Healthcare Provider Fraud Detection Analysis & Prediction

Keerthana Mallavarapu

Submitted to: Dr. Faisal Quader

Date: 6th Dec, 2023

Table of Contents

1. Abstract
2. Introduction
3. Literature survey
4. Methodology
5. Experimental results
    a. Decision tree
    b. Random forest
    c. LightGBM
    d. XGBoost
6. References

1. Abstract

This project aims to enhance integrity within healthcare systems by pioneering a predictive analytical approach to detect and prevent fraud. The healthcare industry, surrounded by an array of fraud types including billing and insurance malpractices, faces significant financial and reputational harm, which ultimately detracts from patient care. Through the development and application of predictive models, this project aims to ingrain out existing fraud and flag potential future deceptions. The implementation of this data-driven approach promises not only to enhance the quality of healthcare and preserve patient trust but also to mitigate financial and reputational damages while optimizing resource allocation.

2. Introduction

The primary objective of this project is to develop and implement data-driven techniques and predictive models to identify and prevent fraudulent activities within the healthcare provider systems. There have been many fraudulent activities happening within the healthcare industry such as billing fraud, insurance fraud, unnecessary medical procedures, and many more. These fraudulent activities divert resources and attention away from providing quality healthcare to patients. Efficient resource allocation is important in healthcare to help improve care quality and allow for medical research. Identifying these frauds and predicting future frauds would help the healthcare industry not only provide quality healthcare to patients but also reduce financial losses, and reputation loss, improve compliance and resource allocation as well and gain patient trust.

Detecting and addressing fraud within healthcare provider organizations using data analytics is of utmost importance for several reasons. Healthcare institutions are responsible for safeguarding sensitive patient information, encompassing medical records, financial billing data, and personal particulars. Any compromise or breach of this confidential data can result in severe consequences, including identity theft, medical identity theft, and emotional distress for patients. Effectively tackling fraud is critical to preserving patient privacy and maintaining trust in healthcare institutions.

Furthermore, healthcare fraud can put a significant financial strain on both patients and providers. Deceptive tactics, such as fraudulent billing for services not rendered or claim inflation, can raise healthcare expenses. Fraudulent activities may also add inaccuracies and discrepancies to healthcare data, affecting the smooth delivery of patient care, resulting in inaccurate billing, and hindering clinical judgment. Leveraging data analytics and cybersecurity measures is essential for achieving these critical objectives.

3. Literature survey

Johnson, J.M., Khoshgoftaar, T.M. Medicare fraud detection using neural networks. *J Big Data* 6, 63 (2019).

The literature survey on "Medicare fraud detection using neural networks" by Justin M. Johnson and Taghi M. Khoshgoftaar explores the application of neural network models for the detection and prevention of fraud in the Medicare system. The study contributes to the growing body of research addressing fraud in healthcare through advanced machine-learning techniques. The literature reveals an increasing interest in leveraging neural networks, a subset of deep learning, for their ability to automatically learn intricate patterns and representations from complex healthcare data. Studies in this domain highlight the advantages of neural networks in capturing non-linear relationships and discerning subtle anomalies that may signify fraudulent activities within Medicare claims. The research likely delves into the design and architecture of neural networks tailored for Medicare fraud detection, potentially exploring recurrent neural networks (RNNs) or convolutional neural networks (CNNs) for sequential and image-based data, respectively. Additionally, the survey may discuss the challenges associated with implementing neural network models in healthcare fraud detection, such as interpretability and the need for labeled training data. Overall, the literature by Johnson and Khoshgoftaar contributes valuable insights into the application of neural networks to enhance the accuracy and efficiency of Medicare fraud detection systems.

Bauder, R., Khoshgoftaar, T.M. & Seliya, N. A survey on the state of healthcare upcoding fraud analysis and detection. Health Serv Outcomes Res Method 17, 31–55 (2017).

This paper explores the state of healthcare upcoding fraud analysis and detection, providing a comprehensive overview of existing research and methodologies. The study acknowledges the increasing prevalence and significance of upcoding fraud within the healthcare sector, emphasizing the need for advanced analytical tools to combat this form of fraudulent activity. The survey encompasses a broad spectrum of literature, incorporating various statistical and machine-learning approaches employed in the analysis and detection of upcoding fraud. Descriptive statistics are employed to understand the patterns and characteristics associated with upcoding, while inferential statistics contribute to identifying statistically significant anomalies. Machine learning models, such as neural networks and ensemble methods, are discussed for their potential to enhance the accuracy of upcoding detection. The survey underscores the importance of data quality, feature engineering, and model interpretability in achieving effective fraud detection outcomes. Real-world case studies and empirical findings presented in the literature surveyed offer valuable insights into the challenges and successes of upcoding fraud analysis and detection. By synthesizing the current state of knowledge, the survey provides a foundation for future research and the continued development of sophisticated strategies to combat healthcare upcoding fraud.

Nabrawi, E., & Alanazi, A. (2023). Fraud detection in healthcare insurance claims using machine learning. *Risks*.

The literature by Eman Nabrawi and Abdullah Alanazi on fraud detection in healthcare insurance claims using machine learning emphasizes the significance of employing advanced techniques to combat fraudulent activities in the healthcare insurance sector. The study likely explores a range of machine learning models, including supervised and unsupervised approaches, to analyze claims data. Key considerations include feature engineering, model interpretability, and the integration of multiple methods to enhance accuracy. Case studies or empirical evidence may illustrate practical applications in real-world scenarios, contributing valuable insights to the ongoing development of effective fraud detection strategies in healthcare insurance.

Hancock, J.T., Bauder, R.A., Wang, H. *et al.* Explainable machine learning models for Medicare fraud detection. *J Big Data* 10, 154 (2023).

Recent research, such as "Explainable machine learning models for Medicare fraud detection," has placed a heavy emphasis on the analysis-driven deployment of advanced machine learning techniques, indicating a paradigm shift in the literature on Medicare fraud detection. Specifically, ensemble techniques like Decision Tree, XGBoost, Random Forest, CatBoost, and LightGBM have become popular for efficiently categorizing imbalanced healthcare datasets. The paper stands out for its innovative ensemble feature selection method, which cleverly leverages the built-in feature importance features of various machine learning methods. This method finds feature subsets that are crucial for accurate fraud detection, which not only lowers the complexity of datasets but also significantly improves interpretability. The research goes deeper into statistical studies to examine the complex effects of feature subsets and classifiers on classification performance. ANOVA and HSD tests are utilized in this process. These results reveal the value of reduced feature sets, challenging traditional evaluation metrics like AUC in favor of more informative measures like AUPRC. This study contributes to the literature by offering insightful information on the complex subject of healthcare fraud detection and guiding it in the direction of more analytically driven approaches by closely analyzing the interactions among classifiers, feature subsets, and classification efficacy.

Li, J., Huang, K.-Y., Jin, J., & Shi, J. (2007). A survey on statistical methods for health care fraud detection. *Health Care Manage Sci.*

The survey of statistical methods for detecting healthcare fraud published provides an insightful exploration into evaluating the effectiveness of binary classifiers. Their distinction between error-based and cost-based evaluation methodologies constitutes a noteworthy contribution to their work. This paper offers a basic framework for classifier evaluation by thoroughly explaining how to create a confusion matrix with test datasets. The extensive evaluation of error-based approaches such as ROC curves and AUC is particularly remarkable, providing a nuanced perspective of classifier efficacy. The use of cost-based techniques, which involve a thorough model taking the effort and case costs into account, provides depth by examining costs related to different classification outcomes. The study's identification of obstacles, such as the issue of erroneous labels in training data and the necessity for evolving approaches, highlights the practical challenges in detecting healthcare fraud. Overall, this paper effectively analyzes the intricacies of classifier assessment techniques, which makes it an invaluable tool for scholars and industry professionals.

Waghade, S. S., & Karandikar, A. M. (2018). A Comprehensive Study of Healthcare Fraud Detection based on Machine Learning. *International Journal of Applied Engineering Research*, *13(6)*, 4175-4178. Research India Publications.

The literature review on healthcare fraud detection emphasizes in-depth analysis as a critical component in identifying and tackling fraudulent activity in the healthcare industry. The survey methodically examines the difficulties presented by complex fraud patterns and the shortcomings of conventional manual detection techniques, especially when dealing with large datasets. It highlights how sophisticated machine learning and data mining techniques may change the game by automating and improving fraud detection through in-depth analysis. The significance of specialized analytical techniques is highlighted by the division of techniques into supervised, unsupervised, and semi-supervised learning categories. The study explores detailed analyses carried out in different research projects, from Rule-based Data Mining for anomaly identification to supervised methods and graph analytics for risk assessment. The commitment to comprehensive investigation in identifying potential fraud indications is demonstrated by the in-depth review of physician practices, educational backgrounds, and specialty-specific analyses. To maintain the integrity, quality, and cost of healthcare systems, the survey's conclusion emphasizes the necessity of continuing study, innovation, and investigation of cutting-edge methods to boost the efficacy of healthcare fraud detection.

4. Methodology

The methodology employed is illustrated in the flowchart below. The process begins with defining the problem of healthcare provider fraud. Data is then collected and subjected to a detailed pre-processing to ensure quality and relevance. The clean data is explored (EDA) to uncover initial insights and is subsequently divided into training and test datasets.

Machine learning models, including Decision Trees, Random Forest, LightGBM, and XGBoost, are trained and evaluated. Decision Trees offer interpretable decision logic and precise variable identification. Random Forest, an ensemble of decision trees, provides a robust evaluation of feature importance. LightGBM uses a leaf-wise growth strategy for enhanced efficiency and reduced overfitting risks. XGBoost, another ensemble method, leverages regularization and effective handling of imbalanced data for detecting subtle fraud indicators [4]. Each model's performance is quantified using accuracy and precision metrics, with the top-performing model being identified based on these measures. The results lead to the identification of patient and provider patterns that are indicative of potential fraud. This comprehensive methodology enables an understanding of fraudulent behavior, informing the development of predictive models that could significantly impact healthcare fraud detection and prevention.

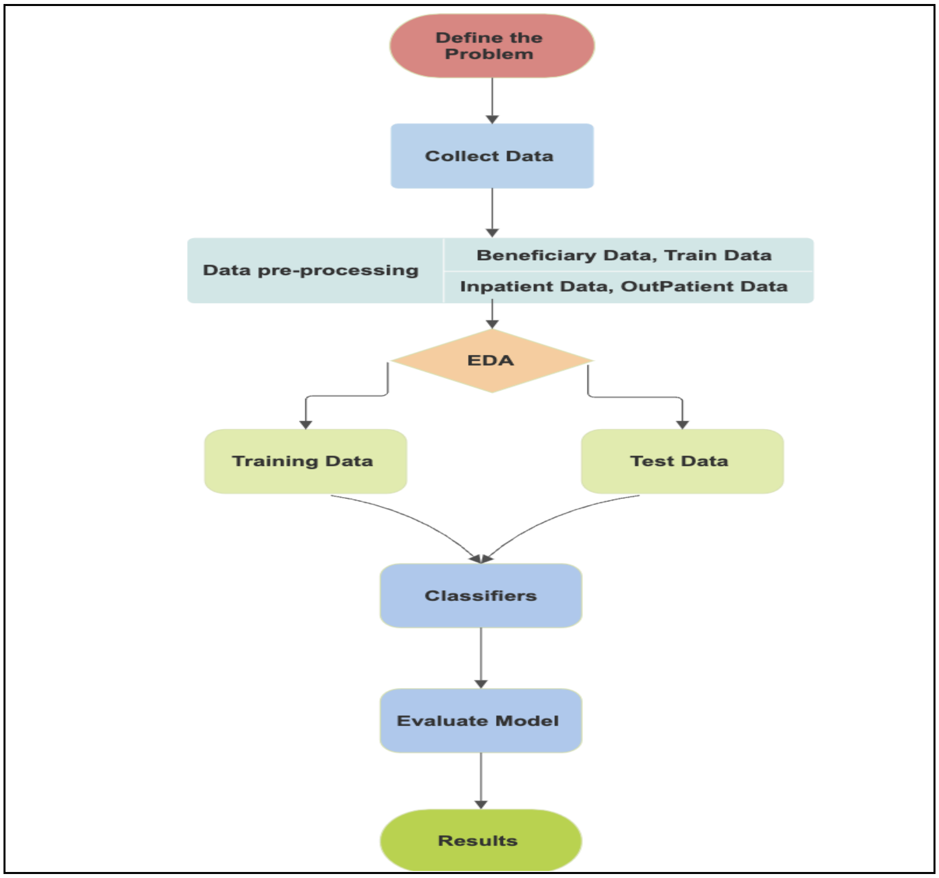


Fig 1 - Methodology

Data preprocessing:

The data we collected had four CSV files which contained the data of Benefeciary, Inpatient, and Output and the result of insurance fraud. Each file had several missing values in columns such as Date of Death (DOD, because the majority of the patients are still alive), ProcedureCode columns (because the majority of patients did not undergo any surgery), OtherPhysician, DiagnosisCode, ChronicDiseases were removed.

The data was also highly biased towards the non-fraud cases, hence, the SMOTE technique was used to sample the data and make it unbiased.

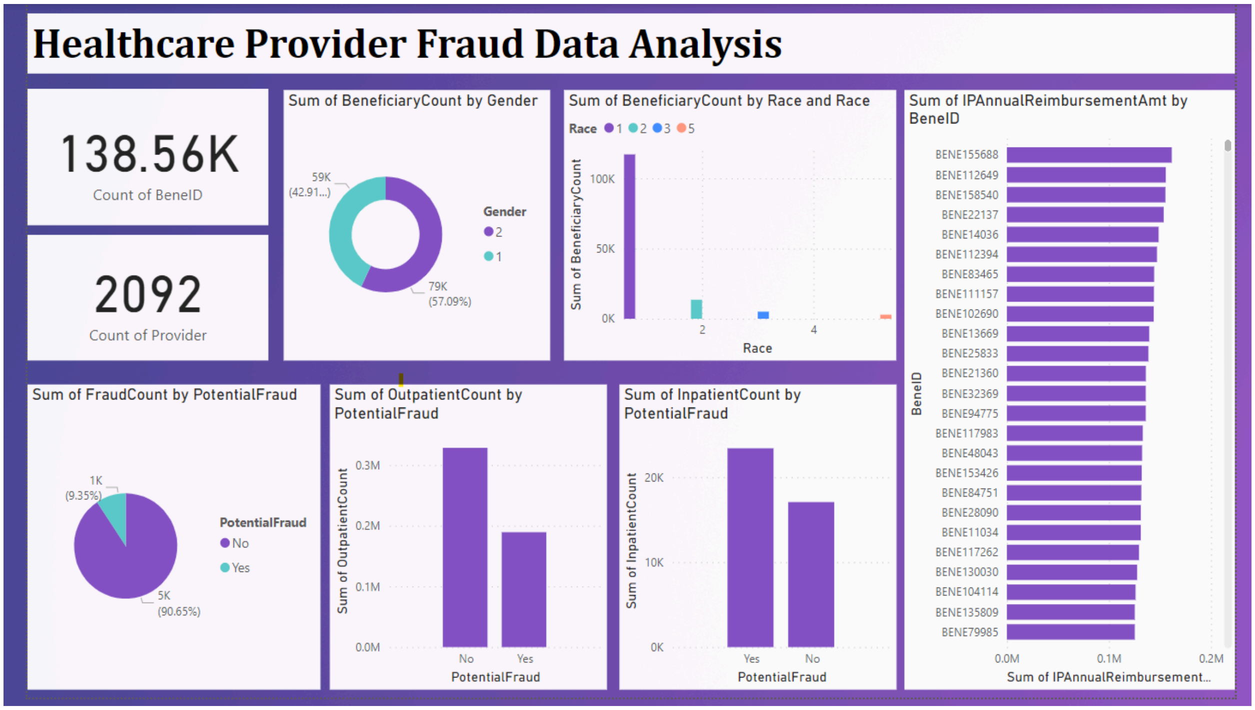


Fig 2 - EDA

5. Experimental results

For effective classification tasks, it is essential to have a balanced distribution of instances for each class, ensuring an equal representation of both fraudulent and non-fraudulent labels. Imbalanced classes can lead to biased models that favor the majority class. Upon careful analysis, it was observed that the pre-processed data exhibits a significant imbalance. This imbalance is visually depicted in the graph below, emphasizing the need for techniques to address and rectify the skewed class distribution.

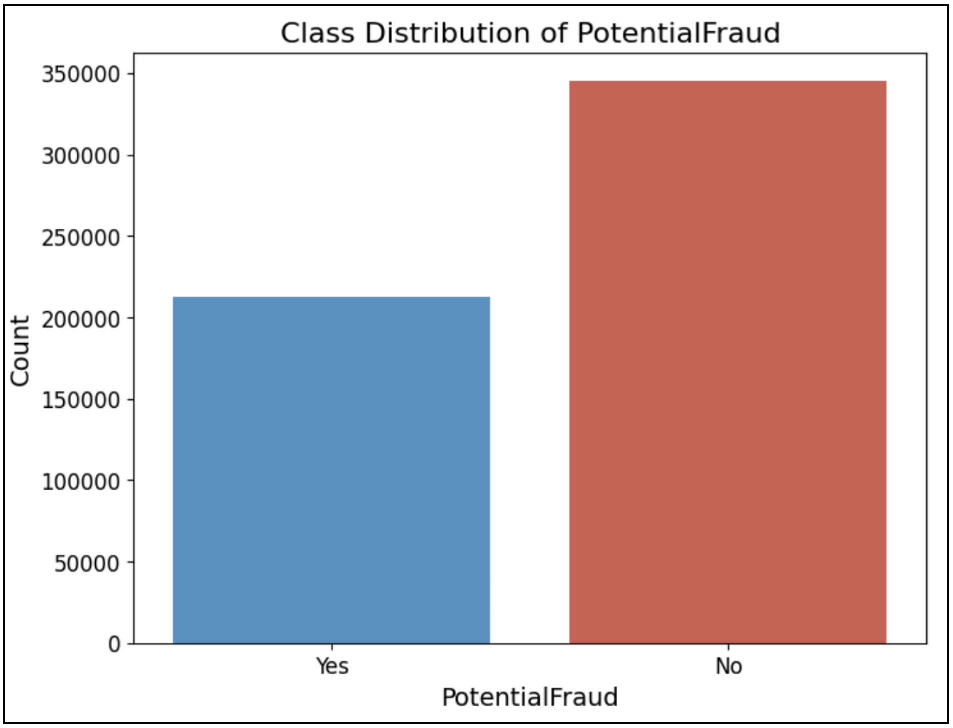


Fig 3 - Class Imbalance

To remove this imbalance SMOTE (Synthetic Minority Over-sampling Technique) technique is used to randomly upsample the data. SMOTE generates synthetic samples for the minority class, thereby balancing the class distribution. The technique involves creating synthetic instances along the line segments that connect existing minority class instances. By introducing diversity into the training data through these synthetic examples, SMOTE helps mitigate the risk of overfitting on the majority class and enhances the model's ability to generalize well to unseen data.

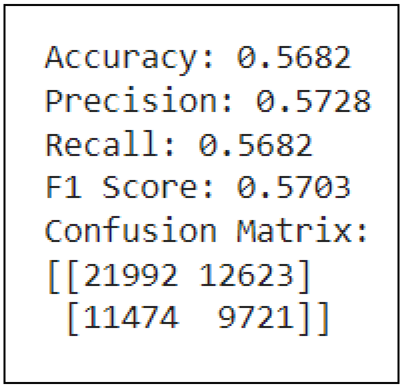
Since predicting fraud is a classification task, data is trained using 4 models. They are Decision Tree, Random Forest, LightGBM, and XGBoost, and evaluated the models' results. Below are the results.

a. ***Decision tree:***

Decision trees provide a clear and interpretable representation of decision-making processes. Each node in the tree represents a decision based on a specific feature.

Decision trees can quantify the importance of each feature in the prediction process. Features that appear higher in the tree and are used for splitting nodes are considered more important in making predictions, providing insights into which variables contribute most to identifying potential fraud.

Results: (Accuracy = 56%)



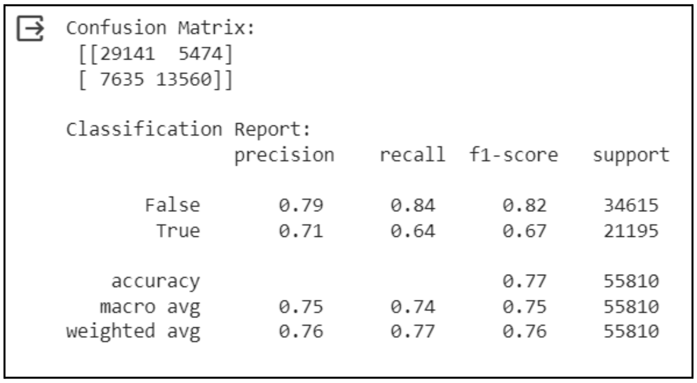
```
Accuracy: 0.5682
Precision: 0.5728
Recall: 0.5682
F1 Score: 0.5703
Confusion Matrix:
[[21992 12623]
 [11474  9721]]
```

Fig 4 - Classification Report for Decision Tree

b. Random forest

Random Forest combines multiple decision trees to form an ensemble. Each tree is trained on a random subset of the data with replacement (bootstrapped samples). It introduces additional randomness by considering only a subset of features at each split in a tree. Each tree in the forest independently makes a prediction, and the final prediction is determined by a majority vote.

Results: (Accuracy = 77%)



```
Confusion Matrix:
 [[29141  5474]
 [ 7635 13560]]

Classification Report:
               precision    recall  f1-score   support

       False       0.79      0.84      0.82     34615
        True       0.71      0.64      0.67     21195

    accuracy                           0.77     55810
   macro avg       0.75      0.74      0.75     55810
weighted avg       0.76      0.77      0.76     55810
```

Fig 5 - Classification Report for Random Forest

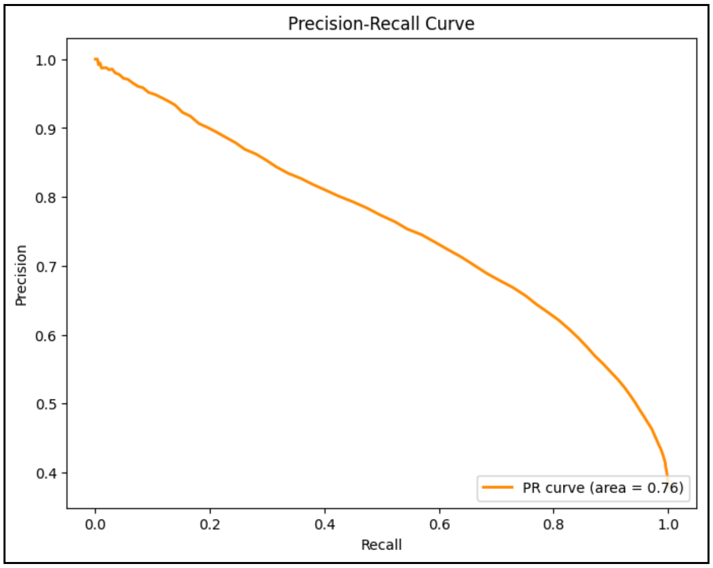


Fig 6 - Precision - Recall Curve for Random Forest

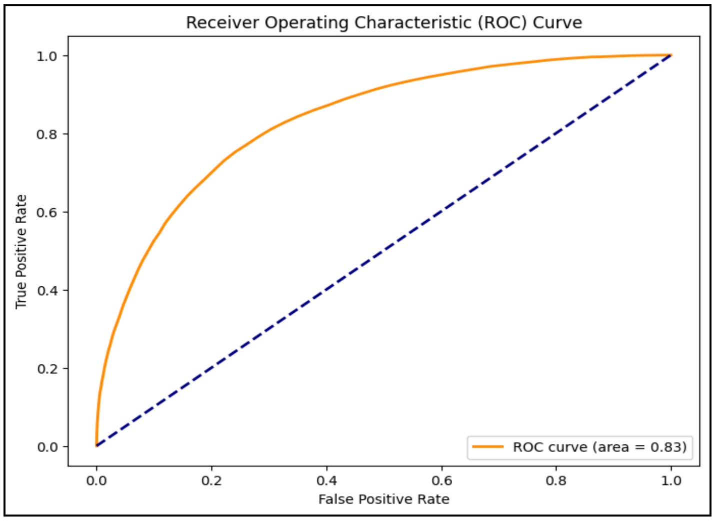
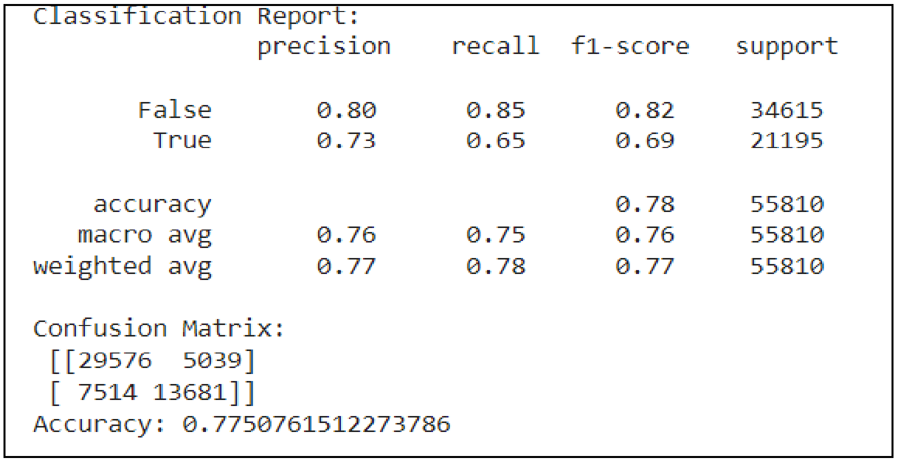


Fig 7 - ROC Curve for Random Forest

c. ***LightGBM***

LightGBM is a gradient-boosting framework. It uses a leaf-wise tree growth strategy. During training, it chooses the leaf with the greatest delta loss, resulting in a more complex tree structure. It can handle categorical features directly without the requirement for one-hot encoding, making it more memory and compute-efficient. However, the data was numerical.

Results: (Accuracy = 77%)



```
Classification Report:
              precision    recall  f1-score   support

       False       0.80      0.85      0.82     34615
        True       0.73      0.65      0.69     21195

    accuracy                           0.78     55810
   macro avg       0.76      0.75      0.76     55810
weighted avg       0.77      0.78      0.77     55810

Confusion Matrix:
 [[29576  5039]
 [ 7514 13681]]
Accuracy: 0.7750761512273786
```

Fig 8 - Classification Report for LightGBM

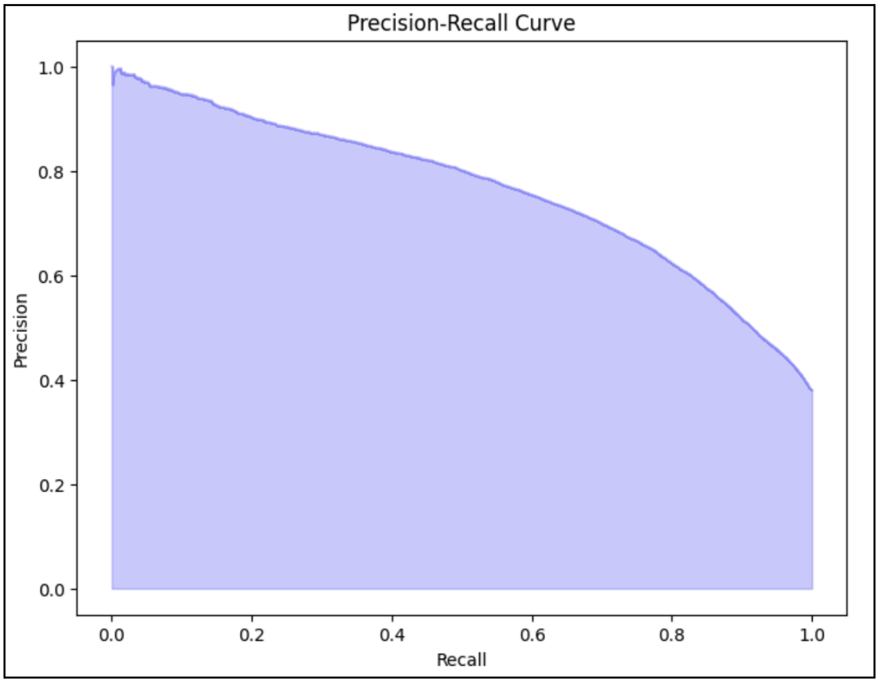


Fig 9 - Precision-Recall Curve for LightGBM

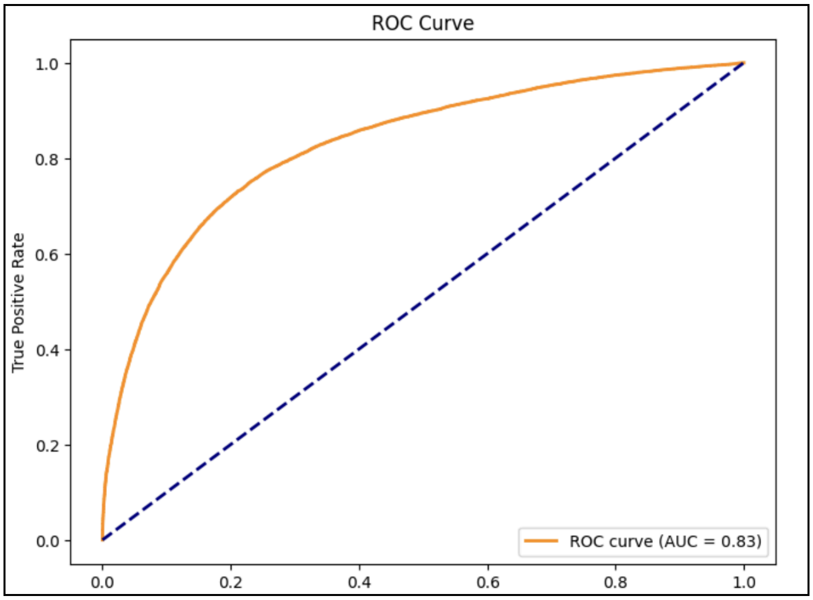
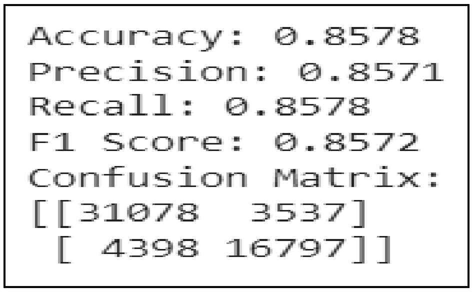


Fig 10 - ROC Curve for LightGBM

d. XGBoost

XGBoost builds trees sequentially, with each tree fixing the errors of the preceding ones. During training, it minimizes a specific loss function. To reduce the loss function, it employs gradient descent optimization. In each iteration, it fits a new tree to the residual errors of the combined predictions of the existing trees. Regularization terms are included in XGBoost's objective function to prevent overfitting. It also employs a shrinkage parameter to regulate each tree's contribution. XGBoost has a mechanism for determining feature relevance based on how frequently features are utilized in trees and how much they contribute to lowering the loss function.

Results: (Accuracy = 85%)



```
Accuracy: 0.8578
Precision: 0.8571
Recall: 0.8578
F1 Score: 0.8572
Confusion Matrix:
[[31078  3537]
 [ 4398 16797]]
```

Fig 11 - Classification Report for XGBoost

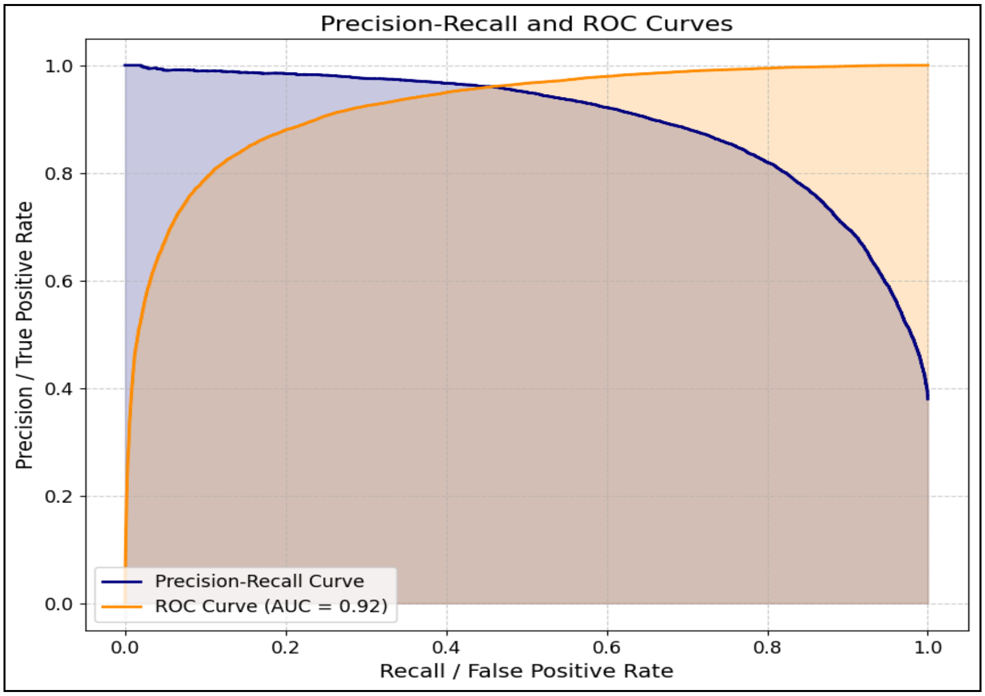


Fig 12 - Precision-Recall and ROC Curve for XGBoost

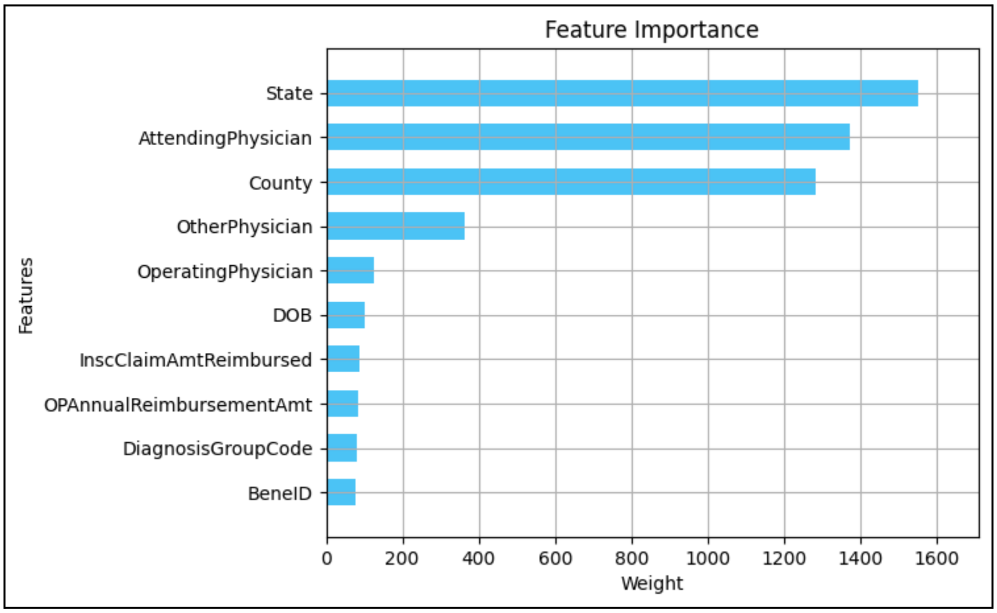


Fig 13 - Feature Importance for XGBoost

The XGBoost algorithm demonstrated the highest level of accuracy in our analysis. Additionally, a feature importance graph was generated to identify the key contributors to label prediction. Notably, attributes such as State, AttendingPhysician, County, OperatingPhysician, and OperatingPhysician played a significant role in predicting whether a provider is fraudulent or non-fraudulent.

To conclude, a thorough examination of healthcare data has yielded valuable insights into possible fraudulent activities among healthcare providers. Noteworthy patient-level patterns include elevated claim amounts and patients with significant chronic conditions being potential targets for fraud. Likewise, fraudulent providers exhibit patterns such as an excessive number of claims, engagement in high-profit procedures, and unconventional billing practices. Consequently, by incorporating relevant features such as geographical information, patient details, and claim amounts, along with employing a suitable machine learning algorithm, it is possible to establish a robust foundation for an efficient fraud detection system.

6. References

1. Johnson, J., & Khoshgoftaar, T. M. (2019b). Medicare fraud detection using neural networks. *Journal of Big Data*, *6*(1).
2. Bauder, R. A., Khoshgoftaar, T. M., & Seliya, N. (2016b). A survey on the state of healthcare upcoding fraud analysis and detection. *Health Services and Outcomes Research Methodology*, *17*(1), 31–55.
3. Nabrawi, E., & Alanazi, A. (2023). Fraud detection in healthcare insurance claims using machine learning. *Risks*, *11*(9), 160.
4. Hancock, J.T., Bauder, R.A., Wang, H. *et al.* Explainable machine learning models for Medicare fraud detection. *J Big Data* 10, 154 (2023).
5. Li, J., Huang, K.-Y., Jin, J., & Shi, J. (2007). A survey on statistical methods for health care fraud detection. *Health Care Manage Sci.*
6. Waghade, S. S., & Karandikar, A. M. (2018). A Comprehensive Study of Healthcare Fraud Detection based on Machine Learning. *International Journal of Applied Engineering Research*, *13(6)*, 4175-4178. Research India Publications.