Report On

“Medicare Fraud Detection”

By

Rohan Singh  
Yash Sarang  
Kshitij Shidore

Under the Guidance of Mrs. Bhavana Chaudhari

Department of Artificial Intelligence and Data Science Vivekanand Education Society’s Institute of Technology 2023-24

**Table Of Contents**

1. **Problem Definition and Scope of Project**
   1. Introduction
   2. Problem Definition and Scope of Project
   3. Users of the system
   4. Dataset
2. **Literature Review**
3. **Conceptual System Design**

3.1 Conceptual System Design Methodology:

* + 1. Data Gathering / Loading
    2. Data Preprocessing
    3. Feature Engineering
    4. Model Statistics
    5. Visualizations

1. **Technology Used**
2. **Implementation**
3. **Results and conclusion**

# Problem Definition and Scope of Project

## Introduction

The project addresses the critical issue of Medicare fraud detection. Medicare, being a vital healthcare program, is susceptible to fraudulent claims and activities that lead to substantial financial losses.

### Problem Definition and Scope of Project.

The project aims to develop a comprehensive fraud detection system that utilizes advanced data analytics, machine learning, and anomaly detection techniques to scrutinize extensive Medicare claims data. It will identify irregular billing patterns, duplicate claims, and other fraudulent activities. This system's primary scope includes data collection, thorough data preprocessing, feature engineering, model training, and the establishment of real-time monitoring and reporting mechanisms. It will also require integration with existing Medicare databases and claim processing systems for seamless operation.

Moreover, the project's ongoing scope extends to regular updates, model enhancements, and adapting to emerging fraud strategies. A successful project will lead to significant cost savings, the early detection and prevention of fraudulent activities, increased trust in the Medicare system, and the efficient allocation of resources to genuine medical services. This will ultimately ensure the program's sustainability and integrity while benefiting both the healthcare industry and the patients it serves.

### Users of the system.

The users of the statement are stakeholders involved in the credit card fraud detection project. This includes financial institutions, their customers, data analysts, machine learning engineers, and regulatory authorities. They rely on the statement to understand the project's purpose, goals, and implementation scope

### Dataset

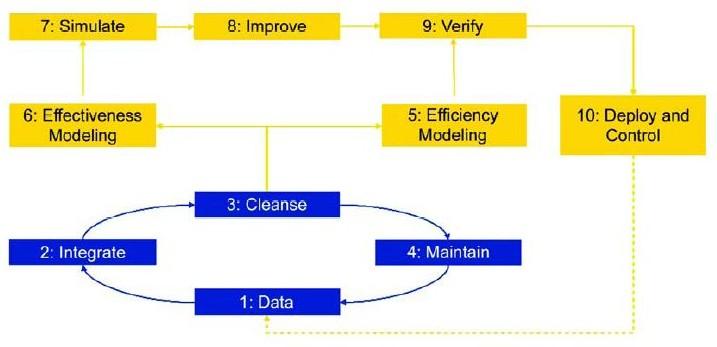
The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

# Literature Review

* Part D Prescriber Data CY 2017. (n.d.). Retrieved June 23, 2020, from https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/PartD2017:
  + This paper presents a realistic modeling approach and a novel learning strategy for credit card fraud detection. It discusses the challenges in real-world fraud detection and proposes a new framework that incorporates both machine learning and feature engineering techniques.
* LEIE Downloadable Databases: Office of Inspector General: U.S. Department of Health and Human Services. (2020, June 10). Retrieved June 23, 2020, from https://oig.hhs.gov/exclusions/exclusions\_list.asp:
  + This paper explores the use of deep learning techniques, specifically deep neural networks, for credit card fraud detection. It investigates the effectiveness of deep learning in identifying fraudulent transactions and compares it to traditional methods.
* Dataset Downloads. (n.d.). Retrieved June 23, 2020, from https://www.cms.gov/OpenPayments/Explore-the-Data/Dataset-Downloads:
  + This paper discusses the application of machine learning algorithms for credit card fraud detection. It explores various machine learning models and their performance in identifying fraudulent activities, emphasizing the importance of consumer protection.

### Conceptual System Design

* 1. **Conceptual System Design**



### Methodology:

The methodology for Medicare fraud detection encompasses several key steps. It begins with data collection from healthcare claims, which is followed by data preprocessing to clean, normalize, and structure the data for analysis. Feature engineering is then employed to extract relevant attributes and patterns from the claims data. Machine learning algorithms, such as anomaly detection and clustering, are applied to identify irregularities, unusual billing patterns, and duplicate claims. Real-time monitoring and reporting systems are established to trigger immediate alerts when potential fraud is detected.

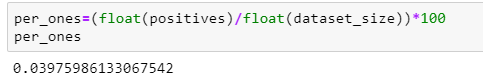
Furthermore, the methodology requires regular updates and model refinements to adapt to evolving fraud tactics. The use of domain knowledge, including known fraud indicators, enhances the accuracy of fraud detection. This holistic approach combines data analytics, machine learning, and domain expertise to create a robust fraud detection system, ultimately safeguarding healthcare funds, improving the Medicare system's integrity, and ensuring that resources are appropriately allocated to legitimate medical services.

### Data Gathering / Loading

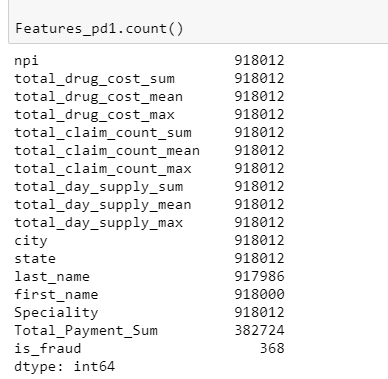
* + 25M+ rows and 21 columns
  + All information related to prescription, drugs, payments and charges by National Provider Identifier (NPI).
  + All information on the physician (NPI, Name, City, Practice, etc.)
  + 11M+ rows and 75 columns
  + Physicians in the US are required to declare all payments received from pharmaceutical companies
  + The sum of general payment
  + Name of drug associated with the payments
  + list of individuals and entities that are excluded from participating in federally funded healthcare programs (i.e. Medicare) due to previous healthcare fraud.
  + Mapped fraud labels

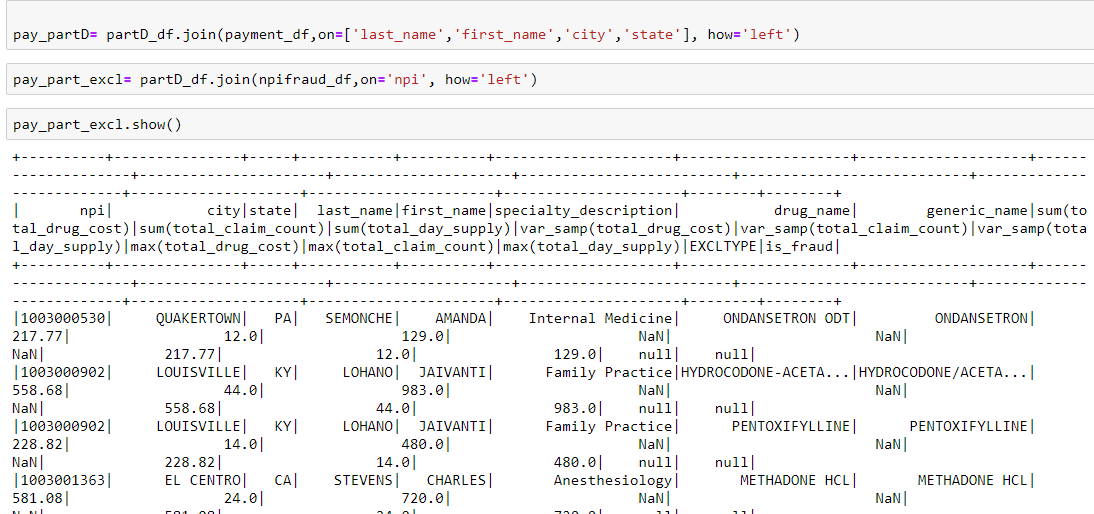
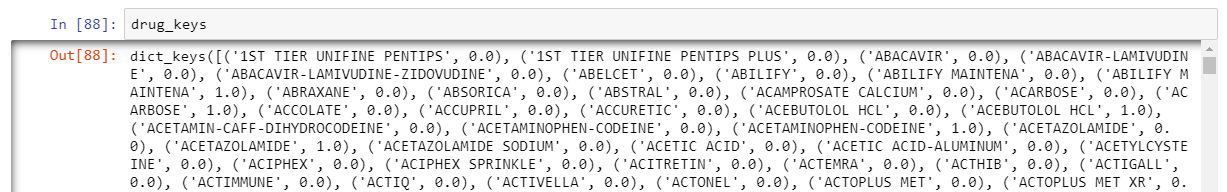
### Data Preprocessing

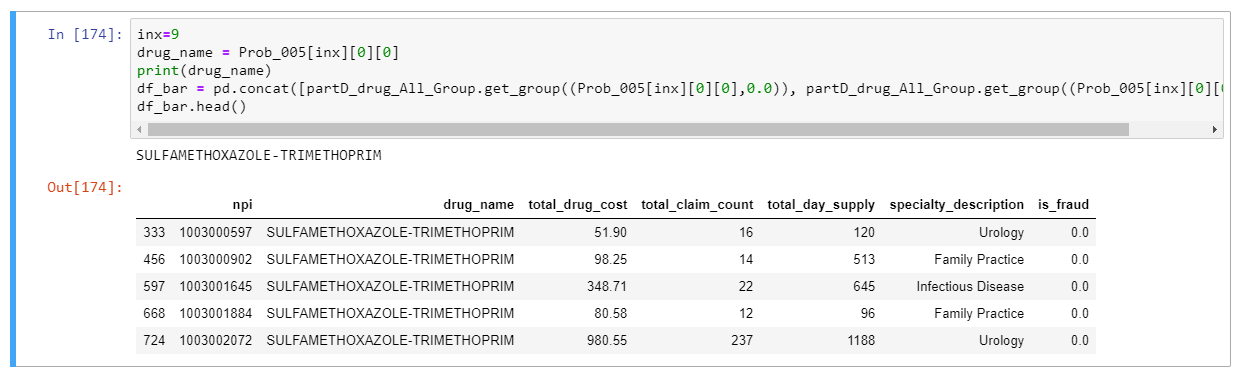
* Impute missing Data
* Removing duplicates
* Removing outliers
* Factoring the categorical data
* Removing data based on general information.
* Data Sampling: The data set is very imbalanced in terms of fraud detection context as it is very skewed (99 % no fraudulent cases and less than 1% fraudulent cases)

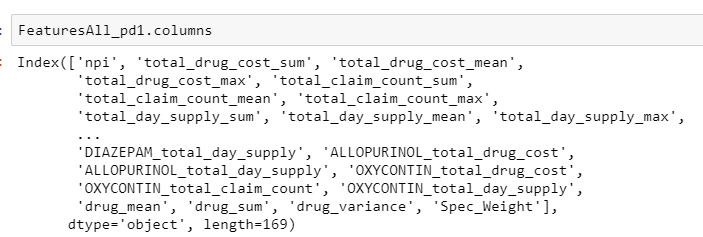


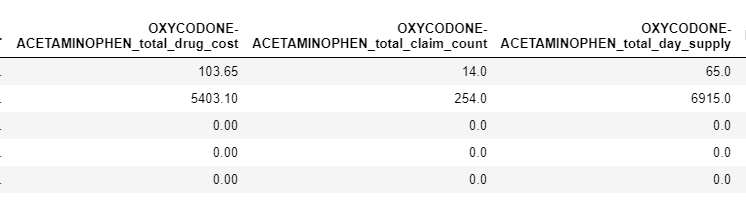
### Feature Engineering

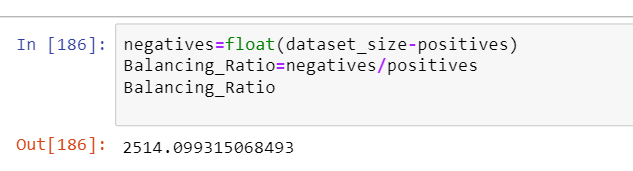
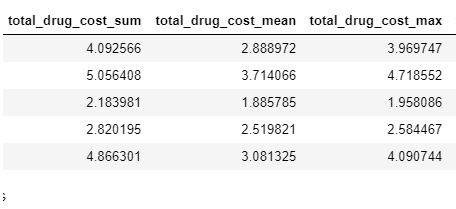






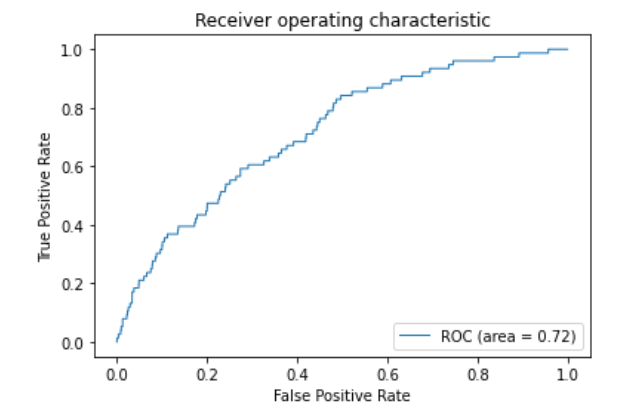
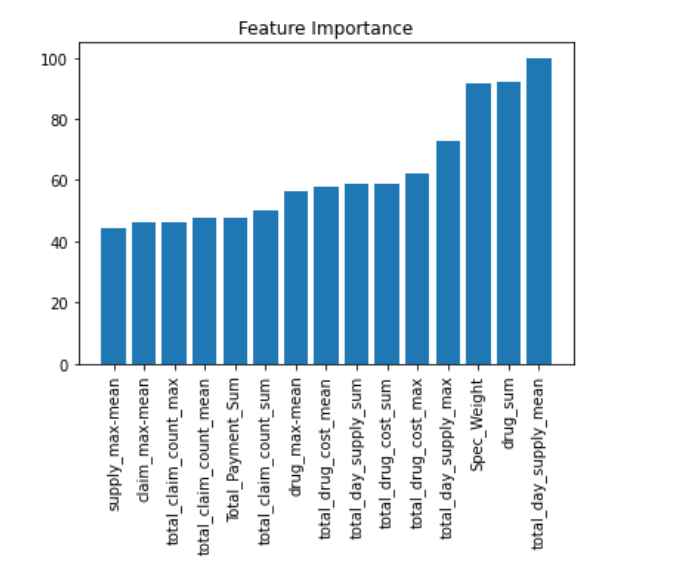




Class

### Model Statistics:

### Visualizations:



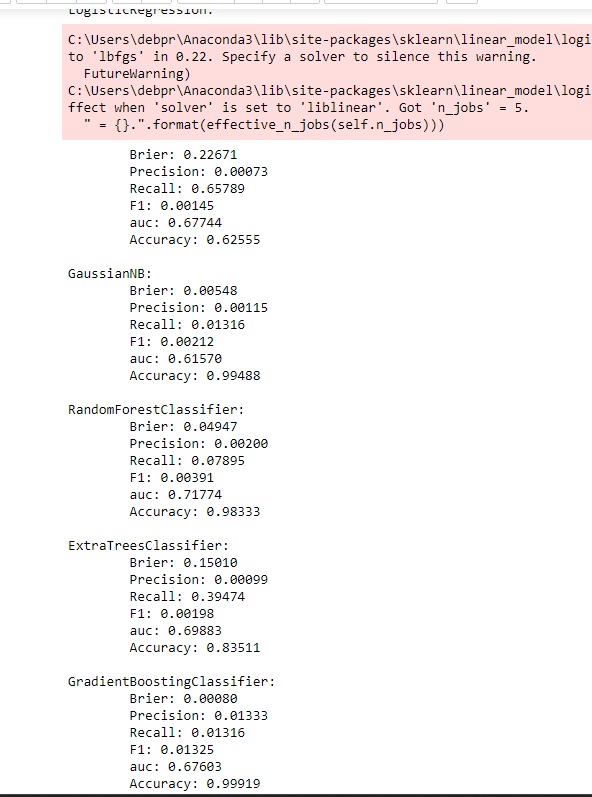
* 1. **Technology Used**

The technology stack for credit card fraud detection involves a combination of data processing tools, machine learning libraries, and potentially cloud services for scalability. Here's an overview of the technologies and their roles:

1. Python
2. Scikit-Learn
3. Pandas
4. NumPy
5. Logistic Regression
6. Decision Trees
7. Naive Bayes
8. ExtraTreesCLassifier
9. GradientBoostingClassifier
10. Matplotlib and Seaborn
11. Jupyter Notebooks

## Implementation

The implementation involves using Python with Scikit-Learn, Pandas, and NumPy for data processing. Logistic Regression, Decision Trees, ExtraTreesCLassifier, GradientBoostingClassifier Naive Bayes algorithms are employed for classification. Data is preprocessed, features engineered, and models trained and evaluated. Matplotlib and Seaborn are used for visualization. Jupyter Notebooks facilitate interactive development. Optionally, cloud services like AWS may be utilized for scalability. The system undergoes rigorous testing before deployment, with regular updates and monitoring for optimal fraud detection.



## Results and Conclusion

Thus, GradientBoostingClassifier was the most efficient at detecting fraudulent data from among all other classifiers.

In conclusion, our analysis of anonymized credit card transactions, with labels distinguishing between fraudulent and genuine cases, underscores the significance of data-driven approaches in addressing financial fraud. Leveraging Python libraries, we have harnessed the power of machine learning to navigate this complex landscape.

Our findings have demonstrated the efficacy of GradientBoostingClassifier, emerging as the most efficient algorithm in this endeavor. These classifiers have exhibited a high degree of accuracy in distinguishing between fraudulent and genuine transactions, providing a robust foundation for real-time fraud detection.

**\*\*\*\*\***