1.7 Documentation

This section provides a complete and structured record of the workflow, decisions, experiments, and reasoning followed throughout the development of the Healthcare Fraud Detection System.

1.7.1 Project Workflow Overview

The project followed a systematic, end-to-end data science pipeline:

1. Data Understanding & Exploration
Raw datasets were explored to understand scope, structure, and relationships between entities.

2. Feature Engineering & Aggregation
Claim-level data was transformed into provider-level records through statistical summarization techniques.

3. Class Imbalance Handling
The minority class (fraud providers) was balanced using oversampling techniques.

4. Modeling & Algorithm Comparison
Multiple models were implemented and evaluated using cross-validation.

5. Evaluation & Error Analysis
Performance was assessed using appropriate metrics and business-driven cost analysis.

1.7.2 Data Exploration and Processing Decisions

Dataset Relationships

The project combined four datasets:

- Provider label dataset (fraud classification)
- Beneficiary table
- Inpatient claims
- Outpatient claims

Primary join keys:

- `Provider` → claim-level to provider-level mapping
- `BeneID` → beneficiary linkage
- `ClaimID` → transactional uniqueness

Data Quality Measures

Observed issues:

- Missing beneficiary fields
- Skewed reimbursement values
- Duplicate claims
- Inconsistent date formats

Solutions:

- Type casting and datetime standardization
- Aggregation-based null handling
- Outlier compression via logarithmic scale
- Dropping non-informative or invalid fields

Feature Engineering Strategy

Provider-level modeling required aggregation of raw claims:

Feature Type	Example
Volume metrics	Total claims
Monetary metrics	Total / mean reimbursement
Diversity metrics	Unique patients
Behavior ratios	Inpatient to outpatient ratio
Risk proxies	Claims per patient

Rationale: Fraud is more visible in behavioral patterns over time than in single transactions.

1.7.3 Class Imbalance Handling Strategy

Fraud cases represent a minority in the dataset.

To correct for this:

- **SMOTE** was used for oversampling the minority class.
- **Stratification** ensured balanced splits.
- Accuracy was deprioritized in favor of:
    - Recall
    - PR-AUC
    - F1-score

Justification:
Missing fraud cases is costlier than reviewing legitimate providers.

1.7.4 Model Selection Rationale

Evaluated models:

Model	Purpose
Logistic Regression	Interpretability
Random Forest	Feature importance
Gradient Boosting	Final selected model

Selection Criteria:

- Robustness to imbalance
- Ability to learn non-linear relationships
- Computational feasibility
- Interpretability
- Stability across validation folds

Final Model Selected: Gradient Boosting
Reason: Best PR-AUC and stable generalization.

1.7.5 Experimental Log & Trials

A structured experiment sequence was maintained:

Experiments Executed:

Experiment	Result
Without SMOTE	Low recall
SMOTE added	Recall increased by ~20%
Feature scaling	SVM & Logistic improved
Tree depth tuning	Reduced overfitting
Metric switch to PR-AUC	Better fraud ranking

Feature Selection Tests:

- Dropped provider IDs
- Removed low-variance metrics

- Eliminated text-based columns
- Introduced ratio-based features

1.7.6 Error Analysis

Observed False Positives

Issue:

- Large but legitimate providers flagged incorrectly

Reason:

- High claim counts resemble fraud patterns

Impact:

- Extra audits but controllable risk

Observed False Negatives

Issue:

- Small providers escaping detection

Reason:

- Behavior blends into normal patterns

Impact:

- More serious due to missed fraud cost

Recommendations for Improvement

- Temporal modeling (rolling windows)
- Unstructured text analysis
- Provider peer-group modeling
- Cost-sensitive learning
- Network-based fraud detection

1.7.7 Model Evaluation Rationale

Metrics selected:

Metric	Why
Recall	Catch fraud
Precision	Reduce false alarms
F1-score	Balance
PR-AUC	Minority classification
ROC-AUC	Separation ability

Business Cost Logic

Error	Cost
False Positive	$500
False Negative	$10,000

Model choice optimized **financial protection**.

1.7.8 Documentation Standards

- Modular notebooks
- Explicit justification
- Visual explanations
- Metric transparency
- Reproducible code
- Consistent formatting

This documentation serves as an audit trail for:

- Decisions
- Trade-offs
- Ethical design
- Risk reasoning