Detecting Fraudulent Healthcare Providers Using Machine Learning

DATA 606 Capstone in Data Science

Professor: Zeynep Kacar

By

Amrita Chandrasekar

NH53017

12/11/2024

Table of Contents

EXECUTIVE SUMMARY

This project aimed to develop a machine learning-based solution for detecting fraudulent healthcare providers using claims data from widely available data source Kaggle. The project used advanced data analysis with an additional focus on addressing class imbalance using SMOTE and machine learning techniques to identify patterns indicative of fraudulent activities. By leveraging multiple machine learning models like Logistic Regression, Random Forest, XGBoost, Gradient Boosting, and LightGBM(gradient boosting model) the project compared models' performance and utilized the best-suited model for the deployment. An interactive Streamlit dashboard was built to visualize predictions and insights, offering healthcare administrators a tool for fraud detection and key decision-making.

Key Findings

- The LightGBM model outperforms with an overall accuracy of 98%, achieving high precision and recall for both classes. Its balanced F1-score of 0.98 for both classes indicates strong predictive capability and minimal bias towards either class.
- The most important features for fraud detection included ProviderFraudRate and ClaimsPerProvider calculated during feature engineering.
- The application of SMOTE significantly improved model performance by addressing the class imbalance in the dataset.
- The Streamlit dashboard provided an intuitive interface for users to interact with the model, explore fraud patterns, and make real-time predictions.

Significance of the Project

This project is highly significant as it tackles the pressing issue of healthcare provider fraud, which results in substantial financial losses and compromises patient care quality. Applying advanced machine learning algorithms enables the early detection of fraudulent activities, ensuring more efficient and accurate identification of suspicious behaviors. The use of Streamlit to create an interactive and user-friendly dashboard enhances accessibility and effective decision-making. Overall, this project contributes to strengthening the integrity of healthcare systems and promotes data-driven solutions for fraud prevention, leading to more sustainable and trustworthy healthcare practices.

Summary of Methodologies and Conclusions

The project involved extensive **data preprocessing**, including merging multiple datasets, handling missing values, and addressing class imbalance using **SMOTE** to enhance model performance. After thorough **Exploratory Data Analysis (EDA)**, key features such as **ProviderFraudRate** and **ClaimsPerProvider**, which were crucial for detecting fraud were identified. **LightGBM** was selected for its high accuracy and efficiency in dealing with imbalanced data, achieving **98% accuracy** in fraud prediction.

The project culminated in the development of a Streamlit dashboard for fraud prediction, enabling interactive exploration of results. In conclusion, the project successfully integrated machine learning and data visualization to create a powerful fraud detection tool for the healthcare industry.

Introduction

This project focuses on leveraging machine learning to detect fraudulent healthcare providers using claims data from widely available data source Kaggle. Healthcare fraud is a significant issue, leading to financial losses, increased premiums, and compromised patient care. Detecting fraudulent activities has traditionally been challenging due to the complexity of claims data and the high volume of submissions. By applying machine learning techniques, this project aims to automate and improve fraud detection processes, providing an efficient tool for identifying potentially fraudulent providers. The project also includes building an interactive Streamlit dashboard, enabling predictions and data visualizations to enhance decision-making.

Problem Statement

In the U.S. healthcare system, fraud costs taxpayers billions of dollars every year. A real-world example is the case of a healthcare provider who submits fraudulent Medicare claims for services that were never provided, such as billing for unnecessary medical tests or overcharging for treatments. In 2019, it was reported that healthcare fraud led to losses of over $60 billion in the U.S. alone. This fraudulent behavior not only strains the financial stability of healthcare systems but also puts patient safety at risk, as resources are diverted away from legitimate care.

Traditional fraud detection methods, which often rely on manual audits and rule-based checks, struggle to keep up with the sheer volume and complexity of claims data, making it difficult to identify subtle patterns of fraud. This project aims to harness machine learning to automate the process, providing a more efficient, scalable, and accurate solution for detecting fraudulent healthcare providers, reducing financial losses, and ensuring the quality and integrity of care provided to patients.

Research Questions

- How effectively can machine learning algorithms identify potentially fraudulent Medicare providers based on claims data?
- Which features in the dataset are most indicative of fraudulent behavior?
- How can we address the class imbalance in the dataset, where fraudulent claims are far fewer than legitimate ones?

Background and Context

Healthcare fraud, involving fraudulent billing and exaggerated claims, significantly impacts both the financial stability of healthcare systems and the quality of patient care. Traditional methods of fraud detection struggle to manage the vast volume of claims and complex fraud patterns. With machine learning, particularly supervised learning algorithms, there is a great potential to analyze historical claims data and identify subtle, non-obvious patterns indicative of fraud. This project aims to apply machine learning models to predict fraudulent providers, using the claims dataset that includes inpatient, outpatient, beneficiary information, and Potential fraud data.

Significance of the Research Question

- Financial Impact
- Patient Safety at Risk
- System Integrity
- Resource Allocation
- Ineffective Traditional Methods

Why Is this question important?

- Public Trust
- Enlightens the need for Initiative-taking Fraud Prevention

What are the key concepts or theories that relate to your project?

- Data Imbalance Handling
- Feature Engineering
- Supervised Classification

Objectives

- Detect fraudulent healthcare providers using machine learning.
- Build an interactive **Streamlit dashboard** for real-time predictions and data visualization.

Data Overview:

Data Collection:

- Dataset is taken from https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis/data .
- It is derived from Medicare claims data, sourced from the Centers for Medicare & Medicaid Services (CMS).

Data Description

The dataset consists of four main components.

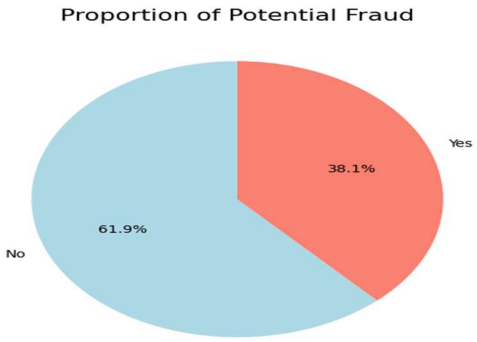
- **Inpatient Data**: Claims filed for patients admitted to hospitals. Includes admission and discharge dates, and diagnosis codes.
- **Outpatient Data**: Claims filed for patients who visit hospitals without admission.
- **Beneficiary Data**: Contains beneficiary KYC details. Includes health conditions and regional information of patients.
- **Train Data(Fraud Labels)**: It is labeled as fraud/non-fraud concerning healthcare providers.It is the target column of project

Data Preparation/Data Cleaning:

- Dropped columns that have more than 50% null values for inpatient, outpatient, and beneficiary datasets individually.
- Merged “inpatient” and “outpatient” as “merged_data” with the help of the outer join function. Then merged “merged_data” with the beneficiary dataset on BeneID column with the outer join function as “final merged” dataset.
- Shape of Final Merged Dataset: (558211, 46).
- Then, combined the “final merged” dataset with the “train dataset” on the provider column with the help of inner join as “df_final.”
- Shape of df_final dataset: (558211, 47).
- Removed columns from df_final that have more than 50%null values as they are not going to be helpful for the analysis.

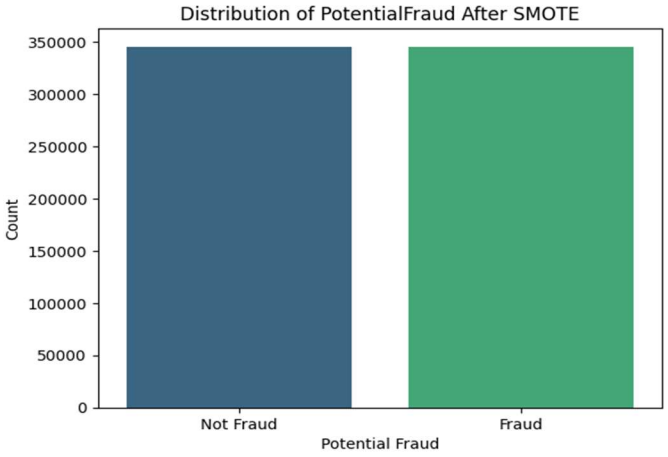
- Shape of df_final dataset:(558211, 34)
- Checked for duplicate values ensuring there were no duplicates in the dataset.
- Checked for null values in all columns. Out of 34 columns in the dataset, only the following 4 columns had null values.
    - AttendingPhysician 1508
    - DeductibleAmtPaid 899
    - ClmDiagnosisCode_1 10453
    - ClmDiagnosisCode_2 195606
- To manage null values, “unknown” is used in place of null values for the column “AttendingPhysician,” while “median” and “mode” is used for columns “DeductibleAmtPaid”,” ClmDiagnosisCode_1” and “ClmDiagnosisCode_2” respectively.
- Converted date to standard datetime format.
- Performed label encoding to convert categorical features to numerical features(For example target column (PotentialFraud) was labeled as Fraud, Non-Fraud. With the help of encoding, it got converted into 0 and 1 indicating not fraud and fraud, respectively. for making data ready for visualizations.

Challenges with Data



This graph shows the class imbalance of fraud and not fraud labels.

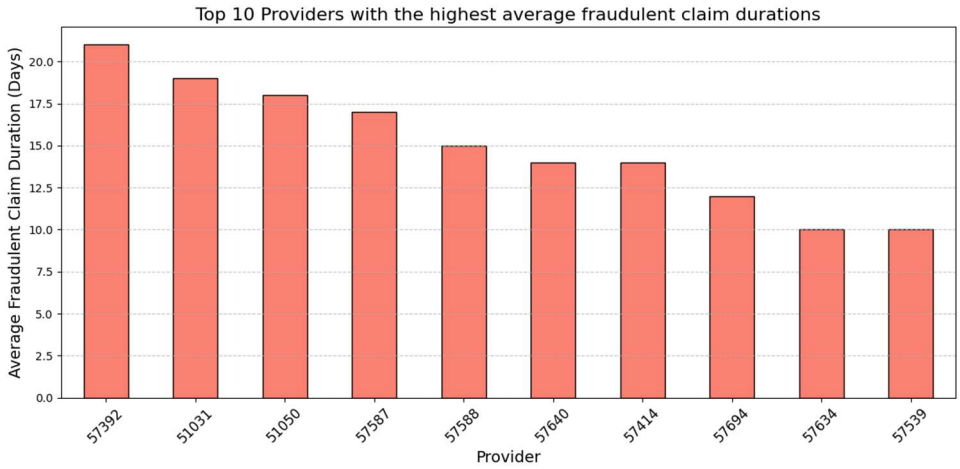
SMOTE was used to manage class imbalance. SMOTE (Synthetic Minority Over-sampling Technique) is a technique used to address class imbalance by generating synthetic samples for the minority class. It works by selecting a sample from the minority class, finding its nearest neighbors, and then creating new samples by interpolating between the original sample and its neighbors. This helps to balance the dataset, preventing the model from being biased toward the majority class.

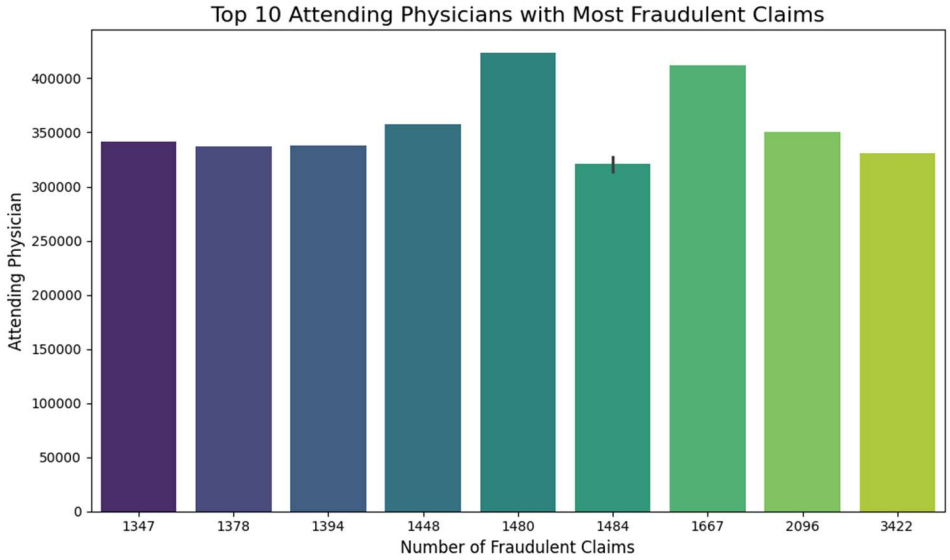


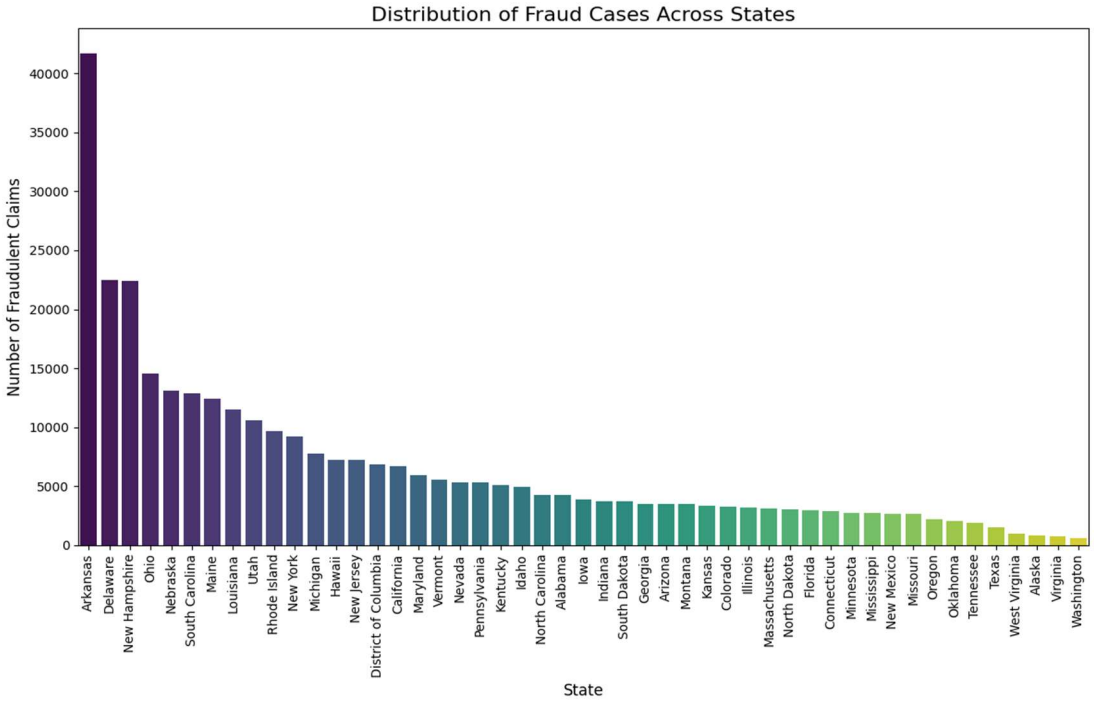
After SMOTE, combined features columns(X_resampled) and target variable column(y_resampled) into a single data frame as “balanced_df” having shape (690830, 32).

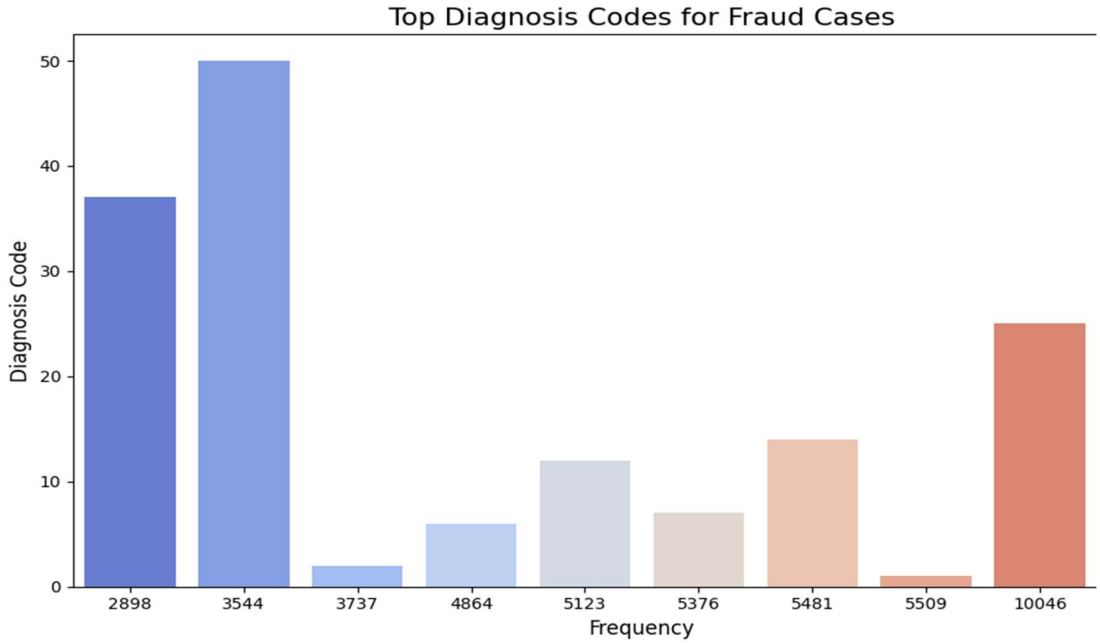
Methodology

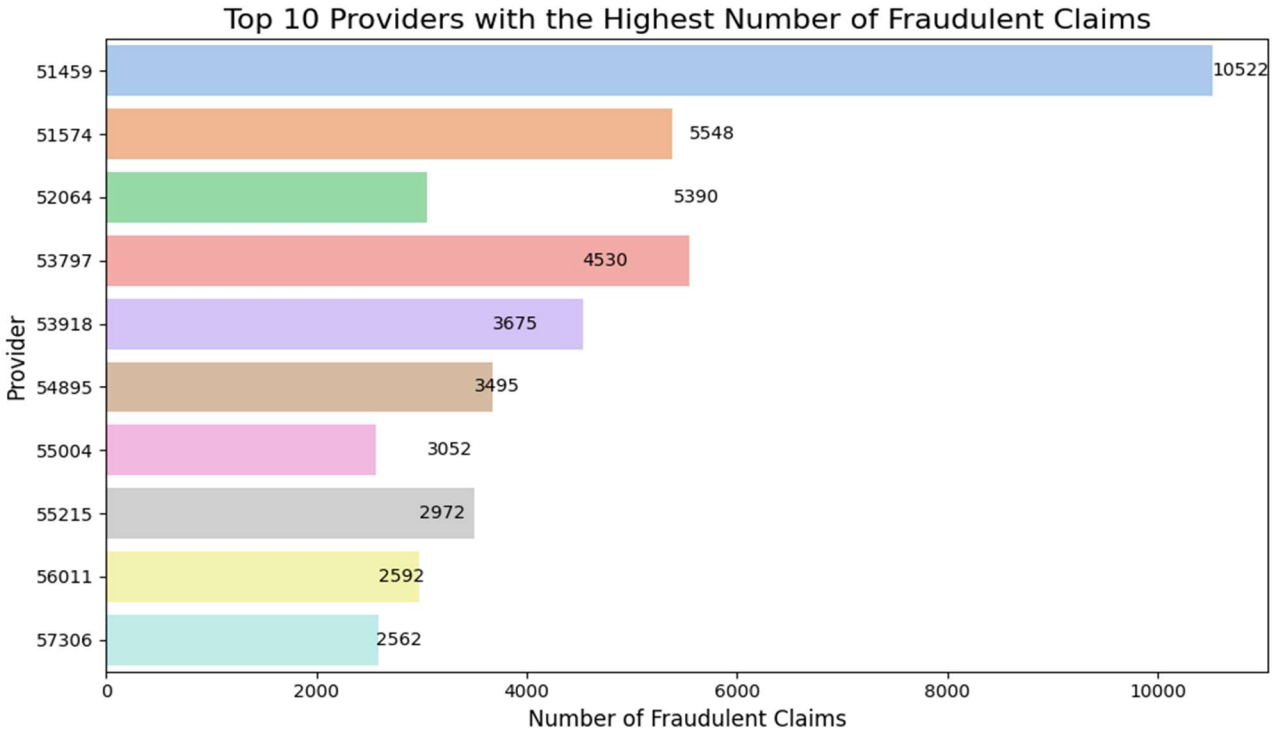
Exploratory Data Analysis

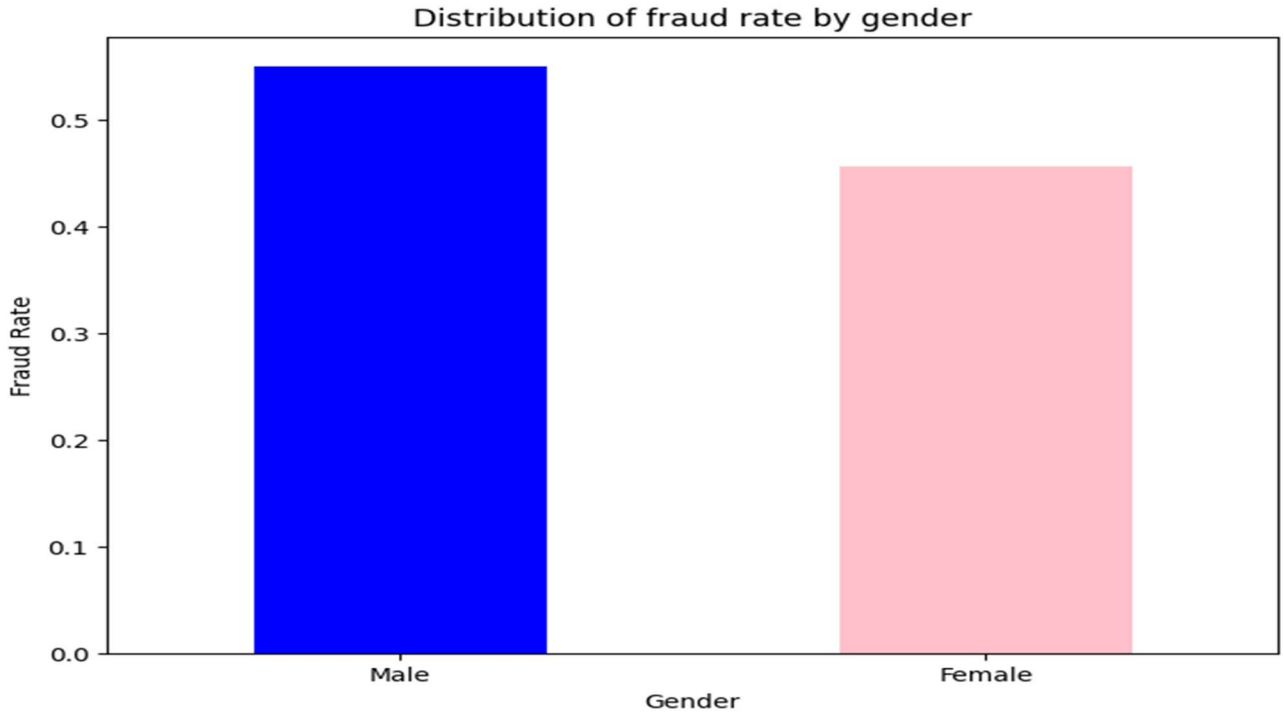




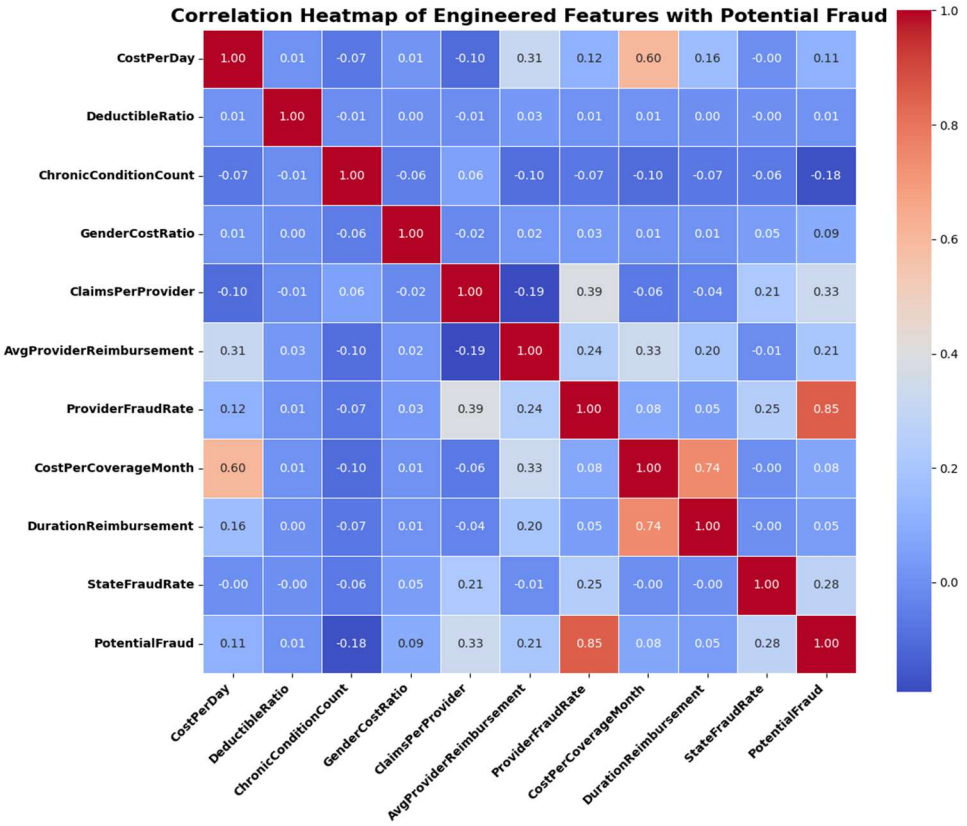

Distribution of Fraud Cases Across States
Number of Fraudulent Claims
0
5000
10000
15000
20000
25000
30000
35000
40000
Arkansas
Delaware
New Hampshire
Ohio
Nebraska
South Carolina
Maine
Louisiana
Utah
Rhode Island
New York
Michigan
Hawaii
New Jersey
District of Columbia
California
Maryland
Vermont
Nevada
Pennsylvania
Kentucky
Idaho
North Carolina
Alabama
Iowa
Indiana
South Dakota
Georgia
Arizona
Montana
Kansas
Colorado
Illinois
Massachusetts
North Dakota
Florida
Connecticut
Minnesota
Mississippi
New Mexico
Missouri
Oregon
Oklahoma
Tennessee
Texas
West Virginia
Alaska
Virginia
Washington
State


Top Diagnosis Codes for Fraud Cases
Diagnosis Code
0
10
20
30
40
50
2898
3544
3737
4864
5123
5376
5481
5509
10046
Frequency


Top 10 Providers with the Highest Number of Fraudulent Claims
Provider
51459
51574
52064
53797
53918
54895
55004
55215
56011
57306
10522
5548
5390
4530
3675
3495
3052
2972
2592
2562
0
2000
4000
6000
8000
10000
Number of Fraudulent Claims


Distribution of fraud rate by gender
Fraud Rate
0.0
0.1
0.2
0.3
0.4
0.5
Male
Female
Gender

Feature Engineering



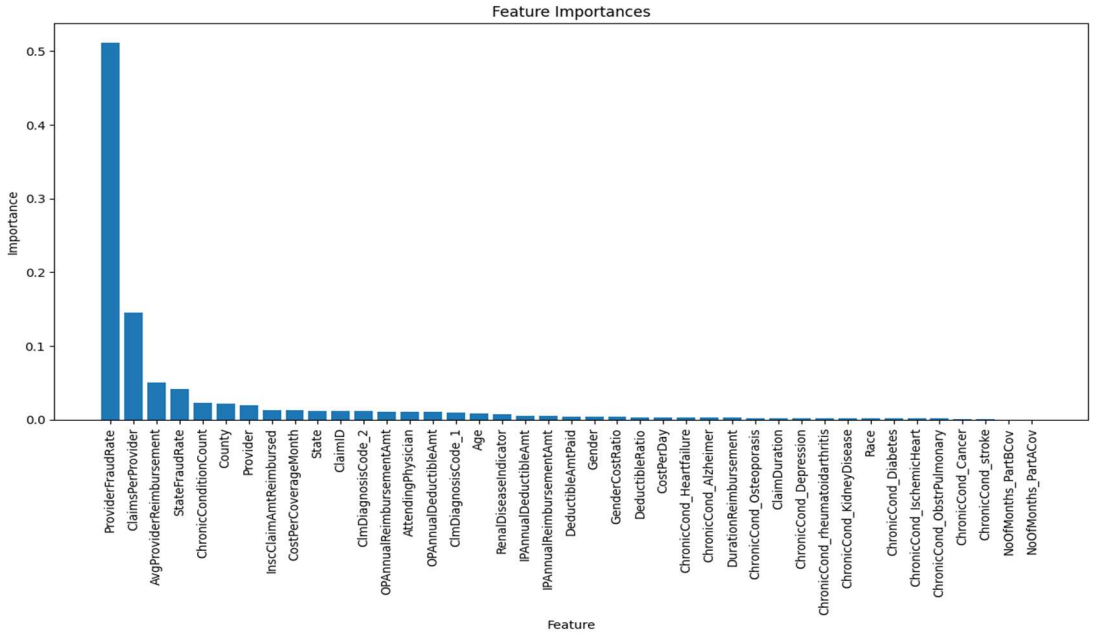
- **CostPerDay:** This feature is moderately correlated with potential fraud (0.11) and can help identify unusually excessive costs per day for providers, which could indicate fraud.
- **DeductibleRatio:** With an exceptionally low correlation to potential fraud, this feature may be less impactful in identifying fraud but could provide additional context on healthcare cost structures.
- **ChronicConditionCount:** This feature has a weak negative correlation with potential fraud, suggesting it might not strongly indicate fraudulent behavior but could be useful in understanding patient risk.
- **GenderCostRatio:** Shows a weak positive correlation with potential fraud (0.09), which might be useful for detecting anomalies in cost allocation across genders.

- **ClaimsPerProvider:** This feature strongly correlates with potential fraud (0.33), as a higher number of claims per provider could signal fraudulent activity.
- **AvgProviderReimbursement:** Correlated at 0.21 with potential fraud, this feature is relevant for detecting discrepancies between expected and actual reimbursements, indicating potential fraud.
- **ProviderFraudRate:** This column is highly correlated with potential fraud (0.85), making it one of the most critical features for detecting fraudulent behavior in healthcare providers.
- **CostPerCoverageMonth:** Strong correlation with potential fraud (0.74), suggesting that unusually excessive costs per coverage month could be a strong indicator of fraudulent activity.
- **DurationReimbursement:** With a weaker correlation to potential fraud (0.05), this feature might not be highly indicative of fraud but could still be valuable in a broader fraud detection model.
- **StateFraudRate:** While not highly correlated with potential fraud, this feature (0.28) provides geographic context, potentially useful for understanding state-level fraud trends.
- **PotentialFraud:** This is the target variable in the analysis, with a correlation of 1, making it central to the project for identifying fraudulent healthcare providers.

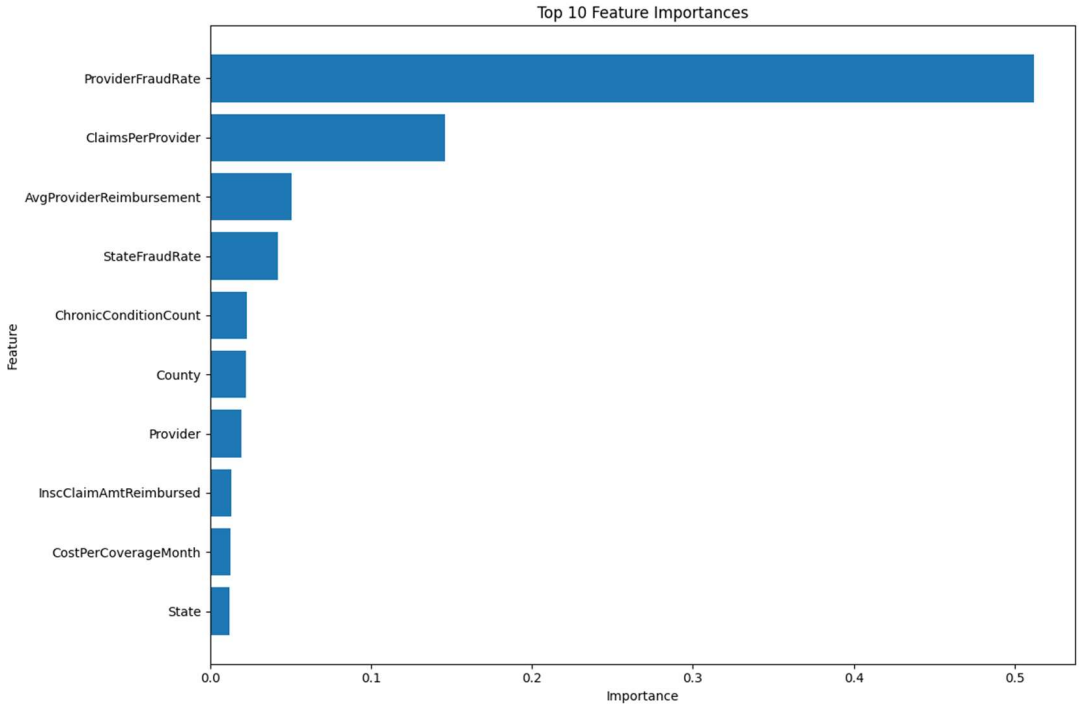
Feature Importance analysis

The process begins by preparing the data, where categorical variables are one-hot encoded, and the dataset is split into training and testing sets. The presence of infinite and NaN values is checked and managed by replacing infinite values with the column maximum. A Random Forest model is then trained on the data, and feature importance is computed to identify the top features most strongly correlated with the target variable "**PotentialFraud**."

The top features are selected for model implementation. This analysis is essential for further model development and integration into a Streamlit dashboard for interactive fraud detection.



The top 10 features used for the model's implementation are:



- Random Forest is implemented again on these top 10 features to prepare the data for the model implementation.
- The dataset is divided into 4 parts namely before model implementation.
- X_train_top.csv

  This file contains the training dataset with only the top features selected based on feature importance. It is used to train machine learning models to detect fraudulent claims.
- X_test_top.csv

  This file contains the testing dataset with only the top features. It is used to evaluate the trained model's performance by evaluating its ability to predict on unseen data.
- y_train.csv

  This file contains the target variable (PotentialFraud) corresponding to the training dataset. It is used as the true labels during the model training process.
- y_test.csv

  This file contains the target variable (PotentialFraud) corresponding to the testing dataset. It is used as the true labels to evaluate the accuracy, precision, recall, and F1-score of the predictions made by the trained model.

Modeling/Analysis Techniques

The following models were implemented on the dataset:

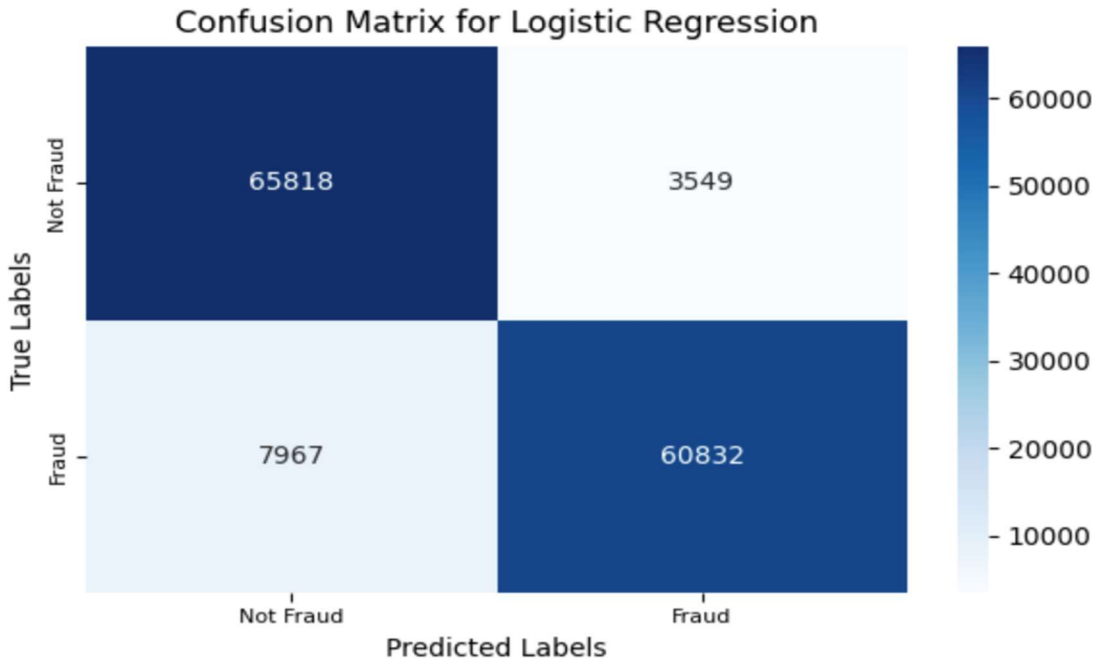
Logistic Regression

```
Model Name: Logistic Regression
Train Accuracy: 0.9175
Test Accuracy: 0.9167
Train Recall: 0.8857
Test Recall: 0.8842
Train F1: 0.9149
Test F1: 0.9135
Train AUC: 0.9717
Test AUC: 0.9708

Classification Report (Test):
              precision    recall  f1-score   support

           0       0.89      0.95      0.92     69367
           1       0.94      0.88      0.91     68799

    accuracy                           0.92    138166
   macro avg       0.92      0.92      0.92    138166
weighted avg       0.92      0.92      0.92    138166
```



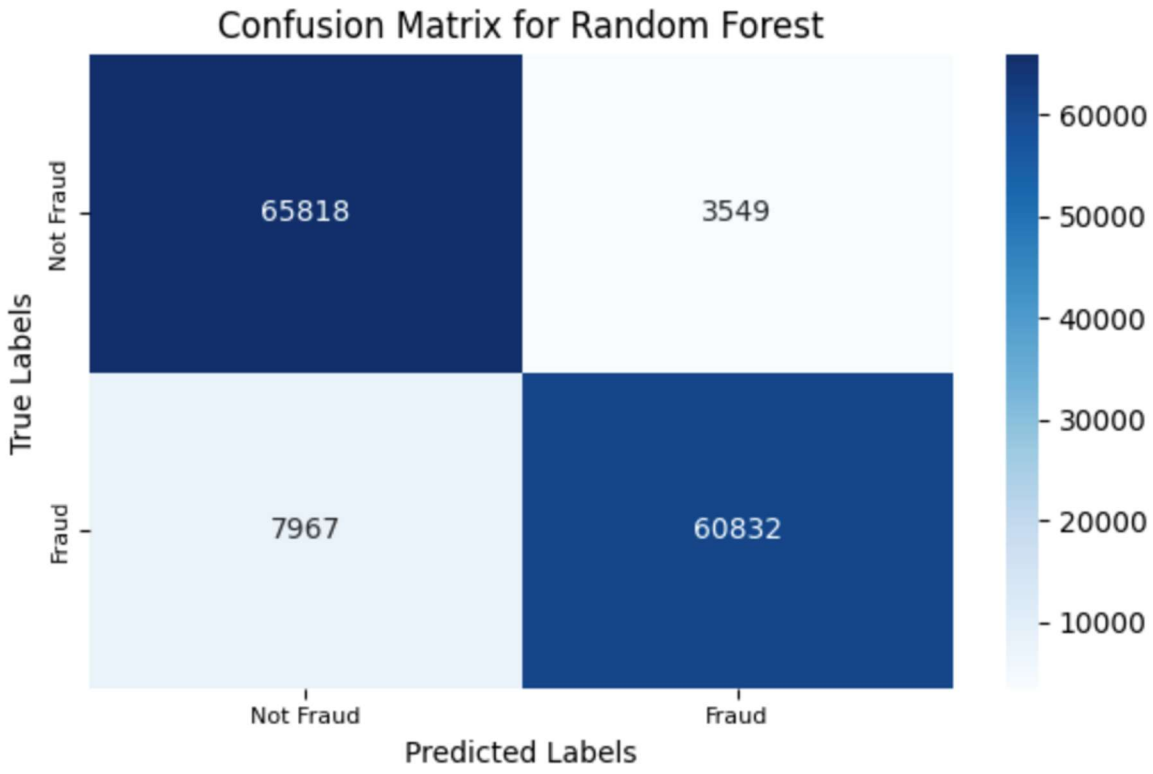
Random Forest

```
Model Name: Random Forest
Train Accuracy: 1.0000
Test Accuracy: 0.9654
Train Recall: 1.0000
Test Recall: 0.9456
Train F1: 1.0000
Test F1: 0.9646
Train AUC: 1.0000
Test AUC: 0.9960

Classification Report (Test):
              precision    recall  f1-score   support

           0       0.95      0.99      0.97     69367
           1       0.98      0.95      0.96     68799

    accuracy                           0.97    138166
   macro avg       0.97      0.97      0.97    138166
weighted avg       0.97      0.97      0.97    138166
```



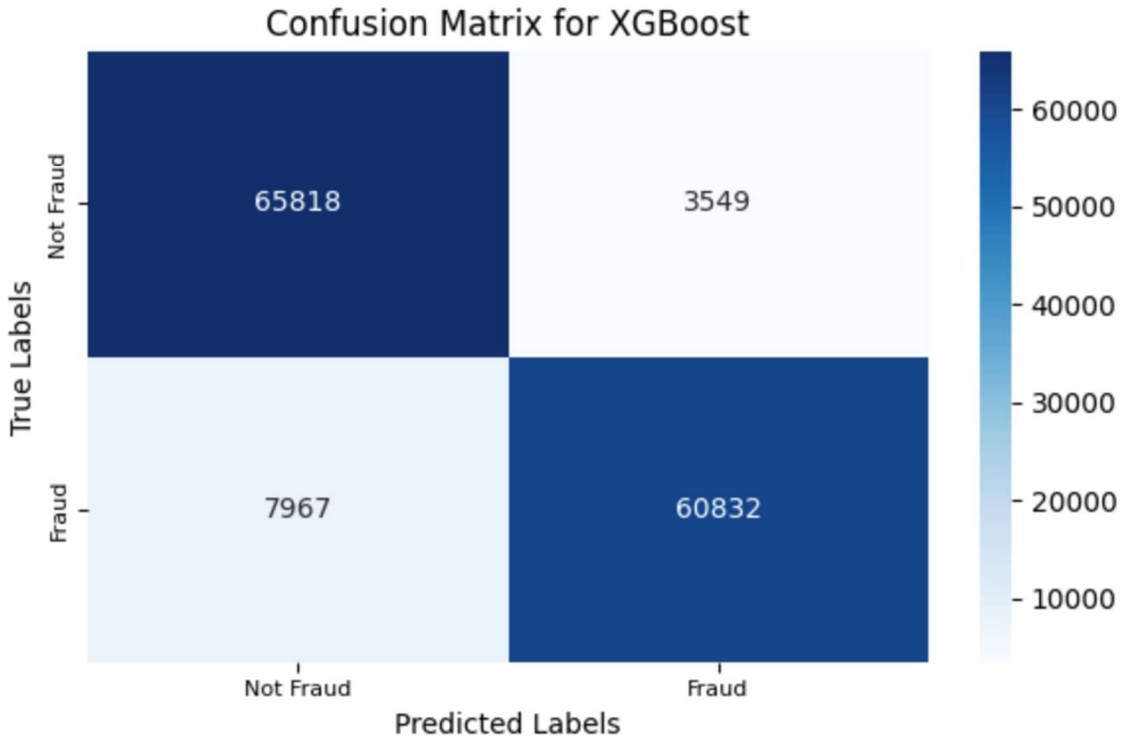
XGBoost

```
Model Name: XGBoost
Train Accuracy: 0.9833
Test Accuracy: 0.9812
Train Recall: 0.9729
Test Recall: 0.9699
Train F1: 0.9832
Test F1: 0.9809
Train AUC: 0.9978
Test AUC: 0.9973

Classification Report (Test):
              precision    recall  f1-score   support

           0       0.97      0.99      0.98     69367
           1       0.99      0.97      0.98     68799

    accuracy                           0.98    138166
   macro avg       0.98      0.98      0.98    138166
weighted avg       0.98      0.98      0.98    138166
```



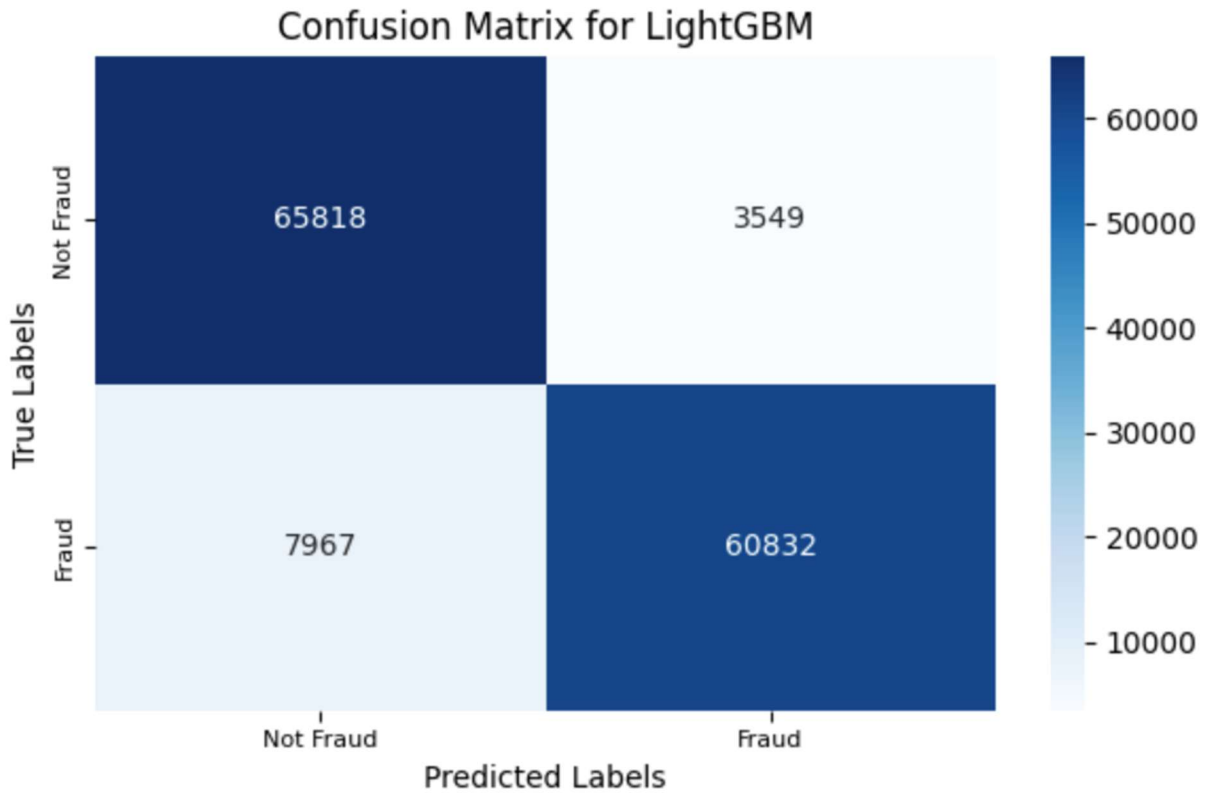
LightGBM

```
Model Name: LightGBM
Train Accuracy: 0.9844
Test Accuracy: 0.9835
Train Recall: 0.9743
Test Recall: 0.9726
Train F1: 0.9842
Test F1: 0.9832
Train AUC: 0.9983
Test AUC: 0.9981

Classification Report (Test):
              precision    recall  f1-score   support

           0       0.97      0.99      0.98     69367
           1       0.99      0.97      0.98     68799

    accuracy                           0.98    138166
   macro avg       0.98      0.98      0.98    138166
weighted avg       0.98      0.98      0.98    138166
```



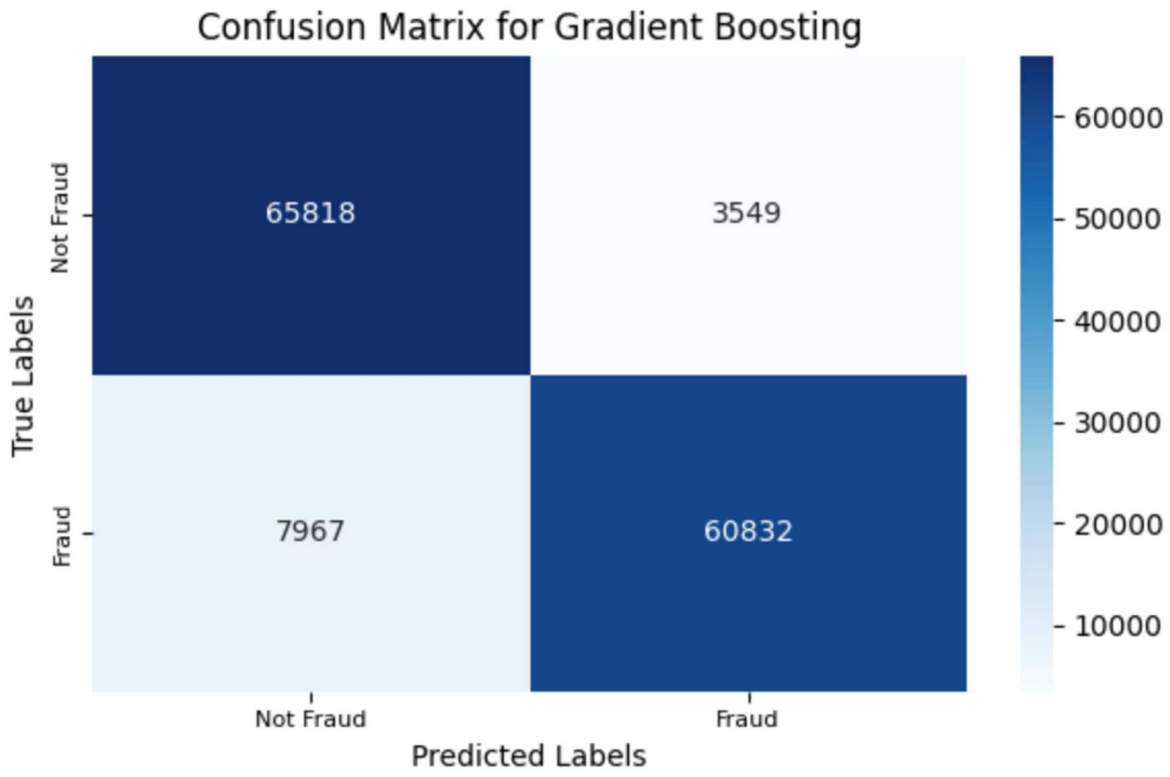
Gradient Boosting

```
Model Name: Gradient Boosting
Train Accuracy: 0.9509
Test Accuracy: 0.9502
Train Recall: 0.9277
Test Recall: 0.9261
Train F1: 0.9498
Test F1: 0.9488
Train AUC: 0.9900
Test AUC: 0.9896

Classification Report (Test):
              precision    recall  f1-score   support

           0       0.93      0.97      0.95     69367
           1       0.97      0.93      0.95     68799

    accuracy                           0.95    138166
   macro avg       0.95      0.95      0.95    138166
weighted avg       0.95      0.95      0.95    138166
```



Models Comparison

Model Name	Train Recall	Test Recall	Train F1	Test F1	Train AUC	Test AUC
Logistic Regression	0.885737	0.884199	0.914863	0.913531	0.971703	0.970783
Random Forest	0.999986	0.945552	0.999993	0.96457	1	0.996013
XGBoost	0.972937	0.969869	0.983181	0.980911	0.997846	0.997306
Light GBM	0.974318	0.972616	0.984211	0.98322	0.998321	0.998073
Gradient Boosting	0.927669	0.926104	0.949797	0.948803	0.990036	0.98964

The LightGBM model outperforms other models due to the following:

High Recall (Test Recall = 0.9726)

Recall is crucial for fraud detection because it ensures most fraudulent cases are correctly identified. LightGBM has the highest test recall among all models, making it dependable for detecting fraud.

High F1-Score (Test F1 = 0.9832):

F1-score balances precision and recall. LightGBM has the highest F1-score, indicating a good balance between avoiding false positives and catching fraud cases.

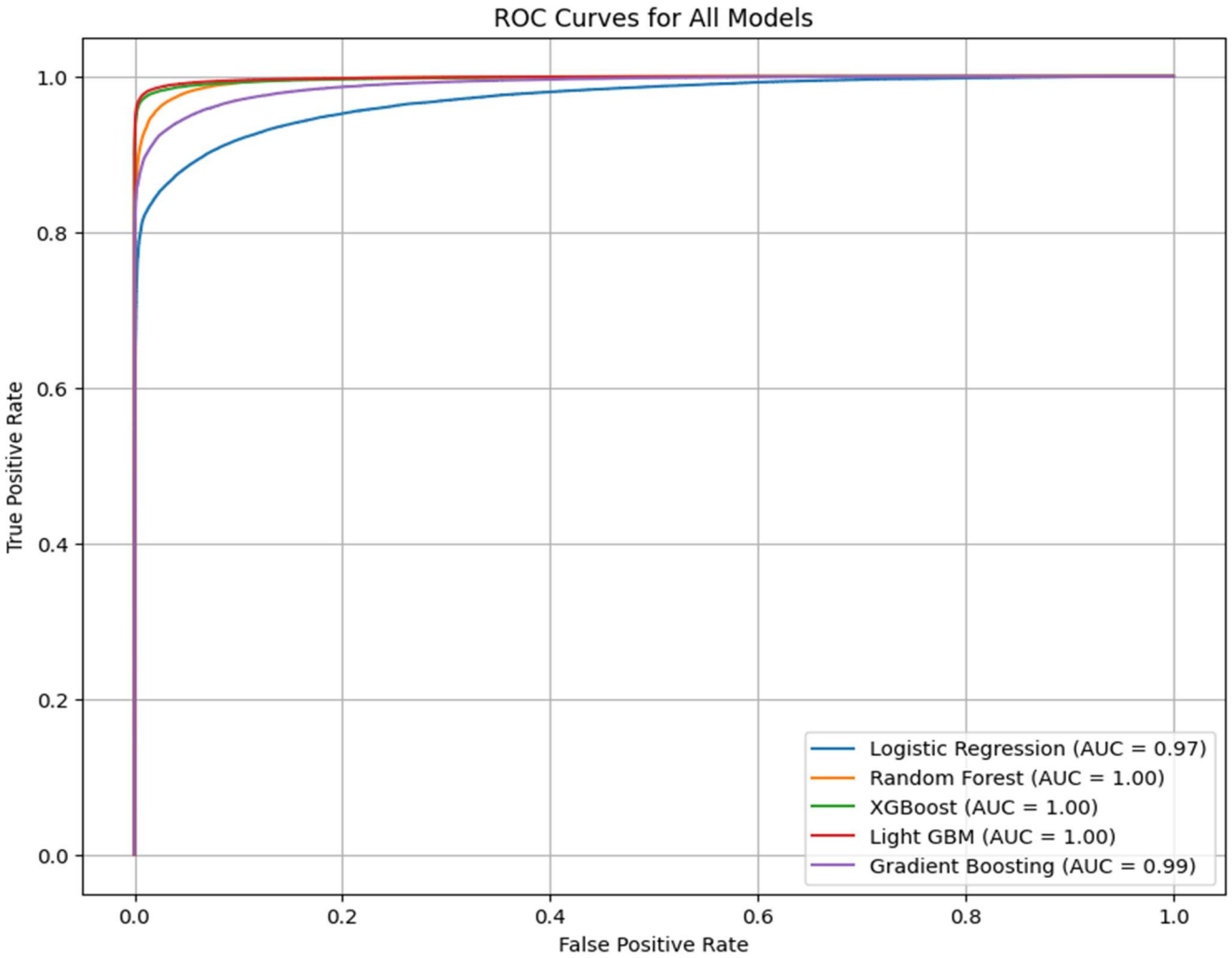
High AUC (Test AUC = 0.9981):

The AUC score reflects the model's ability to distinguish between fraud and non-fraud cases. LightGBM AUC is the highest, showing it is excellent at ranking fraudulent providers.

Generalization:

The gap between Train Recall (0.9743) and Test Recall (0.9726) is small, indicating that LightGBM generalizes well to unseen data without overfitting.

Therefore, the LightGBM model achieves the best balance between accuracy, recall, F1-score, and AUC, making it the most suitable choice for healthcare fraud detection. Its ability to generalize and prioritize recall ensures reliable performance in real-world scenarios.



Limitations and Assumptions:

- The model's performance is highly dependent on the quality and completeness of the dataset. Missing data could lead to inaccurate predictions, limiting the effectiveness of fraud detection.
- The model is trained on historical data, which limits its ability to detect new or evolving fraud patterns do not present in the training data.

Results:

Key Findings:

- The machine learning model, specifically the LightGBM classifier, achieved an overall accuracy of 98%, with a balanced F1-score of 0.98 for both classes, indicating strong predictive capability and minimal bias towards either fraudulent or non-fraudulent healthcare providers.
- The feature importance analysis revealed that key features such as **ProviderFraudRate**, **CostPerCoverageMonth**, and **ClaimsPerProvider** are the most influential in predicting healthcare fraud, offering significant insights into potential fraudulent behavior.

Streamlit Application:

Streamlit is an open-source framework designed to create interactive web applications with minimal code. In this project, Streamlit was used to build a dynamic dashboard that allows users to interact with the machine learning model for real-time fraud prediction and visualize insights from the dataset. The dashboard is designed for healthcare administrators and stakeholders to quickly identify fraudulent healthcare providers based on historical claims data.

Key Features of the Streamlit Dashboard:

- Users can input specific data (e.g., select a healthcare provider) and receive immediate fraud predictions from the trained LightGBM model.
- The dashboard displays visualizations such as fraud distribution, state-wise fraud rates, and a correlation heatmap of key features.
- The application includes widgets such as dropdown menus and buttons, enabling users to explore predictions and insights interactively without any coding experience.

Steps for Building the Streamlit Dashboard:

- **Data Integration and Preprocessing**

  The datasets were loaded and preprocessed before creating the dashboard. This included merging relevant data, managing missing values, and addressing the class imbalance using SMOTE. Once the data was ready, it was integrated into the Streamlit app to provide real-time interaction.

- **Fraud Prediction Model Integration**

  The trained LightGBM model was incorporated into the dashboard using joblib to make predictions about whether a healthcare provider is fraudulent or not. The model is accessed on the backend and the predictions are displayed instantly when users interact with the dashboard.

- **Interactive Features for Prediction**

  Users can select a specific provider from the dataset using a dropdown menu. Based on the selection, the dashboard fetches the relevant provider data and uses the model to

predict fraud, showing the likelihood of fraud and displaying the result (fraudulent or non-fraudulent).

- **Data Visualization and Insights**

  The following visualizations are displayed in the app.

- Fraud vs. Non-Fraud Counts

  A bar chart to show the distribution of fraudulent and non-fraudulent claims.

- State-wise Fraud Distribution

  A bar chart that visualizes the fraud rate across different states.

- Correlation Heatmap

  A heatmap to show the correlations between numerical features in the dataset, helping users understand the relationships between key variables.

- **User-Friendly Interface**

  The interface is designed to be intuitive and user-friendly, enabling users to easily navigate between the different sections of the dashboard. Streamlit's simple and responsive design ensures that users can access important insights without needing technical expertise.

- **Deployment**

  The final Streamlit app was deployed on Streamlit Community Cloud, making it accessible via a web link. This easy access and interaction with the dashboard in real time, help them make informed decisions based on fraud predictions.

- In conclusion, the Streamlit dashboard enhances the usability of the fraud detection system, providing healthcare professionals with a user-friendly interface to explore data, view predictions, and make timely decisions.

Discussion

Interpretation of Results

- The key findings of this project indicate that the LightGBM model is highly effective in detecting fraudulent healthcare providers, with an accuracy of 98%. The model identified ProviderFraudRate and ClaimsPerProvider as the most important features for predicting fraud. These features relate to the likelihood that a provider may engage in fraudulent activities, such as exaggerated billing or misrepresentation of services. The use of **SMOTE (Synthetic Minority Over-sampling Technique)** to address class imbalance significantly improved the model's ability to detect fraudulent cases, as evidenced by higher recall and F1-scores. SMOTE helped balance the dataset by generating synthetic samples of the minority class (fraudulent claims), which allowed the model to better generalize to new data.
- The Streamlit dashboard provided predictions and visual insights, making it a practical tool for healthcare administrators to quickly assess the likelihood of fraud. By allowing users to interact with the data, the dashboard enhanced the decision-making process. It enabled stakeholders to explore visualizations like fraud distribution by state, feature importance, and prediction results, which helped identify trends and make timely decisions.
- This interactive feature of the Streamlit dashboard improved the overall efficiency of fraud detection, reducing the need for manual work and enabling faster identification of potentially fraudulent providers. The combination of the fraud detection model and the Streamlit dashboard serves as a valuable resource for reducing financial losses due to fraud, ensuring that healthcare systems remain efficient and cost-effective.

Comparison with Existing Literature

- The findings of this project are consistent with existing literature on healthcare fraud detection using machine learning. Studies such as **Bauder & Khoshgoftaar (2017)** demonstrate the effectiveness of machine learning techniques, specifically Gradient Boosting Machines in detecting Medicare fraud.
- These studies also emphasize the importance of handling **class imbalance**, an issue we addressed with the **SMOTE** technique in this project. SMOTE is widely recommended for improving model performance in imbalanced datasets, particularly in fraud detection where fraudulent claims are significantly fewer than legitimate ones.
- Similarly, **Garmdareh et al. (2023)** proposed a machine learning-based approach for anomaly detection in medical insurance fraud, which aligns closely with our approach. Their work involved feature engineering and model evaluation to enhance fraud detection systems, confirming the value of using machine learning techniques for such tasks.
- Both studies stress the need for techniques like **oversampling** to manage class imbalance effectively, mirroring our approach with **SMOTE**. Moreover, their focus on **feature engineering** and model performance aligns with the importance of identifying the right features, as demonstrated by **ProviderFraudRate** and **ClaimsPerProvider** in our model.

Unexpected Findings

- While the project achieved promising results with the **LightGBM model**, one unexpected finding was the **significant role of features like ProviderFraudRate**. Initially, it was anticipated that features like **ClaimsPerProvider** would be the most predictive, but

ProviderFraudRate, which measures a provider's fraud history, played an even more significant role than expected.

- This highlights the value of **feature engineering** in uncovering hidden patterns and improving prediction accuracy. The ability to discover such features demonstrates the importance of in-depth data exploration, which can significantly enhance model performance.
- Another surprising aspect was the **impact of SMOTE** on improving **precision** as well as **recall**. Initially expected SMOTE to primarily address recall by helping the model identify more fraudulent cases, but it also improved precision, ensuring that the fraudulent claims identified were indeed fraudulent. This improvement helped in achieving a better balance between recall and precision, which contributed to a higher **F1-score** and overall model reliability.
- Lastly, despite the model's high accuracy, **false positives** remained a challenge. While the model was effective at detecting fraudulent cases, some legitimate claims were flagged as fraudulent. This is a common issue in fraud detection systems, where the cost of missing a fraudulent case is high, but the cost of investigating false positives can also be significant. This underscores the need for **continued refinement** of the model to reduce false positives while maintaining its ability to accurately detect fraud. Future work could focus on fine-tuning the model to improve its precision further.
- In addition, **Streamlit** played an important role in helping users interact with the model's predictions. The ability to quickly explore results and gain insights from the visualizations, such as the **fraud vs. non-fraud distribution** and **feature importance**, gave users actionable information. However, the challenge of **false positives** also

highlighted the need for ongoing model adjustments and further feature exploration, which could be done through iterative analysis and updates to the dashboard.

Conclusion

In conclusion, this project lays the foundation for an efficient, automated healthcare fraud detection system, combining machine learning, data engineering, and interactive tools. With future enhancements, it has the potential to drastically reduce fraud, optimize healthcare resources, and improve overall healthcare delivery.

Results and Insights

1. Model Performance and Accuracy

- **LightGBM Outperforms Other Models**: The LightGBM model achieved an **overall accuracy of 98%**, with a **balanced F1-score of 0.98**, indicating strong predictive capabilities for both fraudulent and non-fraudulent healthcare providers. This suggests that LightGBM is highly dependable for detecting fraud with minimal bias, making it well-suited for real-world deployment in fraud detection systems.
- **Key Metrics**:
    - **Recall**: The LightGBM model achieved a **test recall of 0.9726**, ensuring that most fraudulent claims were correctly identified, which is crucial in fraud detection where false negatives can be costly.
    - **AUC**: The **AUC score of 0.9981** reflects the model's exceptional ability to distinguish between fraud and non-fraud cases, providing an exceptionally reliable fraud detection mechanism.

2. Addressing Class Imbalance

- **Impact of SMOTE on Performance**: The application of **SMOTE (Synthetic Minority Over-sampling Technique)** successfully addressed the class imbalance issue, where fraudulent claims are far fewer than legitimate claims. By generating synthetic samples for the minority class (fraudulent claims), SMOTE helped the model identify fraudulent behavior more effectively, leading to improved **recall and F1-scores**.
    - **Balanced Dataset**: After SMOTE, the dataset was more balanced, which helped prevent the model from being biased towards the majority class (non-fraudulent claims). This ensures that the model can detect fraud with higher sensitivity.

3. Feature Importance and Insights

- **Critical Fraud Indicators**: Feature engineering revealed that certain features, particularly **ProviderFraudRate** (0.85 correlation with fraud), **ClaimsPerProvider** (0.33 correlation with fraud), and **CostPerCoverageMonth** (0.74 correlation with fraud), are the strongest indicators of fraudulent behavior. These features provide significant insights into fraud detection:
    - **ProviderFraudRate**: This feature was one of the most critical for detecting fraudulent behavior. A high **ProviderFraudRate** indicates that a provider has a history of submitting fraudulent claims, making it a key indicator for fraud detection.
    - **ClaimsPerProvider**: A higher number of claims per provider strongly correlates with fraud, suggesting that fraudsters may submit more claims to exploit the system.

- **CostPerCoverageMonth**: Unusually high costs per coverage month were identified as an indicator of fraudulent billing practices.

4. Insights from Streamlit Dashboard

- **Real-Time Fraud Prediction**: The **Streamlit dashboard** enhanced the practical application of the fraud detection model by allowing healthcare administrators to interactively assess fraud risks. This dashboard provides real-time predictions on whether a healthcare provider is potentially fraudulent based on the data entered.
    - **User Interaction**: Healthcare professionals can easily input specific provider data and receive instant predictions, helping them identify suspicious providers swiftly.
    - **Data Visualizations**: The dashboard displays key visualizations like fraud distribution by state, feature importance, and fraud vs. non-fraud counts. These visualizations help stakeholders better understand fraud patterns and make data-driven decisions in a user-friendly format.

5. Practical and Financial Impact

- **Cost Reduction**: By detecting fraudulent healthcare providers early, the model can prevent financial losses from fraudulent claims. The **98% accuracy** ensures that fraudulent claims are flagged with high reliability, allowing for quicker investigations and resource allocation, which could result in significant cost savings for healthcare systems.
- **Operational Efficiency**: The implementation of the machine learning model and the Streamlit dashboard enables automated fraud detection, reducing the need for manual

claim reviews. This leads to faster identification of fraudulent claims, enhancing operational efficiency within healthcare organizations.

Recommendations and Relevance to the Problem

- **Continuous Monitoring and Real-Time Integration**: The model's ability to predict fraud in real-time, through integration into a healthcare provider's claims system, could enable **continuous fraud monitoring**, preventing fraudulent activities before they impact the healthcare system.
- **False Positive Management**: While the model is highly accurate, false positives remain a challenge. Introducing post-prediction analysis and a **scoring system** to prioritize fraudulent claims based on severity could help reduce unnecessary investigations and optimize resource allocation.
- **Anomaly Detection:** Implement unsupervised learning models to detect emerging or unknown fraud patterns that may not be captured by supervised models.

Article1

- This paper is highly relevant to Medicare fraud detection project using machine learning.
- The study uses the 2015 Medicare Provider Utilization and Payment Data that is like Kaggle dataset along with the List of Excluded Individuals/Entities (LEIE) database for fraud labels.
- It compares various machine learning methods, including supervised, unsupervised, and hybrid approaches, to detect fraudulent Medicare providers.
- The authors address the class imbalance problem by using oversampling and under-sampling techniques.

- They evaluate model performance using multiple metrics, including balanced accuracy, F-measure, G-measure, and Matthew's Correlation Coefficient.
- The study also breaks down the analysis by provider types.
- The paper discusses the challenges of working with large-scale Medicare claims.
- data including data cleaning, normalization, and feature engineering.
- It provides insights into which algorithms and techniques might be most effective for Medicare fraud detection, comparing the performance of different machine learning models across various provider specialties.
- The authors' findings on the effectiveness of different sampling methods and the performance of various machine learning algorithms in the specific context of Medicare fraud detection can help in determining the methodology and model selection.

Article2

- This paper highly relevant to project on using machine learning for medical insurance fraud detection.
- The authors propose a machine learning-based approach to detect anomalies and potential fraud in medical insurance claims by predicting the expected claim price and comparing it to the actual price.
- Regression algorithms like decision trees to predict claim prices based on various features of the claim and patient.
- They compare the predicted price to the actual price to identify anomalous claims that deviate significantly.
- They evaluate multiple machine learning algorithms including decision trees, random forests, and deep learning to find the best performing model.

- They use feature engineering and selection techniques to improve the model performance.
- They validate their approach by having human experts review the flagged anomalous claims.
- Overall, aligns closely with goal of using machine learning to detect potential fraud in medical insurance claims.

References

- Dataset Link
- Bauder,R.A.,&Khoshgoftaar,T.M.(2017). Medicare fraud detection using machine learning methods. In 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE. Article1
- Garmdareh,M.S.,Neysiani,B.S.,Nogorani, M. Z., & Bahramizadegan, M. (2023). A Machine Learning-based Approach for Medical Insurance Anomaly Detection by Predicting Indirect Outpatients' Claim Price. In 2023 9th International Conference on Web Research (ICWR) (pp. 129- 134). IEEE. Article2

Appendices

- DataExplorationNotebook
- DataCleaningNotebook
- ExploratoryDataAnalysisNotebook
- MachineLearningModelsImplementation
- Github
- Streamlit