HEALTHCARE PROVIDER FRAUD DETECTION - TECHNICAL REPORT

1. PROBLEM & OBJECTIVE

Problem: Healthcare fraud costs billions annually. Need to identify fraudulent providers from Medicare claims data.

Goal: Build ML model to detect fraudulent providers with high accuracy while minimizing false alarms.

Dataset:

- 5,410 providers total (506 fraudulent = 9% class imbalance)
- ~558,000 claims from inpatient and outpatient data
- ~138,000 beneficiary records with demographics and chronic conditions

2. WHAT WE DID

STEP 1: DATA INTEGRATION & PREPROCESSING

• Combined 4 separate datasets: Beneficiary, Inpatient, Outpatient, Labels

• Merged inpatient + outpatient claims (added ClaimType flag)

• Joined with beneficiary data on patient ID

• Converted dates to datetime and calculated patient age

• Fixed chronic condition encoding (2→0 for "No")

• Handled missing values (median for numeric, constant for categorical)

STEP 2: FEATURE ENGINEERING (KEY TO SUCCESS!)

Created 20+ provider-level features from claim-level data:

Financial Features:

- TotalReimbursed, AvgReimbursed
- TotalDeductible, AvgDeductible

Volume Features:

- TotalClaims (inpatient + outpatient separately)

- UniqueBeneficiaries, UniqueAttendingPhysicians

- ClaimsPerBeneficiary ratio

Patient Demographics:

- AvgPatientAge, MostCommonState

- Sum of chronic conditions (Alzheimer, Diabetes, Heart Failure, etc.)

Practice Patterns:

- InpatientRatio, AvgHospitalStay

STEP 3: EXPLORATORY DATA ANALYSIS

Key Findings:

• Fraudulent providers have higher average reimbursement amounts

• Fraudulent providers submit more total claims

• Strong correlations: TotalReimbursed (0.45), TotalClaims (0.38)

• Severe class imbalance: 91% non-fraud, 9% fraud → need SMOTE

STEP 4: MODEL DEVELOPMENT

Data Split:

- 80% training (4,328 providers), 20% testing (1,082 providers)

- Stratified split to maintain class balance

Preprocessing Pipeline:

- Numeric: Median imputation → StandardScaler

- Categorical: Constant imputation → OneHotEncoder

Class Imbalance Solution: SMOTE (Synthetic Minority Over-sampling)

- Generated synthetic fraud cases to balance training data

- Applied inside pipeline to prevent data leakage

Models Tested:

1. Logistic Regression (baseline, fast, interpretable)
2. Random Forest (ensemble, handles non-linearity, feature importance)
3. Gradient Boosting (sequential boosting, high performance)

Hyperparameter Tuning:

- Used RandomizedSearchCV on Random Forest
- 3-fold cross-validation, F1 scoring
- Best params: n_estimators=300, max_depth=20

STEP 5: EVALUATION

Metrics Used:

- ROC-AUC: Overall class separation ability
- PR-AUC: Performance on minority class (fraud)
- Precision: How often fraud predictions are correct
- Recall: How many fraud cases we catch
- F1-Score: Balance of precision and recall

Cost Analysis:

- False Negative (missed fraud): $10,000 per case
- False Positive (false alarm): $1,000 per case

3. RESULTS

FINAL MODEL: Random Forest (Tuned)

Performance Metrics:

• ROC-AUC: 0.9554 ✓ (Excellent discrimination)

• PR-AUC: 0.7570 ✓ (Strong on imbalanced data)

• Precision: ~0.75 (75% of fraud predictions correct)

• Recall: ~0.70 (Caught 70% of fraud cases)

• F1-Score: 0.72

• Accuracy: 94.6%

Confusion Matrix:

	Predicted Non-Fraud	Predicted Fraud
Actual Non-Fraud:	975 (TN	9 (FP)
Actual Fraud:	30 (FN)	68 (TP)

Interpretation:

✓ Correctly identified 68 out of 98 fraud cases (70% recall)

✓ Only 9 false alarms out of 984 non-fraud providers (0.9% FP rate)

✗ Missed 30 fraud cases (need improvement)

MODEL COMPARISON:

Model	ROC-AUC	PR-AUC	
Logistic Regression	0.9554	0.7570	
Random Forest	0.9429	0.6962	
Gradient Boosting	0.9410	0.7290	
Random Forest (Tuned)	0.9554	0.7570	← WINNER

TOP 10 MOST IMPORTANT FEATURES:

1. TotalReimbursed (18.5%)
2. AvgReimbursed (15.2%)
3. TotalClaims (12.9%)
4. UniqueBeneficiaries (8.9%)
5. AvgPatientAge (6.5%)
6. TotalDeductible (5.5%)

7. ClaimsPerBeneficiary (4.9%)

8. TotalInpatientClaims (4.1%)

9. Sum_ChronicCond_Alzheimer (3.8%)

10. Sum_ChronicCond_HeartFailure (3.4%)

Key Insight: Financial features dominate (top 3 = 47% importance)

COST-BENEFIT ANALYSIS:

Without Model:

- Miss all 98 fraud cases

- Cost: 98 × $10,000 = $980,000

With Model:

- False Negatives: 30 × $10,000 = $300,000

- False Positives: 9 × $1,000 = $9,000

- Total Cost: $309,000

SAVINGS: $671,000 (68.5% cost reduction) 💰

4. KEY FINDINGS

✓ Machine learning works! Achieved >95% ROC-AUC

✓ Feature engineering crucial - aggregating claims to provider level was key

✓ SMOTE successfully handled 9% class imbalance

✓ Random Forest best for interpretability + performance

✓ Financial and volume features most predictive

✓ Model saves ~$671K by catching fraud early

Limitations:

✗ Still misses 30% of fraud cases (room for improvement)

✗ No temporal patterns analyzed (all claims treated equally)

✗ Geographic features weak (state-level too broad)

✗ Model may not detect novel fraud schemes

✗ Requires periodic retraining as fraud evolves

6. CONCLUSION

Successfully built a fraud detection model with 95.54% ROC-AUC that saves $671K (68% cost reduction) by identifying fraudulent providers while keeping false alarms low.

The model is production-ready and provides actionable predictions for investigation teams. Random Forest balances strong performance with interpretability through feature importance.