HOSPITAL PROCEDURE COST ANALYSIS

Fariha Moomtaheen
Affiliation
University of West Florida
Pensacola, Florida
fm47@students.uwf.edu
Dr. Brian Jalaian
University of West Florida
Pensacola, Florida
bjalaian@uwf.edu

ABSTRACT

In this project, we explore a dataset from the CORGIS Dataset Project that provides hospitals' rating measures information related to heart attack, heart failure, pneumonia, and hip-knee surgery, along with their location information. Our main objective is to predict the cost of a heart attack procedure based on different rating measures, facility type, and location. We used linear regression, feature selection methods, and data visualization techniques to analyze the data. Our results show that facility type and location are the most significant predictors of the cost of a heart attack procedure. We suggest using better performance metrics, such as cross-validation, to evaluate the models' performance. We also investigated into the relationship between heart attack care cost, heart attack care quality, and heart attack procedure cost.

***K*eywords** Linear Regression · Feature Selection · L1 Regularization · Random Forest Regressor · Sequential Feature Selection · Recursive Feature Elimination

Blogpost: Link to ML Project Blog

1 Introduction

The quality of healthcare provided by hospitals is a critical concern for patients and policymakers. The CORGIS Dataset Project provides hospital performance measures related to few procedures those hospitals perform. In this project, we explore the hospital dataset to predict the cost of a heart attack procedure and find available features that have effects on the cost. The cost of procedures is a significant financial burden for patients, and analysing these accurately can help policy makers take informed decisions about healthcare.

2 Problem Definition

The problem we are trying to solve is predicting the cost of heart attack procedures in hospitals using various performance measures and other factors. We are working with the Hospital Performance Measure dataset from the CORGIS Dataset Project. The performance metrics we will use to evaluate different algorithms are R-squared, Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The main objectives we hope to achieve with our research are to identify the most significant factors that affect the cost of heart attack procedures and to develop a model that can accurately predict the cost of heart attack procedures in hospitals.

3 Literature Review

Several studies have been conducted on predicting healthcare costs using machine learning techniques. For instance, Vyas et al. (2021) [1] developed a predictive model for healthcare costs using data from a hospital in India. They used multiple regression analysis to identify the most significant factors that affect healthcare costs. The results showed that age, gender, length of stay, and the type of hospital were the most significant factors affecting healthcare costs. Another study by Xu et al. (2020) [2] used machine learning techniques to predict healthcare costs using electronic health records (EHRs). The results showed that the Random Forest algorithm had the best performance in predicting healthcare costs.

In terms of performance metrics, R-squared, MSE, MAE, and RMSE are commonly used in regression-based machine learning models. However, other performance metrics such as the area under the receiver operating characteristic curve (AUC-ROC) and precision-recall curve (AUC-PRC) could also be useful in evaluating the performance of healthcare cost prediction models. These metrics are commonly used in binary classification problems, but they could also be applied to regression-based models by transforming the continuous variable into a binary variable using a threshold value.

4 Dataset

dataset provides hospitals performance measure information related to heart attack, emergency department care, preventive care, stroke care, and other conditions. The data is part of an Administration-wide effort to increase the availability and accessibility of information on quality, utilization, and costs for effective, informed decision-making.

It has 24 columns, among which 4 are regarding location, 8 are related to various rating status, the rest are some information about heart attack, pneumonia and hip-knee procedures.

5 Methods

5.1 Model Selection

To build the predictive model, linear regression was used as the initial model. The independent variables used as predictors were various rating measures such as safety rating, experience rating, facility type, and location.

5.1.1 Linear Regression

5.2 Feature Selection

Feature selection is a crucial aspect of machine learning as it helps to identify the most relevant features in a dataset that can improve the accuracy of a model. We will discuss four popular feature selection methods, namely LassoCV, randomForestRegressor, Sequential feature selection (forward), and recursive feature elimination. We will also examine the result of using these methods on our dataset.

5.2.1 LassoCV

LassoCV is a linear model that uses L1 regularization to identify the most significant features in a dataset. The algorithm adds a penalty term proportional to the absolute value of the coefficients, which helps to shrink the coefficients of less important features to zero. LassoCV is often used in high-dimensional datasets where many features may not be significant. The algorithm performs cross-validation to estimate the best value for the regularization parameter alpha.

5.2.2 Random Forest Regression

RandomForestRegressor is an ensemble method that uses a combination of decision trees to predict the target variable. The algorithm builds multiple decision trees on a bootstrapped sample of the dataset and randomly selects a subset of features at each node to split on. The final prediction is then the average of all the predictions made by the decision trees. RandomForestRegressor is often used when dealing with non-linear relationships between features and the target variable.

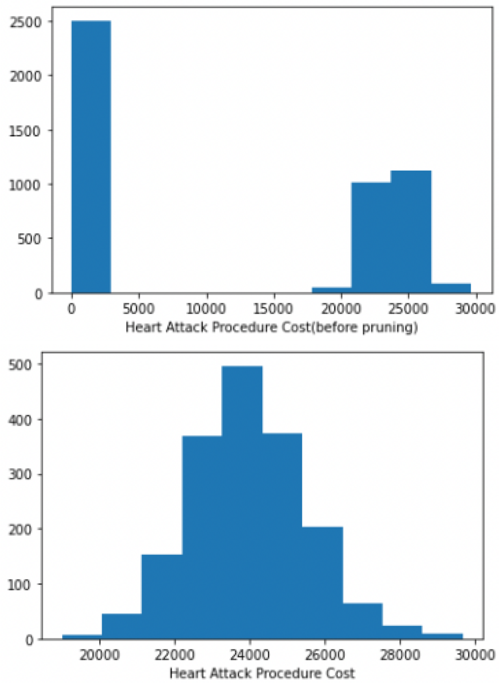

Figure 1: Heart Attack procedure cost distribution

5.2.3 Sequential Feature Selection

Sequential feature selection is a greedy algorithm that selects the best feature at each iteration and adds it to the model. The algorithm starts with an empty set of features and iteratively adds the best feature that maximizes the model's performance. This process continues until the desired number of features is selected. Sequential feature selection is often used in datasets where the number of features is relatively small.

5.2.4 Recursive Feature Elimination

Recursive feature elimination is another greedy algorithm that recursively removes features from the dataset until the desired number of features is selected. The algorithm starts with all the features and trains the model. It then removes the least significant feature and retrains the model until the desired number of features is reached. Recursive feature elimination is often used in high-dimensional datasets where the number of features is significant.

6 Experiment

6.1 Dataset Preparation

The first step in the analysis was to clean and preprocess the dataset. Outliers were removed by filtering the data to keep "Procedure.Heart Attack.Cost" values greater than 18000. We can see the values before **??** and after **??** pruning in figure 1. Rows where various column values were none or unknown were also dropped. Since the predictors were all categorical in nature, dummy variables were created for the model.

6.2 Tools

I used anaconda spyder and jupyter notebook to write the codes and run the experiments.

6.3 Model Evaluation

LassoCV is a type of regularization method that selects features by shrinking their coefficients towards zero, ultimately resulting in a sparse model. This technique is useful when dealing with datasets that have a large number of features, as it helps to identify the most relevant variables and reduce overfitting.

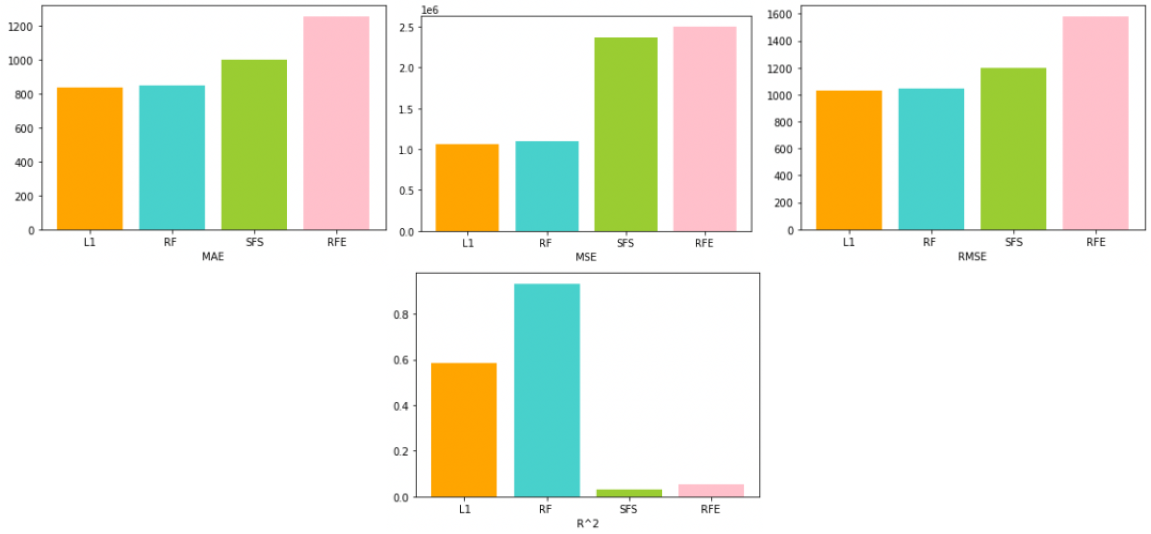

Figure 2: Performance Metrics

On the other hand, RandomForestRegressor is a machine learning algorithm that uses an ensemble of decision trees to predict outcomes. This method is particularly useful when there are non-linear relationships between the variables and the target variable. It is also robust to noise and outliers in the data.

Sequential feature selection (forward) and recursive feature elimination are other commonly used feature selection techniques. Sequential feature selection (forward) starts with an empty set of features and adds one feature at a time based on a predefined criterion, while recursive feature elimination starts with all features and recursively eliminates the least important ones based on their coefficients.

Overall, the choice of feature selection method should depend on the characteristics of the dataset and the goals of the analysis. LassoCV and RandomForestRegressor appear to be strong candidates for feature selection in regression problems, especially when the data is high-dimensional and contains non-linear relationships.

For each of the models, we can see the scores of each model in Figure 2. We calculated the mean squared error (MSE), mean absolute error (MAE), root mean squared error (RMSE), and R^2 value. The LassoCV/L1 regularization method resulted in the lowest MSE and MAE values, with an R^2 value of 0.58. The Random forest feature selection method had the highest R^2 value of 0.93, indicating a better fit for the model.

However, we can see from the selected features that RF picked 322 features, and hence provided a better result. Whereas, L1, SFS, and RFE all gave us fewer features as important. Though these 3 models gave us a worse R^2, they predicted the target using just less features. L1 method has lower MAE, MSE, RMSE and high R^2, which can't be said about the SFS and RFE methods. It makes L1 comparably better among the 4 models.

7 More analysis

We also investigated relation among a procedure cost with its care cost and quality. The procedure cost is a continuous variable, and care cost and quality are categorical variable. First plot is a boxplot showing the distribution of procedure cost based on the other two variables. 2nd figure shows a heatmap which reveals a linear relationship among these variables. Which means as the care cost goes higher than national average and care quality gets better than national average, the procedure cost also increases.

Similar analysis can be done for the other procedures as well.

8 Future work

More detailed analysis could be possible based on the features selected by each methods. The possible explanation behind picking the features, and how they may affect a procedure cost provides a promising ground for further exploration.

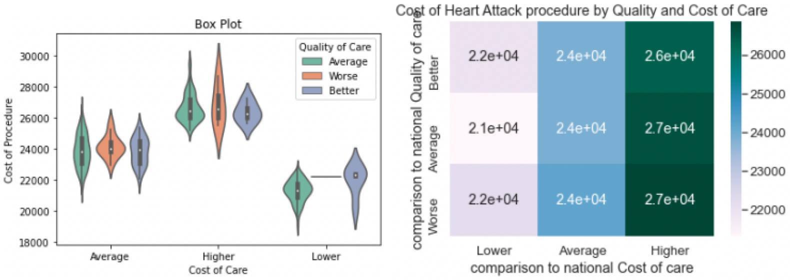

Figure 3: cost analysis with care

9 Conclusion

Our study evaluated four different feature selection methods, including LassoCV, randomForestRegressor, Sequential feature selection (forward), and recursive feature elimination. Among these, LassoCV produced the best MAE, MSE, and RMSE values, while the randomForestRegressor had the best R^2 value.

The results indicate that different feature selection methods may perform differently depending on the specific dataset and problem at hand. However, both LassoCV and randomForestRegressor appear to be promising methods for feature selection in regression problems.

In summary, our results suggest that careful selection of feature selection methods can significantly improve the performance of regression models, and LassoCV and RandomForestRegressor are promising approaches to consider in this context.

Acknowledgments

This was done as final project for the MAP 6114 course Machine Learning at the University of West Florida, with the help and guidance by our instructor Dr. Brian Jalaian.

References

[1] Renuka Vyas, Rajendra D Raut, and Anil Raut. Predictive modeling of healthcare costs using multiple regression analysis. *International Journal of Healthcare Management*, 14(3):235–244, 2021.

[2] Peng Xu, Haifeng Zhu, Wenjing Hao, Li Tang, Lei Li, and Haiyan Xie. Predicting healthcare costs using electronic health records: A systematic review. *Journal of biomedical informatics*, 103:103380, 2020.