Healthcare Provider Fraud Detection Analysis & Prediction

Department of Information Systems, University of Maryland, Baltimore County

Keerthana Mallavarapu, Mouna Chimata, Mounika Cheera, Shivani Gajula , Vrushali Vishal Patil

ABSTRACT

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 industry such as billing & insurance fraud, unnecessary medical procedure etc. These fraudulent activities divert resources away from providing quality healthcare to patients. Identifying these frauds and predicting future frauds would help the healthcare industry not only provide quality healthcare to patients but also reduce financial losses, reputation loss, improve compliance and resource allocation as well as gain patient trust.

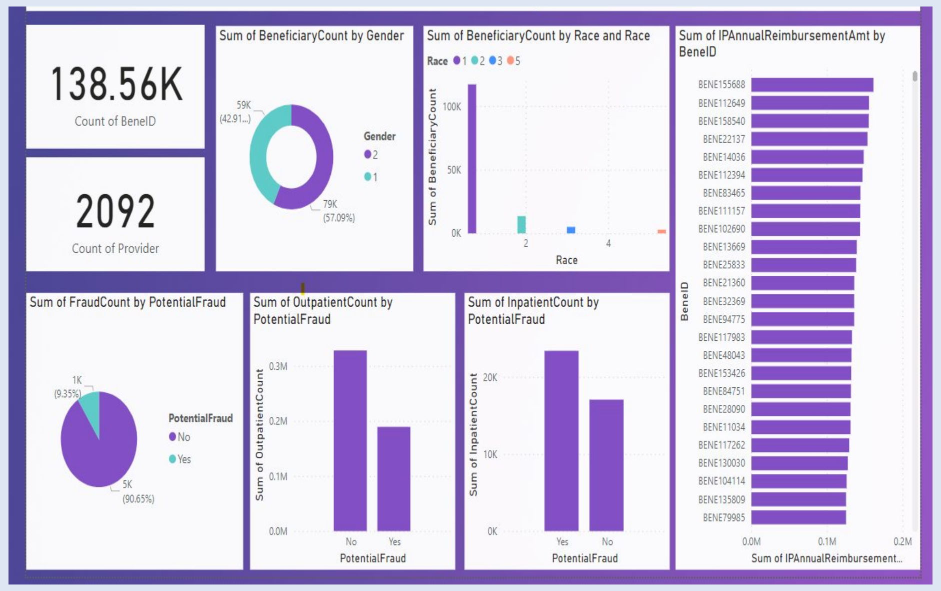
This poster outlines our data collection and analysis techniques, illustrates rigorous data cleaning and preprocessing. We employ four machine learning models for health insurance fraud prediction, meticulously training and evaluating them on a test dataset using diverse metrics. The abstract concludes by identifying the top-performing model and offering insights into its superior performance.

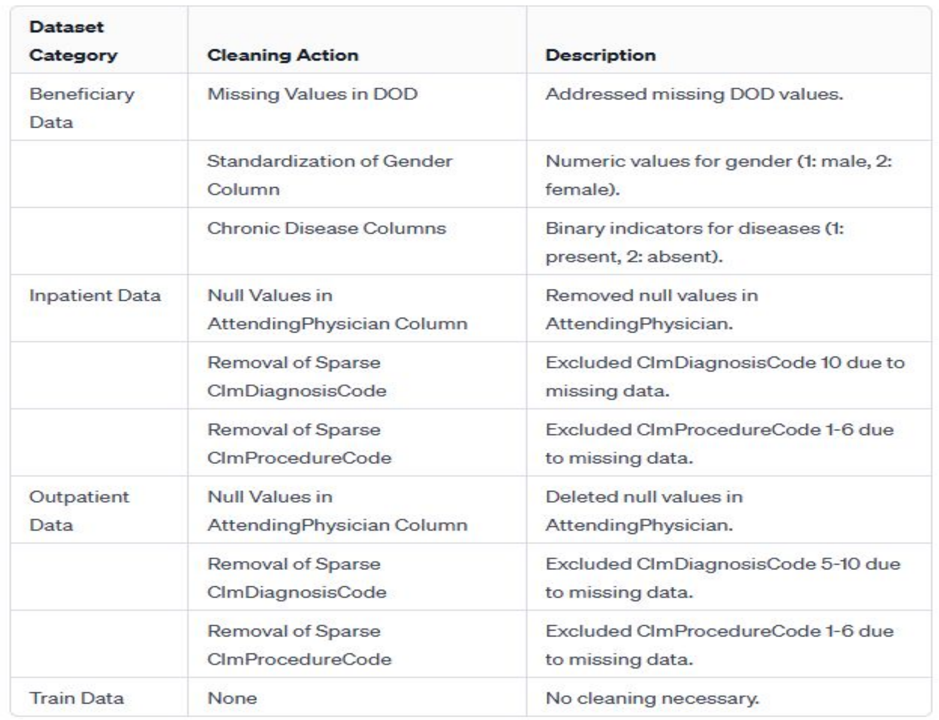
LITERATURE SURVEY

Review 1: “A survey on statistical methods for detecting health care fraud” [3] thoroughly investigates how to evaluate the performance of binary classifiers in health fraud. They differentiate between error-based and cost-based assessment approaches, focusing on establishing a confusion matrix using test datasets. Delving into error-based strategies, like ROC curves and AUC, to gauge classifier effectiveness. Cost-based methods entail a model considering case and effort costs, analyzing expenses linked to true positives, true negatives, false positives, and false negatives. This study also tackles hurdles like inaccurate labels in training data, the necessity for emerging different methods.

Review 2: “Fraud Detection in Healthcare Insurance Claims Using Machine Learning “ [4] .The survey delves into the notable progress made in utilizing ML for healthcare insurance fraud detection, in the research conducted by Nabrawi and Alanazi. Their study utilizes Random Forest, Logistic Regression, and ANN, demonstrating strong accuracy levels and identifying crucial factors like policy type, education, and age that significantly influence fraudulent activities. The survey highlights the versatility of these models, stressing the significance of demographic factors in predicting fraud.

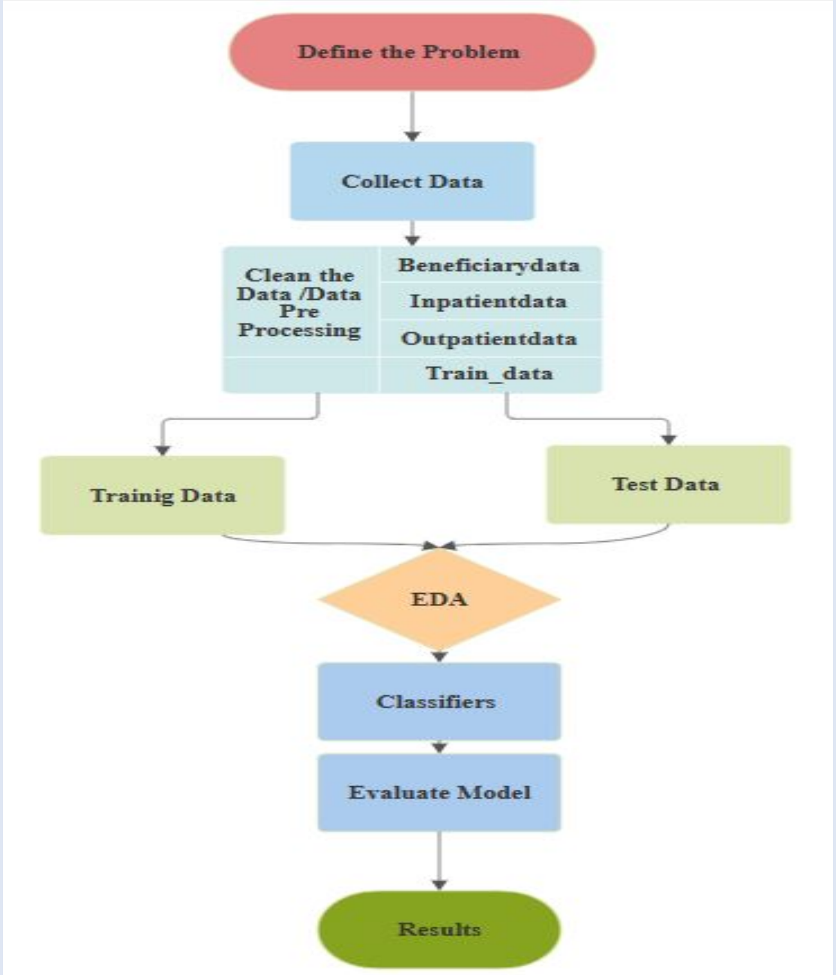
EDA & DATA PREPROCESSING





Dataset Category	Cleaning Action	Description
Beneficiary Data	Missing Values in DOD	Addressed missing DOD values.
	Standardization of Gender Column	Numeric values for gender (1: male, 2: female).
	Chronic Disease Columns	Binary indicators for diseases (1: present, 2: absent).
Inpatient Data	Null Values in AttendingPhysician Column	Removed null values in AttendingPhysician.
	Removal of Sparse ClmDiagnosisCode	Excluded ClmDiagnosisCode 10 due to missing data.
	Removal of Sparse ClmProcedureCode	Excluded ClmProcedureCode 1-6 due to missing data.
Outpatient Data	Null Values in AttendingPhysician Column	Deleted null values in AttendingPhysician.
	Removal of Sparse ClmDiagnosisCode	Excluded ClmDiagnosisCode 5-10 due to missing data.
	Removal of Sparse ClmProcedureCode	Excluded ClmProcedureCode 1-6 due to missing data.
Train Data	None	No cleaning necessary.

METHODOLOGY



Define the Problem

Collect Data

Clean the Data /Data Pre Processing

Beneficiarydata

Inpatientdata

Outpatientdata

Train_data

Trainig Data

Test Data

EDA

Classifiers

Evaluate Model

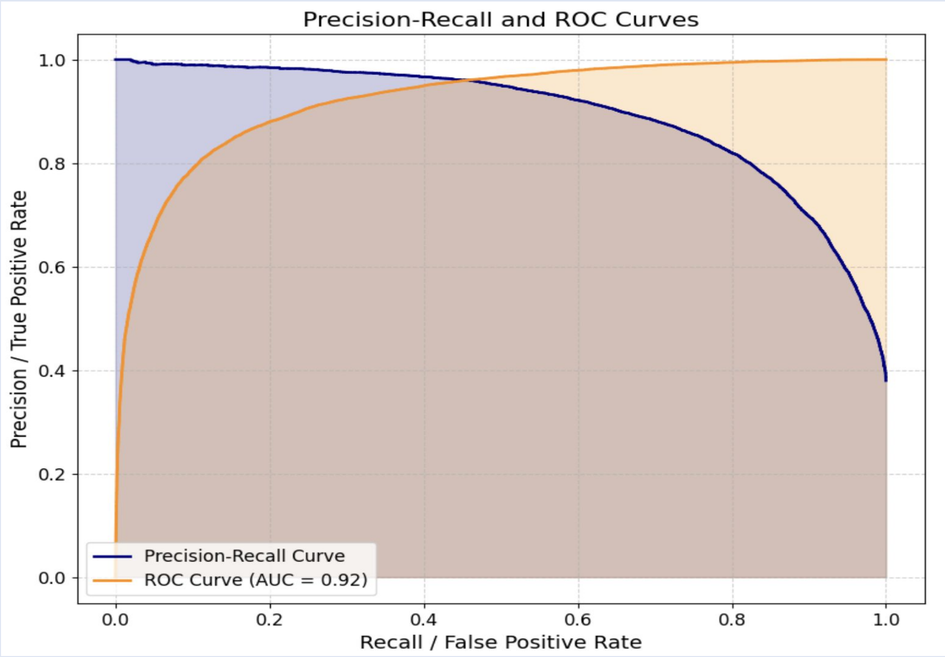
Results

SYSTEM MODELING

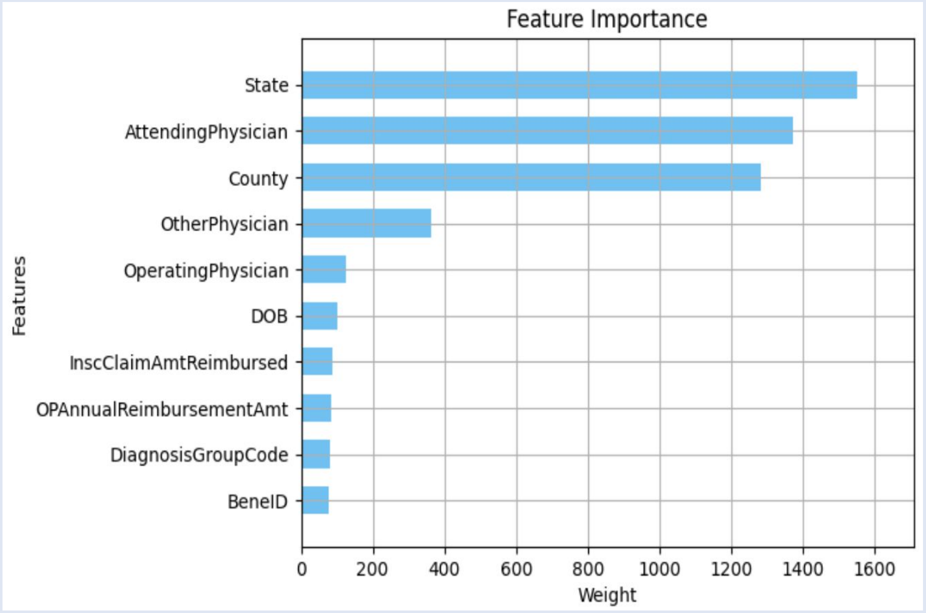
1. **Decision Tree-** The decision tree plot unveils interpretable decision logic, while accuracy and precision metrics quantify its performance. In healthcare fraud detection, the algorithm analyzes feature patterns, offering precise identification of key variables crucial for targeted prediction and mitigation.
2. **Random Forest** - It is an ensemble learning algorithm, merges numerous decision trees to generate predictions. In health fraud detection it assess feature importance, providing a robust evaluation of various variables crucial for identifying and predicting fraudulent activities within healthcare provider systems.
3. **LightGBM -** It is also similar to XGBoost algorithm but uses leaf-wise tree growth strategy. This approach can lead to a more accurate model with fewer leaves, improving efficiency and reducing the risk of overfitting.
4. **XGBoost, or Extreme Gradient Boosting** - It is an ensemble learning algorithm that works by combining the outputs of multiple weak learners. It enhances health fraud prediction by leveraging ensemble learning, regularization, effective handling of imbalanced data, and accurate feature importance assessment, enabling the model to detect subtle patterns indicative of fraudulent activities in insurance claims.

RESULTS & CONCLUSION

Best Model - XGBoost
85 % Accuracy



Class Labels	Class 0	Class 1
Class 0	31038	3537
Class 1	4398	16797



- Patient Patterns:Patients with high costs and numerous chronic conditions are potential fraud targets.
- Provider Insights:Fraudulent providers exhibit unique activity and geographical patterns.
- In conclusion, the thorough examination of healthcare data has revealed insights into potential fraud indicators associated with providers.

FUTURE WORK

- Explore the integration of external databases and open-source intelligence to enrich the dataset used for fraud detection.
- Develop a real-time monitoring system that continuously analyzes incoming data streams to provide instantaneous detection of suspicious activities.
- Investigate behavioral analysis techniques to identify anomalies in provider behavior that may indicate fraudulent practices not captured by traditional models.

REFERENCES

[1]https://nycdatascience.com/blog/student-works/machine-learning/healthcare-fraud-detection-2/

[2]https://www.kaggle.com/code/pavanpyla/fraud-detection-in-health-care/output

[3]https://jhjin.engin.umich.edu/wp-content/uploads/sites/248/2016/01/AP37_HCMS2008_A-survey-on-Statistical-Methods-for-health.pdf

[4] https://www.mdpi.com/2227-9091/11/9/160