effectively.

You are expected to justify all major design choices, including data preparation, model selection, and evaluation.

1.5 Modeling Requirements

1.5.1 Data Understanding & Exploration

- Examine relationships between the four datasets, defining join keys and levels of granularity.
- Assess data quality and completeness, identifying missing values and inconsistencies.
- Conduct exploratory analysis on beneficiaries, claims, and providers to uncover patterns, distributions, and outliers.
- Compare fraudulent and legitimate providers using descriptive statistics and visualizations to detect behavioral differences.
- Define an **aggregation strategy** to consolidate claim-level information into provider-level records, which serve as the modeling unit. Consider appropriate statistical summaries (e.g., counts, means, ratios, and percentages) and ensure consistency across inpatient and outpatient data.
- Produce core plots: target class distribution, claim amount trends, provider-level summaries, correlation heatmaps, and geographic or temporal patterns.

1.5.2 Class Imbalance Strategy

- Address imbalance using approaches such as class weighting, oversampling, undersampling, or cost-sensitive learning.
- Select metrics appropriate for imbalanced data, prioritizing Precision, Recall, F1-score, and PR-AUC over overall accuracy.
- Clearly justify the chosen strategy and discuss trade-offs between performance, fairness, and interpretability.

1.5.3 Algorithm Selection

- Evaluate relevant algorithms (Decision Trees, Random Forest, Gradient Boosting, Logistic Regression, SVM) in the context of fraud detection.
- Consider interpretability, computational feasibility, robustness to imbalance, and suitability for mixed data.
- Justify the primary choice (e.g., Gradient Boosting or Random Forest) and its alignment with the dataset characteristics.

1.5.4 Comparison Models

- Implement at least two additional models for comparison (e.g., Logistic Regression for interpretability and Random Forest for robustness).
- Compare models using standardized metrics and visual analyses (Precision-Recall, ROC curves, and confusion matrices).
- Discuss trade-offs between predictive power and explainability to support your final model recommendation.

1.6 Evaluation Metric

- Apply rigorous validation procedures such as train/validation/test splits or cross-validation (time-based or grouped if appropriate).
- Evaluate models using Precision, Recall, F1, ROC-AUC, and PR-AUC to ensure balanced performance assessment.
- Include confusion matrix and cost-based analyses to interpret real-world implications.
- Conduct an **error analysis** to understand model limitations:
  - Create **case studies** for 2-3 false positives (legitimate providers flagged as fraud) and 2-3 false negatives (fraudulent providers missed).
  - Analyze why the model made these errors and what patterns or features may have contributed.
  - Discuss possible refinements or additional features to mitigate these issues in future iterations.
- Prevent overfitting using appropriate data partitioning, regularization, and validation strategies.

1.7 Documentation

Your documentation should serve as a comprehensive record of your project process and reasoning. It should include:

- A clear explanation of all steps taken - from data exploration through model selection and evaluation.
- The rationale behind each decision, including preprocessing, algorithm choice, and parameter tuning.
- A detailed log of trials and experiments, noting what was tested, why, and what insights were gained.
- An error analysis section discussing false positives, false negatives, and their implications.
- Clear structure and consistent formatting, suitable for submission to a technical review team.

1.8 Deliverables

Teams must create a structured GitHub repository named `fraud_detection_project` containing:

- **README.md** - Project overview, team members, summary of results, and reproduction instructions.
- **data**/ - Dataset storage (with download instructions if needed).
- **notebooks**/ -
  - `01_data_exploration_and_feature_engineering.ipynb`
  - `02_modeling.ipynb`
  - `03_evaluation.ipynb`
- **reports**/ -
  - `technical_report.pdf` (documenting full process and analysis)
  - `presentation.pptx` (approximately 10 minutes)

Evaluation: Grades will be based on both the technical project and the clarity and professionalism of your presentation. Presentations should summarize the problem, methodology, main results, and business impact clearly and concisely for a mixed audience.