Predicting Fraud in Medicare Data - **Final Project Report**
DATA 205
Matvei Shaposhnikov

1. Introduction

Medicare is a critical component of the U.S. healthcare system, but it is also a frequent target of fraud and abuse. Each year, the federal government loses billions of dollars to improper billing, kickback schemes, false patient risk assessments, and other fraudulent behavior. Identifying problematic providers quickly and accurately is central to protecting program integrity and ensuring that healthcare dollars are spent appropriately.
This project uses federal administrative data to explore patterns of fraud-related exclusions and build predictive models that estimate the likelihood that a provider is excluded from federal healthcare programs. “Exclusion” refers to providers listed in the U.S. Department of Health and Human Services Office of Inspector General (OIG) List of Excluded Individuals/Entities (LEIE), generally due to confirmed fraud, abuse, or other serious violations.

Project Goals

The project is organized around three main questions:

1. Which factors are most associated with provider fraud or exclusion?
   I focus on provider type, geography, and Medicare billing patterns.
2. Can I build a model that predicts exclusion with useful accuracy?
   The goal is to develop a screening tool that could help agencies like the OIG prioritize providers for investigation.
3. How competitive is a logistic regression model compared to tree-based predictive models (Random Forest and XGBoost)?

Data Sources

I used two public U.S. government datasets:

1. **CMS Medicare Physician & Other Practitioners – by Provider**
   Publishing agency: Center for Medicare & Medicaid Services (CMS)
   URL:
   https://data.cms.gov/provider-summary-by-type-of-service/medicare-physician-other-practitioners/medicare-physician-other-practitioners-by-provider
   Content (keyed by National Provider Identifier (NPI)):
   - Provider type and specialty
   - State and ZIP code
   - Number of services and beneficiaries
   - Total Medicare payment amount and allowed amount
   - Beneficiary demographics and risk scores
2. **OIG LEIE Exclusions**
   Publishing agency: U.S. HHS Office of Inspector General (OIG)

URL: https://oig.hhs.gov/exclusions/exclusions_list.asp
Content: NPI (when present), name, state
Exclusion date
Exclusion type and statutory code

For modeling, I merged CMS and OIG by NPI to label CMS providers as excluded (1) or not excluded (0).

Tools and Methods

All analysis was conducted in Python using:
Pandas/Numpy, numpy for data cleaning and transformation
Matplotlib, seaborn for visualization
Statsmodels for logistic regression
Scikit-learn for Random Forest and evaluation metrics
Xgboost for gradient boosted trees
GitHub for version control and reproducibility

2. Data Preparation and Descriptive Statistics

OIG LEIE Data

The raw OIG file (cms_data/leie.csv) contained 8,297 rows and 17 columns.
Key steps:

1. Filtered to rows with valid NPI and set NPI as the index.
2. Dropped columns with little use or heavy missingness.
3. Required non-missing SPECIALTY and date of birth (DOB), leaving 7,723 exclusion records and 11 columns.
4. Converted numeric date fields (DOB and EXCLDATE) from “YYYYMMDD” integers to true dates.
5. Computed age_at_exclusion (in years) and year_excl (calendar year of exclusion)
6. Mapped exclusion codes (e.g., 1128a1) to readable categories such as:
   “Conviction – Medicare Fraud”
   “Unlawful Claims”
   “Felony – Healthcare Fraud”

Key descriptive findings (OIG):

1. The most common exclusion types were:
   Conviction – Medicare Fraud (2,921 cases)
   Unlawful Claims (2,388)
   Felony – Healthcare Fraud (982)
   Felony – Drugs (642)
2. Top specialties among excluded providers included Nurse/Nurses Aide, General Practice, and Family Practice. **(Fig. 1)**

3. Top states by number of exclusions were California, Florida, New York, Ohio, and Texas. A bar chart of the top 10 states shows California alone with 984 exclusions. **(Fig. 2)**
4. The age-at-exclusion distribution peaks in the 50–70 range. However, this is misleading as a causal story because investigations and legal processes can take years; many providers are excluded long after the underlying misconduct. **(Fig. 3)**

CMS Provider Data and Merge

The CMS provider files contained roughly one million rows each and 81 core columns.

1. I renamed Rndrng_NPI to NPI and used it as the index.
2. I retained provider identifiers, utilization counts, financial measures, beneficiary demographics, and risk scores.

To define the exclusion label, I joined CMS with the cleaned OIG data on NPI:

1. If a provider had a non-null EXCLDATE in LEIE, excluded = 1.
2. Otherwise, excluded = 0.

In the full CMS population, fewer than 0.1% of providers were excluded, so the raw data is extremely imbalanced.

Rebalancing and Feature Engineering (for ML)

To make supervised learning feasible, I rebalanced the CMS data:

1. Kept all excluded providers.
2. Randomly sampled about 1% of non-excluded providers using a uniform random variable.

This produced year-specific "prep" files (2022, 2023, ect). For modeling, I concatenated them into a combined dataset:

1. Approximate size: 37,000 providers.
2. Excluded providers: 841 (about 2.3% of this sample).

Feature engineering included:

1. Simple billing ratios, such as:
    a. Allowed amount / submitted charges for Medicare
    b. Allowed amount / payment amount for drugs
    c. Drug beneficiaries / total beneficiaries
2. Label-encoded categorical fields:
    a. state from Rndrng_Prvdr_State_FIPS
    b. type from Rndrng_Prvdr_Type
3. Selection of 68 numeric predictors, including age and risk, chronic condition percentages, utilization, financial totals, and ratios.

Descriptive insight (CMS):

A scatterplot of beneficiary average risk score vs. total Medicare payment (log scale) shows:

1. A dense cluster of providers with moderate risk (around 1–3) and modest payments.
2. A small group of high-payment providers with high-risk populations. These are natural candidates for closer scrutiny.

3. Modeling Approaches and Results

I used two families of models:

1. Logistic regression (interpretable odds ratios).
2. Tree-based models (Random Forest and XGBoost) for stronger prediction.

Logistic Regression: Provider Type and Risk

Model 1: Exclusion Odds by Provider Type

The first model examined how exclusion odds differ by provider type alone.

1. Collapsed rare or problematic types (very few providers, or 0%/100% exclusion) into "Other" to avoid numerical instability.
2. Fitted a logistic regression of the form:
   $\text{logit}(P(\text{excluded}=1)) = \alpha + \sum_k \beta_k \cdot I\{\text{provider type}=k\}$

Fit statistics:

1. Observations: 12,496
2. Pseudo $R^2 \approx 0.073$
3. Likelihood ratio test p-value $\approx 2.2\times10^{-69}$ (highly significant)

Key odds ratios (vs. reference type):

1. Pain Management: OR ≈ 12.1
2. General Practice: OR ≈ 5.8
3. Physical Medicine & Rehabilitation: OR ≈ 4.0
4. Psychiatry: OR ≈ 2.9
5. Family Practice: OR ≈ 2.6

Many specialties (e.g., Dermatology, Optometry) had odds ratios well below 1, indicating lower exclusion risk.

This simple model clearly highlights that certain specialties—especially Pain Management and some "family owned" primary-care-like fields—are much more likely to be associated with exclusion. **(Fig. 4)**

Model 2: Provider Type + State + Funding

To develop a more realistic risk model, I included provider type, state, and funding:

1. Response: excluded (0/1).
2. Predictors:
   a. Log-transformed funding: log(1+Tot_Mdcr_Pymt_Amt).
   b. Collapsed provider type (one-hot encoded).
   c. Collapsed state (one-hot encoded).

Fit statistics (representative run):

1. Observations: 12,496
2. Pseudo $R^2 \approx 0.10$
3. LL-Null = −3149.5 Log-Likelihood =−2821.7
4. Likelihood ratio test p-value $\approx 2.2\times10^{-89}$
5. ROC-AUC ≈0.74 on the merged sample.

Interpretation:

1. Even after controlling for state and total Medicare payments, high-risk specialties remain associated with significantly higher exclusion odds.
2. Higher total payments (after logging) are modestly associated with increased exclusion risk.
3. Some states have systematically higher or lower adjusted exclusion odds, likely reflecting utilization patterns and enforcement intensity.

This model is less flexible than the machine learning models but much more interpretable.

Random Forest: Binary Exclusion Classification

For the main predictive task, I trained a RandomForestClassifier on the rebalanced dataset (all years combined):

1. Sample size: ≈ 37,000 providers
2. Excluded providers: 841
3. Features: 68 numeric predictors
4. Settings: 200 trees, max depth 8, class_weight='balanced'

Performance (test set, threshold 0.5):

1. Overall accuracy: 0.90
2. Not excluded (class 0): precision 0.97, recall 0.92
3. Excluded (class 1): precision 0.28, recall 0.54
4. ROC-AUC: AUC ≈ 0.86 **(Fig. 5)**

Most important features:

1. Chronic condition percentages (e.g., cancer, atrial fibrillation, depression, tobacco and alcohol/drug use)
2. Beneficiary average age
3. Provider type (encoded)
4. Total Medicare allowed amount
5. Ratios of allowed to submitted charges and allowed to paid amounts

This suggests that both the clinical mix of patients and billing patterns are informative for fraud risk.

I also trained a reduced Random Forest using the top 10 features:

1. AUC ≈ 0.83
2. Accuracy ≈ 0.85

Thus, a compact, interpretable feature set still gives strong predictive performance.

By sweeping probability thresholds (e.g., 0.2, 0.4, 0.6, 0.8) and examining confusion matrices, I found that a threshold around 0.55 offered a reasonable trade-off between missing true exclusions and over-flagging providers.

XGBoost: Exclusion Type Prediction

To go beyond "excluded vs. not," I used XGBoost to predict exclusion type (e.g., Medicare fraud, unlawful claims, patient abuse/neglect):

1. Same features as the Random Forest model.
2. Target: statutory exclusion code (e.g., 1128a1, 1128b4), with “UNK” for non-excluded/unknown.

For major exclusion categories, ROC curves (one-vs-all style) showed:

1. Conviction – Medicare Fraud (1128a1): AUC ≈ 0.92
2. Unlawful Claims (1128b4): AUC ≈ 0.89
3. Felony – Healthcare Fraud (1128a4): AUC ≈ 0.92
4. Felony – Drugs (1128a3): AUC ≈ 0.93
5. Patient Abuse/Neglect (1128a2): AUC ≈ 0.94

These high AUC values indicate that the model can effectively distinguish providers likely to be excluded for specific types of violations. **(Fig. 6)**

4. Final Data Product and Implications

The final product of this project is a combined analytic toolkit:

1. **Interpretable logistic regression models**
   Quantify how much more (or less) likely each provider type is to be excluded, and how risk varies by state and funding.
   Provide clear, odds-ratio-based explanations suitable for policy and legal audiences.
2. **Random Forest exclusion risk model**
   A high-performing binary classifier (AUC ≈0.86) trained on a balanced sample.
   Uses clinical, demographic, and billing features to score each provider’s risk of exclusion.
   Probability thresholds can be tuned to reflect operational preferences (e.g., catching more fraud vs. minimizing false alarms).
3. **XGBoost exclusion-type model**
   Distinguishes between different types of fraud or misconduct, with AUC values around 0.9 for major categories.
   Offers insight into why a provider may be high-risk, not just whether they are high-risk.

In a realistic workflow, agencies could:

1. Use the Random Forest and XGBoost models to screen large CMS datasets and generate a prioritized list of high-risk providers.
2. Use the logistic regression models to explain those risks in simple terms: high-risk specialty, unusually high payments, or high-risk geography.
3. Direct auditors and investigators to focus first on providers where model risk and interpretability-based signals align.

This combination of predictive power and interpretability can make fraud detection more efficient and transparent.

5. Future Work

Potential extensions include:

1. Cost-sensitive threshold optimization, explicitly trading off the costs of false positives vs. false negatives.
2. Temporal features, such as changes in billing volume or risk scores over time, to detect emerging fraud earlier.
3. Specialty-specific models, tuned for particularly high-risk fields like Pain Management or Behavioral Health.
4. Calibration and fairness analysis, ensuring predicted probabilities are well-calibrated and performance is consistent across regions and specialties.
5. Prototype tools, such as a simple dashboard or script that ingests new CMS data, scores providers, and presents investigators with a ranked list and short explanations.

---

6. References and Acknowledgements

References

1. Centers for Medicare & Medicaid Services (CMS). “Medicare Physician & Other Practitioners – by Provider.” https://data.cms.gov/provider-summary-by-type-of-service/medicare-physician-other-practitioners/medicare-physician-other-practitioners-by-provider
2. U.S. Department of Health and Human Services, Office of Inspector General (OIG). “List of Excluded Individuals/Entities (LEIE).” https://oig.hhs.gov/exclusions/exclusions_list.asp
3. Pedregosa et al. “Scikit-learn: Machine Learning in Python.” JMLR, 2011.
4. Seabold & Perktold. “Statsmodels: Econometric and Statistical Modeling with Python.” Proc. SciPy, 2010.
5. Chen & Guestrin. “XGBoost: A Scalable Tree Boosting System.” KDD, 2016.

Full source code, notebooks, and detailed preprocessing steps are available in my GitHub repository:
https://github.com/matsha2266/DATA205

Acknowledgements

- **Professor Lori Perine and the MC Data Science Faculty Team** – For instruction, guidance, and feedback throughout the course and this project.
- **My father, Nikolai Shaposhnikov** – For sharing practical advice and methods from his experience as a senior data analyst.
- **CMS data analysis team** – For public documentation and context that supported deeper and more accurate analysis.