Technical Report

1. Project Objective:

The goal of this project is to build an end-to-end fraud detection system capable of identifying fraudulent healthcare providers. The system should:

- Detect fraudulent providers from multi-table claims data.
- Address the severe class imbalance with only 10% of providers labeled as fraudulent)
- Ensure explainable predictions, aiding investigators and regulators.
- Prioritize high-risk providers, improving operational efficiency for CMS.

2. Data Understanding & Exploration

2.1 Dataset Description:

The dataset provided by CMS consists of four primary CSV files:

- **Train_Beneficiarydata.csv**: Contains demographics, coverage, and chronic conditions for each Medicare beneficiary.
- **Train_Inpatientdata.csv**: Includes hospital admission claims with financial, procedural, and physician details.
- **Train_Outpatientdata.csv**: Includes outpatient claim data such as visits, tests, and minor procedures.
- **Train_labels.csv**: Contains fraud labels (fraudulent or non-fraudulent) for healthcare providers.

Key Identifiers:

- **BeneID** links patients to claims.
- **Provider** links claims to fraud labels.

2.2 Data Quality and Completeness

Upon reviewing the dataset, we found missing values in several columns, which were handled using imputation methods. Inconsistent data entries (ex: incorrect provider IDs) were cleaned by standardizing formats. The data from multiple sources (inpatient and outpatient) was merged with careful consideration of granularity and consistency across claims.

2.3 Exploratory Data Analysis (EDA)

The EDA phase uncovered several patterns and relationships in the data:

- **Behavioral Differences**: Fraudulent and non-fraudulent providers exhibited distinct patterns in billing amounts, procedure types, and claim frequencies.
- **Outliers**: Certain providers had unusually high claim amounts, which were indicative of potential fraud.
- **Distributions**: Analysis revealed that the fraud label was highly imbalanced, with fraudulent providers representing only 10% of the dataset.

Key Visualizations:

- **Fraud vs Non-Fraud Providers**:
  The distribution of fraudulent versus non-fraudulent providers was highly imbalanced, with fraudulent providers making up only a small portion of the dataset. This

imbalance poses challenges for training a robust fraud detection model.

- **Claim Amount Distribution (Inpatient and Outpatient)**: The distribution of claim amounts showed significant variation. Inpatient claims tended to have higher values, whereas outpatient claims had a smaller spread.
- **Total Claims per Provider**:
  Fraudulent providers were associated with a significantly higher number of claims. The analysis of total claims per provider revealed outliers in both inpatient and outpatient datasets.
- **Average Claim Cost per Provider**:
  Providers associated with fraud had significantly higher average claim amounts, which was consistent across both inpatient and outpatient datasets.
- **Number of Procedures per Provider**:
  The number of procedures was another distinguishing feature between fraudulent and non-fraudulent providers.

2.4 Feature Engineering

To prepare the data for modeling, several features were engineered:

- **Total Procedures per Claim**: Counted the number of non-null procedure codes for each claim.
- **Total Claim Cost per Provider**: Aggregated the total reimbursement amounts for each provider.
- **Beneficiary Features**: Features like age and chronic conditions were aggregated per provider to give more

context about the beneficiaries associated with each provider.

The data was then merged into one comprehensive dataset, combining both inpatient and outpatient features, along with fraud labels.

3. Modeling

3.1 Model Selection and Training

To tackle the fraud detection problem, several classification models were trained, including:

- **Decision Tree Classifier**
- **Logistic Regression**
- **Random Forest Classifier**
- **Gradient Boosting Classifier**
- **Support Vector Machine (SVM)**

Each model was trained using the **SMOTE (Synthetic Minority Over-sampling Technique)** technique to address the **class imbalance** problem. The dataset was preprocessed by performing **Standard Scaling** on numeric features and **One-Hot Encoding** on categorical features.

The final models were evaluated using key metrics including:

- **Precision**
- **Recall**
- **F1 Score**
- **PR AUC**

- **ROC AUC**

3.2 Evaluation Metrics

The following metrics were calculated for each model to assess its performance:

- **Precision**: Measures the proportion of true positive predictions out of all positive predictions (minimizing false positives).
- **Recall**: Measures the proportion of true positive predictions out of all actual positive instances (minimizing false negatives).
- **F1 Score**: The harmonic mean of precision and recall, providing a balance between them.
- **PR AUC**: Area under the Precision-Recall curve, which is more informative for imbalanced datasets.
- **ROC AUC**: Area under the Receiver Operating Characteristic curve, indicating the model's ability to distinguish between classes.

3.3 Results and Comparison

Here are the evaluation results for each model:

Model Comparison Results:

Model	Precision	Recall	F1 Score	PR AUC	ROC AUC
Logistic Regression	0.7581	0.4476	0.5629	0.6813	0.929

Model	Precision	Recall	F1 Score	PR AUC	ROC AUC
Random Forest	0.6885	0.4000	0.5060	0.6340	0.915
Gradient Boosting	0.6667	0.4000	0.5000	0.6608	0.932
Support Vector Machine	0.7660	0.3429	0.4737	0.4979	0.6477

- **Best Performing Model**: The **Logistic Regression** model demonstrated the highest performance, achieving an **ROC AUC of 0.929** and a **PR AUC of 0.6813**, making it a strong candidate for deployment.

3.4 Detailed Evaluation (Random Forest)

Here are the metrics for the **Random Forest** model after training and evaluation:

- **Precision**: 0.53125
- **Recall**: 0.6476190476190476
- **F1 Score**: 0.5938680757472136
- **PR AUC**: 0.617027233399324
- **ROC AUC**: 0.9242838912560317

Confusion Matrix for **Random Forest**:

- **Key Observations**: The **Random Forest** model performs well, with a **Recall** of 0.65, which is essential for identifying as many fraudulent providers as possible, even if it sacrifices some precision.

3.5 ROC Curve and PR Curve

ROC Curve for Random Forest:

PR Curve for Random Forest:

These curves illustrate the trade-off between recall and precision for different thresholds, with **Random Forest** yielding a strong performance in terms of distinguishing fraudulent providers from non-fraudulent ones.

3.6 Conclusion

Based on the evaluation metrics and visual analysis:

- **Logistic Regression** is the most efficient model for fraud detection, with strong scores in both **Precision** and **Recall**, along with the highest **ROC AUC**.
- **Random Forest** also performed well, providing a good balance between **Precision** and **Recall**, with a **ROC AUC** of 0.9243.

Future work will involve fine-tuning these models, potentially incorporating additional features (such as temporal or geographic data) and optimizing hyperparameters to further improve the model's predictive capabilities.

4. Evaluation:

4.1 Cross-Validation Results

In the evaluation phase, we used **Stratified K-Fold Cross-Validation** with 5 folds to evaluate the performance of our models. This method ensures that each fold contains a proportionate number of fraudulent and non-fraudulent cases, providing a balanced assessment. The average cross-validation accuracy across all folds was **91.01%**, indicating that the model performed well in detecting fraud with high consistency.

4.2 False Positive and False Negative Cases

We analyzed the cases where the model made errors by identifying **false positives** and **false negatives**:

- **False Positive Cases**: These are legitimate providers that were incorrectly flagged as fraudulent.
    - Example cases include providers **PRV52296**, **PRV52318**, and **PRV51438**.
- **False Negative Cases**: These are fraudulent providers that were missed by the model.
    - Example cases include providers **PRV51378**, **PRV51480**, and **PRV51170**.

These cases are critical for improving the model, especially in balancing between detecting fraud and avoiding false alarms.

4.3 Cost Analysis of False Positives and False Negatives

We assigned cost values to false positives and false negatives to quantify the financial impact of the model's errors:

- **False Positive Cost**: Each false positive was assigned a cost of **$1,000**, representing the cost of unnecessary fraud investigation or audit.
- **False Negative Cost**: Each false negative was assigned a cost of **$5,000**, reflecting the financial losses from undetected fraudulent activities.

The total cost for false positives and false negatives was calculated, which helped us understand the economic impact of the model’s performance. The total costs for each type of error were as follows:

- **Total False Positive Cost**: $3,000
- **Total False Negative Cost**: $15,000

These costs highlight the importance of reducing false negatives, which have a significantly higher financial impact.

4.4 Hyperparameter Tuning with GridSearchCV

We performed hyperparameter tuning using **GridSearchCV** to find the best configuration for our Random Forest classifier. The optimal hyperparameters were found to be:

- **Max Depth**: None
- **Number of Estimators**: 200

This process improved the model’s performance, leading to an increase in the **ROC AUC** score to **0.77**, indicating a good balance between true positive and false positive rates.

4.5 Model Evaluation Metrics

The models were evaluated using several performance metrics, including **Precision**, **Recall**, **F1 Score**, **PR AUC**, and **ROC AUC**. The following results were obtained for each model:

Logistic Regression:

- **Precision**: 0.4583
- **Recall**: 0.8381
- **F1 Score**: 0.5925
- **PR AUC**: 0.6849

- **ROC AUC**: 0.9366

Random Forest:

- **Precision**: 0.544
- **Recall**: 0.6476
- **F1 Score**: 0.5913
- **PR AUC**: 0.6125
- **ROC AUC**: 0.9231

Gradient Boosting:

- **Precision**: 0.4907
- **Recall**: 0.7619
- **F1 Score**: 0.5970
- **PR AUC**: 0.6265
- **ROC AUC**: 0.9167

4.6 Model Comparison

A comparison table of the evaluation metrics for each model is shown below:

Model	**Precision**	**Recall**	**F1 Score**	**PR AUC**	**ROC AUC**
Logistic Regression	0.4583	0.8381	0.5925	0.6849	0.9366
Random Forest	0.544	0.6476	0.5913	0.6125	0.9231
Gradient Boosting	0.4907	0.7619	0.5970	0.6265	0.9167

4.7 Conclusion

From the results, **Logistic Regression** showed the best performance in terms of **Precision** and **ROC AUC**, indicating it was particularly effective in minimizing false positives while maintaining a high detection rate. However, **Gradient Boosting** demonstrated the highest **Recall**, making it more effective at detecting fraudulent providers, though it came with a tradeoff in **Precision**.

The **Random Forest** model showed a balanced performance, performing moderately across all metrics. Based on this, further improvements in model tuning or feature engineering could help enhance the detection accuracy, particularly for minimizing **false negatives**, which have a higher financial impact.

Further Investigation

We also performed a deeper analysis of the **false positive** and **false negative** cases by examining the features associated with these errors. This analysis helps to understand why the model made certain misclassifications and could inform feature engineering efforts to improve the model's performance in future iterations.

5. Error Analysis:

False Positives

False positives occur when legitimate providers are incorrectly flagged as fraudulent. These misclassifications can lead to unnecessary investigations, audits, and resource allocation, ultimately resulting in operational inefficiencies. In the context of healthcare fraud detection, false positives

might cause undue distress to legitimate providers who could face unnecessary scrutiny or delays in reimbursement.

The cost of false positives was assigned a value of $1,000 per instance, representing the financial burden of performing unnecessary audits or reviews on providers incorrectly labeled as fraudulent. For this analysis, we identified three false positive cases, such as PRV52296, PRV52318, and PRV51438, which contributed to a total False Positive Cost of $3,000.

Business Implications:

- **These cases could affect provider relationships, leading to frustration and possible disengagement from the insurance system.**
- **The healthcare system incurs additional costs due to unnecessary investigations and audits.**

False Negatives

False negatives, on the other hand, occur when fraudulent providers are incorrectly classified as legitimate. This is a more severe error in the context of fraud detection, as it allows fraudulent activities to go undetected, leading to significant financial losses. Fraudulent claims can result in overpayments, abuse of the system, and exploitation of vulnerable populations.

The cost of false negatives was higher, assigned at $5,000 per case, reflecting the substantial financial impact of undetected fraud. Three false negative cases, such as

PRV51378, PRV51480, and PRV51170, were identified, resulting in a total False Negative Cost of $15,000.

Business Implications:

- **Failure to detect fraudulent providers can result in large-scale financial losses, undermining the integrity of the healthcare system.**
- **Persistent false negatives could erode trust in the system, especially if fraudulent activities are not consistently detected and addressed.**

Cost Analysis Summary

- **Total False Positive Cost: $3,000**
- **Total False Negative Cost: $15,000**

The disproportionate impact of false negatives highlights the importance of prioritizing the reduction of these errors in future model improvements. The substantial cost of undetected fraud presents a strong business case for refining the detection capabilities to minimize these errors.