Cost Analysis and Forecasting for Hospital Financial Performance

Charlotte Xu, Ruiyang Dong, Xuyuan Zhang, Zihan Wang (Contributions are in Github ReadMe)

17 December, 2024

Github Link: https://github.com/sergiozxy/BIOSTAT625-Project

Abstract

The hospital financial report can provide valuable insights into hospital operations. With the advent of nationwide healthcare report databases, we are able to explore more aspects of the data utilizing a series of modern data mining techniques. This project analyzes and forecasts hospital financial performance using the CMS Hospital Provider Cost Report dataset. By integrating various statistical and machine learning models including geographically weighted regression, random forest, k-nearest neighbors and interactive visualization tools via R Shiny, this project aims to provide actionable insights into hospital operating efficiency and revenue generation. The analysis spans from 2011 to 2022, focusing on critical metrics such as the Cost-to-Revenue Ratio and Revenue per Bed. Parallelization is applied for some of the model training to mitigate the computational burden.

Introduction

The financial health of hospitals is a critical concern in the U.S. healthcare system, influencing both patient care quality and the operational sustainability of institutions. This report aims to analyze historical operating costs and revenue trends of hospitals using the CMS Hospital Provider Cost Report dataset. The project employs statistical models and machine learning tools to guide hospital administrators in making informed decisions, optimizing budgets, and improving resource allocation efficiency.

The study focuses on two key dependent variables: Cost-to-Revenue Ratio, which measures operating efficiency by comparing operating costs to total revenue, and Revenue per Bed, which evaluates revenue generation relative to hospital capacity. These metrics provide a comprehensive view of hospital financial performance and allow for deeper insights into factors that influence profitability.

The dataset spans from 2011-2022, includes a wide range of variables, such as total discharges, hospital total days, total salaries, inpatient and outpatient charges, total income, liabilities, current and fixed assets, and inventory. This comprehensive set of variables enables robust analysis and modeling to capture the diverse factors impacting hospital financial health. By incorporating both operational and financial measures, the study provides actionable insights into optimizing resource allocation and improving decision-making processes.

This research comes at a critical time, as healthcare financial challenges continue to rise. In 2023, healthcare bankruptcies reached their highest level in five years, with 79 Chapter 11 bankruptcy filings recorded. This marked a significant increase from 51 cases in 2019, primarily driven by large liabilities and pandemic-related economic shifts (Payerchin, 2024). These challenges underline the importance of understanding and addressing hospital profitability determinants.

Revenue cycle management (RCM) also plays a pivotal role in improving hospital profitability by streamlining claim submission, billing, and reimbursement processes. Effective RCM improves cash flow, enhances patient experience, and reduces billing errors (Chandawarkar et al., 2024). Previous research has identified factors affecting hospital profitability, including organizational structure, reimbursement mechanisms, and patient mix (Nevola et al., 2016; Ly & Cutler, 2018). This study focuses on these established variables to assess their influence on financial performance.

By leveraging these dependent variables and exploring various financial and operational covariates, this report aims to provide actionable insights for healthcare leaders and policymakers striving to maintain profitability while delivering high-quality care.

Data Cleaning

A thorough data cleaning process was conducted to prepare the CMS Hospital Provider Cost Report dataset for analysis. This step ensured the data's reliability for understanding hospital financial performance and supporting decision-making. The key dependent variables are Cost-to-Revenue Ratio = Operating Costs / Total Revenue and Revenue per Bed = Total Revenue / Number of Beds. All other independent variables are listed in the linear regression model table.

Duplicates were identified based on the Provider CMS Certification Number (CCN) and year. For numeric variables, the mean was retained, while for non-numeric variables, the first occurrence was preserved. Numeric columns fully missing within duplicate groups were removed.

To address outliers, data points with a Cost-to-Revenue Ratio greater than 100 and Revenue per Bed exceeding $100 million (scaled by dividing by 1,000,000) were excluded. These entries were deemed extreme and could distort the analysis.

Missing numeric values were interpolated using the zoo::na.approx() function when sufficient data points (more than two) were available. Sparse variables, where interpolation was unreliable, were left untouched to avoid introducing artificial patterns. Several financial and operational metrics, such as Total Salaries, Charges, Income, and Liabilities, were scaled to millions for consistency. Two key metrics—Cost-to-Revenue Ratio and Revenue per Bed—were calculated to measure operating efficiency and revenue generation.

This cleaning process ensured the dataset was optimized for robust analysis. By removing inconsistencies, addressing outliers, and filling gaps, it became suitable for deriving actionable insights into hospital profitability, addressing financial challenges, and improving decision-making as outlined in the introduction. We can see that after data cleaning, the dependent variables are following a normal distribution and without significant outliers.

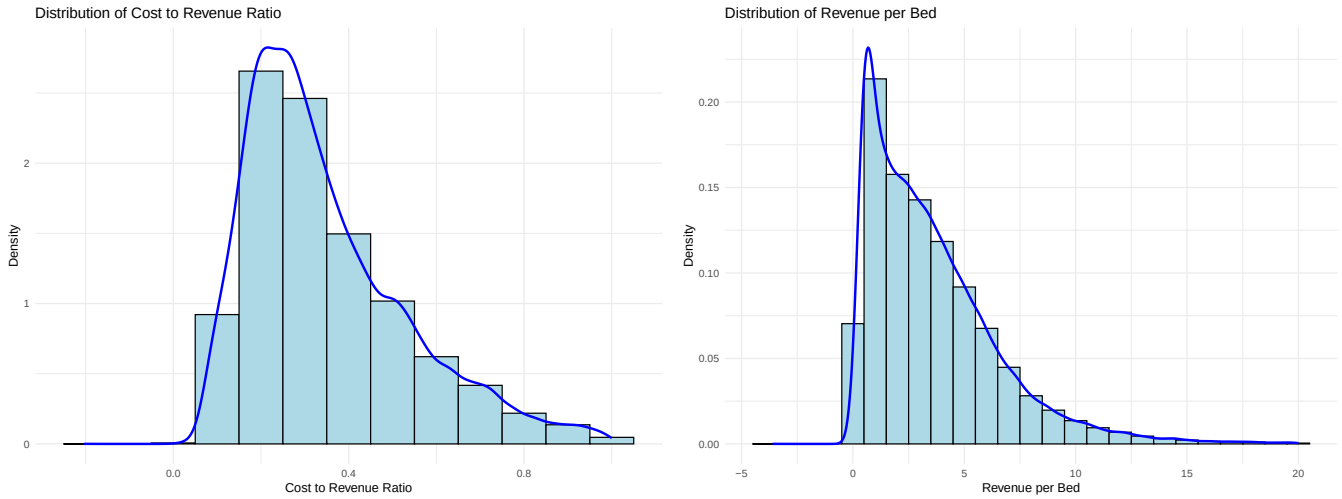

Figure 1: Distribution of Dependent Variables (cost to revenue ratio) and (revenue per bed)

Statistical Summary

R Shiny Application

The CMS Hospital Provider Cost Report dataset, with its wealth of information, provides a unique opportunity to explore financial performance metrics at both granular (hospital-level) and aggregated (state-level) scales. Effectively communicating these insights requires an intuitive and customizable tool, which R Shiny provides through its interactive and dynamic visualization capabilities.

The application includes two key visualization tabs. The first tab, the dot distribution map, presents hospital-level financial metrics for a sample of 4,000 hospitals. Each hospital is represented by a dot, where the color gradient indicates revenue per bed, with darker colors reflecting higher values, and the dot size scales with the cost-to-revenue ratio. Users can interact with the map through pop-ups that display details such as the hospital name, location, revenue per bed, and cost-to-revenue ratio. A temporal slider allows users to explore trends across the dataset's time range from 2011 to 2022. This feature is particularly useful for identifying temporal patterns and evaluating changes in financial performance over time. By adopting the hospital-level information like this, this tab effectively combines geographical distribution with intuitive visual elements, enabling users to gain quick insights. The pop-up functionality allows users to access additional details on demand, offering both high-level patterns and granular information when needed.

The second tab aggregates financial metrics at the state level, providing a broader view of regional financial trends. The state-level map uses color gradients to highlight variations in metrics such as average revenue per bed and average cost-to-revenue ratios. Interactive pop-ups displays detailed summaries for each state, including aggregated metrics and comparisons across time periods. This functionality helps users quickly identify financial outliers, highlighting states that are either excelling or struggling financially. The balance between summarized state-level insights and detailed pop-up information ensures users can explore both high-level trends and specific state data efficiently.

This application not only empowers stakeholders to make evidence-based decisions but also showcases the versatility of R Shiny in handling large and complex datasets. By combining statistical analysis with user-friendly visualization tools, the application serves as a valuable resource for understanding and improving hospital financial performance at multiple levels.

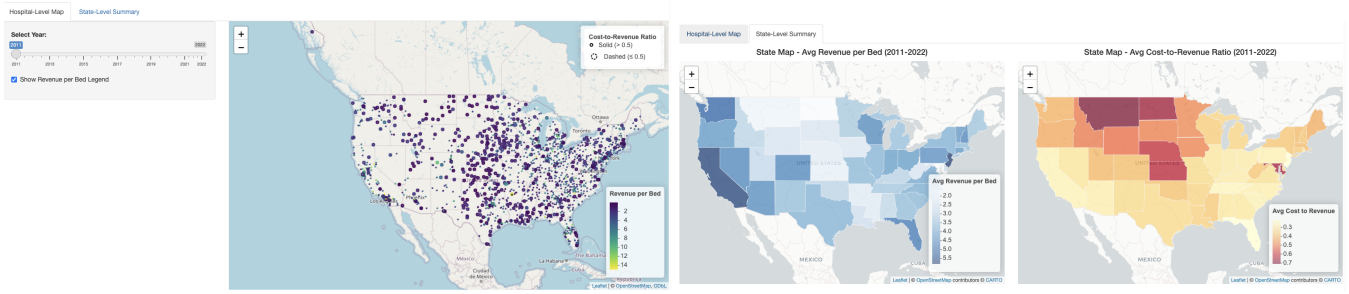

Figure 2: Shiny Results Illustrations (details can be found at https://xxchar.shinyapps.io/hospital-financial-analysis-ui/)

Table 1: Regression Results

Variable	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-318	7.34	-43.4	0
Total.Discharges	0.000133	4.68e-06	28.5	2.78e-177
Hospital.Total.Days	-6.07e-05	1.22e-06	-49.9	0
Total.Salaries	-0.00226	0.00028	-8.07	7.27e-16
Inpatient.Total.Charges	-0.00294	0.000113	-26.1	6.27e-150
Outpatient.Total.Charges	0.000124	0.000117	1.06	0.29
Total.Income	0.000751	0.000186	4.03	5.49e-05
Total.Other.Income	-0.00158	0.000228	-6.92	4.51e-12
Total.Liabilities	-0.00028	3.53e-05	-7.93	2.15e-15
Accounts.Payable	0.000312	0.000233	1.34	0.182
Total.Current.Assets	0.000146	5.9e-05	2.48	0.013
Total.Fixed.Assets	0.000257	0.000124	2.08	0.038
General.Fund.Balance	-5.27e-05	5.3e-05	-0.993	0.321
Inventory	6.08e-10	1.26e-09	0.484	0.628
Total.Patient.Revenue	0.00359	0.000108	33.2	6.04e-239
Number.of.Beds	-3.8e-06	1.9e-06	-2	0.046
year	0.16	0.00364	43.9	0

Regression and Geospatial Analysis

Linear Regression Model

We start with a basic linear regression model and the model can be listed as the follows:

$$Y_{it} = \beta_0 + \beta \mathbf{X_{it}} + \mu_\mathbf{i} + \tau_\mathbf{t} + \epsilon_\mathbf{it}$$

where Y_{it}: Target variable (e.g., total revenue or operating costs). β_i: Coefficients for predictors. $\mathbf{X_{it}}$ are the key independent variables that are reported in the statistics summary table, μ_i is the state level fixed effect and the τ_t is the time fixed effect (year). ϵ_{it}: Error term. The result can be summarized to the Table 1. We can see that most of the factors are significant below 1%, which means that they are of great influence to our dependent variables.

Geographically Weighted Regression

Geographically Weighted Regression (GWR) is a spatial regression technique designed to capture spatial heterogeneity in relationships between dependent and independent variables. Unlike global models, which assume uniform relationships across space, GWR allows for local parameter estimation, enabling better insights into spatially varying patterns.

Given the Shiny visualization, we can see that there is clearly a different trend in the geographical distribution, and therefore, we conducted a GWR model that incorporates the geographical kernel to capture the spatial heterogeneity. The model can be listed as the follows:

$$Y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i) x_{ik} + \epsilon_i$$

where Y_i is the dependent variable for observation i, $\beta_0(u_i, v_i)$ is the intercept that varies at location (u_i, v_i), where u_i and v_i are the spatial coordinates of observation i. $\beta_k(u_i, v_i)$ is the k-th regression coefficient varies at location (u_i, v_i); x_{ik} is the value of independent variables, ϵ_i is the random error; p is the total number of independent variables.

We carried out the county level GWR and the state level GWR, and due to the page limit, we only reported the county level GWR to the result. Starting with a Moran's I index, the index shows that the result is significantly greater than 0.5, which indicates a strong correlation in spatial heterogeneity. The data were analyzed separately for each year using the multi-bandwidth selection method, which determines the optimal bandwidth for each individual dataset. Following the bandwidth selection, the GWR analysis was conducted, incorporating multiprocessing techniques to improve computational efficiency. Without multiprocessing,

the code execution time exceeded 10 minutes; however, with parallel processing, the runtime was significantly reduced. For each annual GWR analysis, coefficient plots were generated and visualized using the ggplot2 package in R. Due to space constraints, we present two of the most significant contrasts for illustration, selected based on their statistical importance and relevance to the study.

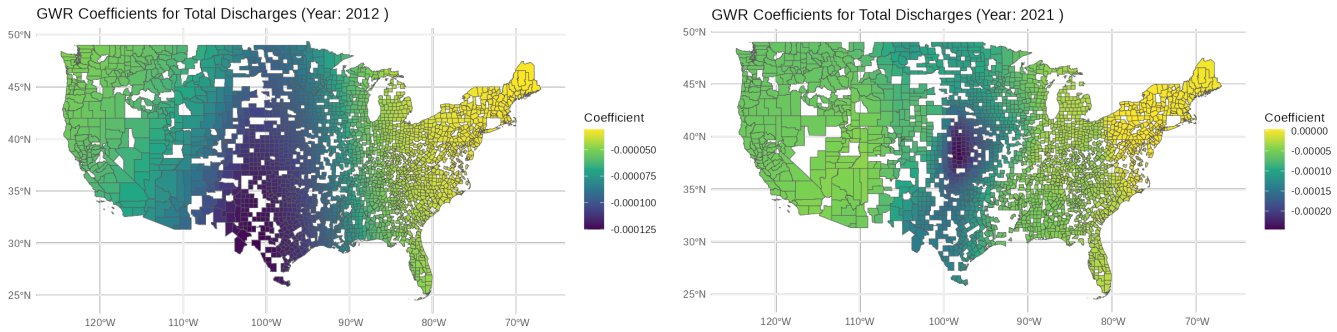

Figure 3: GWR Coefficient Maps for 2012 and 2021

We can see that In 2012, the effect of the independent variable on the dependent variable was not statistically significant in certain areas. However, a clear spatial pattern emerges, with the northeastern region of the United States exhibiting the most significant influence. By 2021, the effect becomes statistically significant across a larger portion of the study area, indicating a broader spatial impact. Nevertheless, the northeastern region continues to demonstrate the most pronounced influence of the independent variable on the dependent variable. This persistence highlights a consistent spatial pattern over time.

Machine Learning Implementation and Analysis

Since some financial indicators, like total cost and total revenue are usually calculated by taking the sum of different subcategories, they are more prone to errors compared with other variables regarding the scale of the health care providers like the total number of beds, the total number of discharges and the inventory of hospital supplies. We aimed to apply a few machine learning methods to predict the Cost to Revenue Ratio using variables related to the scale of providers rather than directly utilizing the cost and revenue data. These models could generate rough but reasonable estimates of the profitability with which the hospital management may be able to reevaluate the correctness of the financial report and more importantly, predict the financial metric with only partial information.

We selected two supervised learning methods—Support Vector Machine (SVM) and Random Forests, and one unsupervised learning method, k-Nearest Neighbors (k-NN). In order to assess the performance of different methods more comprehensively, we used 5-fold cross validation. The cross validation posed significant computational challenges due to the substantial size of the dataset (a training set with 24340 observations and 14 variables), particularly for the SVM training, where we implemented three different kernel functions. Our solution to expedite the SVM training process was exploiting the parallel and doParallel packages to achieve multi-cores parallel computing.

As for the model results, the unsupervised knn outperformed the svm and random forests in terms of both R^2 and RMSE. A possible explanation for this could be that hospitals with similar scale may have alike Cost to Revenue Ratio. We know that k-NN leverages information from the local neighborhood for prediction, and when the data exhibits consistent patterns within local regions, kNN can deliver relatively strong predictive performance. It is noteworthy that the svm with polynomial kernel performed poorly with an R^2 over -1, highlighting the potential problem of overfitting.

Although the k-NN achieved the highest R^2 and lowest RMSE, we did not know the importance of different features due to its unsupervised nature. Given that the performance of the random forest was quite close to the k-NN, we would refer to the feature importance from the random forest model. The graph below shows top 10 important features in the Cost to Revenue Ratio Prediction. The total discharges of patients ranked the first which was reasonable as more patients indicate more profits; The total number of beds and the cost of all buildings ranked the second and third showing that the hospital scale plays a vital role concerning the profitability and efficiency.

There is also some room for future improvements. We can consider experimenting with more machine learning methods. It is also feasible to include data of proceeding years when it becomes available to continue enlarging the sample size. Besides, when it comes to a series of hyperparameters (the number of trees in random forest, etc.) in our model training, we mainly adopted the

default settings without further fine-tuning. There might be some enhancement by doing so but we would not anticipate it to be considerable since the current model performance is already satisfying (R^2 over 0.8).

Table 2: Model Performance Comparison

Model	Mean R^2	Mean RMSE
knn_gaussian	0.7975	0.0930
knn_rectangular	0.7712	0.0989
knn_triangular	0.8265	0.0861
svm_linear	0.3406	0.1680
svm_polynomial	-83.8643	1.5151
svm_radial	0.5126	0.1444
rf_impurity	0.8001	0.0925
rf_permutation	0.8007	0.0923

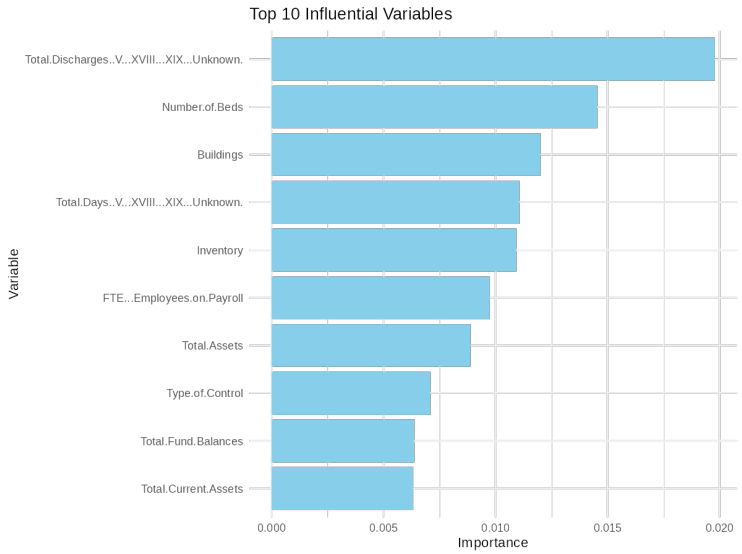

Figure 4: Top Ten Influential Features of the Random Forest Model

Conclusion and Discussion

This project analyzed and forecasted hospital financial performance using the CMS Hospital Provider Cost Report dataset (2011–2022), focusing on Cost-to-Revenue Ratio and Revenue per Bed. Key findings show that hospital scale factors—total discharges, number of beds, and fixed assets—significantly influence financial outcomes. We created Shiny App to visualize the spatial heterogeneity and conducted Geographically Weighted Regression (GWR) to reveal spatial variations, particularly highlighting persistent trends in the Northeastern U.S. We also carried out the machine learning method to forecast the model, k-Nearest Neighbors (k-NN) achieved the best performance ($R^2 = 0.83$), while Random Forest provided insights into feature importance, emphasizing hospital scale and capacity as key drivers of profitability.

Potential exists for us to further enhance the utility of the Shiny app. Expanding the range of financial and operational metrics, such as patient satisfaction scores, staffing ratios, or service line profitability, would provide a more comprehensive view of hospital performance, enabling users to analyze the interplay between financial and non-financial factors. Incorporating predictive analytics, such as models we've currently proposed for forecasting revenue trends, bed occupancy rates, or cost-to-revenue ratios, could offer stakeholders a forward-looking perspective. Visualizing these predictions alongside historical data would enrich the application's insights. Additionally, enabling dynamic filtering by hospital size, type, or region and allowing users to export tailored dashboards would enhance interactivity and flexibility.

References

- Chandawarkar, R., Nadkarni, P., Barmash, E., Thomas, S., Capek, A., Casey, K., & Carradero, F. (2024). Revenue Cycle Management: The Art and the Science. Plastic and reconstructive surgery. Global open, 12(7), e5756.
- Nevola, A., Pace, C., Karim, S. A., & Morris, M. E. (2016). Revisiting 'The Determinants of Hospital Profitability' in Florida. Journal of Health Care Finance.
- Ly, D. P., & Cutler, D. M. (2018). Factors of U.S. Hospitals Associated with Improved Profit Margins: An Observational Study. Journal of General Internal Medicine, 33(7), 1020–1027.
- Payerchin, R. (2024,). Health Care Bankruptcies in 2023 reach highest level in five years. MedicalEconomics. https://www.medicaleconomics.com/view/health-care-bankruptcies-in-2023-reach-highest-level-in-five-years