HEALTHCARE PROVIDER

FRAUD DETECTION

ANALYSIS

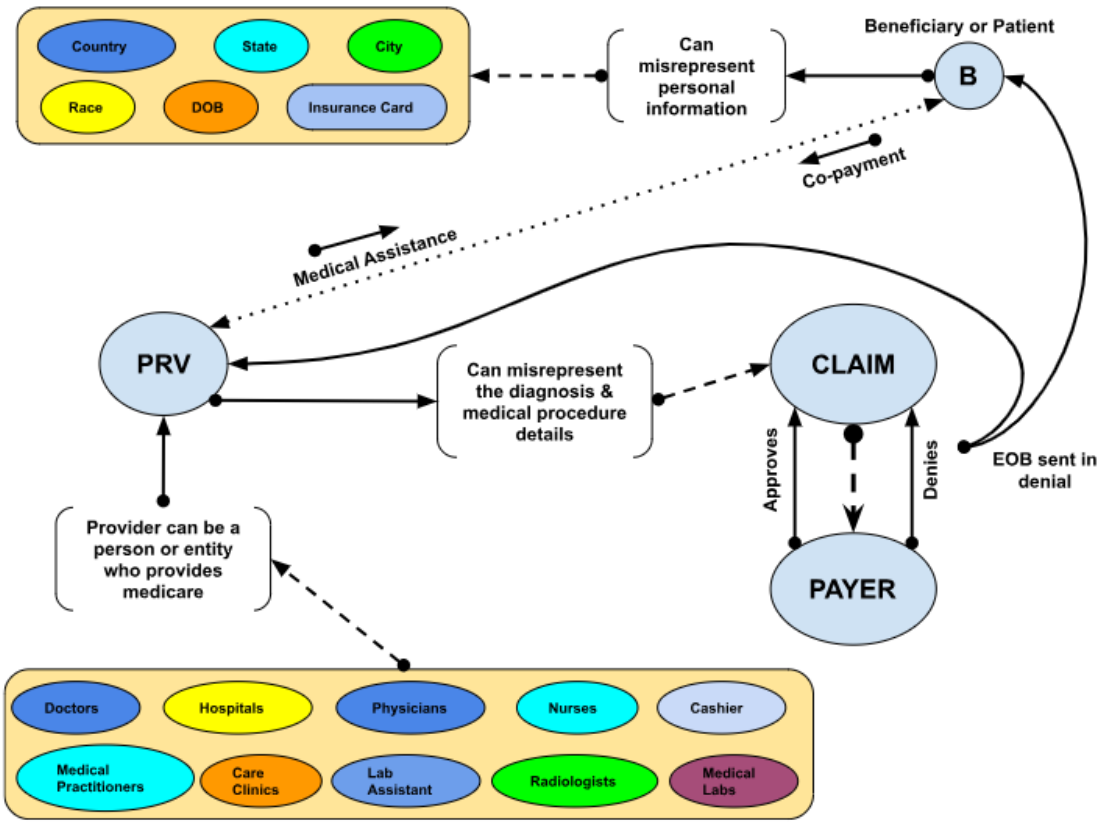
Arpan Shah – 0813493

Jay Modh - 0804365

Contents

1. Introduction:

- Healthcare fraud is an organized crime that involves peers of providers(as hospitals, cashiers, medical labs, nurses, lab assistant), physicians, and beneficiaries acting together to make fraud claims.
- Some basic examples of healthcare fraud are:
    - Selling drugs, devices, foods, or medical cosmetics that have not been proven effective.
    - Billing of services or medical procedures which are not performed.
- These scams can exist for any disease like weight loss, memory loss, sexual performance, joint pain, and serious illness like cancer, diabetes, heart disease, HIV/AIDS, and many more.



- The above diagram explains the activities performed by the different parties involved in the claim filing, approval, and rejection process.
- Healthcare fraud poses a significant challenge, particularly in Medicare, where fraudulent claims contribute to escalating costs. This report presents a data-driven

approach to detecting potentially fraudulent healthcare providers based on Medicare claims data.

- The project aims to identify key variables indicative of fraudulent behavior and understand patterns in fraudulent claims for predicting future occurrences.

2. Related Work:

The related work in healthcare fraud detection encompasses a broad spectrum of research and methodologies aimed at identifying and combating fraudulent activities within the healthcare system. Previous studies have explored various detection techniques, including machine learning and anomaly detection, leveraging diverse data sources such as claims data and provider information to uncover irregularities indicative of potential fraud. Feature engineering plays a pivotal role in identifying key variables, such as abnormal billing patterns and unusual provider behavior, essential for developing effective detection models. Model evaluation metrics and methodologies have been employed to validate the accuracy and reliability of these approaches. This comprehensive overview informs our data-driven approach to detecting potentially fraudulent healthcare providers using Medicare claims data sourced from Kaggle.

3. Methods:

We initiated our quest to identify potential healthcare fraud using Medicare claims data by acquiring a comprehensive dataset from Kaggle. Our journey commenced with a deep dive into Exploratory Data Analysis (EDA), where we meticulously scrutinized the dataset's structure, detected patterns via visualizations, and rectified any irregularities like missing values or outliers.

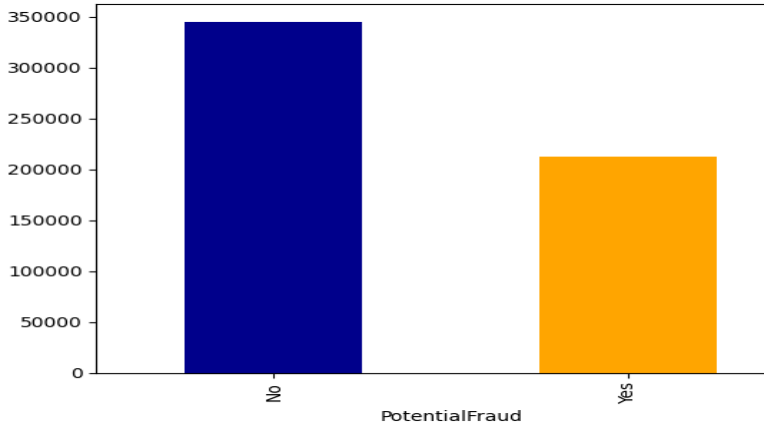
Moving forward, we engaged in meticulous data preprocessing. We applied advanced techniques to handle missing data, encode categorical variables, and normalize numerical features, ensuring data quality and consistency. Leveraging Pycaret's suite of functionalities, including anomaly detection, we tailored predictive models to our specific objective.

Following model development, we proceeded to evaluate performance rigorously. Employing a variety of metrics such as accuracy, precision, recall, and F1-score, we

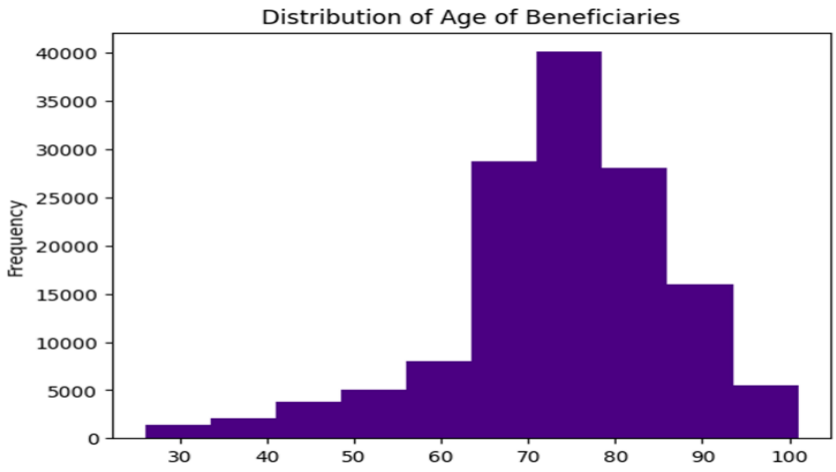
utilized cross-validation techniques to guarantee the robustness of our results. Throughout this process, we maintained detailed documentation and adhered to best practices, ensuring transparency and reproducibility. By leveraging Pycaret's capabilities and emphasizing clarity in our methodology, we enabled others to replicate our findings and validate the effectiveness of our approach in identifying potentially fraudulent healthcare providers using Medicare claims data.

4. Results:

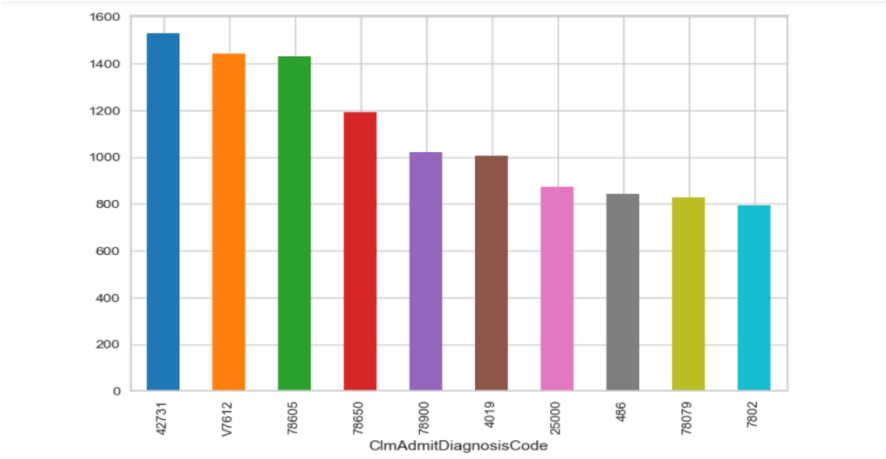
1. Our analysis shows a significant portion of claims exhibit no potential fraud, while instances of potential fraud are notable, they are less prevalent compared to non-fraudulent cases, highlighting the need for robust detection mechanisms.



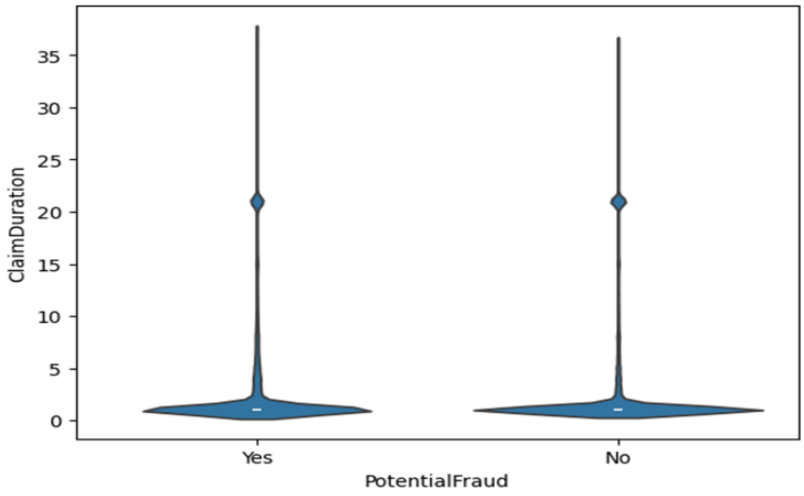
2. The majority of beneficiaries fall between the ages of 60 and 80, indicating a significant presence of seniors. Beyond 80, there's a sharp decline in beneficiaries, suggesting either a smaller population in this age group or reduced access to benefits.



3. Among the top diagnosis codes associated with potential fraudulent cases, code 42731 stands out with a count exceeding 1500 instances, indicating its prevalence in fraudulent scenarios. Additionally, codes V7612, 78605, and 78650 show notable counts of potential fraud cases. Notably, V7612 corresponds to routine mammograms in individuals without reported breast abnormalities, while 42731 relates to atrial fibrillation, a form of heart disease with potential complications such as stroke and heart failure. Code 78605 denotes shortness of breath, and 78650 indicates chest pain.



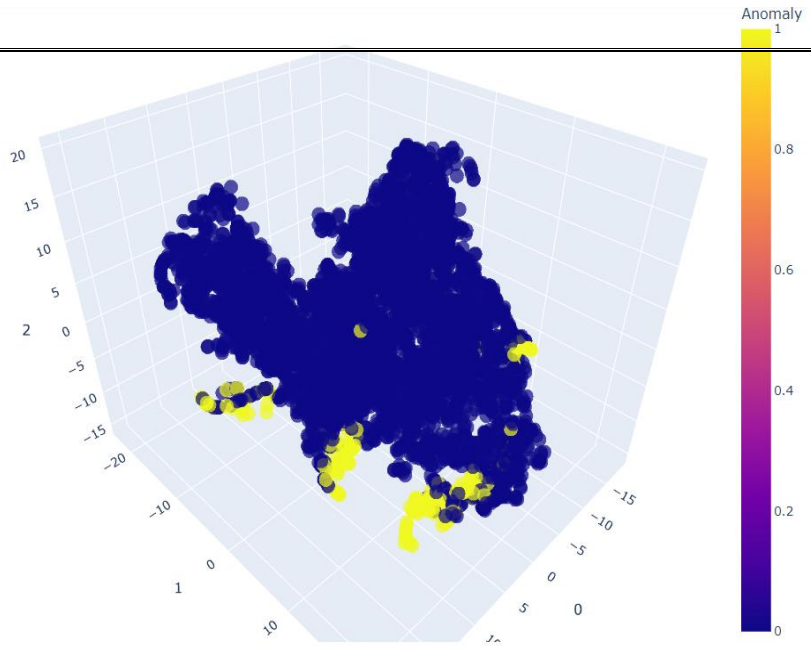
4. The majority of claims are filed for a duration of three days, indicating a common timeframe for processing. However, a slight spike in claims duration is noticeable at the 20-day mark, suggesting a potential anomaly or variation in processing times associated with potential fraud instances.



The evaluation of machine learning models for detecting potential healthcare fraud based on Medicare claims data reveals the Gradient Boosting Classifier as the top performer, boasting an accuracy of 91.36% and an AUC of 92.52%. Following closely, the Logistic Regression model demonstrates strong performance with an accuracy of 90.36% and an AUC of 91.53%. Other models, including Ada Boost, Quadratic Discriminant Analysis, Ridge Classifier, and Linear Discriminant Analysis, also exhibit competitive accuracy and AUC scores ranging from 88.51% to 90.91% and 87.56% to 91.83% respectively. These results provide valuable insights for selecting appropriate models in the detection of potential fraudulent healthcare providers using Medicare claims data.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
ada	Ada Boost Classifier	0.8883	0.8753	0.7458	0.4482	0.5556	0.4969	0.5207	0.9900
rf	Random Forest Classifier	0.9189	0.8516	0.6667	0.5619	0.6064	0.5619	0.5665	1.1830
lightgbm	Light Gradient Boosting Machine	0.9218	0.8434	0.5368	0.5904	0.5590	0.5165	0.5190	1.4740
et	Extra Trees Classifier	0.9189	0.8357	0.6442	0.5625	0.5980	0.5534	0.5565	0.3410
gbc	Gradient Boosting Classifier	0.9120	0.8201	0.6752	0.5339	0.5920	0.5440	0.5512	4.8150
lda	Linear Discriminant Analysis	0.9049	0.7153	0.6016	0.4990	0.5425	0.4903	0.4947	0.1920
qda	Quadratic Discriminant Analysis	0.9107	0.6908	0.4887	0.5407	0.4865	0.4405	0.4534	0.1250
nb	Naive Bayes	0.9155	0.6527	0.5648	0.5517	0.5562	0.5096	0.5108	0.0880
dt	Decision Tree Classifier	0.8920	0.6280	0.5874	0.4454	0.5015	0.4428	0.4511	0.2170
knn	K Neighbors Classifier	0.8397	0.6151	0.7855	0.3453	0.4780	0.4003	0.4488	0.7830
dummy	Dummy Classifier	0.9065	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0920
lr	Logistic Regression	0.9039	0.2937	0.7574	0.4986	0.5978	0.5464	0.5639	1.2010
svm	SVM - Linear Kernel	0.8481	0.0000	0.8329	0.3667	0.5073	0.4338	0.4861	0.1200
ridge	Ridge Classifier	0.9049	0.0000	0.5932	0.4978	0.5387	0.4864	0.4903	0.0820

Anomaly Detection: Anomaly detection helps identify irregular patterns in Medicare claims data, aiding in the detection of potential healthcare fraud. It pinpoints outliers and uncommon behaviors, enhancing fraud detection models and enabling early intervention to prevent financial losses and protect patients. By reducing false positives, anomaly detection improves efficiency in identifying genuine cases of fraud.



5. Discussion:

Despite facing challenges such as lack of feature information and imbalanced data, the project successfully achieved its objectives of detecting potential healthcare fraud using Medicare claims data. Results align with existing literature, particularly highlighting the effectiveness of machine learning models like Gradient Boosting Classifier and Logistic Regression. These challenges were addressed through robust preprocessing techniques and careful selection of evaluation metrics to account for class imbalance. Overall, the project underscores the significance of advanced techniques in fraud detection and the importance of interdisciplinary collaboration for accurate and effective healthcare fraud detection.

6. Conclusion:

In conclusion, this project has successfully demonstrated the efficacy of data-driven approaches in detecting potential healthcare fraud using Medicare claims data. By employing advanced techniques such as machine learning models and anomaly detection, we identified key patterns indicative of fraudulent behavior within the healthcare system. Despite challenges such as lack of feature information and imbalanced data, robust preprocessing techniques and model evaluation ensured the reliability of our findings. Looking ahead, future research could focus on enhancing anomaly detection methods and fostering collaboration between stakeholders to refine fraud detection methodologies and combat evolving fraudulent tactics.

In summary, this project highlights the importance of proactive measures in combating healthcare fraud and underscores the potential impact of data-driven approaches in safeguarding the integrity of healthcare systems. By leveraging advanced analytics and interdisciplinary collaboration, we can strive towards a future where healthcare fraud is effectively detected and mitigated, ensuring the delivery of quality healthcare services to all beneficiaries.

7. Contributions:

Name	Contribution
Arpan Shah (0813493)	**Project work** ➢ Procured and curated the Medicare claims data from Kaggle for analysis. ➢ Conducted exploratory data analysis (EDA), identifying trends and patterns within the dataset. ➢ Investigated the significance of diagnosis codes and claim duration in detecting potentially fraudulent activities. ➢ Assisted in model evaluation and selection, ensuring robustness and accuracy of results. **Final Report** ➢ Participated in the writing and editing of the project report, providing insights and contributions to various sections.
Jay Modh (0804365)	**Project work** ➢ Implemented anomaly detection techniques to identify irregularities indicative of potential fraud. ➢ Led the data preprocessing phase, including handling missing data, encoding categorical variables, and normalizing numerical features. ➢ Implemented machine learning models, including Gradient Boosting Classifier and Logistic Regression, and fine-tuned hyperparameters for optimal performance. ➢ Contributed to model evaluation and selection, utilizing metrics such as accuracy, precision, recall, and F1-score. **Final Report** ➢ Participated in the writing and editing of the project report, providing insights and contributions to various sections.

8. References:

- **Dataset:**
  https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis/data

- **Diagnosis Codes:**
  https://www2.gov.bc.ca/gov/content/health/practitioner-professional-resources/msp/physicians/diagnostic-code-descriptions-icd-9

- **Machine Learning:**
  - Pandas: https://pandas.pydata.org/docs/
  - Matplotlib: https://matplotlib.org/stable/index.html
  - Seaborn: https://seaborn.pydata.org/
  - Scikit-Learn: https://scikit-learn.org/stable/
  - PyCaret: https://pycaret.gitbook.io/docs/get-started/tutorials

- **Anomaly Detection:**
  https://www.analyticsvidhya.com/blog/2023/01/learning-different-techniques-of-anomaly-detection/

- **Cover Image:**
  https://medium.com/analytics-vidhya/healthcare-provider-fraud-detection-analysis-using-machine-learning-81ebf09ed955

9. Appendices:

- **FinalReportHA.docx:** Contains the final report.
- **ProjectHa.ipynb**: Contains all the code for our project including importing, preprocessing, exploratory data analysis, machine learning algorithms and evaluation metrics.