Predictive Analytics for Hospital Costs

Cameron Ervin

Executive Summary

This project leverages 2015 de-identified New York inpatient discharge data to predict hospital patients' healthcare costs and charges using predictive analytics. The analytics were conducted using R and Tableau, focusing on linear regression, regression trees, and clustering to predict Total Charges as the dependent variable. The methodology includes data loading, feature engineering, exploratory data analysis, model creation and fine-tuning, and experimentation with sample sizes. The best-performing linear regression model used a 10% sample size of the original dataset, achieving an R^2 value of 0.7064, indicating a strong ability to predict Total Charges based on the feature set used. Also, my regression tree and hierarchical clustering model findings indicate that the most effective regression tree models utilize a mini bucket size of 10 and criteria of Gini or information gain and accounted for approximately 72.48% of the variability in total charges. The hierarchical clustering identified four distinct patient clusters, offering insights into varying lengths of stay, total charges, and other demographic and health-related characteristics.

Furthermore, this research provides a comprehensive analysis of healthcare costs using advanced statistical methods. Understanding the factors of hospital charges can help in developing strategies for cost management and in designing targeted interventions for patient care. The application of these analytical models demonstrates the potential of predictive analytics in enhancing healthcare.

Table of Contents

[Figures 4](#_Toc159787305)

[Tables 7](#_Toc159787306)

[Introduction 12](#_Toc159787307)

[Methods 13](#_Toc159787308)

[Analysis & Findings 14](#_Toc159787309)

[Summary 14](#_Toc159787310)

[Figures 16](#_Toc159787311)

[Tables 18](#_Toc159787312)

[Conclusion 21](#_Toc159787313)

[References 22](#_Toc159787314)

# Figures



Figure 1.

A graph with a line and a line

Description automatically generated

Figure 2.

A graph of various numbers and colors

Description automatically generated with medium confidence A graph of a number of costs

Description automatically generated

Figure 3a. Figure 3b.

A graph showing a number of different colored lines

Description automatically generated with medium confidence A graph showing a number of costs

Description automatically generated

Figure 3c. Figure 3d.

A graph of different colored lines

Description automatically generated with medium confidence A graph of different colored lines

Description automatically generated with medium confidence

Figure 3e. Figure 3f.

A graph showing different colored shapes

Description automatically generated A graph showing different colored columns

Description automatically generated with medium confidence

Figure 3g. Figure 3h.

A graph showing different colored shapes

Description automatically generated with medium confidence A chart of different colored shapes

Description automatically generated with medium confidence

Figure 3i. Figure 3j.

A graph of a number of individuals

Description automatically generated with medium confidence

Figure 4.

A red line graph with numbers

Description automatically generated

Figure 5.

A diagram of a number of data

Description automatically generated with medium confidence

Figure 6.

A diagram of a number of data

Description automatically generated with medium confidence

Figure 7.

A graph of a number of clusters

Description automatically generated

Figure 8.

A diagram of a cluster

Description automatically generated

Figure 9.

# Tables

Residuals:

Min 1Q Median 3Q Max

-256067 -10067 -445 7814 382919

Coefficients: (1 not defined because of singularities)

Estimate Std. Error

(Intercept) 40585.97 11681.67

`Health Service Area`Central NY -5090.02 9689.46

`Health Service Area`Finger Lakes 3729.19 11337.27

`Health Service Area`Hudson Valley 38664.03 9825.40

…

`Payment Typology 3`Medicaid -1.113 0.265607

`Payment Typology 3`Other -1.142 0.253569

`Payment Typology 3`Private Health Insurance 2.545 0.010969 \*

`Payment Typology 3`Self-Pay 4.094 4.34e-05 \*\*\*

`Emergency Department Indicator`Y 2.221 0.026419 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24560 on 3594 degrees of freedom

Multiple R-squared: 0.7066, Adjusted R-squared: 0.7

F-statistic: 106.9 on 81 and 3594 DF, p-value: < 2.2e-16

[1] "Model"

[1] "R^2:"

[1] 0.7116446

Table 1.

Length of Stay Other Provider License Number Total Charges Total Costs

Length of Stay 1.00000000 -0.01386228 0.72464142 0.79164058

Other Provider License Number -0.01386228 1.00000000 -0.05799892 0.06946813

Total Charges 0.72464142 -0.05799892 1.00000000 0.82615443

Total Costs 0.79164058 0.06946813 0.82615443 1.00000000

Table 2.

[1][1] "HospitalTree50 R^2:" "HospitalTree50 R^2:"

[1][1] 0.5160279 0.5160279

[1][1] "HospitalTree20 R^2:" "HospitalTree20 R^2:"

[1][1] 0.7106319 0.7106319

[1][1] "HospitalTree10 R^2:" "HospitalTree10 R^2:"

[1][1] 0.7248439 0.7248439

[1][1] "HospitalTree5 R^2:" "HospitalTree5 R^2:"

[1][1] 0.6972827 0.6972827

[1][1] "HospitalTree50b R^2:" "HospitalTree50b R^2:"

[1][1] 0.5160279 0.5160279

[1][1] "HospitalTree20b R^2:" "HospitalTree20b R^2:"

[1][1] 0.7106319 0.7106319

[1][1] "HospitalTree10b R^2:" "HospitalTree10b R^2:"

[1][1] 0.7248439 0.7248439

[1][1] "HospitalTree5b R^2:" "HospitalTree5b R^2:"

[1][1] 0.6972827 0.6972827

Table 3.

clusterGroups

1 2 3 4

348 431 404 391

1 2 3 4

2.801724 5.993039 4.373762 2.664962

1 2 3 4

26538.909 29096.519 22318.528 7914.031

1 2 3 4

1.485632 1.030162 2.000000 1.000000

1 2 3 4

1.000000 1.951276 1.992574 1.000000

1 2 3 4

3.724138 3.081206 3.891089 1.749361

1 2 3 4

9.008621 8.389791 8.549505 9.864450

Table 4.

# Introduction

The goal of this project is to leverage the 2015 de-identified New York inpatient discharge data to predict the health care costs and charges of hospital patients. The analytics for this project are performed in R and Tableau. Additionally, the modeling used for this project was linear regression in R. Our aim was to maximize our R^2 value while using our feature space to predict our dependent variable Total Charges.

Predictive analytics largely impacts the field of health care. According to “What Are the Benefits of Predictive Analytics in Healthcare?” by Xtelligent Healthcare Media, predictive analytics supports clinical decision making and advances value-based care. Risk scoring is a large component of clinical decision making. The impact of making the right decision can cost human life and money. Machine learning and data science models can help improve the accuracy of clinical decision making. Furthermore, the article states, “[health care providers] gather data on its member populations to help predict healthcare needs contributing to adverse outcomes, which the payer then uses to connect members with necessary services.” The ability to predict adverse health conditions allows hospitals and other healthcare providers to more quickly and cost effectively attend to the needs of their patients.

As for the use case of my model, according to “Deep learning for prediction of population health costs” by Drewe-Boss et al., “Deep learning models have also shown promise in predicting healthcare costs, with methods evaluated on their ability to predict the summed cost per patient. These models can identify patients whose health status or treatment needs may change, thereby benefiting from preventive interventions.” Predictive analytics can be used as a tool to more accurately predict or forecast healthcare costs for patients, and my goal is to do just that.

Lastly, there were two potential dependent variables for the model, Total Costs and Total Charges. I used Total Charges as my dependent variable and dropped Total Costs. To explain the difference between the two – total costs are the costs incurred by healthcare providers when delivering services to patients, and total charges is the amount billed to patients for the services provided. Often, there is a large difference in total costs and total charges with total charges being much greater. This is to account for other costs the healthcare provider must cover as well as their profit margin.

# Methods

All code can be found in the attached R file. To summarize my methods at a high level:

1. Load in and read data
2. Feature Engineering
3. Exploratory Data Analysis
4. Model creation
   1. Linear regression
   2. Regression trees
   3. Clustering
5. Further feature engineering
6. Model fine tuning
7. Experimentation with sample size

# Analysis & Findings

## Summary

After performing the methods listed above, the best linear regression model was the fine-tuned model using a sample of 1/10th of our original dataset. The R^2 values of the different models are as follows:

|  |  |
| --- | --- |
| Sample Size | R-Squared |
| 10% | 0.7064 |
| 8% | 0.6751 |
| 7.5% | 0.6947 |
| 6% | 0.6654 |
| 5% | 0.6391 |

Although there is slight variation in the R^2 value based on sample size, the best performing model had the largest sample size. Also, one could argue that the R^2 values of 10% and 7.5% are clearly not significantly different statistically. However, there does appear to be a trend, as the sample size goes up, the R^2 value also goes up.

Additionally, the best regression tree models used a mini bucket of 10 and a criterion of either Gini or information gain. Regression trees are a form of decision tree that is used for predicting continuous outcomes. In this context, the outcome is "Total Charges," which represents the total costs associated with a patient's hospital stay. Decision trees work by splitting the data into subsets based on the value of input variables, attempting to isolate groups with distinct outcomes. Notably, the best performing regression tree model achieved an R-squared (R^2) value of 0.7248439, indicating that approximately 72.48% of the variability in Total Charges could be explained by the model. An R^2 value closer to 1 indicates a model with a better fit. To find the optimal mini bucket size and evaluation criterion, I manually tuned each of these parameters by running different trees at varying mini bucket sizes with both Gini and information gain as the criterion. The best two trees can be seen in Figure 6 and Figure 7. Furthermore, for a deeper analysis of the important features of these regression trees, please see the descriptions of Figure 6 and Figure 7 down below.

As for the hierarchical clusters, hierarchical clustering groups the data into distinct clusters. Hierarchical clustering creates a tree-like structure of the data, known as a dendrogram, which shows how each cluster is composed by merging smaller clusters based on their similarity. This technique can reveal patterns or relationships in the data that might not be immediately apparent, such as grouping patients with similar cost drivers or healthcare needs.

Using the WSS method, I found that the optimal number of clusters was 4 since that was the elbow of the WSS graph (Figure 8). To describe the WSS method in more detail, the WSS method is a heuristic used in cluster analysis, particularly in determining the optimal number of clusters in k-means clustering. The technique involves running the clustering algorithm multiple times with a different number of clusters (k) and calculating the within-cluster sum of squares (WCSS) for each k. WCSS measures the compactness of the clusters, and ideally, it decreases as k increases. The method gets its name from the shape of the plot of WCSS as a function of k, which often resembles an arm's elbow. The "elbow" point, where the rate of decrease sharply changes, suggests a suitable number of clusters for the data. In this case, the elbow point was at k = 4, making it the optimal point. This point is considered optimal because adding more clusters beyond this point does not provide significant improvement in the compactness of the clusters, indicating a diminishing return on the cost of complexity (Kassambara). For a greater explanation of the 4 clusters, please see the description of Table 4 down below.

## Figures

Figure 1 – An illustration of total charges by New York State County created in Tableau. Illustrates how the majority spend comes from Manhattan Island and the surrounding area.

Figure 2 – A scatterplot showing the correlation between Total Charges and Total Costs. Total Charges appears to increase at three times the rate of total costs according to our line of best fit. Additionally, Total Charges and Total Costs have the highest correlation in our correlation matrix (see Table 2).

Figures 3a to 3b – Violin boxplots of various categorical variables based on Total Charges (left) and Total Costs (right). One note is the y-axes of the graphs. The values on the y-axes on the Total Charges graphs are much larger than the values on the y-axes on the Total Costs graphs. These categorical variables were chosen because after initial plotting, these variables had the least amount of noise in their graphs. Additionally, outliers have been dropped from these graphs and the graphs have been rescaled. Lastly, it appears most of the distributions of Total Charges and Total Costs have a lower median but are skewed right. They are most likely pulled by outliers and patients with high charges or high costs.

Figure 4 – Paralleled histograms for Length of Stay, Other Provider License Number, Total Charges, and Total Costs. Outliers impact the scaling of the x-axes. After engineering our features, these were the only remaining numeric features.

Figure 5 – Residual plot for our best performing model showing the residuals of y-ypred around zero.

Figure 6 and Figure 7 – Figure 6 and Figure 7 are identical. The two figures are the plots of regression trees created in R using the cleaned hospital data. Both trees use a mini bucket size of 10. However, Figure 6 uses the Gini Criterion, and Figure 7 uses the information gain criterion. The mini bucket size is a parameter that controls the number of leaf nodes in a tree. As the mini bucket decreases, the number of leaf nodes grow exponentially. Based on the regression tree plot, the most important features of the tree are `Length of Stay` since it is the root node and appears twice later in the tree, `Facility Id` since it appears twice as a splitting/leaf node, and `APR MDC Code` since it appears twice as a leaf node.

Figure 8 – WSS graph displaying the number of clusters 1-10. Based on the graph, the optimal number of clusters is 4 since this is the elbow point of the graph. Four clusters balance the return and compactness of the clusters.

Figure 9 – Cluster dendrogram of the hierarchical clustering performed using k = 4 clusters. A cluster dendrogram is a tree-like diagram that displays the arrangements of the clusters produced by hierarchical clustering. The branches of the dendrogram represent clusters, which are grouped together based on their similarity. Each branch can be split into smaller branches, which represent smaller clusters formed at a higher similarity level. The height of the branches reflects the distance or dissimilarity between clusters: the higher the branch, the greater the dissimilarity. In our case, the cluster dendrogram is hard to read which is why I created Table 2 which helps illustrate the makeup of each of the clusters in a more viewer-friendly manner (Bock).

## Tables

Table 1 – An abbreviated output from the best performing model. This model samples 1/10th of the original dataset to reduce time and space complexity. Additionally, this model’s R-Squared value is 0.7116.

Table 2 – A correlation matrix between our numeric features. There seems to be a correlation between Length of Stay and Total Charges, Length of Stay and Total Costs, and Total Charges and Total Costs.

Table 3 – The R^2 output from each of our regression tree models. Again, HospitalTree10 and HospitalTree10b tied for the highest R^2 value of 0.7248439.

Table 4 – Table 2 shows the output of the 4 hierarchical clusters created by our model. Four distinct clusters are created by our model. Here is a deeper analysis of each of the clusters:

Cluster 1

Size: 348 observations

Length of Stay: On average, patients in this cluster have a relatively shorter length of stay in the hospital, averaging 2.8 days.

Total Charges: This cluster has an average total charge of approximately $26,539, which suggests moderate hospitalization costs.

Emergency Department Indicator: The mean value is 1.49, indicating a mix of visits and non-visits to the emergency department.

Severity of Illness: The average severity of illness code is 1.0, indicating that patients in this cluster generally have a lower severity of illness.

Age Group: The mean age group is 3.72, suggesting that patients in this cluster tend to be in the older age groups.

Geographic Location (Zip Code): The average is around 9, indicating some geographic concentration, but the significance of this would depend on the specific zip code distribution.

Cluster 2

Size: 431 observations

Length of Stay: Patients have a longer length of stay compared to Cluster 1, with an average of about 6 days.

Total Charges: The average total charges are around $29,097, which are higher than those in Cluster 1, reflecting longer hospital stays or more intensive care.

Emergency Department Indicator: The average of 1.03 indicates that most patients in this cluster likely did not visit the emergency department.

Severity of Illness: With an average code of 1.95, patients in this cluster have a moderate to high severity of illness.

Age Group: The average age group is lower than in Cluster 1, at 3.08, indicating a younger demographic.

Geographic Location (Zip Code): The average zip code value is slightly lower than in Cluster 1, at around 8.39.

Cluster 3

Size: 404 observations

Length of Stay: This cluster has an average length of stay of 4.37 days, which is between the values of Clusters 1 and 2.

Total Charges: Average total charges are the lowest among all clusters, at about $22,319, which could reflect less intensive or shorter treatments.

Emergency Department Indicator: The average value is exactly 2.0, indicating that all patients in this cluster visited the emergency department.

Severity of Illness: The severity of illness is high with an average code of 1.99, similar to Cluster 2.

Age Group: The mean age group is 3.89, indicating that this cluster also tends to include older patients.

Geographic Location (Zip Code): Similar to Cluster 1, with an average around 8.55.

Cluster 4

Size: 391 observations

Length of Stay: Similar to Cluster 1, with a short average length of stay of 2.66 days.

Total Charges: This cluster has the lowest average total charges, at $7,914, suggesting less intensive or shorter hospital treatments.

Emergency Department Indicator: The mean value is 1.0, suggesting that most patients did not visit the emergency department.

Severity of Illness: The average severity of illness code is 1.0, indicating lower severity conditions similar to Cluster 1.

Age Group: The average age group is significantly lower at 1.75, suggesting that this cluster predominantly consists of younger patients.

Geographic Location (Zip Code): The average zip code value is the highest at around 9.86, possibly indicating a different geographic area compared to the other clusters.

# Conclusion

After created, tuning, and testing our models, the top performing linear regression model sampled 10% of the original dataset and scored an R^2 of 0.7064, the regression trees scored an R^2 metric of 0.7248439, and the hierarchical clustering of the hospital discharge data using k = 4 clusters created 4 clustered groups within the dataset. The successful application of regression tree models and clustering could provide valuable insights for healthcare providers and policymakers. For example, understanding the factors that contribute to higher hospital charges can inform strategies to manage costs more effectively. Similarly, identifying distinct groups of patients based on their healthcare spending can help in designing targeted interventions or personalized care plans to improve outcomes and reduce unnecessary expenses.

# References

Bock, Tim. "What is a Dendrogram? - Hierarchical Cluster Analysis." Displayr, www.displayr.com/what-is-dendrogram/.

Drewe-Boss, Philipp, Dirk Enders, Jochen Walker, and Uwe Ohler. “Deep Learning for Prediction of Population Health Costs - BMC Medical Informatics and Decision Making.” BioMed Central, May 16, 2022. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-021-01743-z.

Kassambara, Alboukadel. "Determining The Optimal Number Of Clusters: 3 Must Know Methods." Datanovia, 2021, www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/.

Staff, Editorial. “What Are the Benefits of Predictive Analytics in Healthcare?” HealthITAnalytics, February 6, 2024. https://healthitanalytics.com/news/what-are-the-benefits-of-predictive-analytics-in-healthcare.