Healthcare Provider Fraud Detection Analysis

Predicting Potentially Fraudulent Providers and Analyzing Fraud Patterns

Naveen Singh Rawat

Problem Statement

Challenges in Healthcare Fraud Detection

- **Significant Financial Impact**: Health insurance fraud leads to major monetary losses through fake or inflated claims.
- **Types of Fraud**: Includes billing for unprovided services, duplicate claims, and charge exaggeration.
- **Manual Detection Issues**: Traditional fraud detection is costly and labor-intensive.
- **Detection Challenges**: Avoiding false positives and need for automated systems.

Project Goal

Objective and Outcomes of Fraud Detection Initiative

- **Primary Objective**: Predict potentially fraudulent healthcare providers using historical claims data.
- **Variable Identification**: Analyze data to uncover key indicators of fraudulent behavior.
- **Actionable Insights**: Generate insights to guide fraud mitigation strategies.
- **Outcome**: Developed a fraud detection dashboard using Power BI for monitoring.

Data Overview

Datasets and Key Variables Used in Analysis

Inpatient Claims
Includes admission details, diagnosis codes, procedures, and reimbursement data.

Outpatient Claims
Captures non-admitted patient interactions and claim specifics.

Beneficiary Data
Demographics, chronic conditions, and KYC identifiers.

Target Variable
Provider-level fraud flag (Yes/No) for supervised prediction.

Understanding the Data

Relationships, Characteristics, and Data Challenges

- **Entity Relationships**: BeneID links beneficiaries to claims; Provider links claims to entities.
- **Inpatient vs. Outpatient**: Inpatient includes admission data; Outpatient emphasizes outpatient visits.
- **Demographic Enrichment**: Beneficiary dataset reveals age, gender, and chronic conditions.
- **Data Quality Challenges**: Not complete data, Not properly defined, not easily understandable, mostly based on ID’s.

Fraud Risk Score Calculation

Outlier-Based Quantification of Suspicious Claims

- **Purpose**: Fraud risk score helps quantify provider-level risk using claims data.
- **Objective**: Prioritize suspicious providers to streamline investigations.
- **Technique**: Z-Score method detects outliers in volume and reimbursement.
- **Focus Metrics**: Claims volume and reimbursement per provider form the risk basis.

Logic Behind Fraud Risk Score

Z-Score Methodology and Risk Classification

- **Z-Score Formula**: $Z = (X - \mu) / \sigma$; where X is provider's value, μ is mean, and σ is standard deviation.
- **Medium Risk**: $Z > 2$ identifies moderate outliers warranting attention.
- **High Risk**: $Z > 3$ highlights significant anomalies likely linked to fraud.
- **Risk Categories**: Providers classified by volume, reimbursement, or both.

Steps to Calculate Fraud Risk Score

From Aggregation to Risk Classification

- **Aggregate Metrics**: Sum claims and average reimbursement for each provider.
- **Population Statistics**: Compute mean and standard deviation for baseline comparison.
- **Z-Score Calculation**: Use statistical formula to measure deviation from the mean.
- **Assign Risk Levels**: Classify providers into Normal, Medium Risk, and High Risk.

Photo by ABRAHAM POPOCATL on Unsplash

Why Fraud Risk Score is Important

Strategic Benefits and Operational Impact

Outlier Identification

Flags unusual claim behaviors missed by manual reviews.

Investigation Prioritization

Enables focused audits on high-risk providers.

Early Detection

Identifies fraud before it escalates financially.

Cost Efficiency

Reduces unnecessary investigations and optimizes resources.

Dashboard Walkthrough

Fraud Monitoring Dashboard in Power BI

KPI Cards

Shows total high-risk providers and percentage of fraudulent claims.

Scatter Plot

Visualizes claim volume and reimbursement, colored by risk score.

Provider Matrix

Sortable matrix with fraud impact by amount.

Key Findings

Fraud Patterns, Geographic Insights, and Risky Procedures

Key Insights	Findings
Total Providers	5410
High Risk Providers	146
Total Claim Outliers	12
Total Reimbursement Outlier	134
Total Reimbursement Amount	557 M
Reimbursement By High-Risk Providers	35 M
% Reimbursement High-Risk Providers	6.28%

Real-World Examples

Case Studies of High-Risk Providers

Case 1: Provider PRV51459
8,240 claims @ $281 avg; Z-score (volume) = 11.72; Billed 5+ visits/day/patient.

Case 2: Provider PRV52173
119 claims @ $11608 avg; @ $1381400 total, Z-score (reimbursement) = 3.5; coded High Risk Reimbursement.

Pattern Recognition
Volume anomalies and reimbursement spikes signal high fraud risk.

Visual Evidence
Charts show claim distribution and anomaly detection across providers.

Limitations

Challenges in Current Fraud Detection Framework

- **False Positives**: Some legitimate high-volume providers (e.g., ERs) might be flagged incorrectly.
- **Undetected Collusion**: Collaboration between providers and beneficiaries may evade detection.
- **Data Gaps**: Lack of standardized benchmarks for procedure costs limits analysis.
- **Model Assumptions**: Z-score method assumes normal distribution, which may not hold in real data.

Next Steps & Impact

Strategic Enhancements and Long-Term Goals

- **Short-Term Actions**: Audit top 20 high-risk providers and implement beneficiary-level scoring.
- **System Integration**: Embed fraud scoring into claims processing systems.
- **Advanced Modeling**: Adopt time-series and anomaly detection for predictive fraud analytics.
- **Operational Impact**: Enhanced fraud prevention, reduced financial loss, and improved trust.