Healthcare Fraud Detection

CS - 226, Big Data Management, Fall 2023

Team Hector - 22

Yash Kathe

Devaki Kalyan Chandra Yadav Podila

Rohith Reddy Kancharakuntla

Shubham Mishra

Omkar Kadam

Agenda

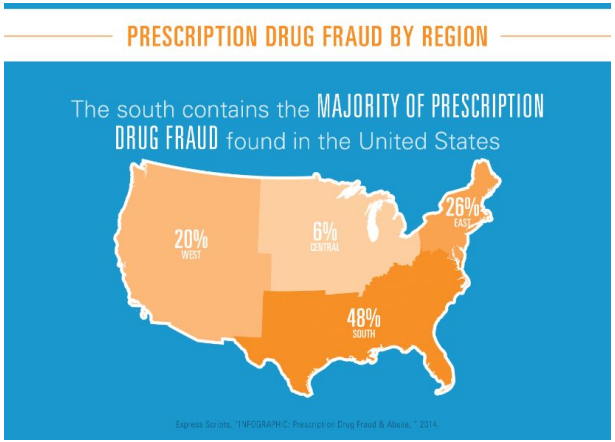
- Background
- Motivation
- Problem Statement
- Relevance to Big Data
- Related Work
- Methodology
- Visualization
- Evaluation
- Conclusion

Background

- Healthcare spending in the US accounts for 17.8% of GDP, and expenses are rising as a result of fraud which affects the government, insurance, and patients.
- National Health Care Fraud Enforcement Action results in charges involving over $1.4 billion in alleged losses
- Such stats highlight the vital role of preventing healthcare fraud for patient well-being and system integrity.

Health Care Fraud Stats

National Health Care Fraud Enforcement Action Results in Charges Involving over $1.4 Billion in Alleged Losses



Prescription drug fraud by region: The South leads with 48%, followed by the East with 26%, the West with 20%, and the Central region with 6%

Motivation

- Fraud detection traditionally relies on random sample examination by human experts.
- This approach may lead to misleading results, as random samples might not represent the full scope of fraud.
- Manual detection is often costly in terms of time and resources, making it an inefficient approach in modern fraud prevention
- Our project helps to identify fraudulent doctors, safeguarding patient well-being and preserving the integrity of the system.

Problem Statement

Our project aims to address

- Abuse of drug prescription
- Questionable billings by the pharmacies
- Identifying geographical hotspots for fraud abuse

Relevance to Big Data

Implementation of the Healthcare Fraud Detection Framework involves several key stages

- Vast Data is collected from Part-D Prescribers (Provider & Drug), Excluded Individuals and Entities Database (LEIE) and payments dataset which needed pre-processing.
- Spark SQL For achieving the query analysis-based objectives
- MLlib is used to develop the predictive models.
- Evaluation of Performance of the predictive machine learning models

Related Work - Paper 1

Title - Framework for Fraud Detection in Health Insurance
Authors - Nirmal Rayan

- This paper introduces a hybrid framework merging rule-based systems, supervised machine learning, and unsupervised techniques for detecting anomalies in extensive medical insurance claims data.
- Through clustering services and diseases, followed by fraud classification, the framework demonstrates a remarkable 209% improvement in fraud hit rates in case studies with a national insurer.
- Leveraging the strengths of supervised classification and unsupervised anomaly detection, the hybrid system automates the identification of suspicious claims, addressing challenges in manual fraud evaluation and scalability.
- The system not only automatically detects fraud patterns but also provides flagged claims with relevant remarks, aiding investigation teams in their analysis.

Related Work - Paper 2

Title - Fraud detection in health insurance using data mining techniques
Authors - Hossein Joudaki, Arash Rashidian, Behrouz Minaei-Bidgoli

- The paper employs big data analytics, including support vector machines and association rules mining, to detect healthcare insurance fraud by identifying anomalies in age, gender, and service availing patterns using a real hospital transaction dataset.
- A proposed framework integrates domain expertise rules, supervised machine learning (SVM), and unsupervised statistical anomaly detection to effectively identify fraudulent claims in large medical datasets.
- Addressing the challenge of rising healthcare costs due to fraud, especially in developing countries with government subsidies for underprivileged populations, the research emphasizes scalable and adaptive big data mining methods for cost management and improved healthcare delivery

Related Work - Paper 3

Title - Medicare Fraud Detection using Machine Learning Methods
Authors - Richard A. Bauder, Taghi M. Khoshgoftaar

- Applied machine learning techniques (Bayesian classifiers, support vector machines, decision trees, neural networks) for Medicare fraud detection in extensive medical claims datasets.
- Conducted an evaluation on a Medicare dataset with fraud labels, using a comparative analysis addressing class imbalance through oversampling and undersampling.
- Found under-sampling to perform better, with supervised methods significantly outperforming unsupervised methods. Argued that rule-based healthcare fraud systems struggle with the complexity of modern health insurance programs like Medicare

Methodology - Data Description

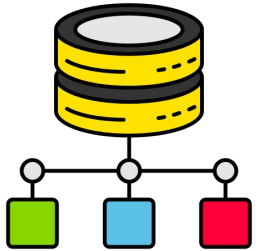
Data.CMS.gov

Medicare Part-D Prescribers (Provider & Drug)

List of Excluded Individuals and Entities Database (LEIE)

Payments Received by Physicians from Pharmaceuticals

Methodology - Data Description



Medicare Part-D Prescribers (Provider & Drug)

The Part D Prescriber dataset provides comprehensive information on prescription utilization, payments, and prescriber details

List of Excluded Individuals and Entities Database (LEIE)

The LEIE includes information on both individual healthcare professionals (e.g., doctors, nurses) and healthcare entities (e.g., hospitals, clinics) that have been excluded.

Payments Received by Physicians from Pharmaceuticals

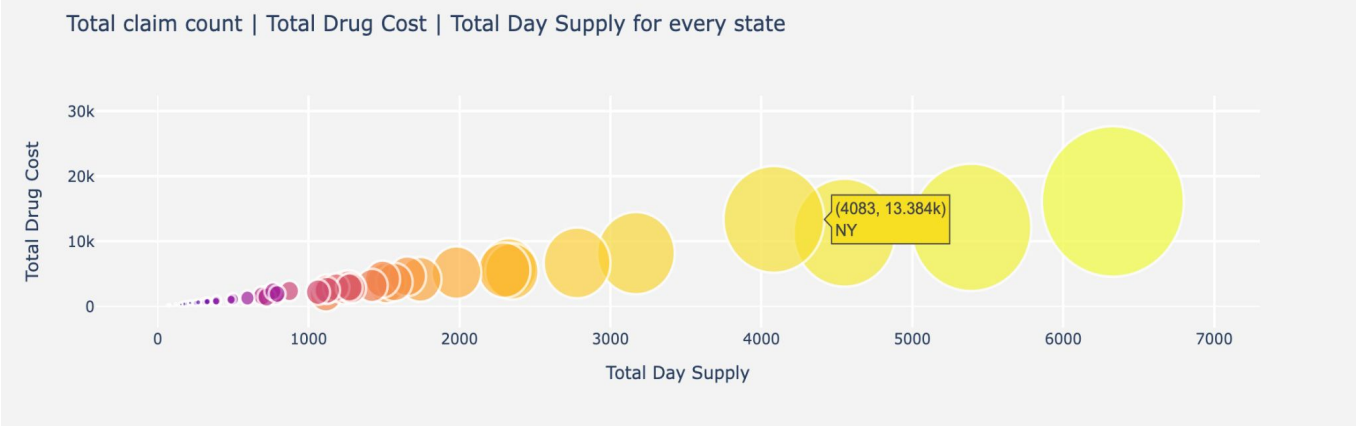
Payment Received dataset tracks payments to physicians from pharmaceutical companies, including drug associations.

Core Methodology

- Performed aggregations, column transformations and filtering of data using PySpark
- Joining Part-D prescriber on Payment using first name, last name, city and state to get total amount of prescribed drugs in USD
- Further this data is joined with LEIE with ‘NPI-National Provider Identifier’ to get the fraud NPIs
- Using this data to predict target variable ‘is_fraud’ using Machine Learning Classification algorithms

Visualization

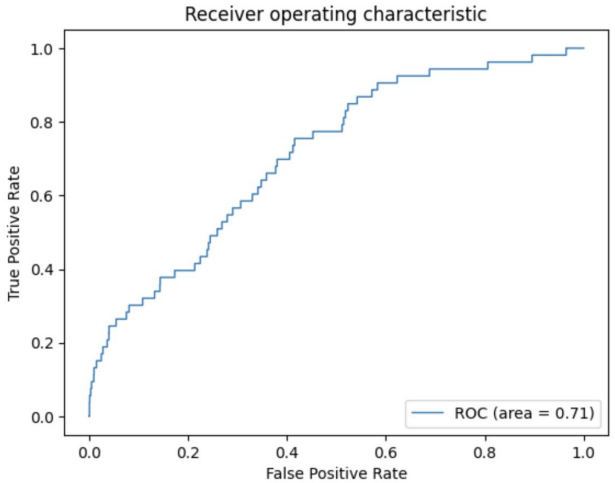
Highlights the relationship between total day supply and total drug cost for different states. The size of the bubbles represent the total claim count. This visualization helps in understanding how drug supply and costs vary across states, with larger bubbles indicating a higher number of claims.



Visualization

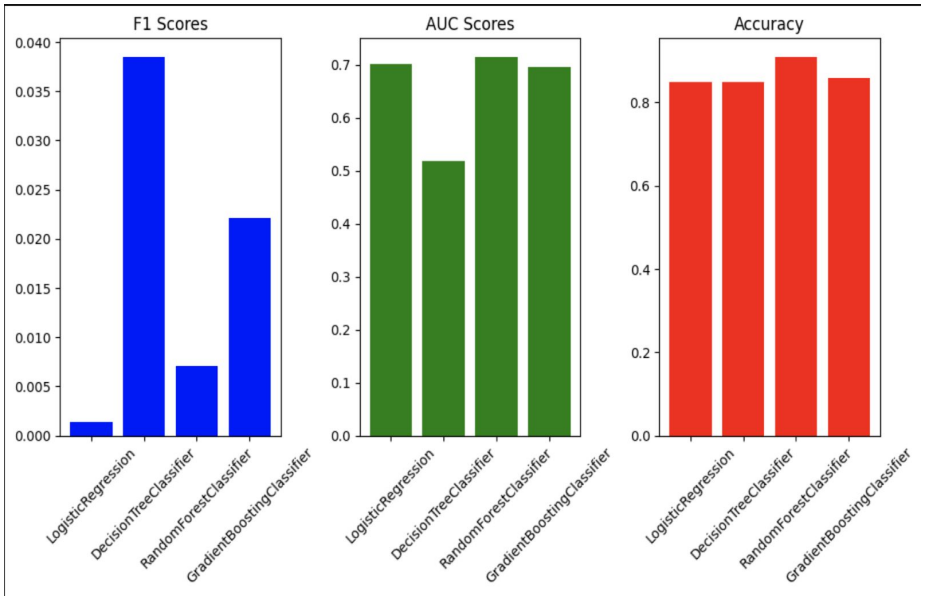
The Map of the United States showing the total number of claims in different states, measured in millions. The color gradient represents the scale of total claims, with darker colors indicating a higher number of claims.

Evaluation



- Receiver Operating Characteristic (ROC) curve with an area under the curve (AUC) of 0.71.
- The ROC curve plots the true positive rate against the false positive rate at various threshold settings.
- An AUC of 0.71 indicates a good level of model performance in distinguishing between the positive and negative classes.
- Generally, an AUC closer to 1 signifies a better-performing model, while an AUC closer to 0.5 suggests no discriminative power.

Evaluation



Evaluating them on the basis of F1 Scores, AUC Scores, and overall Accuracy.

- Logistic Regression emerges with the smallest F1 score, whereas the Random Forest Classifier leads with the highest, indicating a superior balance between precision and recall.
- The AUC Scores are relatively uniform across the models, suggesting that each has a comparable ability to differentiate between classes.
- When considering Accuracy, the Random Forest Classifier performs better than others, marking it as the most precise in this evaluation.

Conclusion

- We have successfully identified fraudulent activities within the healthcare sector by analyzing three different datasets i.e. Part-D, payment, and excluded individual records.
- We have performed data transformation, feature extraction and engineering processes and applied machine learning classification models on top of it and successfully demonstrated different evaluation metrics such as accuracy, f1 and roc curve

THANK YOU