Report for project 2 machine learning

Report for 01_data_exploration_and_feature_engineering:

The dataset contains four distinct tables:

- Train_Beneficiarydata.csv: Patient demographics, coverage months, and chronic conditions.
- Train_Inpatientdata.csv: Hospital admission claims and financial information.
- Train_Outpatientdata.csv: Outpatient visit and procedure records.
- Train_labels.csv: Provider-level fraud labels (Fraud / Not Fraud).

To permit supervised learning, all datasets were merged using the appropriate keys:

- `BeneID` to link patients to their claims
- `Provider` to map providers to fraud labels

Missing values and cleaning:
The raw data contained missing values across several variables, particularly in:

- Death dates (which are expected to be missing for living patients)
- Chronic disease indicators
- Financial fields in claims

Approaches implemented:

- Meaningful defaults were used where medically reasonable.
- Missing financial values were replaced conservatively with zeros.
- Beneficiary records with missing coverage months were explicitly handled.

These choices ensured:

- No row loss due to missing data
- No artificial signal introduced by incorrect imputation

The following cleaning steps were applied:

- Removal of illegal or malformed values
- Conversion of date fields to consistent datetime formats
- Normalization of categorical encodings
- Validation of numeric ranges (e.g., reimbursement values must be non-negative)

Aggregation:

Claims exist at transaction level, while fraud labels exist at provider level. To train a provider-level classifier, claims had to be summarized per provider.

A systematic aggregation approach was adopted:

For each provider:

- Claim counts
- Mean / sum / max of reimbursement amounts
- Average length of stay
- Frequency of diagnoses
- Procedure diversity
- Beneficiary count per provider
- Cost per beneficiary
- Total inpatient vs outpatient billing ratios

This design transforms raw transactional logs into stable behavioral profiles.

Feature engineering:
More than **60 derived features** were created to improve discriminative power:
ex:

Patient Risk Profile

- Prevalence of chronic conditions
- Clinical burden per provider

- Severity patterns

The final modeling dataset:

- Contains **one row per provider**
- Excludes identifiers from training
- Includes:
    - engineered numeric features
    - the fraud label
    - metadata saved separately for reproducibility

•Visualizations:
The following exploratory plots were produced:

- Fraud class imbalance visualization
- Provider claim volume distribution
- Reimbursement distributions
- Correlation heatmaps between cost features

These confirmed:

- Strong class imbalance (~10% fraud)
- Financial skewness and non linear distributions
  Conclusion:
  This stage transformed raw administrative data into a structured intelligence view of provider behavior. The aggregation strategy preserved financial realism while producing interpretable features.

This preparation made the dataset:

- Predictive
- Explainable
- Stable under imbalance
- Business-relevant

The prepared dataset in order to procced with the modeling.

Report for 02_modeling:

Our goal in the modeling phase:

•is to design and train machine learning models capable of identifying potentially fraudulent healthcare providers based on structured provider-level features.

Since fraud detection is a highly imbalanced classification problem (fraud ≈ 10%), the modeling phase

Train–Test Splitting Strategy:
• we used 70 – 30 for better accuracy.
• dataset was split at the provider level to prevent leakage between training and testing samples. This ensures that claims from the same provider do not appear in both sets.

A fixed random seed was used for reproducibility.

Final dataset structure:

- Input: Engineered provider predictors
- Output: Binary fraud label

•Handling the class imbalance:
Healthcare fraud datasets exhibit natural imbalance, which skews classifiers toward predicting "Not Fraud".

Three imbalance strategies were applied:

• Class Weighting
Models such as Logistic Regression and Random Forest were configured to penalize fraud misclassifications more heavily.

This improves sensitivity without modifying original distributions.

• Oversampling (SMOTE)
Synthetic Minority Over-sampling Technique (SMOTE) was used to generate realistic fraud-like synthetic samples.

Its purpose:

- Prevent model bias
- Encourage fraud pattern learning
- Avoid overfitting through trivial duplication

• Metric Selection
Accuracy was intentionally avoided.

Primary metrics:

- Precision
- Recall
- F1-score
- PR-AUC
- ROC-AUC

Algorithms implemented:

Model	Purpose
Logistic Regression	Baseline & interpretability
Random Forest	Robust ensemble
Gradient Boosting	High precision learning
Support Vector Machine	Margin-based classification

Hyperparameter Tuning

Each model was calibrated using grid search or reasonable defaults:

Examples:

- Logistic: Regularization tuning

- Random Forest: Number of estimators, max depth
- Gradient Boosting: Learning rate, estimators

Model Interpretation

Feature Importance

Tree-based models provided natural interpretability through:

- Feature importance ranking
- Cost drivers influencing fraud probability

Feature Scaling

Normalization and standardization were applied where required:

- Logistic Regression
- SVM

Scaling was excluded for tree-based models since it offers no benefit.

Final Model Selection

After evaluating all trained models using ROC-AUC as the primary selection criterion, Logistic Regression achieved the best overall performance.
Logistic Regression was selected as the final production model due to its:

- **Highest ROC-AUC score among all tested models**
- **Strong discrimination between fraudulent and legitimate providers**
- **High model stability with less variance between training and test results**

Conclusion:

The modeling phase produced a balanced fraud detection system optimizing:

Detection power

Cost sensitivity

Report for 03_evaluation:

Evaluation Objective

The evaluation phase assesses the quality, reliability, and business usability of the trained fraud detection models.

Because provider fraud is highly imbalanced, evaluation emphasizes:

- Fraud detection capability
- Cost-sensitive decision making
- Model explainability

•Metrices used:

Metric	Importance
ROC-AUC	Primary discriminative measure
PR-AUC	Minority-class sensitivity
Precision	Reduces false investigations
Recall	Captures hidden fraud
F1-score	Balance measure
Confusion Matrix	Error understanding

•Results:

Best Model: Logistic Regression (by ROC-AUC)

Logistic Regression achieved the highest ROC-AUC score, outperforming more complex models in separating:

- Fraudulent providers
- Legitimate providers
  This indicates:

•Linear behavior patterns dominate fraud separation.
•Complex non-linear models added no measurable benefit.
•Confusion matrix interpetaion:

Category	Insight
True Positives	Fraud successfully detected
False Positives	Legitimate providers wrongly flagged
False Negatives	Fraud cases missed
True Negatives	Correctly cleared providers

•What we interpreted from the confusion matrix:

- False positives tended to occur among high-cost specialists.
- False negatives mainly correspond to low-volume, stealth fraud

•Roc curve: Logistic Regression achieved the largest area under the curve, reflecting excellent global discrimination.

Error Analysis

False Positives

Causes:

- Rare medical procedures
- High-risk patient populations

Interpretation:
These providers are legally expensive, not fraudulent.

False Negatives

Causes:

- Fraud camouflage
- Conservative billing behavior
- Low-volume manipulation

Interpretation:
These cases evade statistical detection.

Finally;

Logistic Regression demonstrated:

- No overfitting
- Stable precision-recall trade-offs
- Clear coefficient behavior

Supports:

- Legal accountability
- Auditable decisions

Our conclusion:

The final system reliably identifies suspicious providers, balancing:

- Detection
- Fairness
- Interpretability
- Cost