Medicare Fraud Detection with Machine Learning

N. Nikitha Reddy[1], P. Mahidhar Reddy[2], K. Rithik Reddy[3], G. Kadirvelu[4]

[1,2,3]*Department of Artificial Intelligence and Machine Learning, Sphoorthy Engineering College, Hyderabad, India*

[4]*AP, Department of Artificial Intelligence and Machine Learning, Sphoorthy Engineering College, Hyderabad, India*

***Abstract*: This project is dedicated to building big data solutions with tangible applications at the intersection of healthcare and insurance industry. This Capstone project will build a Medicare Fraud Detection model to analyze open data and predict/detect the fraudulent Medicare providers based on fraud patterns, anomaly analysis and geo-demographic metrics, with FDA drug data's help this model is also trying to figure out the Opiate prescriptions and Overdoses related fraudulence. All datasets will be based solely on publicly available Medicare data from the Centers for Medicare and Medicaid Services (CMS), LEIE and other open data resources.**

Keywords: Machine learning, Class, imbalance Medicare fraud, Anomaly detetion CatBoost XGBoost Light GBM Gradient boosted machines Sampling

I.INTRODUCTION

Healthcare fraud poses a significant challenge, resulting in substantial financial losses within the Medicare/Medicaid and insurance industry. The Centers for Medicare and Medicaid Services (CMS) have implemented Medicare Part D programs since 2006 to detect and prevent fraud, waste, and abuse. However, traditional detection methods relying on random samples and manual expert analysis have limitations. The consequences of relying solely on these methods include potential misinformation and high detection costs.

This project seeks to address these challenges by leveraging machine learning methods to detect fraudulent Medicare claims. The aim is to extract insights from CMS open datasets, providing a proactive and automated approach to fraud detection, overcoming the limitations of current methods.

II. EXISTING SYSTEM

The existing systems for Medicare fraud detection predominantly rely on traditional methods, which have several limitations and drawbacks:

A.Random Sample Inspections

Traditional methods often involve random sample inspections conducted by human experts. This approach has inherent limitations as the samples may not be representative of the overall dataset, leading to potential inaccuracies in fraud detection.

B.Human Expert Analysis

Detection of fraudulent activities is heavily dependent on manual analysis by human experts. This process is time-consuming, resource-intensive, and susceptible to errors. Moreover, it may not effectively keep pace with the evolving and surge in healthcare fraud, coupled with the emergence of new fraud patterns and schemes, poses challenges that traditional methods are ill equipped to handle. The inadequacy of these methods in addressing modern complexities necessitates a more sophisticated and automated approach. sophisticated fraud patterns.

III. LITERATURE SURVEY

A.Survey of Major Area Relevant to Project

The literature survey provides an insightful exploration of the major areas relevant to the Medicare Fraud Detection project. It involves a comprehensive review of existing research, methodologies, and technologies applied in the domain of healthcare fraud detection, anomaly analysis, and machine learning. The survey serves as the foundation for understanding the current landscape and identifying gaps that the proposed project aims to address.

B.HealthCare Fraud Detection

Healthcare fraud detection has garnered significant attention in recent years due to the escalating financial losses within the Medicare/Medicaid and insurance industry. Traditional methods of fraud detection often involve manual inspection and random sampling, leading to limitations in accuracy and cost-effectiveness. Various studies have explored the challenges and shortcomings of these conventional approaches, emphasizing the need for automated, data-driven solutions.

C.Challenges in handling BigData

Projects dealing with healthcare data, particularly Medicare datasets, face challenges related to the sheer volume of data. The literature survey explores different strategies for efficient data handling, querying, and analysis. Techniques such as Google Cloud BigQuery API, PostgreSQL/MySQL, and Apache Spark with
PySpark are discussed as potential solutions.

D.Evaluation Metrics in Fraud Detection

The evaluation of fraud detection models involves the use of performance metrics, with Area Under the Receiver Operating Characteristic Curve (AUC) being a widely adopted measure. The literature survey delves into the significance of AUC in assessing the capabilities of binary classification methods, especially in scenarios with severe class imbalances.

IV. PROPOSED SYSTEM

The proposed system envisions the development of an advanced Medicare Fraud Detection model, leveraging machine learning techniques and big data solutions. The key components of the proposed system include:

A. Machine Learning Model

Implementing a sophisticated machine learning model capable of analyzing vast datasets to detect fraudulent Medicare providers. The model will utilize anomaly analysis, fraud patterns and geodemographic metrics for accurate predictions.

B. Big Data Solutions

Employing big data solutions to efficiently process and handle the large volumes of healthcare and insurance data. This includes leveraging technologies like Apache Spark and PySpark for in-memory computation and distributed computing to enhance speed and scalability.

C. Automation

Automating the fraud detection process to reduce reliance on manual inspections. Automation enhances the speed of detection, reduces operational costs, and allows for continuous monitoring of evolving fraud patterns.

D. Adaptability to Modern Challenges

Designing the system to adapt to evolving fraud patterns and challenges in the healthcare and insurance industry. This adaptability ensures that the model remains effective in identifying new and sophisticated fraud schemes.

E. Anomaly Analysis

Incorporating unsupervised learning techniques for anomaly analysis, aimed at identifying unusual behavior and deviations from the norm. This approach is crucial for detecting fraudulent activities without the need for labeled historical data.

F. Integration of FDA Drug Data

Exploring FDA drug data to enhance fraud detection capabilities, specifically focusing on opioid prescriptions and related fraudulence. This integration aims to provide a comprehensive understanding of prescription patterns and potential fraudulent activities.

G. Scalability and Efficiency

Addressing the challenges of handling large datasets by utilizing Google Cloud BigQuery API, PostgreSQL, or MySQL for efficient querying and analysis. The proposed system aims to overcome the limitations of traditional methods by embracing scalable and efficient technologies.

V. SOFTWARE AND HARDWARE REQUIREMENTS

A. Software Requirements:

The successful execution and development of the Medicare Fraud Detection project necessitate the following software components

Jupyter Notebook: The project will be developed using Jupyter Notebook, providing an interactive and collaborative environment for code development.

Python: The primary programming language for implementing machine learning algorithms, data analysis, and model development.

Google Cloud Platform (GCP): Leveraging GCP for cloud-based services, including BigQuery for efficient querying and processing of large datasets.

BigQuery: Google Cloud's fully-managed, serverless data warehouse for analytics. It will be used for handling and querying large-scale Medicare datasets.

Plotly: A Python graphing library that will be utilized for creating interactive and visually informative plots and charts.

Pandas: A powerful data manipulation and analysis library in Python, essential for handling datasets and performing data exploration.

B. Hardware Requirements:

Computational Resources: Adequate computational resources are necessary to handle significant data processing tasks. The exact specifications depend on the size of the datasets and the complexity of the machine learning models.

Memory: Sufficient RAM to support the efficient processing of large datasets and machine learning model training.

Storage: Adequate storage capacity to store datasets, intermediary results, and model files.

Processor: A multi-core processor to facilitate parallel processing and enhance overall performance.

Internet Connection: A stable internet connection is required for accessing cloud-based services and resources.

VI. TECHNIQUES AND ALGORITHMS

A. Supervised Learning:

Logistic Regression: Logistic Regression is a widely used supervised learning algorithm in fraud detection. It is particularly effective when the target variable is binary, making it suitable for classifying instances as fraudulent or non-fraudulent

B. Unsupervised Learning

Nearest-Neighbor-Based Techniques: Unsupervised learning, often referred to as anomaly detection, aims to find unusual behavior deviating from the norm. Nearest-neighbor-based techniques, such as k-Nearest Neighbors (k NN), are utilized to identify anomalies by measuring the distance between data points.

Clustering-Based Methods: Clustering algorithms, such as K-Means and DBSCAN, group data points based on similarity.

C. Ensemble Methods

Random Forest Classification: Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and control overfitting.

Boosting Models: Boosting is a machine learning ensemble technique that combines weak learners to create a strong learner.

VII. SYSTEM ARCHITECTURE

The system architecture of the Medicare Fraud Detection project is designed to integrate various components and ensure seamless interactions for efficient fraud detection in the healthcare and insurance industry. The architecture comprises the following key components:

A. Data Sources

The project relies on multiple data sources to gather comprehensive information for analysis. These sources include:

B. Medicare Provider Utilization and Payment Data (Part D Prescriber):

Contains aggregated data on prescriptions, utilization, payments, and charges by National Provider Identifier (NPI), Healthcare Common Procedure Code, and Place of Service.

C.Data Processing

The collected data undergoes preprocessing and cleansing to handle issues such as missing values, outliers, and inconsistencies.

D.Feature Engineering

Feature engineering involves selecting and transforming relevant features for different fraud patterns. The features include prescribing patterns, drug patterns, provider location, specialty, and more.

E.Machine Learning Model

The heart of the system is the machine learning model, which is trained to detect various fraud patterns in Medicare claims. Two main types of analytics are applie

F.Descriptive Analytics (Unsupervised Learning):

Involves anomaly detection methods such as nearest-neighbor-based techniques and clustering to find unusual behavior deviating from the norm.

G.Predictive Analytics (Supervised Learning): Utilizes logistic regression and random forest classification to build a predictive model for fraud detection. Ensemble and boosting models may be explored to optimize the predictive model

H.System Flow:

System Flows are system models that show the activities and decisions that system execute. They are useful for understanding complex system interactions because they visually show the back-and-forth interactions between systems and complex branching.

UML DIAGRAM

A UML diagram is based on Unified Modeling Language with the purpose of visually representing a system along with its main actors, roles, actions, artifacts and classes in order to understand, alter, maintain the document information about the system.

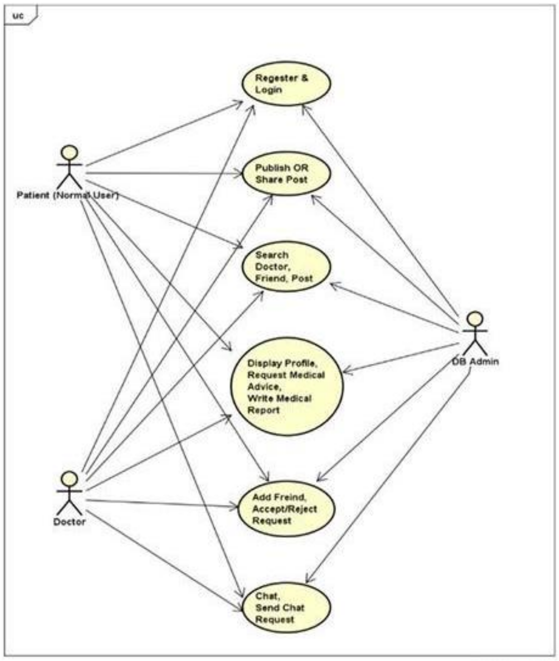
uc
Regester & Login
Publish OR Share Post
Search Doctor, Friend, Post
Patient (Normal User)
Display Profile, Request Medical Advice, Write Medical Report
DB Admin
Add Freind, Accept/Reject Request
Doctor
Chat, Send Chat Request

Fig 1. Use Case Diagram

SEQUENCE DIAGRAM

Sequence Diagrams are used for the documentation of various system's requirements andto flush out a system's design.

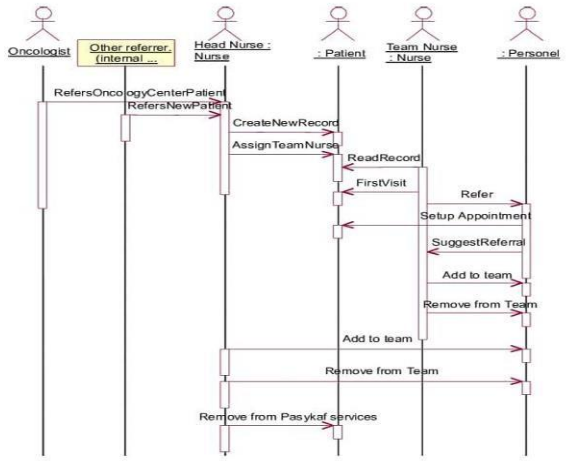

Fig 2. Sequence Diagram

XIII. IMPLEMENTATION

A.Environmental Setup Anaconda Distribution

Anaconda Distribution is a free and open-source software distribution that simplifies the process of working with Python and its associated scientific computing libraries. It aims to provide a comprehensive environment for data science, machine learning, and scientific computing tasks. Anaconda Distribution includes the core Python programming language along with commonly used libraries such as NumPy, Pandas, Matplotlib, SciPy, and scikit-learn, among others. These libraries are essential for data manipulation, numerical computation, data visualization, and machine learning.

One of the key components of Anaconda is the Anaconda Prompt. The Anaconda Prompt is a command-line interface (CLI) that comes bundled with the Anaconda distribution. It provides an environment specifically designed for managing and working with Anaconda environments, packages, and conda commands. The Anaconda Prompt is available on Windows, macOS, and Linux operating systems. Anaconda uses the conda package manager, which simplifies the installation and management of packages and dependencies. The Anaconda Prompt allows you to search for packages, install, update, and remove them using conda commands. It also facilitates the creation of environment specific package installations, making it easier to reproduce and share projects. The Anaconda Prompt provides direct access to the conda commands, allowing users to leverage the full power of the conda package manager.

B.Configuration and Setup:
The implementation of the Medicare Fraud Detection project requires a well-configured development environment. The following steps outline the environmental setup:

C.Jupyter Notebook
Ensure Jupyter Notebook is installed.
Use the following command to install Jupyter Notebook using pip:
pip install notebook

Launch Jupyter Notebook using the command:
jupyter notebook

D.Python Libraries
Install essential Python libraries using the following commands:
pip install bq-helper plotly pandas matplotlib plotly google-cloud

E.Google Cloud Platform Integration Obtain Google Cloud credentials (JSON file). Set the environment variable for credentials: Python

IX. EVALUATION

A. Datasets
Datasets used in the project include:
Part D Prescriber dataset:
This dataset contains information about healthcare providers, their specialties, and prescription details, including opioids.

B.List of Excluded Individulas and Entities (LEIE) datasets
The LEIE dataset provides information about individuals and entities excluded from participation in Medicare, Medicaid, and other federal healthcare programs.

C.Payments Received by Physician from Plarmaceuticals
This dataset contains details about payments made by pharmaceutical companies to physicians, providing insights into potential financial conflicts of interest.

D.FDA Drug Ingredients Data and Opioid Drug list
These datasets offer information about drug ingredients and specifically focus on opioids, aiding in the identification and analysis of opioid prescriptions.

E.Test Cases
Ensure that both Part D Prescriber and LEIE datasets are successfully loaded into the system.
Verify that the system correctly identifies opioid drugs from the Part Prescriber dataset.
Validate the accuracy of calculated metrics such as the number of opioids prescribed, total prescriptions, and the opioid prescription ratio.
Confirm that generated visualizations, including scatter plots and bar charts, accurately represent the data distribution

X. RESULT

The project aims to achieve the following results:
A.Insights into Opioid Prescriptions: Provide insights into the distribution and characteristics of opioid prescriptions, including the identification of top-prescribed opioids.

B.Geographical Analysis:
Analyze and visualize the distribution of opioid prescribers across different states.

C.Specialty Analysis:
Examine and visualize the distribution of opioid prescribers across medical specialties.

D.Integration Accuracy:
Ensure the seamless integration of modules and functionalities to create a comprehensive and informative system.

XI. CONCLUSION

The project successfully identified healthcare providers who prescribe opioid drugs, shedding light on the extent of opioid prescriptions within the Medicare and Medicaid systems. Through comprehensive data analysis and visualization, the project provided insights into the geographical distribution of opioid prescribers across different states. Additionally, it analyzed the distribution of opioid prescriptions among various medical specialties.

REFERENCE

[1]CMS partD Datasets https://www.cms.gov/Research-Statistics-Dataand-Systems/Statistics-Trends-and- Reports/Medicare-Provider-Charge-Data/Part-DPrescriber.html.

[2]LEIE Datasets https://oig.hhs.gov/exclusions/exclusions_list.asp

[3] Offices of Inspector General Reports https://oig.hhs.gov/reports-and publication/portfolio/index.asp

[4]FDA Datasets https://www.fda.gov/Drugs/InformationOnDrugs /ucm079750.htm#collapseOne

[5] Google Cloud Query-Part D Data https://bigquery.cloud.google.com/table/bigquery