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 modeling to predict Total Charges as the dependent variable. Predictive analytics in healthcare supports clinical decision-making and value-based care, with risk scoring playing a crucial role. The model aims to accurately predict or forecast healthcare costs, improving clinical decision accuracy, and facilitating quicker, more cost-effective patient care. The methodology includes data loading, feature engineering, exploratory data analysis, model creation and fine-tuning, and experimentation with sample sizes. The best-performing 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.

Table of Contents

[Figures 4](#_Toc159787305)

[Tables 7](#_Toc159787306)

[Introduction 9](#_Toc159787307)

[Methods 10](#_Toc159787308)

[Analysis & Findings 11](#_Toc159787309)

[Summary 11](#_Toc159787310)

[Figures 11](#_Toc159787311)

[Tables 12](#_Toc159787312)

[Conclusion 13](#_Toc159787313)

[References 14](#_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.

# 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.

# 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
5. Further feature engineering
6. Model fine tuning
7. Experimentation with sample size

# Analysis & Findings

## Summary

After performing the methods listed above, the best 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.

## 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.

## 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.

# Conclusion

In summary, this investigation highlights the effectiveness of predictive analytics within healthcare, particularly in predicting hospital-related costs. By employing thorough analytical methods, the model demonstrated accuracy in forecasting Total Charges, emphasizing the importance in improving predictive analytics for health care costs.

# References

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.

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.