Predicting ED Utilization Ratio using Machine Learning

Shiv Maharajh

3/18/2024

1 Introduction

In this report, we discuss a predictive modeling project undertaken for Jamaica Hospital, with the primary objective of predicting the Emergency Department Utilization Ratio (ED_Utilization_Ratio). This metric is critical for hospital administration and planning, as it directly impacts emergency department operations, resource allocation, and patient care quality.

2 Dataset Overview

The dataset utilized in this analysis was sourced from the Emergency Department Volume and Capacity dataset available on the Data.gov catalog (`https://catalog.data.gov/dataset/emergency-department-volume-and-capacity`). It encompasses various features related to hospital operations, geographical information, and emergency department metrics. Notably, it includes the total number of ED visits and the number of ED stations, allowing for the derivation of the ED_Utilization_Ratio, which serves as the target variable for our predictive models.

3 Feature Selection and Engineering

The initial dataset featured a broad range of variables, from which the following were selected or engineered based on their relevance and potential predictive power for the ED_Utilization_Ratio: Facility Name, County Name, System, LICENSED_BED_SIZE, Hospital Ownership, Urban/Rural Designation, Teaching Designation, Category, EDDX Count, Primary Care Shortage Area, Mental Health Shortage Area. These features encompass both categorical and numerical data types, providing a holistic view of each hospital's operational context.

A new feature, ED_Utilization_Ratio, was engineered by dividing the total number of ED visits by the number of ED stations to quantify each emergency department's capacity utilization.

4 Model Selection and Evaluation

Two regression models were chosen for this task: the XGBoost Regressor and the Random Forest Regressor. These models were selected due to their proven track record in handling both categorical and numerical data effectively, their capability to manage complex interactions between features, and their robustness in generalizing to new data.

4.1 Hyperparameter Tuning

Each model underwent a process of hyperparameter tuning, with at least three hyperparameters adjusted to optimize performance. The XGBoost Regressor was fine-tuned using GridSearchCV, focusing on n_estimators, learning_rate, and max_depth. Similarly, the Random Forest Regressor was optimized through a combination of manual selection and RandomizedSearchCV, adjusting n_estimators, max_features, and max_depth.

4.2 Evaluation Metrics

The primary metric used to evaluate model performance was the Root Mean Squared Logarithmic Error (RMSLE). This metric was chosen because it penalizes underestimations more than overestimations, which is particularly important in the context of emergency department utilization prediction. I have also calculated both R Squared and RMSE to help evaluate both models.

5 Recommendations and Future Work

Based on the performance metrics and the analysis conducted, we recommend Jamaica Hospital consider implementing the Random Forest or XGBoost models for predicting ED_Utilization_Ratio.

For future work, we propose further investigation into outliers and additional feature engineering to potentially uncover more nuanced relationships within the data.

6 Conclusion

While the current models provide a solid foundation for predicting ED utilization, continuous refinement and exploration of additional data sources could further enhance their accuracy and utility for Jamaica Hospital's operational planning and decision-making. From this study, I would choose the XGBoost Regressor has a similar R Squared to the Randomforest Regressor model but a lower RMSLE.