PROJECT PROPOSAL

Big-Data-Medicare-Prescriber-Fraud-Detection Tool

- Siddharth Shah - ss16130
- Kunal Thadani - kmt9501
- Soumili Nandi - sn3699

Big Data, Section D
Spring 2024

By Professor: Dr. Amit Patel

Abstract:

The healthcare industry in the United States faces significant challenges due to the increasing costs of medical services, particularly in Medicare, which represents a substantial portion of healthcare spending. Healthcare fraud, estimated to be between 3% to 10% of total healthcare spending, has a significant impact on Medicare's financial health. To address this issue, this project focuses on developing a Medicare Fraud Detection model using big data analytics. By analyzing open data and employing anomaly detection and geo-segmentation metrics, the model aims to predict and detect fraudulent Medicare providers. The project utilizes public datasets such as the Part D Prescriber Dataset, Excluded Individuals and Entities (LEIE) dataset, and Payments Received dataset to build a robust data model. The methodology involves data exploration, cleansing, and preparation, followed by feature engineering to select effective feature sets for different fraud patterns. Machine learning models, including Random Forest, variational autoencoders are then employed to detect fraud patterns. The project's ultimate goal is to create a market-ready product that can help reduce healthcare fraud, leading to cost savings and improved compliance with government regulations.

Project Statement:

The "Big-Data-Medicare-Fraud-Detection" project aims to develop a robust Medicare Fraud Detection model using big data analytics to predict and detect fraudulent Medicare providers. The project will leverage public datasets such as the Part D Prescriber Dataset, Excluded Individuals and Entities (LEIE) dataset, and Payments Received dataset to build a comprehensive data model. By analyzing these datasets and employing anomaly detection and geo-segmentation metrics, the model will identify patterns indicative of fraud.

The project will follow a methodology that includes data exploration, cleansing, and preparation, followed by feature engineering to select effective feature sets for different fraud patterns. Machine learning models will be utilized to detect fraud patterns based on the selected features.

The project aims to contribute to reducing healthcare fraud, improving cost management in healthcare, and enhancing the overall credibility of the healthcare industry.

Objective:

The main objective of this project is to develop a Medicare Fraud Detection model using big data analytics to predict and detect fraudulent Medicare providers. Specific objectives include:

- Collect data from multiple trustworthy sources for building a basic data model to show the relationships among different datasets and identify key features for fraud detection.

- Apply Big data tools like Apache Spark/PySpark to visualize the data and feature selection and use map reduce for feature transformation if required. All of this is accountable in feature engineering.
- Developing a comprehensive AI model to identify fraud patterns based on features such as service providers (doctors, pharmacies), insurance subscribers (patients), geo-segment, and commonly misused drug prescriptions.
- Training and comparing multiple classifiers like Random Forest, Regression, autoencoders on the dataset to select the best model along with validation and testing.
- Establishing a benchmark metric to measure and evaluate the model's performance.
- Creating a UI for real time user query for detecting fraud and getting model results to develop a market-ready product.
- Developing a last layer of applying big data tools to visualize the result and show it to the user.
- Making fraud prevention easier by providing additional information for assessing fraud risks and identifying possible patterns.
- Helping payers maintain compliance with government prompt payment regulations by reducing the need for manual claims processing and investigation.
- Ultimately, contributing to the reduction of healthcare fraud, the improvement of healthcare costs management, and the enhancement of overall industry credibility.

Dataset :

1) Part D Prescriber Dataset
2) Excluded (LEIE) dataset
3) FDA datasets
4) LEIE Datasets: Office of Inspector General Reports

Number of records ~ 1 million records
Size : ~ 1GB

Tools:

1) Apache spark
2) Python
3) Deep Learning Models/ML
4) Tableau
5) ReactJs