# Healthcare Provider Fraud Detection

## Using Machine Learning to Combat Medicare Fraud

**DataOrbit Team Project**

Protecting $68B+ annually from fraudulent claims

Presenter 1

## The Problem & Objectives

### The Challenge

- **$68 billion** lost to healthcare fraud annually
- CMS can investigate only a **fraction** of suspicious cases
- **~10%** of providers are fraudulent
- Limited resources for manual investigation

### Our Objectives

- Detect fraudulent providers from multi-table claims data
- Handle severe class imbalance
- Provide explainable predictions for investigators
- Demonstrate business value by prioritizing high-risk cases

## Data Understanding

### Beneficiary Data

Patient demographics, coverage info, chronic conditions

### Inpatient Claims

Hospital admissions with financial & procedural data

### Outpatient Claims

Outpatient visit claim data

### Fraud Labels

Provider-level fraud indicators (Yes/No)

### Key Findings from EDA

- ✓ Fraudulent providers show **higher claim volumes**
- ✓ Unusual **billing patterns** detected
- ✓ Evidence of **upcoding** and **unbundling**
- ✓ Linked via **BeneID** (patient) and **Provider**

## Feature Engineering

**Goal:** Aggregate patient-level and claim-level data into **provider-level features** for modeling

### Count Features

- Total claims per provider
- Inpatient claim counts
- Outpatient claim counts
- Unique patients served

### Statistical Features

- Mean claim amounts
- Median claim amounts
- Standard deviation
- Min/max values

### Ratio Features

- % high-cost claims
- Chronic condition rates
- Inpatient/outpatient ratio
- Avg claims per patient

**Output:** provider_features.csv → Ready for modeling

## Modeling Approach

### Models Tested

- Logistic Regression (baseline)
- Random Forest (tuned)
- Gradient Boosting (sklearn)
- XGBoost (interpretable)

### Evaluation Metrics

- ✓ **Precision** - Accuracy of fraud predictions
- ✓ **Recall** - % of fraud cases caught
- ✓ **F1-Score** - Harmonic mean
- ✓ **ROC-AUC** - Overall discrimination
- ✓ **PR-AUC** - Primary metric (imbalance)

### Winner: Logistic Regression

Selected based on **highest PR-AUC (0.7077)** — most suitable for imbalanced fraud detection

Also provides **interpretability** through feature coefficients for investigators

## Results & Evaluation

### Model Performance

| Precision | Recall |
|---|---|
| 0.445 | 0.881 |
| **F1-Score** | **ROC-AUC** |
| 0.591 | 0.933 |

PR-AUC (Primary Metric)

**0.7077**

### Key Insights

**High Recall (88.1%)**
Catches most fraudulent providers

**Excellent ROC-AUC (93.3%)**
Strong overall discrimination

**Moderate Precision (44.5%)**
Some false positives - acceptable trade-off

### Top Fraud Indicators (Feature Importance)

- High inpatient claim counts
- Elevated average claim amounts
- High claim amount variance
- Unusual billing ratios
- More chronic condition patients
- Abnormal procedure patterns

## Business Impact & Value

### Reduced Financial Loss

- Reduce part of **$68B annual fraud**
- Focus on highest-risk providers
- Prevent future fraudulent claims
- ROI through recovered funds

### Investigation Efficiency

- **Prioritize** limited investigation resources
- **Ranked risk scores** for each provider
- Reduce time-to-detection
- Focus on actionable cases

### Explainability

- **Interpretable** model for auditors
- Clear feature importance
- Justifiable decisions in court
- Build trust with stakeholders

### Minimize False Positives

- Balance fraud detection vs. provider burden
- Reduce unnecessary investigations
- Maintain provider satisfaction
- Optimize investigation costs

## Future Enhancements

### Temporal Features

- Claim frequency over time periods
- Sudden behavioral changes
- Rolling averages & trends
- Seasonal pattern analysis

### Graph-Based Detection

- Network analysis of providers
- Detect collusion rings
- Referral pattern analysis
- Graph Neural Networks (GNNs)

### Anomaly Detection

- Isolation Forest for outliers
- Autoencoders for patterns
- Discover new fraud types
- Unsupervised learning methods

### Enhanced Explainability

- SHAP values for interpretability
- Investigator dashboards
- Individual risk explanations
- Audit trail support

### Current Limitations

⚠ No temporal/timestamp data
⚠ Limited clinical context (diagnosis codes)
⚠ Class imbalance challenges remain
⚠ Missing geographic/specialty data

Presenter 4

## Conclusion

### Project Achievements

- ✓ Multi-table data integration
- ✓ Effective feature engineering
- ✓ Multiple models evaluated
- ✓ Class imbalance handled
- ✓ Interpretable predictions
- ✓ Business value demonstrated

### Final Model Performance

| Precision | Recall | F1 | ROC-AUC | PR-AUC |
|---|---|---|---|---|
| 0.445 | 0.881 | 0.591 | 0.933 | 0.708 |

An **effective, explainable, and operationally viable** solution

Prioritize high-risk providers → Reduce fraud losses → Protect Medicare

Presenter 4