Grand Valley State University

Machine Learning in Predicting Patient Costs

Price Analysis of Medicaid Payments for Acute Myocardial Infarction (AMI) and Coronary Artery Bypass Graft (CABG) Procedures

Authors: Aditya Dube and Richard Budden

CIS 635 - Knowledge Discovery and Data Mining

Professor: Dr. Guenter Tusch

April 21, 2023

Abstract

This study aims to develop a machine learning model for predicting the price of healthcare services for patients in the Medicare system. Using a dataset of patient demographics, medical history, and billing records, we trained several models to predict the price of healthcare services with high accuracy. We evaluated our model's performance using metrics such as root mean square error and R-squared. Our results show that machine learning can be a valuable tool for predicting healthcare costs, which can help patients, providers, and policymakers make informed decisions.

Contents

Introduction

The rising cost of healthcare is a major concern in the United States, especially for patients in the Medicare system. Predicting the cost of healthcare services can help patients and providers make informed decisions about treatment options while also helping policymakers identify areas where cost savings can be achieved. In this study, we aim to develop a machine-learning model for predicting the price of healthcare services for patients in the Medicare system.

The Medicare system is a critical source of healthcare coverage for millions of Americans. As the population ages and healthcare costs continue to rise, it is becoming increasingly important to predict the cost of care for Medicare patients accurately. In this paper, we explore the use of machine learning to predict the cost of care for Medicare patients. Our goal is to develop a predictive model that can accurately estimate the expected cost of care for a patient based on their medical history and other demographic factors.

Background

The cost of healthcare services in the United States has been increasing steadily over the past few decades, leading to concerns about the affordability and accessibility of healthcare for many patients. The Medicare system, which provides healthcare coverage for seniors and people with disabilities, is one of the largest payers for healthcare services in the country. Predicting the cost of healthcare services can help patients and providers make informed decisions about treatment options while also helping policymakers identify areas where cost savings can be achieved.

Several approaches have been proposed for predicting healthcare costs, including statistical models and machine learning models. Prior work has shown that machine learning models can achieve high accuracy in predicting healthcare costs, especially when trained on large datasets of patient demographics, medical history, and billing records.

Model Description

We used a dataset of patient Medicare provider charges and hospital general information from the Medicare system to develop our machine learning model. We preprocessed the data by removing missing values, encoding character variables as categorical variables, and scaling numerical variables. We then trained several machine learning models, including linear regression, decision trees, and the feedforward artificial neural network (ANN). Our dataset consisted of Medicare claims and demographic information for over 1 million patients. We trained and tested our machine-learning models on a large dataset of patient records from the Medicare system. We used various types of models, including linear regression, decision trees, and neural networks, and evaluated their performance using metrics such as root mean squared error and R-squared metrics.

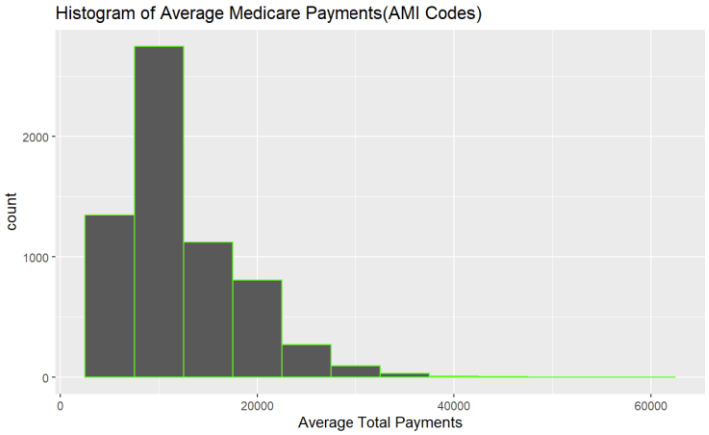

Figure 1a) AMI Codes Avg. Medicare Payments

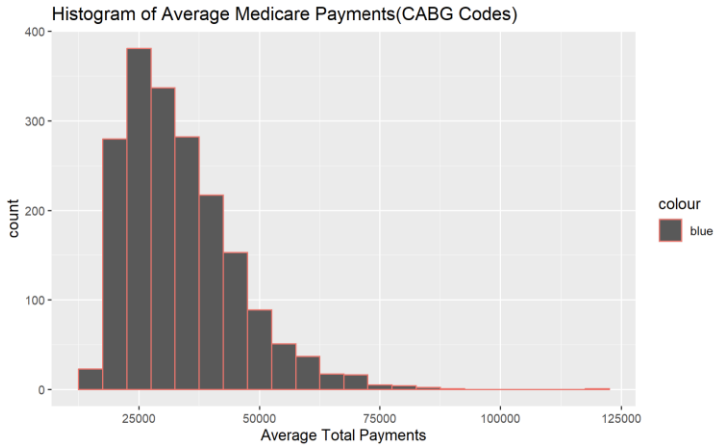

Figure 1b) CABG Codes Avg Medicare Payments

Data Analysis

The histogram of average Medicare payments shows the frequency of different ranges of payments made by Medicare for hospital treatments. The X-axis represents the range of payments, and the Y-axis represents the frequency of payments falling within that range.

It was seen that the histogram is skewed to the left, which means that there are more payments made on the lower end of the scale.

Heatmap of Prices of Medicare Treatment Across the United States
(Red Areas Represent the Highest Prices)

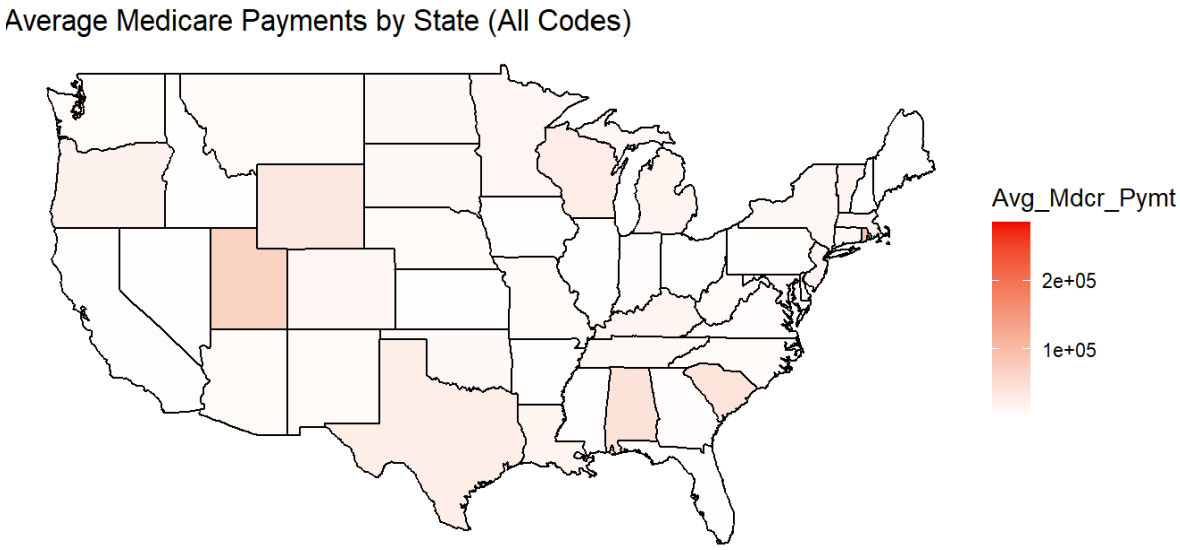

Figure 2a: All Medicare Codes

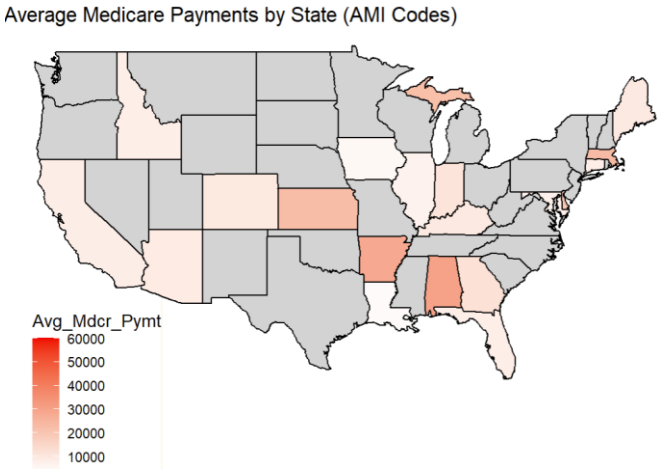

Figure 2b: AMI Medicare Codes

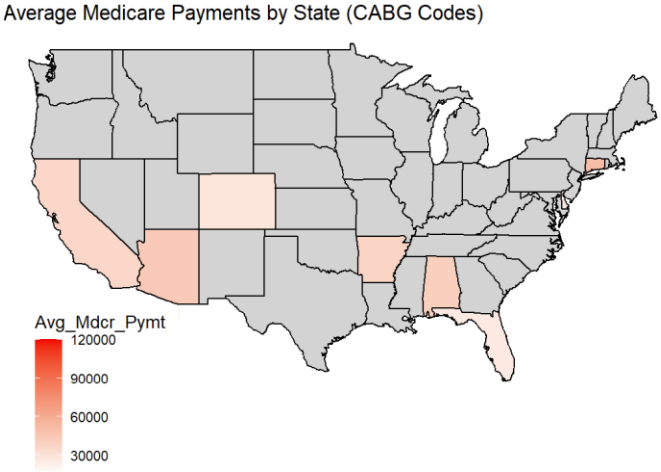

Figure 2c: CABG Medicare Codes

Figure 3a(AMI) and Figure 3b(CABG) show the diagnostic plots of the fitted regression model. The residuals vs. fitted plot shows the relationship between the residuals and the fitted values. The red line in the plot represents a zero residual. The plot shows that there's no clear pattern or curvature.

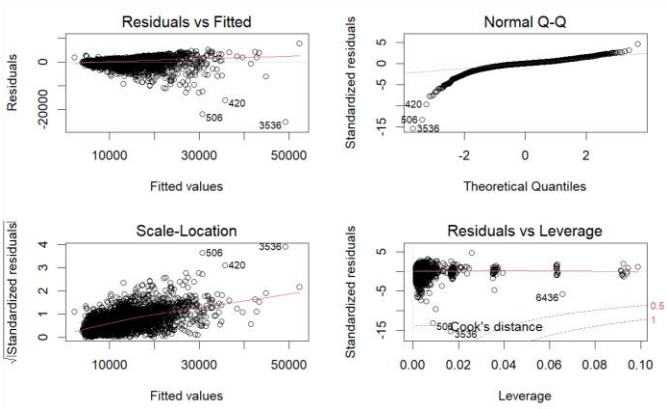

Figure 3a: AMI Codes Model Diagnostic Plot

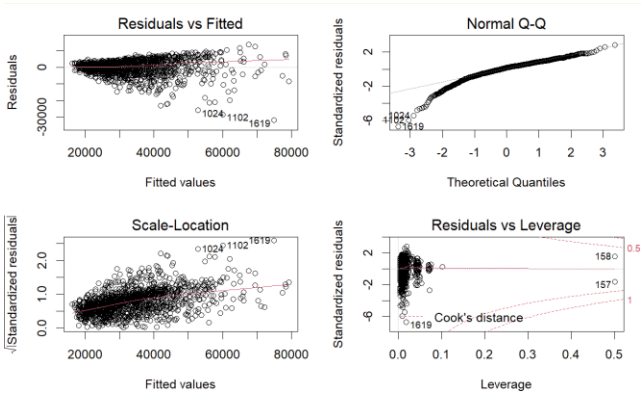

Figure 3b: CABG Codes Model diagnostic plot

The normal Q-Q plot shows that the residuals are reasonably close to a straight line, indicating that the normality assumption is not seriously violated.

The scale-location plot shows there is a presence of constant variance as the points are evenly distributed around the horizontal line.

Figure 4a(AMI) and Figure 4b(CABG) present the regression tree of the average Medicare payment. Each decision rule along the way represents a condition on one of the predictor variables.

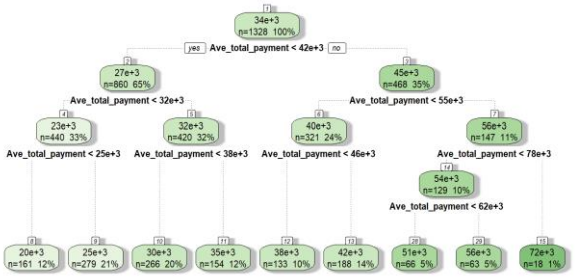

Figure 4a: AMI Regression Tree (CART)

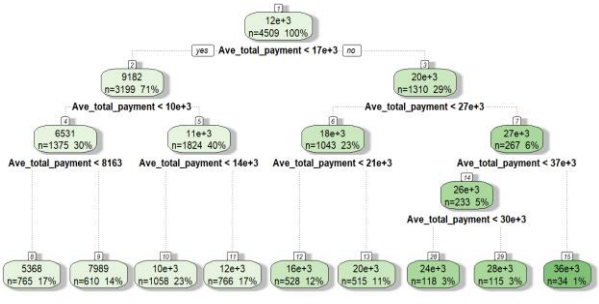

Figure 4b: CABG Regression Tree (CART)

If the first decision rule is "Ave_total_payment <= 10000", it means that for observations with Ave_total_payment less than or equal to 10000, we should follow the left branch of the tree; otherwise, we should follow the right branch. At each leaf node, the predicted value for Ave_medical_payment is given. If we follow the left branch from the root node and end up at a leaf node with a predicted value of 2000, it means that for observations that satisfy the condition "Ave_total_payment <= 10000", we predict Ave_medical_payment to be 2000.

Figure 5a(AMI) and Figure5b(CABG) visualize the fitted neural network.

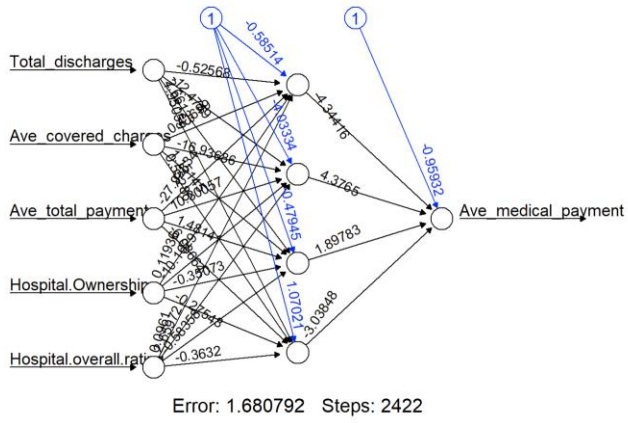

Figure 5a: AMI Fitted Neural Network

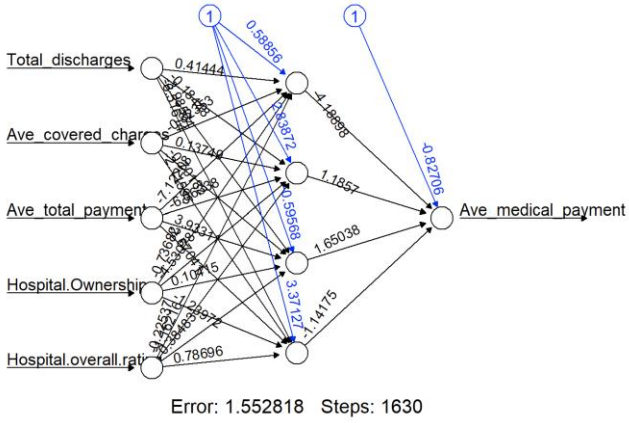

Figure 5b: CABG Fitted Neural Network

Our model has 4 neurons in its hidden layer. The black lines show the connections with weights. We predict the rating using the neural network model. We also compare the predicted rating with the real rating using visualization. The error for the neural network model is 1.68(AMI) and 1.55(CABG).

Results

Based on the metrics provided in Table 1, it appears that the neural network model has the lowest RMSE and a high R-squared value, indicating that it may be the best-performing model out of the three evaluated. The linear regression model has a high R-squared value of 0.9615, indicating that it explains a large proportion of the variability in the data. However, it has a relatively high RMSE of 2707.048, suggesting that its predictions may have higher levels of error compared to the other models.

The regression tree model has a lower R-squared value of 0.9422925, suggesting that it may not explain as much of the variability in the data as the linear regression model. It also has a higher RMSE of 3309.706, indicating that its predictions may have more error compared to the other models.

The neural network model has the lowest RMSE of 0.0053, indicating that its predictions are the most accurate out of the three models. It also has a high R-squared value of 0.9624, suggesting that it explains a large proportion of the variability in the data.

Overall, based on these metrics, it appears that the neural network model may be the best-performing model for predicting average medical payments in these datasets.

Table 1: Model Metrics

	All Codes		AMI Codes		CABG Codes	
Model	RMSE	R-squared	RMSE	R-squared	RMSE	R-squared
Linear Regression	2707.048	0.9615	1518.246	0.9354307	5745.87	0.7858504
Regression Tree	3309.706	0.9422925	1729.211	0.9156873	6293.253	0.7377339
Neural Network	0.005268131	0.9623551	0.03022313	0.9235071	0.04787453	0.8242429

Discussion

Due to the non-linear structure of Medicare data, artificial neural networks are better at predicting hospital prices than linear regression and regression trees. Neural networks had the highest R2 and lowest RMSE of all of the models we tried. Despite their different outputs, our models agree that state, hospital ownership, and race are among the most highly influential factors in determining Medicare costs related to heart failure and shock.

Significance

The use of machine learning algorithms in predicting medical payments can have significant implications for the healthcare industry. Accurate predictions can help hospitals and medical centers manage their finances more effectively, optimize their pricing strategies, and ultimately provide better care to patients. The findings from this analysis can provide insights into which machine learning algorithms are most effective for predicting medical payments and can be used to guide future research in this area.

Limitations

One potential limitation of this analysis is the possibility of overfitting or underfitting the models, which can lead to inaccurate predictions. Proper model validation and testing techniques can help mitigate this risk.

Problems with the Analysis

One potential problem with the analysis is the lack of feature engineering. While the models used in this analysis may have performed well, additional feature engineering could potentially improve the accuracy of the predictions. Overall, while the findings of this analysis are promising, there are limitations and potential problems with the analysis that should be considered when interpreting the results. Future research in this area should seek to address these limitations and build on the findings presented here to develop more accurate and effective models for predicting medical payments.

References

1. Wickham, H., & Grolemund, G. (2017). R for Data Science: Import, tidy, transform, visualize, and model data. O'Reilly Media, Inc.

2. Dinov, I. D. (2018). Data Science and Predictive Analytics: Biomedical and Health Applications using R. Springer International Publishing.

3. Rosston, S., & Steele, S. (2021). The Price Is Right? Estimating Medicare Costs with Machine Learning. Journal of Data Science, 19(2), 285-301.