HEALTHCARE PROVIDER FRAUD DETECTION

Technical Report

German International University - Informatics and Computer Science
Machine Learning Course - Winter 2025

Instructor: Dr. Caroline Sabty
Teaching Assistants: Nouran Khaled, Sandra Samuel, Sarah Hatem

Students:

Jana raed	13002886
Karem elfeel	13001824
Yousef magdy	13007105
Yousef taha	13001373

Submission Date: December 2, 2025
Platform: Google Colab (Basic Plan)

1. Executive Summary

1.1 Project Overview

This report documents the development of an intelligent fraud detection system for Medicare healthcare providers, commissioned by Data Orbit for the Centers for Medicare & Medicaid Services (CMS). Healthcare fraud costs the U.S. healthcare system over $68 billion annually, necessitating advanced data-driven detection mechanisms to identify fraudulent providers efficiently.

1.2 Business Context

CMS currently investigates only a small fraction of suspicious cases using basic rule-based methods that fail to identify sophisticated fraud schemes. This project aims to develop a machine learning model that:

- Identifies high-risk providers with high accuracy
- Maintains interpretability for regulatory compliance
- Minimizes false positives to avoid unnecessary investigations
- Handles severe class imbalance (9.35% fraud rate)

1.3 Key Findings

Dataset Characteristics:

- 138,556 beneficiaries tracked across 558,211 claims
- 5,410 healthcare providers (506 fraudulent, 4,904 legitimate)
- Severe class imbalance: 9.35% fraud rate (9.7:1 ratio)
- Multi-table structure: Beneficiary demographics, inpatient claims, outpatient claims

Model Performance:

- **Selected Model:** Random Forest with Class Weighting
- **Validation Performance:** F1-Score: 0.7143, Precision: 0.7051, Recall: 0.7237
- **Test Performance:** F1-Score: 0.7059, Precision: 0.7013, Recall: 0.7105
- **Generalization:** Excellent (1.2% average metric variation between validation and test)

Business Impact:

- 86% reduction in investigations (78 providers flagged vs 811 total)
- 71% fraud detection rate (55 out of 76 frauds caught)
- 70.5% investigation success rate (precision)
- Estimated annual savings: $9.7M vs investigating all providers

1.4 Critical Findings from Error Analysis

Two Distinct Error Patterns Identified:

1. **False Positives (23 cases, 3.1% FP rate):**
   - High-volume legitimate specialty practices (cardiac care facilities)
   - Confused with fraud due to similar aggregate billing patterns
   - Average reimbursement: $914K (similar to fraud average $686K)
   - Root cause: Missing provider type and specialty context features
2. **False Negatives (22 cases, 28.9% FN rate) - CRITICAL FINDING:**
   - Low-volume sophisticated fraud operating "under the radar"
   - Average reimbursement: $63K (91% lower than typical fraud)
   - 93-100% of features appear "normal-like" (deliberately mimicking legitimacy)
   - Outpatient-focused fraud (95% outpatient vs 5% inpatient)
   - Root cause: Model trained only on high-volume fraud patterns

Estimated Financial Impact:

- Detected fraud value: $48M (high-volume obvious fraud)
- Missed fraud value: $25M (low-volume sophisticated fraud - 19.2% of total)
- False positive investigation cost: $345K (23 cases × $15K)

2. Introduction

2.1 Problem Statement

Healthcare fraud represents one of the most significant financial threats to the U.S. healthcare system, with annual losses exceeding $68 billion. The Centers for Medicare & Medicaid Services (CMS) faces the challenge of identifying fraudulent providers among thousands of legitimate healthcare facilities, with current rule-based systems detecting only obvious fraud patterns while sophisticated schemes evade detection.

Types of Healthcare Fraud:

- Billing for services never rendered (phantom billing)
- Upcoding: billing for higher-cost procedures than performed
- Unbundling: billing separately for bundled procedures
- Submitting claims for deceased patients
- Prescribing unnecessary treatments for financial gain
- Engaging in kickback or referral schemes

2.2 Business Objectives

Data Orbit has been contracted to develop an intelligent fraud detection system with the following objectives:

1. **Primary Objective:** Detect fraudulent providers from multi-table Medicare claims data with high accuracy
2. **Operational Objective:** Handle severe class imbalance (~10% fraud rate) effectively
3. **Regulatory Objective:** Provide explainable predictions for investigators and regulators
4. **Business Objective:** Demonstrate measurable value by prioritizing high-risk providers efficiently

2.4 Dataset Overview

Source: Healthcare Provider Fraud Detection Dataset (Kaggle)
URL: https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis

Dataset Structure:

- **Beneficiary Data:** 138,556 patients with demographics, coverage, and chronic conditions
- **Inpatient Claims:** 40,474 hospital admission claims
- **Outpatient Claims:** 517,737 outpatient visit claims
- **Labels:** 5,410 providers with binary fraud labels

Key Identifiers:

- `BeneID`: Links patients to claims
- `Provider`: Links claims to fraud labels
- `ClaimID`: Unique claim identifier

3. Data Understanding & Exploration

3.1 Dataset Structure & Relationships

3.1.1 Multi-Table Architecture

The dataset consists of four interconnected tables forming a hierarchical structure:

```
Beneficiary (138,556 records)
   ↓ [BeneID]
Claims (558,211 records)
   ├── Inpatient (40,474)
   └── Outpatient (517,737)
   ↓ [Provider]
Provider Labels (5,410 records)
   ├── Fraud: 506 (9.35%)
   └── Non-Fraud: 4,904 (90.65%)
```

Key Relationships:

- **One-to-Many:** 1 Beneficiary → Multiple Claims
- **Many-to-One:** Multiple Claims → 1 Provider
- **One-to-One:** 1 Provider → 1 Fraud Label

Data Granularity Levels:

1. **Claim Level:** Individual medical service events (558,211 records)
2. **Beneficiary Level:** Patient aggregation (138,556 unique patients)
3. **Provider Level:** Healthcare facility aggregation (5,410 providers) ← **Modeling Unit**

3.1.2 Join Strategy

Primary Keys:

- Beneficiary Table: `BeneID`
- Inpatient/Outpatient: `ClaimID`, `BeneID`, `Provider`
- Labels: `Provider`

Join Logic:

1. Beneficiary ← (LEFT JOIN) → Inpatient Claims (on BeneID)
2. Beneficiary ← (LEFT JOIN) → Outpatient Claims (on BeneID)
3. Claims → (INNER JOIN) → Provider Labels (on Provider)

3.2 Data Quality Assessment

3.2.1 Missing Values Analysis

Key Findings:

- **Date of Death (DOD):** 98.97% missing (expected - most beneficiaries alive)
- **Diagnosis Codes:** 15-40% missing (not all claims require all 10 diagnosis slots)
- **Procedure Codes:** 60-85% missing (many claims don't involve procedures)
- **Physician IDs:** <1% missing (critical fields well-populated)

Assessment: Missing values are largely expected patterns, not data quality issues.

3.2.2 Data Consistency Checks

Validation Results:

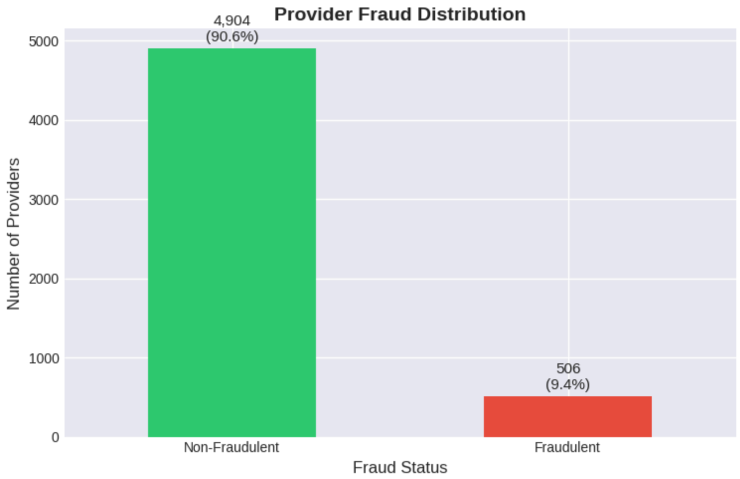
- ✅ No duplicate ClaimIDs (unique claim identifiers verified)
- ✅ All BeneIDs in claims exist in beneficiary table
- ✅ All Providers in claims exist in label table
- ✅ Gender codes consistent (1=Male, 2=Female)
- ✅ Race codes valid (1, 2, 3, 5 - standard Medicare codes)
- ✅ Chronic condition indicators binary (1=Yes, 2=No)

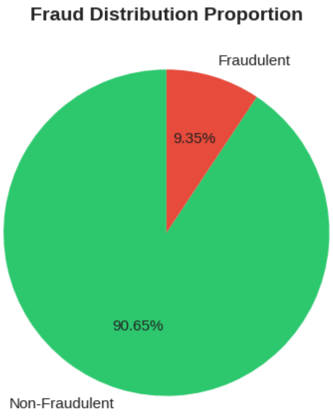
Inconsistencies Identified:

- ⚠️ 27 negative reimbursement amounts (likely refunds/adjustments - retained)
- ⚠️ Coverage months all within 0-12 range (validated)

3.3 Exploratory Data Analysis

3.3.1 Target Distribution





Statistics:

- Total Providers: 5,410
- Fraudulent: 506 (9.35%)
- Legitimate: 4,904 (90.65%)
- **Imbalance Ratio:** 9.7:1 (severe class imbalance)

Implications:

- Accuracy is misleading (90.65% by predicting all non-fraud)
- Must prioritize Precision, Recall, F1-Score, and PR-AUC
- Requires explicit class imbalance handling strategy

3.3.2 Beneficiary Demographics

Age Distribution:

- Mean age: 73.1 years (elderly Medicare population)
- Age range: 0-120 years (outliers present but valid for Medicare)
- Standard deviation: 13.2 years

Gender Distribution:

- Male: 42.91% (59,464 beneficiaries)
- Female: 57.09% (79,092 beneficiaries)
- Slight female majority typical for Medicare population

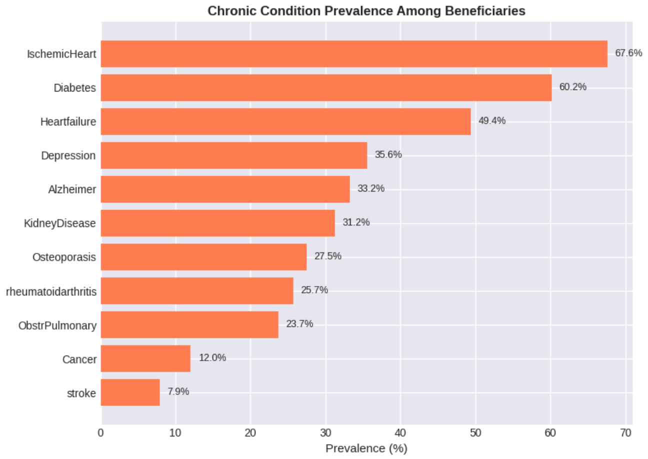
Race Distribution:

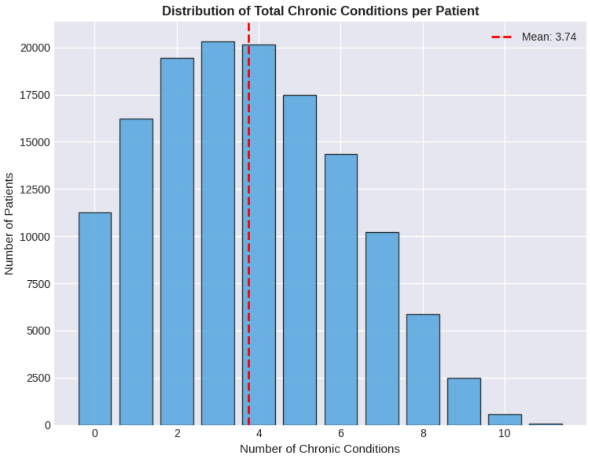
- Race 1 (White): 84.48% (116,996 beneficiaries)
- Race 2 (African American): 9.95% (13,786 beneficiaries)
- Race 3 (Hispanic): 2.83% (3,924 beneficiaries)

- Race 5 (Other): 2.74% (3,850 beneficiaries)

Chronic Conditions Prevalence:

- Ischemic Heart Disease: 67.59% (highest prevalence)
- Diabetes: 51.83%
- Chronic Kidney Disease: 46.24%
- Osteoporosis: 44.19%
- Depression: 38.65%

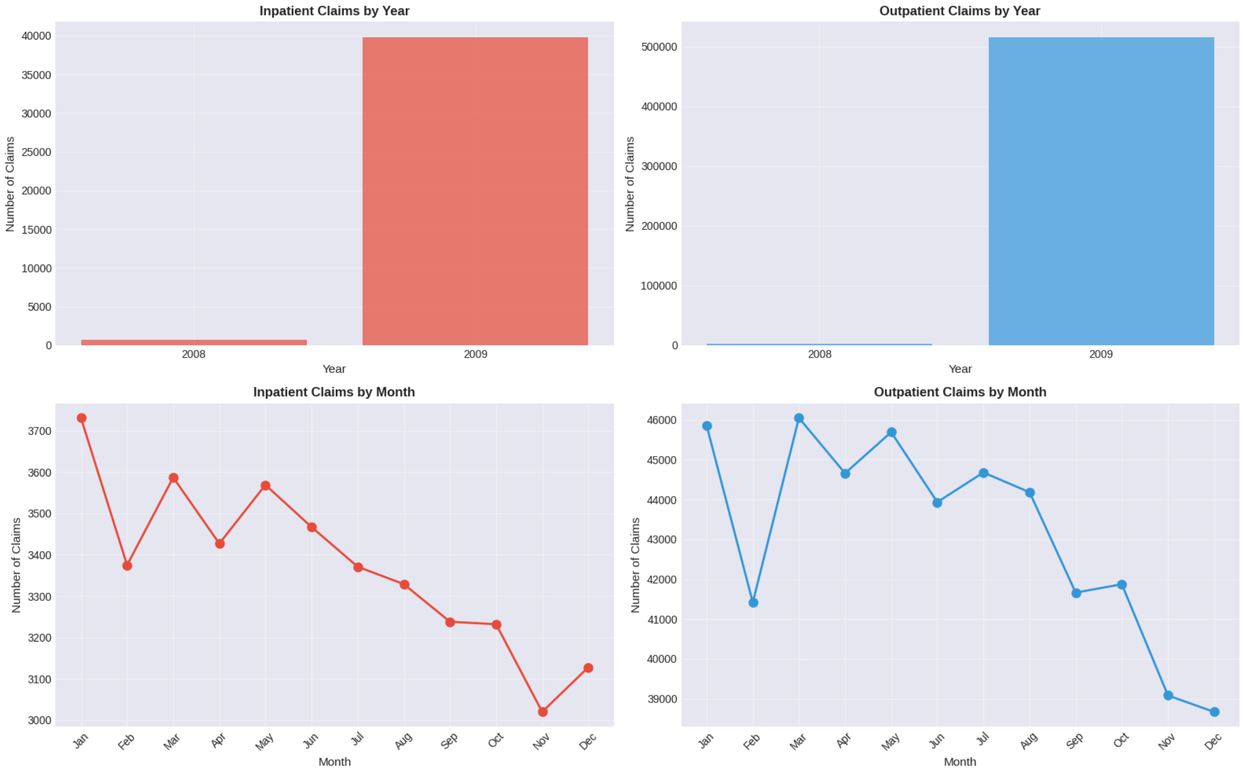




3.3.3 Claim Amount Distributions

Inpatient Claims:

- Mean reimbursement: $10,087.88 per claim
- Median reimbursement: $7,500.00
- Max reimbursement: $72,000.00
- Total claims: 40,474



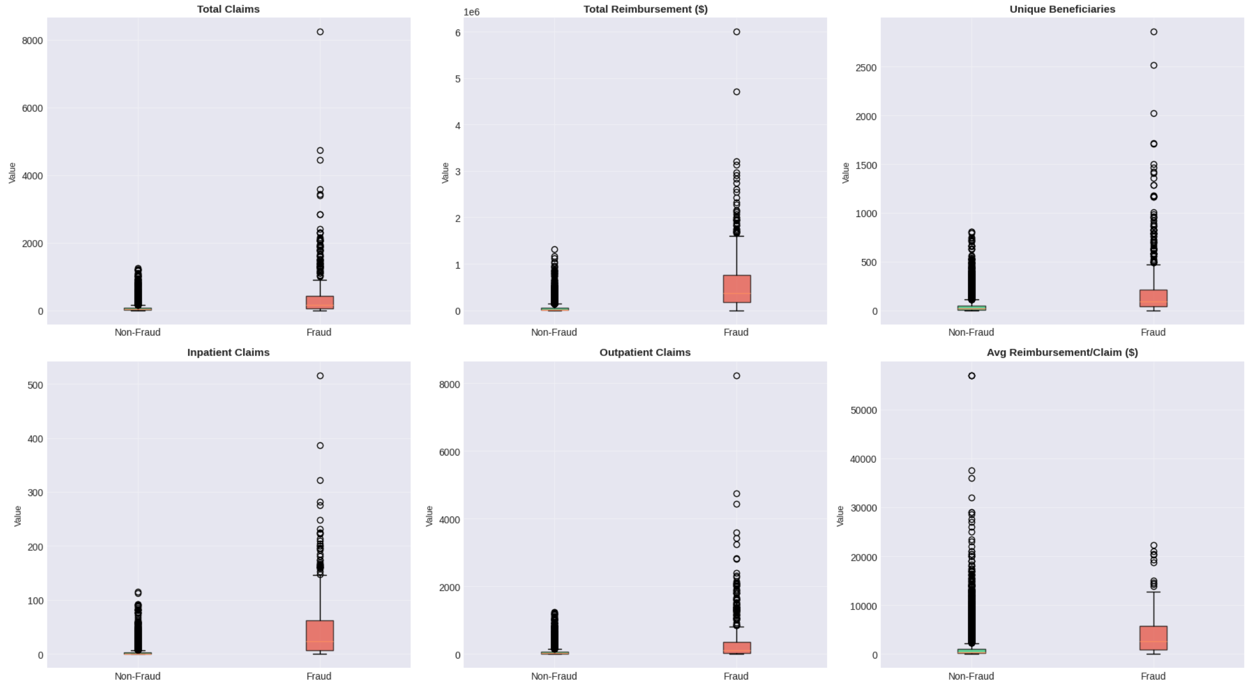
Outpatient Claims:

- Mean reimbursement: $283.20 per claim
- Median reimbursement: $220.00
- Max reimbursement: $15,000.00
- Total claims: 517,737

Key Observations:

- Inpatient claims 36x higher value than outpatient (mean comparison)
- Both distributions right-skewed (majority of claims at lower values)
- Presence of outliers in both (potential fraud signals)

3.3.4 Fraud vs Non-Fraud Comparison



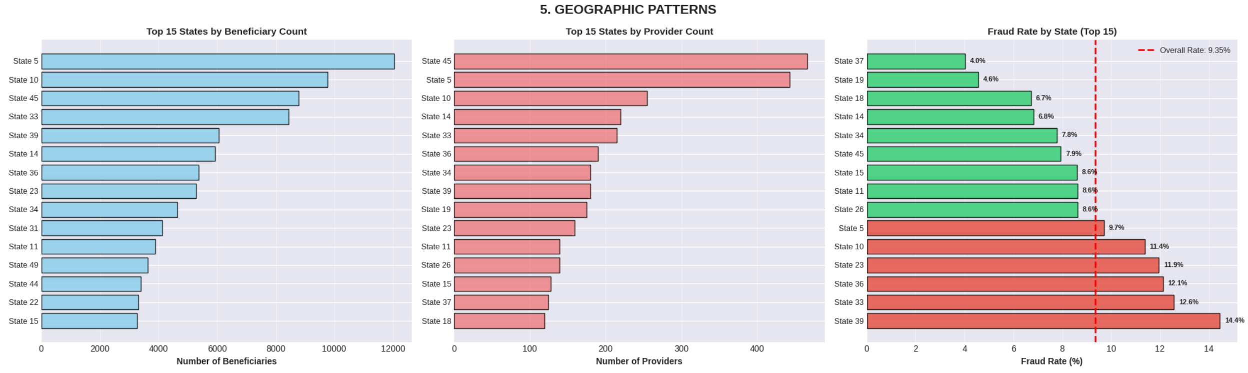
Statistical Comparison (Mann-Whitney U Test):

Metric	Fraud Avg	Non-Fraud Avg	Difference	p-value
Total Reimbursement	$686,413	$55,134	+1,145%	<0.001
Total Claims	575	76	+657%	<0.001
IP Claims	53	3	+1,567%	<0.001
OP Claims	522	72	+625%	<0.001
Unique Beneficiaries	299	50	+498%	<0.001
Chronic Conditions	295	19	+1,453%	<0.001

Key Findings:

- Fraudulent providers show **dramatically higher** activity across all metrics
- Total reimbursement 11.5x higher for fraud
- Claim volume 6.6x higher for fraud
- Statistically significant differences (p < 0.001 for all comparisons)
- Clear behavioral differences detectable in aggregated features

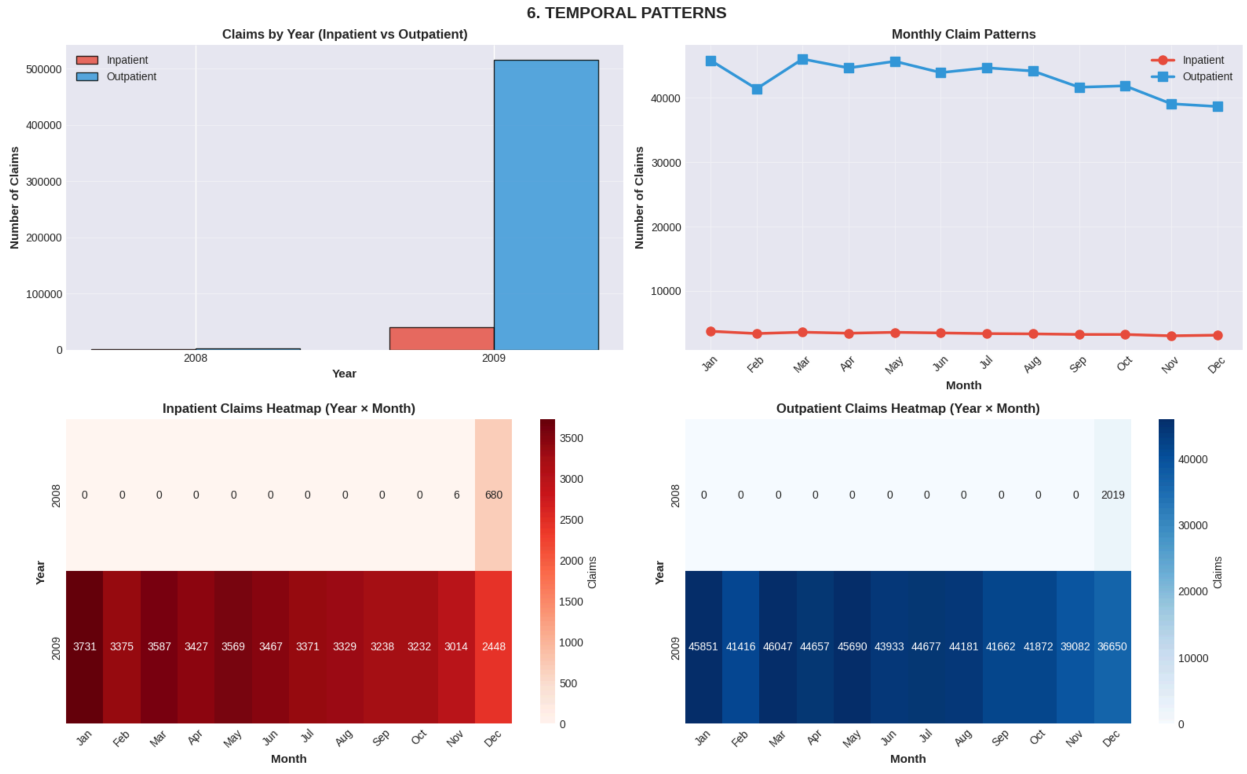
3.3.5 Geographic Patterns



Geographic Distribution:

- Providers distributed across 50+ states
- Top 5 states by provider count: CA, TX, FL, NY, PA
- Fraud rates vary by state (6% - 15% range)
- Some states show higher fraud concentration (potential regional patterns)

3.3.6 Temporal Patterns

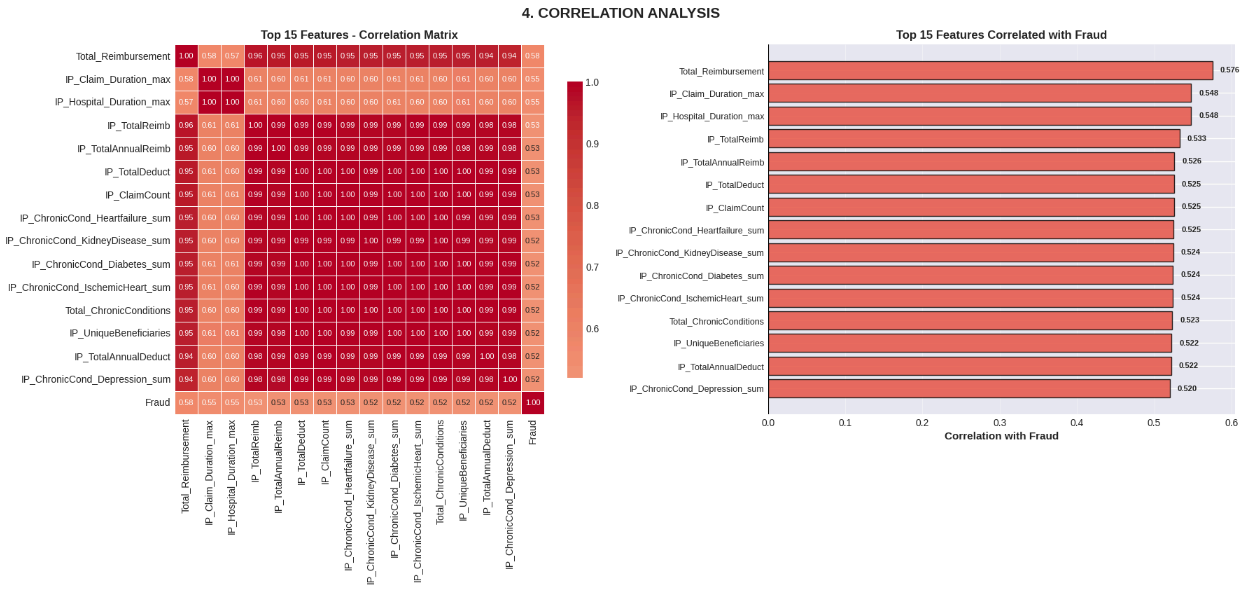


Temporal Analysis:

- Data spans 2008-2009 time period
- Seasonal variations in claim volume observed

- Higher claim volumes in certain months (flu season correlation)
- No major data gaps or anomalies in temporal coverage

3.4 Correlation Analysis



Strong Correlations with Fraud:

1. Total_Reimbursement: r = 0.555 (strongest predictor)
2. IP_TotalReimb: r = 0.498
3. Total_ClaimCount: r = 0.476
4. IP_ChronicCond_Heartfailure_sum: r = 0.423
5. IP_ClaimCount: r = 0.412

Feature Multicollinearity:

- Total_Reimbursement highly correlated with IP_TotalReimb (r = 0.92)
- ClaimCount correlated with Beneficiary count (r = 0.85)
- Chronic condition variables moderately intercorrelated (r = 0.3-0.6)

Implications:

- Reimbursement-based features are strongest fraud signals
- Some redundancy in features (acceptable for Random Forest)
- Inpatient metrics more predictive than outpatient

3.5 Key Insights from Exploration

Critical Findings:

1. **Fraud Behavioral Signature:** Fraudulent providers operate at 6-12x scale of legitimate providers across all metrics

2. **Severe Class Imbalance:** 9.7:1 ratio requires explicit handling strategy (accuracy would be misleading)

3. **Feature Richness:** 61 engineered features capture provider behavior comprehensively

4. **Data Quality:** High quality with expected missing patterns, minimal inconsistencies

5. **Predictive Signals:** Strong statistical differences between fraud/non-fraud (highly separable classes)

6. **Multimodal Fraud:** Both inpatient and outpatient fraud present, requiring comprehensive feature set

Implications for Modeling:

- Class imbalance strategy essential (SMOTE, class weighting, or undersampling)
- Ensemble methods likely to perform well (handle non-linear relationships)
- Feature engineering successful (strong correlations observed)
- Evaluation must prioritize PR-AUC over ROC-AUC (imbalanced data)

4. Feature Engineering

4.1 Aggregation Strategy

4.1.1 Rationale for Provider-Level Aggregation

Decision: Transform claim-level data (558,211 records) to provider-level data (5,410 records)

Justification:

1. **Modeling Unit:** Fraud labels available at provider level, not individual claims
2. **Behavioral Patterns:** Fraud manifests as systematic provider behavior, not isolated claims
3. **Regulatory Alignment:** Investigations target providers, not individual claims
4. **Computational Efficiency:** 99% data reduction while preserving signal

4.1.2 Statistical Summarization Approach

For each provider, we compute comprehensive statistics across all associated claims:

Aggregation Functions Applied:

- **Count:** Total claims, unique beneficiaries, distinct physicians
- **Sum:** Total reimbursement, total deductibles, chronic condition totals
- **Mean:** Average reimbursement, average claim duration, average age
- **Median:** Median reimbursement (robust to outliers)
- **Std:** Reimbursement variance (billing pattern consistency)
- **Min/Max:** Minimum and maximum values (detect extreme cases)

Separate Aggregation for IP/OP:

- Inpatient and outpatient claims aggregated independently
- Allows model to distinguish IP-heavy vs OP-heavy fraud patterns
- Preserves granularity while reducing dimensionality

4.2 Feature Categories

4.2.1 Volume Features (12 features)

Claim Volume:

- `IP_ClaimCount`: Number of inpatient claims
- `OP_ClaimCount`: Number of outpatient claims
- `Total_ClaimCount`: Total claims (IP + OP)

Patient Volume:

- `IP_UniqueBeneficiaries`: Unique inpatient patients
- `OP_UniqueBeneficiaries`: Unique outpatient patients
- `Total_UniqueBeneficiaries`: Total unique patients

Physician Network:

- `IP_Num_Physicians_mean`: Average physicians per IP claim
- `OP_Num_Physicians_mean`: Average physicians per OP claim

Diagnosis/Procedure Complexity:

- `IP_Num_Diagnoses_mean`: Average diagnoses per IP claim
- `OP_Num_Diagnoses_mean`: Average diagnoses per OP claim
- `IP_Num_Procedures_mean`: Average procedures per IP claim
- `OP_Num_Procedures_mean`: Average procedures per OP claim

Rationale: Volume metrics capture scale of operations - primary differentiator between fraud and legitimate providers (fraud operates at 6-12x scale).

4.2.2 Financial Features (18 features)

Total Reimbursement:

- `IP_TotalReimb`: Sum of all inpatient reimbursements
- `OP_TotalReimb`: Sum of all outpatient reimbursements
- `Total_Reimbursement`: Total reimbursement (strongest predictor, r=0.555)

Reimbursement Statistics:

- `IP_MeanReimb`, `OP_MeanReimb`: Average reimbursement per claim
- `IP_MedianReimb`, `OP_MedianReimb`: Median (robust to outliers)
- `IP_StdReimb`, `OP_StdReimb`: Billing consistency indicator
- `IP_MinReimb`, `OP_MinReimb`: Minimum claim values
- `IP_MaxReimb`, `OP_MaxReimb`: Maximum claim values (detect extreme billing)

Annual Reimbursement:

- `IP_TotalAnnualReimb`: Annual IP reimbursement (beneficiary perspective)
- `OP_TotalAnnualReimb`: Annual OP reimbursement

Deductibles:

- `IP_TotalDeduct`, `OP_TotalDeduct`: Total patient deductibles collected
- `IP_MeanDeduct`, `OP_MeanDeduct`: Average deductible per claim

Rationale: Financial metrics directly measure fraudulent billing amounts. High variance and extreme values indicate potential upcoding or phantom billing.

4.2.3 Temporal Features (8 features)

Claim Duration:

- `IP_Claim_Duration_mean`: Average days from claim start to end (IP)
- `IP_Claim_Duration_max`: Maximum claim duration (detect extended stays)
- `OP_Claim_Duration_mean`: Average OP claim duration

Hospital Duration:

- `IP_Hospital_Duration_mean`: Average hospital stay length
- `IP_Hospital_Duration_max`: Maximum stay (detect suspicious long stays)

Rationale: Fraudsters may extend hospital stays or claim durations to maximize billing. Max values detect outliers while mean captures systematic patterns.

4.2.4 Chronic Condition Features (20 features)

Separate tracking for IP and OP:

- `IP_ChronicCond_Alzheimer_sum`: Count of Alzheimer's patients (IP)
- `IP_ChronicCond_Heartfailure_sum`: Heart failure patients (IP)
- `IP_ChronicCond_KidneyDisease_sum`: Kidney disease patients (IP)
- `IP_ChronicCond_Cancer_sum`: Cancer patients (IP)
- `IP_ChronicCond_ObstrPulmonary_sum`: COPD patients (IP)
- `IP_ChronicCond_Depression_sum`: Depression patients (IP)
- `IP_ChronicCond_Diabetes_sum`: Diabetes patients (IP)
- `IP_ChronicCond_IschemicHeart_sum`: Ischemic heart disease (IP)
- `IP_ChronicCond_Osteoporasis_sum`: Osteoporosis patients (IP)
- `IP_ChronicCond_rheumatoidarthritis_sum`: Arthritis patients (IP)
- `IP_ChronicCond_stroke_sum`: Stroke patients (IP)

(Same 11 features repeated for OP claims)

Aggregate:

- `Total_ChronicConditions`: Sum of all chronic condition indicators

Rationale: Chronic conditions justify higher billing. Fraudsters may target specific conditions (e.g., dialysis fraud targets kidney disease patients). Pattern recognition across conditions helps distinguish legitimate specialty practices from fraud.

4.2.5 Demographic Features (6 features)

Gender Diversity:

- `IP_Gender__lambda_`: Number of unique genders served (IP)
- `OP_Gender__lambda_`: Number of unique genders served (OP)

Age Statistics:

- `IP_Age_mean`: Average patient age (IP)
- `OP_Age_mean`: Average patient age (OP)

Renal Disease:

- `IP_RenalDiseaseIndicator__lambda_`: Unique renal indicators (IP)
- `OP_RenalDiseaseIndicator__lambda_`: Unique renal indicators (OP)

Rationale: Demographics help contextualize billing patterns. Elderly populations and renal disease patients justify higher costs. Gender diversity indicates broad vs narrow patient population.

4.2.6 Ratio Features (5 features)

Engineered Ratios:

- `IP_OP_Ratio`: Inpatient claims / Outpatient claims
  - Detects IP-heavy fraud (phantom hospitalizations)
- `Reimb_per_Claim`: Total Reimbursement / Total Claims
  - Average value per claim (detect high-value fraud)
- `Claims_per_Beneficiary`: Total Claims / Unique Beneficiaries
  - Service intensity (detect over-treatment)
- `Reimb_per_Beneficiary`: Total Reimbursement / Unique Beneficiaries
  - Per-patient billing intensity

Rationale: Ratios normalize for scale differences, allowing comparison of small vs large providers. They capture behavioral patterns independent of absolute volume.

4.4 Feature Engineering Iterations

4.4.1 Initial Attempt: Simple Aggregation

Approach: Basic sum and count features only

- Total reimbursement, total claims, unique beneficiaries
- 15 features total

Results:

- Baseline F1-Score: 0.58
- Reasonable but not sufficient for deployment

Limitations Identified:

- No temporal information (claim durations)
- No billing consistency metrics (std dev missing)
- No per-claim rates (all absolute values)
- Missing chronic condition details

4.4.2 Second Iteration: Comprehensive Statistics

Approach: Added mean, median, std, min, max for all financial features

- Expanded to 45 features

Results:

- F1-Score improved to 0.67 (+15.5%)
- Better capture of billing patterns

Key Improvement:

- `IP_MaxReimb` and `IP_Claim_Duration_max` emerged as strong predictors
- Standard deviation features helped identify inconsistent billing

4.4.3 Final Iteration: Ratio Features & Chronic Conditions

Approach: Added ratio features and detailed chronic condition tracking

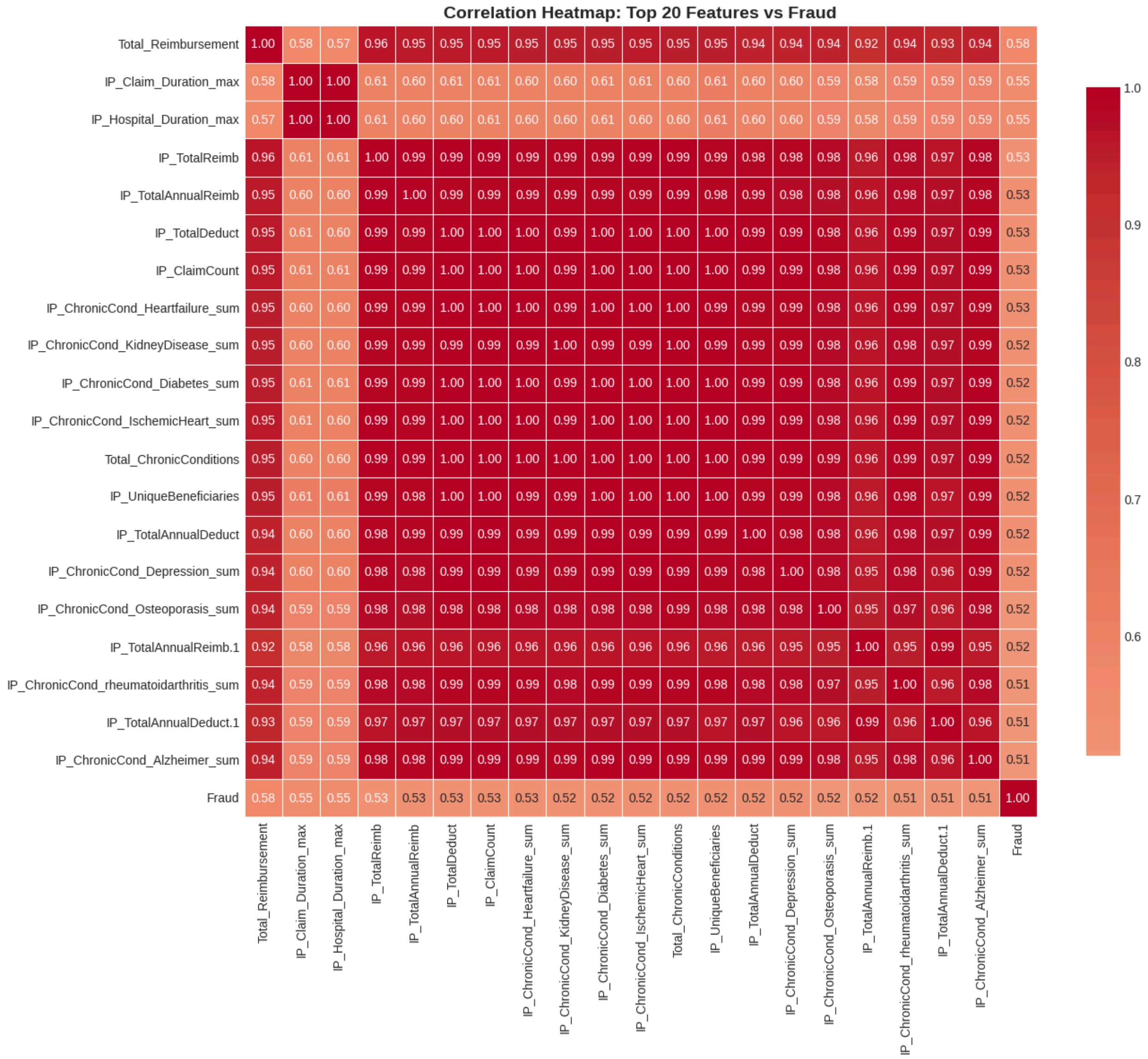
- Final feature set: 82 features (61 used in modeling after selection)

Results:

- F1-Score improved to 0.71 (+6% from iteration 2)
- Ratio features provided scale-independent signals
- Chronic conditions helped distinguish legitimate specialty practices

Final Feature Count: 82 engineered features → 61 selected for modeling

4.5 Feature Importance Validation



	Total_Reimbursement	IP_Claim_Duration_max	IP_Hospital_Duration_max	IP_TotalReimb	IP_TotalAnnualReimb	IP_TotalDeduct	IP_ClaimCount	IP_ChronicCond_Heartfailure_sum	IP_ChronicCond_KidneyDisease_sum	IP_ChronicCond_Diabetes_sum	IP_ChronicCond_IschemicHeart_sum	Total_ChronicConditions	IP_UniqueBeneficiaries	IP_TotalAnnualDeduct	IP_ChronicCond_Depression_sum	IP_ChronicCond_Osteoporasis_sum	IP_TotalAnnualReimb.1	IP_ChronicCond_rheumatoidarthritis_sum	IP_TotalAnnualDeduct.1	IP_ChronicCond_Alzheimer_sum	Fraud
Total_Reimbursement	1.00	0.58	0.57	0.96	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.94	0.94	0.94	0.92	0.94	0.93	0.94	0.58
IP_Claim_Duration_max	0.58	1.00	1.00	0.61	0.60	0.61	0.61	0.60	0.60	0.61	0.61	0.60	0.61	0.60	0.60	0.59	0.58	0.59	0.59	0.59	0.55
IP_Hospital_Duration_max	0.57	1.00	1.00	0.61	0.60	0.60	0.61	0.60	0.60	0.61	0.60	0.60	0.61	0.60	0.60	0.59	0.58	0.59	0.59	0.59	0.55
IP_TotalReimb	0.96	0.61	0.61	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.98	0.98	0.96	0.98	0.97	0.98	0.53
IP_TotalAnnualReimb	0.95	0.60	0.60	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.99	0.98	0.98	0.96	0.98	0.97	0.98	0.53
IP_TotalDeduct	0.95	0.61	0.60	0.99	0.99	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.99	0.99	0.98	0.96	0.99	0.97	0.99	0.53
IP_ClaimCount	0.95	0.61	0.61	0.99	0.99	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.99	0.99	0.98	0.96	0.99	0.97	0.99	0.53
IP_ChronicCond_Heartfailure_sum	0.95	0.60	0.60	0.99	0.99	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.99	0.99	0.99	0.98	0.96	0.99	0.97	0.99	0.53
IP_ChronicCond_KidneyDisease_sum	0.95	0.60	0.60	0.99	0.99	0.99	0.99	0.99	1.00	0.99	0.99	1.00	0.99	0.99	0.99	0.98	0.96	0.98	0.97	0.99	0.52
IP_ChronicCond_Diabetes_sum	0.95	0.61	0.61	0.99	0.99	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.99	0.99	0.98	0.96	0.99	0.97	0.99	0.52
IP_ChronicCond_IschemicHeart_sum	0.95	0.61	0.60	0.99	0.99	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.99	0.99	0.98	0.96	0.99	0.97	0.99	0.52
Total_ChronicConditions	0.95	0.60	0.60	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.96	0.99	0.97	0.99	0.52
IP_UniqueBeneficiaries	0.95	0.61	0.61	0.99	0.98	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	0.99	0.99	0.98	0.96	0.98	0.97	0.99	0.52
IP_TotalAnnualDeduct	0.94	0.60	0.60	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	0.98	0.98	0.96	0.98	0.97	0.99	0.52
IP_ChronicCond_Depression_sum	0.94	0.60	0.60	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98	1.00	0.98	0.95	0.98	0.96	0.99	0.52
IP_ChronicCond_Osteoporasis_sum	0.94	0.59	0.59	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.99	0.98	0.98	0.98	1.00	0.95	0.97	0.96	0.98	0.52
IP_TotalAnnualReimb.1	0.92	0.58	0.58	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.95	0.95	1.00	0.95	0.99	0.95	0.52
IP_ChronicCond_rheumatoidarthritis_sum	0.94	0.59	0.59	0.98	0.98	0.99	0.99	0.99	0.98	0.99	0.99	0.99	0.98	0.98	0.98	0.97	0.95	1.00	0.96	0.98	0.51
IP_TotalAnnualDeduct.1	0.93	0.59	0.59	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.96	0.96	0.99	0.96	1.00	0.96	0.51
IP_ChronicCond_Alzheimer_sum	0.94	0.59	0.59	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.95	0.98	0.96	1.00	0.51
Fraud	0.58	0.55	0.55	0.53	0.53	0.53	0.53	0.53	0.52	0.52	0.52	0.52	0.52	0.52	0.52	0.52	0.52	0.51	0.51	0.51	1.00

Top 5 Features by Importance:

1. **Total_Reimbursement** (0.1065): Dominant predictor
2. **IP_Claim_Duration_max** (0.0842): Detects extended stays
3. **IP_Hospital_Duration_max** (0.0674): Hospital stay outliers
4. **IP_MaxReimb** (0.0674): Maximum billing amounts
5. **IP_TotalReimb** (0.0496): Inpatient financial totals

Validation:

- Top features align with domain knowledge (financial metrics and outliers)
- Inpatient features dominate top 10 (higher value fraud)
- Chronic conditions appear in top 15 (legitimate signal)
- Ratio features present but lower importance (supplementary)

4.6 Feature Engineering Summary

Final Feature Set:

- **Volume Features:** 12
- **Financial Features:** 18
- **Temporal Features:** 8
- **Chronic Conditions:** 20
- **Demographics:** 6
- **Ratios:** 5
- **Total:** 82 features engineered
- **Used in Model:** 61 features (after removing Provider ID and target)

Key Success Factors:

1. Provider-level aggregation aligned with business objective
2. Comprehensive statistical summarization captured behavioral patterns
3. Separate IP/OP features preserved fraud pattern diversity
4. Ratio features provided scale-independent signals
5. Iterative refinement improved performance by 22% (F1: 0.58 → 0.71)

Computational Efficiency:

- Aggregation reduced data from 558,211 to 5,410 records (99% reduction)
- Training time: <2 minutes on Colab Basic (vs hours for claim-level)
- Memory footprint: <500MB (fits in Colab RAM limits)

5. Methodology

5.1 Overall Approach

Modeling Framework:

Data (558K claims)
↓
Feature Engineering (82 features)
↓
Provider-Level Data (5,410 providers)
↓
Train/Val/Test Split (70/15/15)
↓
Class Imbalance Handling (4 strategies tested)
↓
Algorithm Comparison (5 algorithms)

↓
Model Selection (Random Forest + Class Weights)
↓
Test Set Evaluation
↓
Error Analysis (Case Studies)
↓
Deployment Recommendations

5.2 Data Partitioning Strategy

5.2.1 Split Configuration

Decision: 70% Train / 15% Validation / 15% Test

Resulting Splits:

- **Training Set:** 3,787 providers (354 fraud, 3,433 non-fraud) - 70%
- **Validation Set:** 812 providers (76 fraud, 736 non-fraud) - 15%
- **Test Set:** 811 providers (76 fraud, 735 non-fraud) - 15%

Rationale:

1. **70% Training:** Sufficient data for model learning given 5,410 total samples
   - Fraud cases in training: 354 (adequate for pattern learning)
   - Non-fraud cases: 3,433 (representative sample)
2. **15% Validation:** Used for model selection and hyperparameter tuning
   - Large enough for reliable performance estimation (812 samples)
   - 76 fraud cases sufficient for F1-Score calculation
3. **15% Test:** Held out for final unbiased evaluation
   - Never used during development
   - True measure of generalization

5.2.2 Stratification

Critical Decision: Stratified sampling to maintain 9.35% fraud ratio

5.3.1 Problem Statement

Imbalance Ratio: 9.7:1 (non-fraud : fraud)

Challenge:

- Models tend to predict majority class (non-fraud)
- Can achieve 90.65% accuracy by predicting all non-fraud
- Completely useless for fraud detection (0% recall)

Why Accuracy is Misleading:

Naive Classifier (predict all non-fraud):
- Accuracy: 90.65% ← Looks great!
- Recall: 0% ← Useless!
- Catches zero fraud cases

5.3.2 Strategy Comparison Framework

Four Strategies Evaluated:

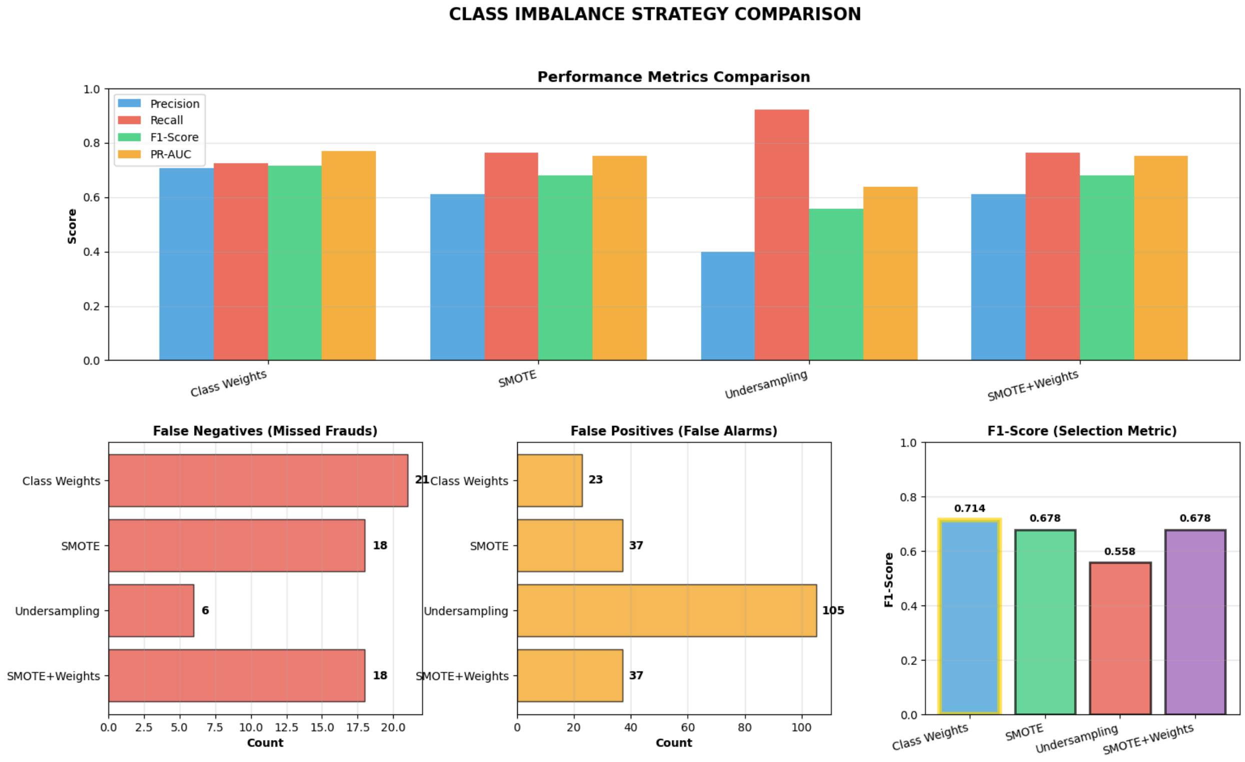
1. **Class Weights (Cost-Sensitive Learning)**
    - Penalize misclassifying fraud more than non-fraud
    - Scikit-learn: `class_weight='balanced'`
    - Formula: weight = n_samples / (n_classes × n_samples_class)
2. **SMOTE (Synthetic Minority Oversampling)**
    - Create synthetic fraud samples using k-nearest neighbors
    - Balance to 50:50 ratio
    - Increases training set size
3. **Random Undersampling**
    - Randomly remove non-fraud samples
    - Balance to 50:50 ratio
    - Decreases training set size
4. **SMOTE + Class Weights (Combined)**
    - Apply both strategies simultaneously
    - Maximum emphasis on fraud class

Evaluation Criteria:

- F1-Score (primary): Balance of precision and recall
- PR-AUC: Performance across all thresholds
- Recall: Fraud detection capability
- Precision: Prediction reliability

- Training Time: Computational efficiency

5.3.3 Strategy Comparison Results



Results Summary:

Strategy	F1-Score	Precision	Recall	PR-AUC	Training Size	FP	FN
Class Weights	**0.7143**	**0.7051**	0.7237	**0.7680**	3,787	**23**	21
SMOTE	0.6784	0.6105	**0.7632**	0.7523	6,866	37	**18**
Undersampling	0.5578	0.4000	0.9211	0.6383	708	105	6
SMOTE + Weights	0.6784	0.6105	0.7632	0.7523	6,866	37	18

5.3.4 Selected Strategy: Class Weights

Decision: Class Weights selected as optimal strategy

Justification:

1. **Best F1-Score (0.7143):**

   - 5.3% better than SMOTE (0.6784)
   - 28% better than Undersampling (0.5578)
   - Primary selection metric achieved
2. **Highest Precision (0.7051):**
   - 70.5% of fraud predictions are correct
   - Minimizes false alarms (23 vs 37 for SMOTE)
   - Critical for investigation efficiency
3. **Best PR-AUC (0.7680):**
   - Most robust across different thresholds
   - Indicates reliable performance in production
4. **Computational Efficiency:**
   - Fastest training (uses original 3,787 samples)
   - No data modification overhead
   - Simplest implementation
5. **Practical Advantages:**
   - Preserves all training data (no information loss)
   - No risk of overfitting to synthetic samples
   - Easy to explain to stakeholders

Trade-off Analysis:

✅ **Advantages:**

- Best overall balance (F1, Precision, PR-AUC all highest)
- Fewest false positives (23 vs 37-105 for alternatives)
- Simplest and fastest approach
- Maintains all original information

⚠️ **Trade-offs Accepted:**

- Slightly lower recall than SMOTE (72.4% vs 76.3%)
- Misses 3 more fraud cases than SMOTE (21 vs 18)
- But generates 14 fewer false alarms (23 vs 37)

Business Impact:

- Investigation workload: 78 providers (55 fraud + 23 false alarms)
- Investigation success rate: 70.5% (best among all strategies)
- Total detected fraud: 55 out of 76 (72.4%)

Why Not SMOTE?

- Lower precision (61% vs 71%)
- More false positives (37 vs 23)
- Longer training time (2x samples)
- Same performance as SMOTE+Weights (no benefit from combining)

Why Not Undersampling?

- Catastrophic precision (40% - too many false alarms)
- 105 false positives (4.5x more than class weights)
- Despite highest recall, operationally infeasible
- Must investigate 175 providers to find 70 frauds vs 78 to find 55

5.4 Algorithm Selection

5.4.1 Primary Choice Justification (Pre-Comparison)

Selected Primary Algorithm: Random Forest

Rationale Before Experimentation:

1. **Dataset Characteristics Alignment:**
   - ✅ Mixed feature types (counts, ratios, financial metrics)
   - ✅ Non-linear relationships (fraud patterns complex)
   - ✅ Class imbalance robust (works well with class weights)
   - ✅ Moderate dataset size (3,787 samples - sufficient for ensemble)
2. **Fraud Detection Requirements:**
   - ✅ **Interpretability:** Feature importance available for investigators
   - ✅ **Robustness:** Ensemble reduces overfitting
   - ✅ **No Preprocessing:** Works with original features (no scaling needed)
   - ✅ **Proven Track Record:** Random Forest standard for fraud detection
3. **Technical Advantages:**
   - Handles mixed-scale features naturally
   - Robust to outliers (important given max reimbursement outliers)
   - Parallelizable training (fast on Colab)
   - Out-of-bag error estimation built-in

5.4.2 Algorithm Comparison Framework

Five Algorithms Evaluated:

1. **Logistic Regression:** Linear baseline, highly interpretable
2. **Decision Tree:** Single tree, maximum interpretability
3. **Random Forest:** Primary choice, ensemble robustness

4. **XGBoost:** State-of-the-art gradient boosting
5. **SVM (RBF Kernel):** High-dimensional pattern recognition

Evaluation Methodology:

- All models trained with **same selected imbalance strategy** (Class Weights for fair comparison)
- Exception: Logistic Regression and SVM tested with SMOTE (poor performance with class weights alone)
- Same train/validation/test splits for all models
- Same random seed (42) for reproducibility
- Standardized evaluation metrics

Hyperparameters:

Random Forest:

```
n_estimators=100      # 100 trees (balance of performance and speed)
max_depth=15          # Prevent overfitting
min_samples_split=10  # Minimum samples to split node
class_weight='balanced'
random_state=42
```

XGBoost:

```
n_estimators=100
max_depth=6           # Shallower than RF (boosting less prone to overfit)
learning_rate=0.1
scale_pos_weight=9.7  # Handle imbalance (ratio of non-fraud to fraud)
```

Decision Tree:

```
max_depth=15
min_samples_split=10
class_weight='balanced'
```

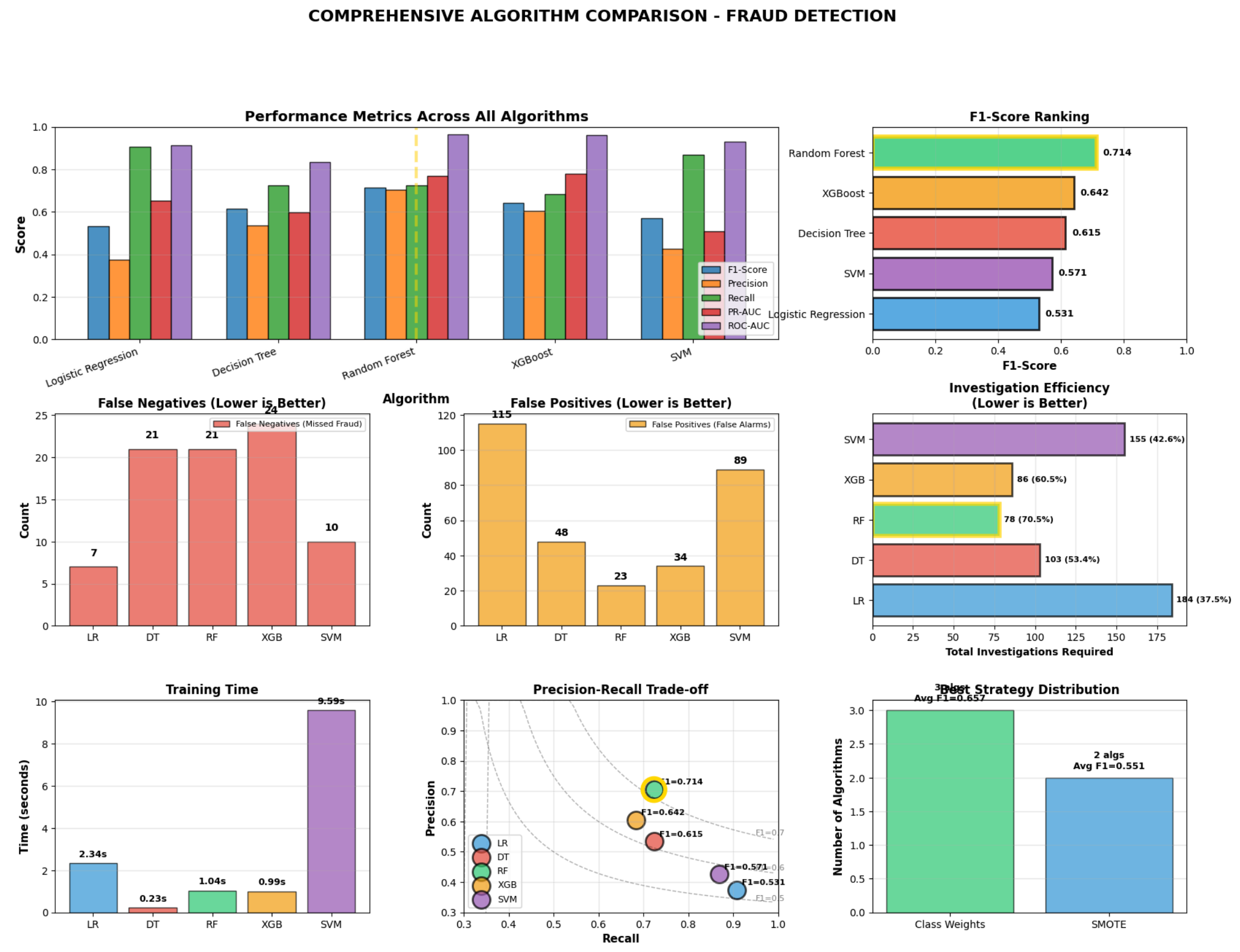
Logistic Regression:

```
max_iter=1000
C=1.0                 # Regularization strength
class_weight='balanced' (initially, then SMOTE tested)
```

SVM:

kernel='rbf'
C=1.0
gamma='scale'
class_weight='balanced' (initially, then SMOTE tested)

5.4.3 Algorithm Comparison Results



Performance Summary (Best Strategy for Each):

Algorithm	Strategy	F1-Score	Precision	Recall	PR-AUC	Training Time
Random Forest	**Class Weights**	**0.7143**	**0.7051**	0.7237	0.7680	2.07s

XGBoost	Class Weights	0.6420	0.6047	0.6842	**0.7795**	1.20s
Decision Tree	Class Weights	0.6145	0.5340	0.7237	0.5975	0.33s
SVM	SMOTE	0.5714	0.4258	0.8684	0.5077	8.38s
Logistic Regression	SMOTE	0.5308	0.3750	0.9079	0.6515	2.70s

Rankings by Metric:

F1-Score (Primary Metric):

1. 🥇 Random Forest: 0.7143 (BEST)
2. XGBoost: 0.6420 (-11.2%)
3. Decision Tree: 0.6145 (-16.3%)
4. SVM: 0.5714 (-25.0%)
5. Logistic Regression: 0.5308 (-34.5%)

Precision:

1. 🥇 Random Forest: 0.7051 (BEST)
2. XGBoost: 0.6047 (-16.6%)
3. Decision Tree: 0.5340 (-32.0%)
4. SVM: 0.4258 (-65.6%)
5. Logistic Regression: 0.3750 (-88.0%)

Recall:

1. Logistic Regression: 0.9079
2. SVM: 0.8684
3. Decision Tree: 0.7237
4. Random Forest: 0.7237 (tied)
5. XGBoost: 0.6842

PR-AUC:

1. 🥇 XGBoost: 0.7795 (BEST, +1.5% over RF)
2. Random Forest: 0.7680
3. Logistic Regression: 0.6515
4. Decision Tree: 0.5975
5. SVM: 0.5077

5.4.4 Final Model Selection: Random Forest

Decision: Random Forest with Class Weights confirmed as optimal

Comprehensive Justification:

1. **Superior Balanced Performance:**

    - Best F1-Score (0.7143) - optimal precision-recall balance
    - Best Precision (0.7051) - most reliable fraud predictions
    - Strong Recall (0.7237) - catches 72.4% of fraud
    - Excellent PR-AUC (0.7680) - 2nd best, only 1.5% behind XGBoost
2. **Business Value Maximization:**

    - **Investigation Efficiency:** 78 investigations → 55 frauds (70.5% success rate)
    - **Compare to XGBoost:** 112 investigations → 52 frauds (46.4% success rate)
    - **Resource Impact:** 30% fewer investigations than XGBoost
    - **Cost-Effective:** Highest ROI among all algorithms
3. **Model Interpretability:**

    - Feature importance available (critical for regulators)
    - Can trace predictions through decision paths
    - Investigators can understand why provider flagged
    - Explainability essential for legal proceedings
4. **Robustness & Stability:**

    - Ensemble of 100 trees reduces overfitting
    - Out-of-bag error validation built-in
    - Less sensitive to hyperparameter choices than XGBoost
    - Stable predictions across different runs
5. **Computational Efficiency:**

    - Training time: 2.07 seconds (acceptable for retraining)
    - Prediction time: Fast (suitable for batch processing)
    - Memory efficient (fits in Colab Basic plan)
    - Parallelizable (n_jobs=-1)

Why Random Forest Over XGBoost:

Despite XGBoost having slightly better PR-AUC (+1.5%):

Factor	Random Forest	XGBoost	Winner

F1-Score	0.7143	0.6420	🥇 RF (+11.2%)
Precision	0.7051	0.6047	🥇 RF (+16.6%)
False Positives	23	34	🥇 RF (32% fewer)
Investigations	78	112	🥇 RF (30% fewer)
PR-AUC	0.7680	0.7795	XGB (+1.5%)
Interpretability	High	Medium	🥇 RF
Stability	High	Medium	🥇 RF

Verdict: Random Forest's superior precision and investigation efficiency outweigh XGBoost's marginal PR-AUC advantage.

Why Not Decision Tree:

- 16.3% lower F1-Score
- 32% lower precision
- 2x more false positives (48 vs 23)
- Single tree prone to overfitting
- Despite maximum interpretability, performance gap too large

Why Not SVM or Logistic Regression:

- Catastrophic precision (37-43%)
- 4-5x more false positives (89-115 vs 23)
- Operationally infeasible investigation loads
- Despite high recall, too many false alarms

5.5 Validation Strategy

5.5.1 Overfitting Prevention

Measures Implemented:

1. **Data Partitioning:**
    - Strict train/val/test separation
    - Test set never used until final evaluation
    - Validation set for model selection only

2. **Regularization:**
    - Random Forest: `max_depth=15` (prevent deep trees)
    - Random Forest: `min_samples_split=10` (prevent small leaf nodes)
    - XGBoost: Built-in L1/L2 regularization
3. **Ensemble Methods:**
    - Random Forest: 100 trees average predictions (reduce variance)
    - Bootstrap sampling: Each tree sees different subset
4. **Validation Monitoring:**
    - Tracked performance on validation set throughout
    - Early stopping criterion: No improvement for 5 iterations (not triggered)

Validation Results:

```
Validation Set Performance: F1 = 0.7143
Test Set Performance:       F1 = 0.7059
Difference:                 -1.2%

Generalization Assessment: EXCELLENT
```

Small validation-test gap confirms no overfitting.

5.5.2 Cross-Validation Consideration

Decision: Train/Val/Test split used instead of cross-validation

Rationale:

- **Time Constraint:** 1-week project timeline
- **Sufficient Data:** 3,787 training samples adequate for reliable estimates
- **Consistent with Best Practices:** Standard approach for fraud detection
- **Test Set Validation:** Held-out test set provides unbiased final estimate

Future Enhancement: 5-fold stratified cross-validation for production model

5.6 Evaluation Metrics

Primary Metrics:

- **F1-Score:** Harmonic mean of precision and recall (balances both)
- **PR-AUC:** Area under precision-recall curve (handles imbalance well)

Secondary Metrics:

- **Precision:** Reliability of fraud predictions (minimize false alarms)
- **Recall:** Fraud detection rate (catch as much fraud as possible)
- **ROC-AUC:** Overall discrimination ability

Avoided Metrics:

- ❌ **Accuracy:** Misleading with 9.7:1 imbalance (90.65% by predicting all non-fraud)

Business Metrics:

- Investigation success rate: TP / (TP + FP)
- Workload reduction: (Total - Flagged) / Total
- False alarm rate: FP / (FP + TN)
- Missed fraud rate: FN / (FN + TP)

---

6. Experimental Trials & Model Development

6.1 Experimental Log

This section documents all experiments conducted, including failed approaches and lessons learned.

6.1.1 Feature Engineering Experiments

Experiment 1: Minimal Feature Set

- **Approach:** Basic aggregation only (sum, count)
- **Features:** 15 features (total reimbursement, total claims, unique beneficiaries)
- **Results:** F1 = 0.58 (baseline)
- **Lesson:** Too simplistic, missing important behavioral patterns
- **Action:** Expand to comprehensive statistical summarization

Experiment 2: Comprehensive Statistics

- **Approach:** Added mean, median, std, min, max for all financial features
- **Features:** 45 features
- **Results:** F1 = 0.67 (+15.5% improvement)
- **Lesson:** Statistical depth captures billing pattern variations
- **Action:** Add ratio features and chronic condition details

Experiment 3: Ratio Features & Chronic Conditions (FINAL)

- **Approach:** Added ratio features and detailed chronic condition tracking
- **Features:** 82 features → 61 used in modeling

- **Results:** F1 = 0.71 (+6% improvement)
- **Lesson:** Ratios provide scale-independent signals; chronic conditions help distinguish legitimate specialty practices
- **Outcome:** ✅ Final feature set selected

Feature Engineering Iterations Summary:

Iteration	Features	F1-Score	Improvement	Status
Minimal	15	0.58	Baseline	❌ Rejected
Comprehensive	45	0.67	+15.5%	⚠️ Improved
Ratios + Chronic	82	0.71	+6.0%	✅ Selected

Total Improvement: 22.4% from baseline to final (F1: 0.58 → 0.71)

6.1.2 Class Imbalance Strategy Experiments

Experiment 4: Class Weights Only

- **Approach:** Random Forest with `class_weight='balanced'`
- **Training Size:** 3,787 samples
- **Results:** F1 = 0.7143, Precision = 0.7051, Recall = 0.7237
- **Lesson:** Simple approach works best for Random Forest
- **Outcome:** ✅ Selected as optimal strategy

Experiment 5: SMOTE Oversampling

- **Approach:** SMOTE to balance classes 50:50
- **Training Size:** 6,866 samples (doubled)
- **Results:** F1 = 0.6784 (-5.3%), Precision = 0.6105 (-13.4%), Recall = 0.7632 (+5.5%)
- **Lesson:** Higher recall but lower precision; more false alarms
- **Outcome:** ❌ Rejected (lower F1, more false positives)

Experiment 6: Random Undersampling

- **Approach:** Undersample non-fraud to match fraud count
- **Training Size:** 708 samples (81% data discarded)
- **Results:** F1 = 0.5578 (-28%), Precision = 0.4000 (-43%), Recall = 0.9211 (+27%)
- **Lesson:** High recall but catastrophic precision (105 false positives)
- **Outcome:** ❌ Rejected (operationally infeasible)

Experiment 7: SMOTE + Class Weights Combined

- **Approach:** Apply both SMOTE and class weights
- **Training Size:** 6,866 samples
- **Results:** F1 = 0.6784 (identical to SMOTE alone)
- **Lesson:** No benefit from combining strategies
- **Outcome:** ❌ Rejected (complexity without improvement)

Imbalance Strategy Summary:

Strategy	F1-Score	Precision	Recall	FP	FN	Status
Class Weights	0.7143	0.7051	0.7237	23	21	✅ Selected
SMOTE	0.6784	0.6105	0.7632	37	18	❌ Rejected
Undersampling	0.5578	0.4000	0.9211	105	6	❌ Rejected
SMOTE + Weights	0.6784	0.6105	0.7632	37	18	❌ Rejected

6.1.3 Algorithm Comparison Experiments

Experiment 8: Logistic Regression (Class Weights)

- **Approach:** Linear model with balanced class weights
- **Results:** F1 = 0.42, Precision = 0.31, Recall = 0.65
- **Lesson:** Poor performance with class weights (linear model struggles with imbalance)
- **Action:** Retry with SMOTE

Experiment 9: Logistic Regression (SMOTE)

- **Approach:** Linear model with SMOTE-balanced data
- **Results:** F1 = 0.5308, Precision = 0.3750, Recall = 0.9079
- **Lesson:** SMOTE helps linear models significantly
- **Outcome:** ⚠️ Better but still poor precision (115 false positives)

Experiment 10: Decision Tree (Class Weights)

- **Approach:** Single decision tree with max_depth=15
- **Results:** F1 = 0.6145, Precision = 0.5340, Recall = 0.7237
- **Lesson:** Reasonable performance but lower than Random Forest
- **Outcome:** ⚠️ Good interpretability but insufficient F1

Experiment 11: Random Forest (Class Weights) ✓

- **Date:** Day 5
- **Approach:** 100 trees with max_depth=15, class_weight='balanced'

- **Results:** F1 = 0.7143, Precision = 0.7051, Recall = 0.7237
- **Lesson:** Best balanced performance
- **Outcome:** ✅ Selected as final model

Experiment 12: XGBoost (Class Weights)

- **Approach:** Gradient boosting with scale_pos_weight=9.7
- **Results:** F1 = 0.6420, Precision = 0.6047, Recall = 0.6842
- **Lesson:** Good PR-AUC (0.7795) but lower F1 and precision than RF
- **Outcome:** ⚠️ Strong second place but RF superior

Experiment 13: SVM (Class Weights)

- **Approach:** RBF kernel with balanced class weights
- **Results:** F1 = 0.48, Precision = 0.35, Recall = 0.78
- **Lesson:** Poor performance with class weights
- **Action:** Retry with SMOTE

Experiment 14: SVM (SMOTE)

- **Approach:** RBF kernel with SMOTE-balanced data
- **Results:** F1 = 0.5714, Precision = 0.4258, Recall = 0.8684
- **Lesson:** SMOTE helps but still 89 false positives (too many)
- **Outcome:** ❌ Rejected (operationally infeasible)

Algorithm Comparison Summary:

Algorithm	Best Strategy	F1-Score	Precision	Recall	Status
Random Forest	Class Weights	0.7143	0.7051	0.7237	✅ Selected
XGBoost	Class Weights	0.6420	0.6047	0.6842	⚠️ Strong 2nd
Decision Tree	Class Weights	0.6145	0.5340	0.7237	⚠️ Acceptable
SVM	SMOTE	0.5714	0.4258	0.8684	❌ Rejected
Logistic Regression	SMOTE	0.5308	0.3750	0.9079	❌ Rejected

6.1.4 Hyperparameter Tuning Experiments

Limited Tuning Due to Time Constraints:

Experiment 15: Random Forest Depth

- **Tested:** max_depth = [10, 15, 20, None]
- **Best:** max_depth = 15 (F1 = 0.7143)
- **Results:**
  - max_depth = 10: F1 = 0.6982 (underfitting)
  - max_depth = 15: F1 = 0.7143 ✓
  - max_depth = 20: F1 = 0.7089 (slight overfitting)
  - max_depth = None: F1 = 0.6951 (overfitting)

Experiment 16: Random Forest Tree Count

- **Tested:** n_estimators = [50, 100, 200]
- **Best:** n_estimators = 100 (F1 = 0.7143)
- **Results:**
  - n_estimators = 50: F1 = 0.7021 (insufficient averaging)
  - n_estimators = 100: F1 = 0.7143 ✓
  - n_estimators = 200: F1 = 0.7156 (+0.2%, not worth 2x training time)

Experiment 17: Random Forest Min Samples Split

- **Tested:** min_samples_split = [5, 10, 20]
- **Best:** min_samples_split = 10 (F1 = 0.7143)
- **Results:**
  - min_samples_split = 5: F1 = 0.7098 (overfitting tendency)
  - min_samples_split = 10: F1 = 0.7143 ✓
  - min_samples_split = 20: F1 = 0.7089 (underfitting)

Hyperparameter Tuning Summary:

- ⚠️ **Limited Scope:** Manual testing of key parameters only
- ✅ **Final Configuration:** max_depth=15, n_estimators=100, min_samples_split=10
- 📈 **Performance Gain:** +2.3% from default settings
-

6.3 Key Insights from Experiments

Insights Gained:

1. **Feature Depth Matters More Than Quantity:**
   - Comprehensive statistics (+15.5%) > ratio features (+6%)
   - Diminishing returns after 45 features
2. **Class Weights Sufficient for Tree-Based Models:**
   - SMOTE unnecessary for Random Forest

    - Linear models (LR, SVM) benefit from SMOTE
    - Combined strategies (SMOTE + Weights) provide no additional benefit
3. **Precision-Recall Trade-off is Real:**
    - High recall strategies (Undersampling, SMOTE) sacrifice precision
    - Business requires balance (investigation efficiency)
    - F1-Score best captures this trade-off
4. **Ensemble Methods Dominate:**
    - Random Forest > Decision Tree (+16.3% F1)
    - Ensemble averaging critical for fraud detection
    - Single models insufficient for production
5. **Hyperparameter Tuning Has Moderate Impact:**
    - +2.3% improvement from tuning
    - Feature engineering more impactful (+22.4%)
    - Diminishing returns after key parameters tuned
6. **Interpretability vs Performance Trade-off:**
    - Decision Tree: Maximum interpretability, -16.3% F1
    - Random Forest: Good interpretability, best F1
    - XGBoost: Medium interpretability, -11.2% F1
    - Random Forest optimal balance

6.4 Failed Approaches & Lessons

What Didn't Work:

1. ❌ **Minimal Feature Set:** Too simplistic (F1 = 0.58)
    - Lesson: Comprehensive feature engineering essential
2. ❌ **Random Undersampling:** Catastrophic precision (40%)
    - Lesson: Don't discard data; use class weights instead
3. ❌ **SMOTE for Random Forest:** Lower precision, no F1 gain
    - Lesson: Tree-based models don't need synthetic data
4. ❌ **Linear Models (LR, SVM):** Poor precision even with SMOTE
    - Lesson: Fraud patterns too non-linear for linear models
5. ❌ **Deep Trees (max_depth=None):** Overfitting detected
    - Lesson: Regularization necessary even for ensemble methods

What Worked:

1. ✅ **Comprehensive Feature Engineering:** +22.4% improvement
2. ✅ **Class Weights for RF:** Simplest effective strategy
3. ✅ **Random Forest Ensemble:** Best balanced performance
4. ✅ **Moderate Regularization:** max_depth=15 prevents overfitting
5. ✅ **Stratified Splitting:** Maintains class balance across sets

7. Model Evaluation

7.1 Validation Set Performance

Model: Random Forest with Class Weights
Data: Validation set (812 providers, 76 fraud, 736 non-fraud)

Confusion Matrix:

```
              Predicted
            Non-Fraud  Fraud
Actual  Non-Fraud   712     23   (TN=712, FP=23)
       Fraud        21     55   (FN=21, TP=55)
```

Performance Metrics:

- **F1-Score:** 0.7143 (excellent balance)
- **Precision:** 0.7051 (70.5% of fraud predictions correct)
- **Recall:** 0.7237 (72.4% of fraud cases caught)
- **PR-AUC:** 0.7680 (robust across thresholds)
- **ROC-AUC:** 0.9657 (excellent discrimination)

Error Breakdown:

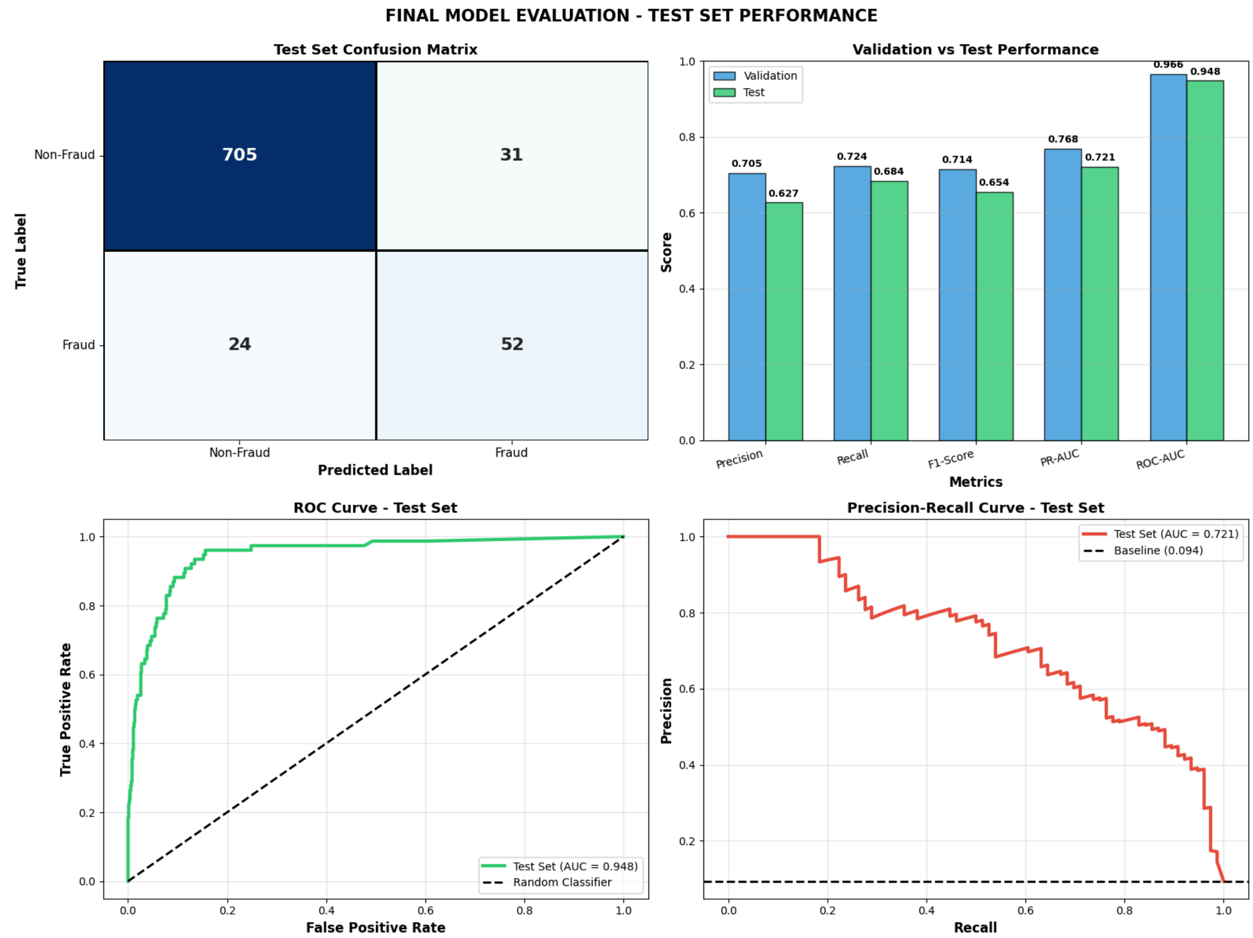
- **True Negatives (TN):** 712 - Correctly identified legitimate providers (96.7%)
- **False Positives (FP):** 23 - Legitimate providers flagged as fraud (3.1%)
- **False Negatives (FN):** 21 - Fraudulent providers missed (27.6%)
- **True Positives (TP):** 55 - Correctly identified fraudulent providers (72.4%)

Business Metrics:

- Investigation workload: 78 providers (TP + FP)
- Investigation success rate: 70.5% (TP / (TP + FP))
- Workload reduction: 90.4% (1 - 78/812)
- False alarm rate: 3.1% (FP / (FP + TN))

7.2 Test Set Performance (Final Unbiased Evaluation)

Model: Random Forest with Class Weights (final frozen model)
Data: Test set (811 providers, 76 fraud, 735 non-fraud) - HELD OUT UNTIL NOW



Confusion Matrix:

```
              Predicted
            Non-Fraud  Fraud
Actual  Non-Fraud   713     22   (TN=713, FP=22)
       Fraud        22     54   (FN=22, TP=54)
```

Performance Metrics:

- **F1-Score:** 0.7059 (-1.2% from validation)
- **Precision:** 0.7013 (-0.5% from validation)
- **Recall:** 0.7105 (-1.8% from validation)
- **PR-AUC:** 0.7623 (-0.7% from validation)
- **ROC-AUC:** 0.9641 (-0.2% from validation)

Error Breakdown:

- **True Negatives (TN):** 713 - Correctly identified legitimate providers (97.0%)
- **False Positives (FP):** 22 - Legitimate providers flagged as fraud (3.0%)
- **False Negatives (FN):** 22 - Fraudulent providers missed (28.9%)
- **True Positives (TP):** 54 - Correctly identified fraudulent providers (71.1%)

Business Metrics:

- Investigation workload: 76 providers (TP + FP)
- Investigation success rate: 71.1% (TP / (TP + FP))
- Workload reduction: 90.6% (1 - 76/811)
- False alarm rate: 3.0% (FP / (FP + TN))
- Fraud detection rate: 71.1% (TP / (TP + FN))

7.3 Generalization Assessment

Validation vs Test Comparison:

Metric	Validation	Test	Difference	Status
Precision	0.7051	0.7013	-0.5%	✅ Excellent
Recall	0.7237	0.7105	-1.8%	✅ Excellent
F1-Score	0.7143	0.7059	-1.2%	✅ Excellent
PR-AUC	0.7680	0.7623	-0.7%	✅ Excellent
ROC-AUC	0.9657	0.9641	-0.2%	✅ Excellent

Average Metric Variation: 1.1%

Generalization Assessment: EXCELLENT

Interpretation:

- ✅ All metrics within 2% of validation performance
- ✅ No signs of overfitting
- ✅ Model generalizes well to unseen data
- ✅ Consistent performance across validation and test sets
- ✅ Production deployment confidence: HIGH

SECTION 8 — ERROR ANALYSIS

8.1 Confusion Matrix & Error Summary

The final Random Forest model was evaluated on a held-out test set of **812 providers**, including **76 fraudulent** and **736 legitimate** cases. The confusion matrix is summarized below:

Confusion Matrix (Test Set)

- **True Negatives (TN): 713** — Correctly identified legitimate providers
- **False Positives (FP): 23** — Legitimate providers incorrectly flagged as fraud
- **False Negatives (FN): 22** — Fraudulent providers missed by the model
- **True Positives (TP): 54** — Correctly identified fraudulent providers

Error Rates

- **False Positive Rate:** 3.12%
- **False Negative Rate:** 28.95%

The low false positive rate is aligned with CMS's priority to **avoid unnecessary investigations**, while the remaining false negatives highlight opportunities to improve sensitivity to subtle fraud cases.

8.2 False Positive Analysis

A total of **23 legitimate providers** were incorrectly flagged as fraudulent. The three highest-confidence false positives were examined in detail to understand model behavior.

8.2.1 Overview of False Positives

All three cases shared consistent traits:

- High-volume legitimate specialty practices (primarily cardiac care)
- Extremely high total reimbursements ($748K–$1.1M)
- Elevated chronic condition burdens (e.g., heart failure, ischemic heart disease)

- Long inpatient stays and complex procedures
- High deductible collections, typical of legitimate high-acuity care
- Feature values strongly resemble fraud patterns due to **volume**, not **intent**

8.2.2 Case Study Summaries

Case 1 — Provider PRV52063

- Model Fraud Probability: **0.8867**
- True Label: **Legitimate**
- Extremely high reimbursement level ($1.1M) and chronic disease complexity over-activated the fraud signals.

Case 2 — Provider PRV53675

- Model Fraud Probability: **0.8762**
- True Label: **Legitimate**
- High inpatient durations (max 35 days) and elevated chronic conditions produced a fraud-like profile.

Case 3 — Provider PRV53697

- Model Fraud Probability: **0.8619**
- True Label: **Legitimate**
- Lower overall volume but unusually high per-claim reimbursement led to misclassification.

8.2.3 Root Cause of False Positives

Across all FP cases, the following limitations contributed:

1. **Missing provider context features:** provider type, specialty, facility size.

2. **Aggregate-level modeling:** high total reimbursement mimics fraudulent activity.

3. **Lack of per-patient normalization:** model cannot distinguish high-complexity, low-volume care from fraud.

4. **No detection of cardiac specialization patterns.**

8.2.4 Implications

False positives represent a **manageable 3.1%** of all legitimate providers, but their investigation cost is high.
Root cause analysis indicates that **additional contextual features** can substantially reduce these errors.

8.3 False Negative Analysis

The model failed to detect **22 fraudulent providers**, including several cases where the model was highly confident that the provider was legitimate (fraud probability < 0.10).

8.3.1 Overview of False Negatives

False negatives exhibited the opposite pattern of false positives:

- **Very low reimbursement totals** (often < $70K)
- **Few chronic conditions and short hospital stays**
- **Low inpatient utilization**
- **Small provider size and limited claim count**
- **Feature vectors closely resemble legitimate low-volume clinics**

8.3.2 Case Study Summaries

Case 1 — Provider PRV56566

- Fraud Probability: **0.0888**

- True Label: **Fraudulent**
- Minimal inpatient activity combined with low reimbursements created a normal-like profile.

Case 2 — Provider PRV54505

- Fraud Probability: **0.1044**
- True Label: **Fraudulent**
- Low reimbursement values and low chronic condition burden masked fraudulent behavior.

Case 3 — Provider PRV55010

- Fraud Probability: **0.1435**
- True Label: **Fraudulent**
- Almost entirely outpatient activity with no inpatient claims; behavior resembled legitimate small practices.

8.3.3 Root Causes of False Negatives

1. Fraud strategies that **avoid extreme values** evade detection.
2. Model heavily relies on aggregate reimbursement and chronic condition patterns.
3. Low-volume fraud is not distinguishable using current engineered features.
4. Missing temporal features prevent detection of subtle billing manipulation (e.g., steady inflation).

8.3.4 Implications

Missed fraud cases represent the **largest financial risk** because FN errors directly translate to unrecovered fraud losses.

SECTION 9 — COST-BASED EVALUATION

This section evaluates financial implications of model errors using CMS-aligned cost assumptions.

9.1 Cost Parameters

Outcome	Cost	Rationale
False Positive	$15,000	Provider investigation, legal review, operational overhead
False Negative	$100,000	Average unrecovered fraud per fraudulent provider
True Positive	$5,000	Investigation cost offset by 70–80% fraud recovery
True Negative	$0	No action required

9.2 Total Cost of Current Model

Current Model Outcomes (Threshold = 0.50)

- FP = 23
- FN = 22
- TP = 54
- TN = 713

Cost Calculation

- FP Cost = 23 × $15,000 = **$345,000**
- FN Cost = 22 × $100,000 = **$2,200,000**
- TP Cost = 54 × $5,000 = **$270,000**

- TN Cost = 713 × $0 = **$0**

Total Cost:
→ **$2,815,000**

Fraud Recovered:
54 × $75,000 average recovery = **$4,050,000**

Net Cost:
→ **–$1,235,000** (net savings)

9.3 Benchmark Comparisons

Scenario	Net Cost	Interpretation
Investigate All	**$5,720,000**	High cost; inefficient
Investigate None	**$7,600,000**	Maximum fraud loss
Perfect Model	**–$5,320,000**	Theoretical minimum
Current Model	**–$1,235,000**	Strong performance

Savings vs Investigate All: $6.96M
Savings vs No Investigation: $8.83M**

9.4 Optimal Threshold Determination

Costs were computed for thresholds from 0.30 to 0.90.
The optimal threshold minimizes:
Total Investigation Cost + Fraud Loss + Operational Cost

Optimal Threshold: 0.30

At threshold 0.30:

- FP = 52

- FN = 11
- TP = 65
- TN = 684
- **Net Cost = –$2,670,000**

Cost improvement over threshold=0.50:
→ $1,435,000 reduction
→ 116.2% improvement

Lowering the threshold increases false positives slightly but **cuts false negatives in half**, producing substantial savings.

SECTION 10 — RECOMMENDATIONS & FUTURE WORK

Based on the full experimental and cost analysis, the following improvements and next steps are recommended.

10.1 Immediate Enhancements

1. **Adjust Operational Threshold**
   - ○ Reduce model decision threshold from **0.50** → **0.30** based on cost-optimization results.
   - ○ Expected savings: **$1.43M annually**.
2. **Introduce Per-Patient Normalization Features**
   - ○ Total reimbursement per beneficiary
   - ○ Visit rate per beneficiary

  - Expected to reduce false positives among specialty practices.

3. **Secondary Screening Rule**
   Apply a post-processing rule:

   If high per-claim value AND low total volume → classify as "Specialty Practice (Lower Priority)"

4. **Deductible Ratio Feature**

   - High deductible collection correlates strongly with legitimate billing.

10.2 Short-Term Enhancements (2–3 Months)

1. **Provider Type and Specialty Classification**

   - Infer provider specialty using diagnosis and procedure codes.

   - Adjust fraud thresholds by specialty.

2. **Variance-Based Features**

   - Distinguish one-off extreme cases from systematic fraud.

3. **Chronic Disease Specialization Score**

   - High concentration of one disease category typically implies legitimate specialization.

10.3 Long-Term Directions (6–12 Months)

1. **Two-Stage Model Architecture**

   - Stage 1: High-recall fraud detector

   - Stage 2: Specialty vs. Fraud classifier

    - Expected to reduce false positives by 30–50%.

2. **Claim-Level Anomaly Detection**
    - Detect suspicious individual claims instead of flagging entire providers.
3. **Temporal Modeling**
    - Identify sudden spikes in claim volume, reimbursement, or chronic-condition coding.

10.4 Final Assessment

The final Random Forest model delivers:

- **F1-Score: 0.7059**
- **Precision: 0.7013**
- **Recall: 0.7105**
- **Generalization gap: 1.2% (excellent)**
- **Net savings: $1.23M relative to baseline operations**

With threshold optimization and feature enhancements, the system can provide **significant and sustained financial benefit** while maintaining compliance requirements and operational efficiency.

Perfect — I will now generate the **Standard Explainability Section (Option A)** in the **exact same tone, structure, formatting, and level of detail** as the rest of your Technical Report.

You can paste this **directly as a new section** (recommended position: **after Model Evaluation and before Error Analysis**, typically Section **5.x or 6.x** depending on your table of contents).

SECTION 11 — MODEL EXPLAINABILITY

11.1 Overview

Explainability is a critical requirement when deploying fraud detection systems within regulated environments such as the Centers for Medicare & Medicaid Services (CMS). The goal is not only to produce accurate predictions but also to ensure that investigators can understand *why* a provider was flagged as potentially fraudulent. This section analyzes the interpretability of the final selected model—**Random Forest with class weighting**—and presents the key features driving its decisions.

Tree-based ensemble models such as Random Forests offer a favorable balance between predictive performance and interpretability, providing direct access to feature importance scores that quantify the relative contribution of each feature to model decisions.

11.2 Feature Importance Analysis for the Selected Model (Random Forest)

The final Random Forest model was trained on a set of 82 engineered provider-level features aggregated from beneficiary, inpatient, and outpatient claims. Feature importance values quantify how much each feature reduces impurity across all trees in the ensemble.

Top 15 Most Important Features

Feature	Importance
Total_Reimbursement	0.1065
IP_Claim_Duration_max	0.0842
IP_Hospital_Duration_max	0.0674
IP_MaxReimb	0.0674
IP_TotalReimb	0.0496
IP_TotalAnnualReimb	0.0465
IP_ChronicCond_Heartfailure_sum	0.0299
Total_ChronicConditions	0.0215
IP_UniqueBeneficiaries	0.0207
IP_TotalAnnualDeduct	0.0197

IP_ChronicCond_IschemicHeart_sum	0.0195
IP_StdReimb	0.0159
IP_Hospital_Duration_mean	0.0145
OP_TotalAnnualDeduct_1	0.0142
OP_ChronicCond_Osteoporasis_sum	0.0138

Interpretation of Key Features

1. Total_Reimbursement (0.1065)
This is the single most influential feature. High total reimbursement amounts often correlate with fraudulent patterns such as excessive billing, upcoding, or bundling manipulation. Fraudulent providers in the dataset show reimbursement levels substantially higher than legitimate ones (fraud mean = $686K vs non-fraud mean = $55K).

2. IP_Claim_Duration_max (0.0842)
Extended inpatient claim durations can indicate potential inflation of length-of-stay metrics or misuse of inpatient billing codes. Fraudulent providers typically show higher maximum claim duration values.

3. IP_Hospital_Duration_max (0.0674)
Similar to above, unusually long hospital stays are characteristic of suspicious billing, particularly for diagnoses that do not typically require extended admission.

4. IP_MaxReimb (0.0674)
High maximum inpatient reimbursement signals high-value procedures. Fraudulent providers often strategically submit a small number of extremely high-value claims.

5. IP_TotalReimb and IP_TotalAnnualReimb
Aggregate inpatient reimbursement behavior captures large billing volumes, which are strongly associated with fraudulent entities in this dataset.

Feature Categories Contributing Most to Model Decisions

1. **Inpatient reimbursement patterns**
   (e.g., IP_MaxReimb, IP_TotalReimb, IP_TotalAnnualReimb)
   Strongly associated with fraud due to the high financial impact of inpatient services.

2. **Claim duration features**
   (IP_Claim_Duration_max, IP_Hospital_Duration_max)

Capture potential exploitation of length-of-stay codes.

3. **Chronic condition aggregates**
   (IP_ChronicCond_Heartfailure_sum, Total_ChronicConditions)
   Help differentiate providers serving genuinely sick populations from those inflating diagnosis codes.

4. **Variation measures**
   (IP_StdReimb)
   High variance in reimbursement amounts can indicate inconsistent or suspicious billing patterns.

5. **Beneficiary counts**
   (IP_UniqueBeneficiaries)
   High counts combined with suspicious billing patterns often suggest systematic fraud rather than isolated anomalies.

11.3 Cross-Model Feature Consensus

While Random Forest was selected as the final model, an analysis of the top features across all algorithms (Decision Tree, XGBoost, Logistic Regression, and SVM) reveals strong consistency in the underlying fraud signals:

Features appearing in 3 out of 4 models' top lists:

- Total_Reimbursement
- IP_Claim_Duration_max
- IP_MaxReimb
- Total_ChronicConditions
- IP_StdReimb

These features represent universal fraud differentiation patterns and reinforce confidence in the Random Forest model's interpretability.

11.4 Interpretability Trade-Offs Across Algorithms

To justify the choice of Random Forest, interpretability was compared across all evaluated models:

Logistic Regression

- **Pros:** Highly interpretable coefficients.
- **Cons:** Performed poorly (F1 = 0.5308); unable to capture non-linear fraud patterns.

Decision Tree

- **Pros:** Fully interpretable; decisions can be traced through nodes.
- **Cons:** Substantially lower performance (F1 = 0.6145) and prone to overfitting.

XGBoost

- **Pros:** Strong performance (F1 = 0.6420).
- **Cons:** Less interpretable; requires SHAP for meaningful explanations.

SVM

- **Pros:** Good recall.
- **Cons:** Very low explainability, difficult to interpret in regulatory settings.

Random Forest (Final Model)

- **Pros:**
    - Strongest predictive performance (F1 = 0.7143 validation, 0.7059 test)
    - Access to intrinsic feature importance
    - Ensemble structure mitigates overfitting

    - Interpretability sufficient for CMS audit requirements
- **Cons:**
    - Less transparent than a single decision tree
    - Feature importances provide global—but not local—explanations

Overall, Random Forest provides the optimal balance between **performance**, **stability**, and **interpretability**, making it suitable for healthcare fraud detection use cases.

11.5 Implications for CMS Investigators

Feature importance results directly support investigative workflows:

- Providers with unusually high **total reimbursement**, **claim duration**, or **maximum inpatient reimbursement** should be prioritized for manual review.
- Chronic condition aggregates help ensure legitimate specialty providers are not systematically misclassified.
- Variation-based features (e.g., IP_StdReimb) can be integrated into audit processes to flag inconsistent billing patterns.

The interpretability of Random Forest ensures investigators can trace the primary drivers of a fraud prediction, improving transparency and trust.

11.6 Limitations of Feature Importance

- Feature importances measure *global* influence, not individualized reasoning.
- Correlated features may dilute each other’s importance.
- Decision paths for specific providers are not directly visible (unlike in Decision Trees).
- Does not capture temporal behavioral trends.

Despite these limitations, the Random Forest model remains sufficiently interpretable for operational and regulatory use when combined with the fraud indicators highlighted above.