DSC680 Applied Data Science

Week Three – Project One White Paper

Author: Alysen Casaccio

Title: Predicting Cost Categories Related to Length of Stay Using Unsupervised and Supervised Learning

Introduction

Business Problem

Accurately estimating the cost of excess inpatient days is a persistent challenge for hospital administrators (Morid MA, 2018 April). In practice, a uniform cost estimate across all patients often fails to capture the variability due to patient characteristics and the complexity of care (Daniel B. McLaughlin, 2009). I took a hybrid approach utilizing unsupervised and supervised learning techniques to address this. This approach categorized patients into distinct cost clusters and predicted the excess day cost within each cluster, offering a more precise and nuanced cost estimation.

Background

Effective cost management is crucial for healthcare providers aiming to optimize resource allocation and improve patient outcomes (Leusder M, 2022 Dec). Traditional cost estimation methods, which apply a uniform rate to all excess days, overlook patient-specific factors influencing costs (Morid MA, 2018 April). This project leveraged publicly available data to explore more sophisticated modeling techniques for cost estimation (Health, 2024).

Data Description

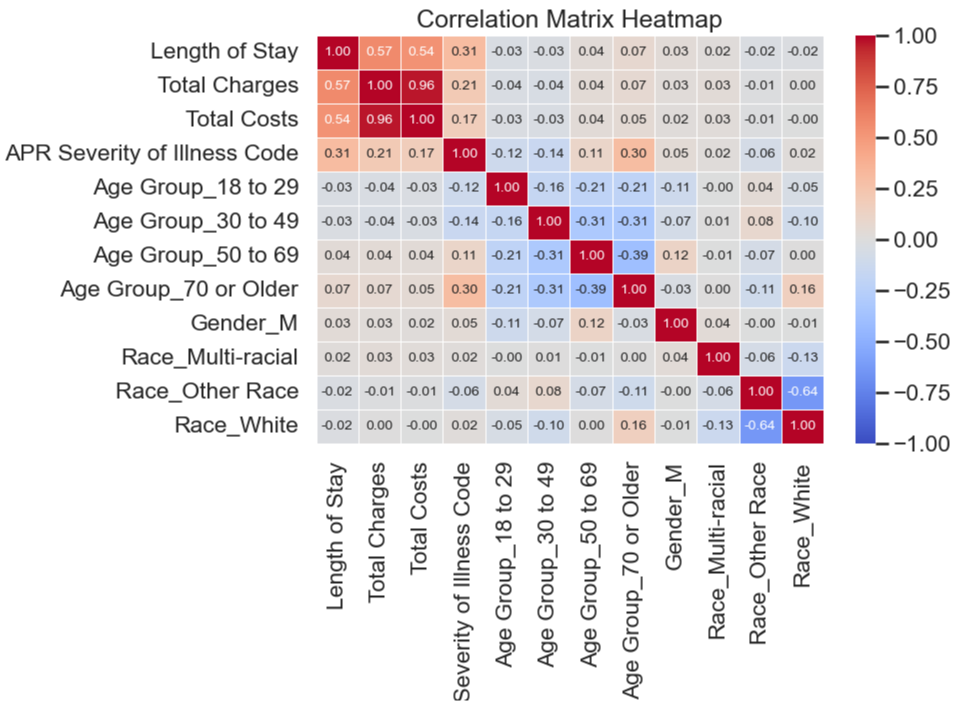
Data Source

The analysis utilized the publicly available 2017 inpatient discharge dataset from New York State, including comprehensive hospital stay information. This dataset was obtained from the Statewide Planning and Research Cooperative System (SPARCS), which currently collects patient-level details on patient characteristics, diagnoses and treatments, services, and charges for each hospital inpatient stay and outpatient visit (Health, 2024).

Data Preparation

Feature Selection

Relevant columns were initially chosen based on domain knowledge, and later, a correlation matrix was examined to understand better which features had the largest impact on cost. Features selected for the dataset Facility Name, Age Group, Gender, Race, Ethnicity, Length of Stay, Type of Admission, Patient Disposition, CCS Diagnosis Code, CCS Diagnosis Description, APR DRG Code, APR DRG Description, APR Severity of Illness Code, APR Severity of Illness Description, Payment Typology 1, Total Charges, and Total Costs.



Data Cleaning

Data cleaning involved managing missing values, converting data types, and correcting anomalies. The amount of data used for the project was also reduced from over 500,000 rows to only 5,000 for improved performance and a more focused project. A sampling methodology was used to perform this reduction to ensure the rows were selected fairly across the dataset.

Normalization

Features were scaled to ensure they contributed equally to the clustering algorithm.

Data Dictionary

A comprehensive data dictionary was maintained, detailing the purpose and content of each feature used in the analysis. A shortened form of the dictionary is included in Appendix 1.

Methodology

Unsupervised Learning:

Clustering Technique

K-Means Clustering was employed to group patients into clusters based on features like Length of Stay, Total Charges, and demographic information. The optimal number of clusters was determined using silhouette scores and the elbow method.

Cluster Analysis

Each cluster was analyzed to understand the common characteristics and patterns that define it, helping to interpret cost categories.

Supervised Learning:

Regression Models

Two iterations of Random Forest Regression were used to predict the cost of an excess day within each identified cluster.

Model Tuning

Hyperparameters were tuned using Grid Search, and regularization techniques were applied to improve model generalization.

Model Evaluation:

Metrics

The models were evaluated using R-squared, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) to assess predictive accuracy.

Cross-Validation:

Approach

5-fold cross-validation was employed to validate the models, ensuring robustness and mitigating overfitting.

Analysis

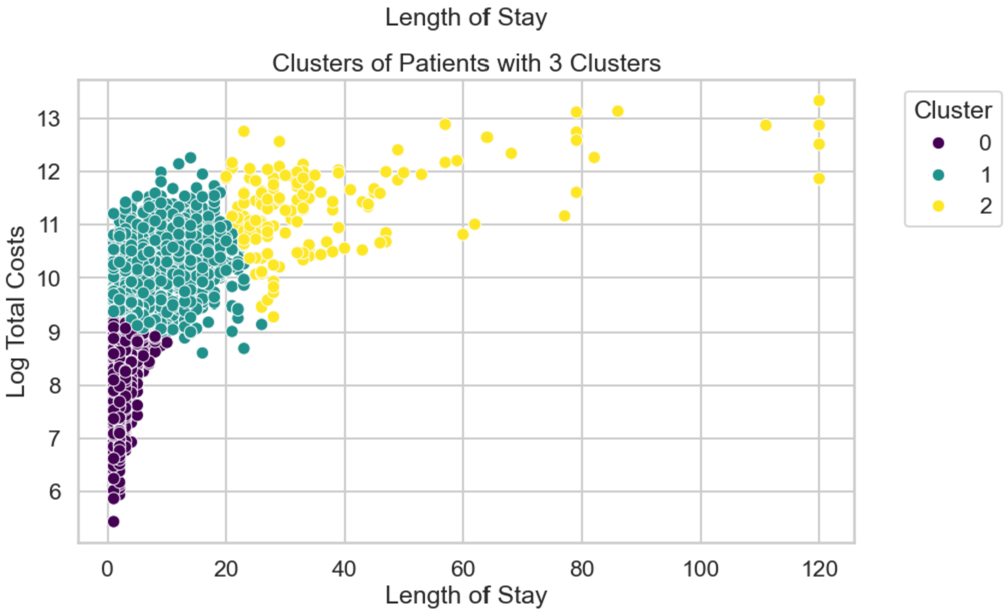
Clustering Results:

Optimal Clusters

Three distinct clusters were identified as optimal based on silhouette scores and visual examination of the elbow method.

Cluster Characteristics

Each cluster represented patients with different lengths of stay and cost patterns, providing meaningful categorizations for further analysis.



Regression Results:

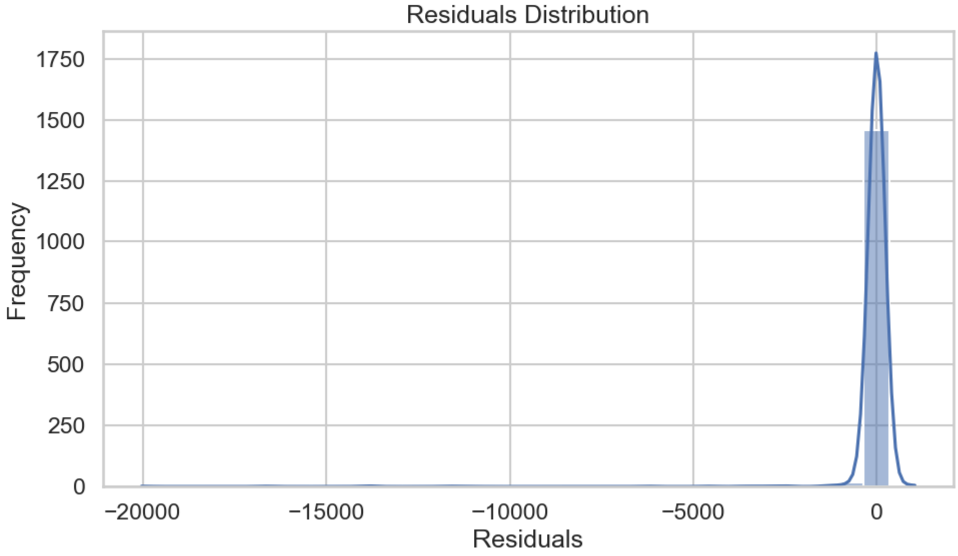
Performance Metrics

The Random Forest Regression model, with hyperparameter tuning and regularization, achieved the best balance between bias and variance. Test set metrics included a Mean Squared Error (MSE) of 958,120.37, a Root Mean Squared Error (RMSE) of 978.84, a Mean Absolute Error (MAE) of 104.45, and an R-squared (R^2) of 0.9985.

Residual Analysis:

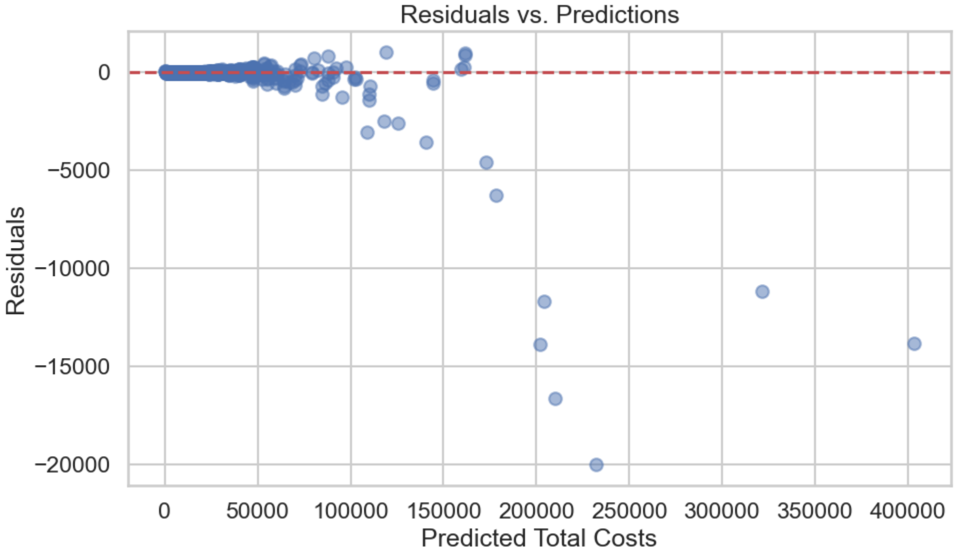
Residual Distribution

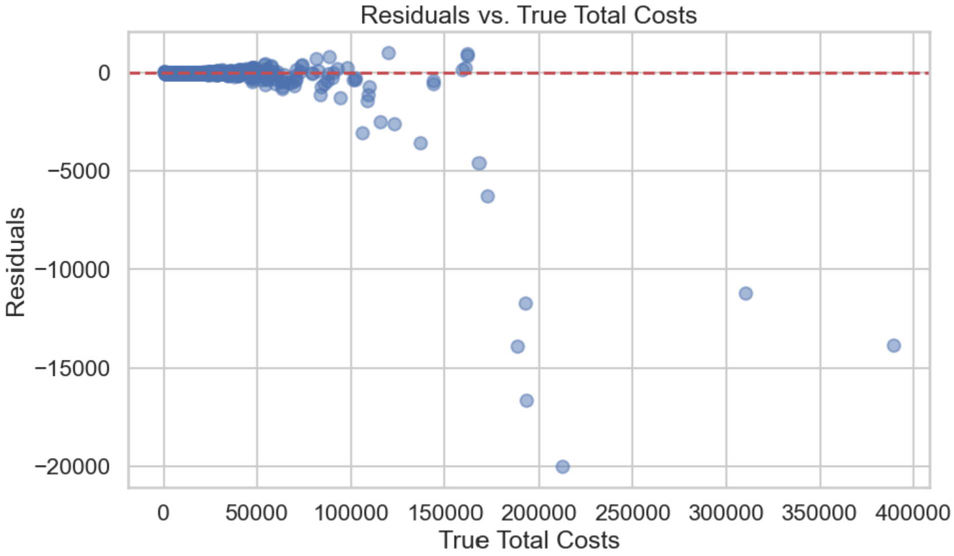
Most residuals were tightly clustered around zero, indicating high predictive accuracy.



Outliers

Some residuals deviated significantly, highlighting cases where the model's predictions were less accurate, warranting further investigation.





Cluster-Based Cost Analysis:

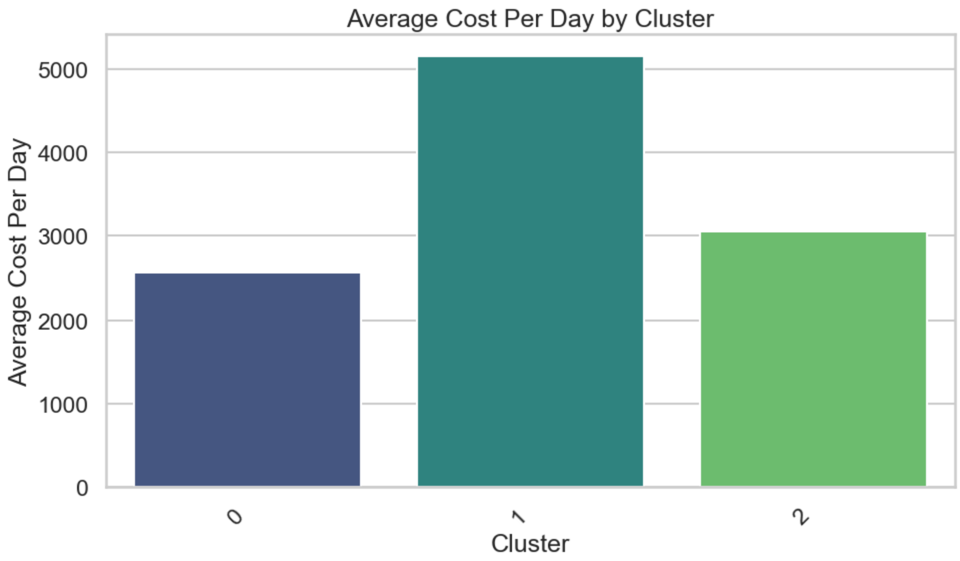
Average Cost Per Day

The average cost per day for each cluster was calculated and compared against the overall average cost per day ($3737.95):

Cluster 0: $2562.45 per day

Cluster 1: $5160.01 per day

Cluster 2: $3056.45 per day



Implications:

Alignment with Overall Estimate

These results show significant variability in daily costs among the clusters. While the average daily cost is $3737.95, the cost within clusters varies from $2562.45 to $5160.01. This variation underscores the inadequacy of a uniform cost estimate and highlights the value of a more detailed approach to cost estimation.

Conclusion

Findings:

The hybrid approach effectively categorized patients into cost clusters and provided accurate cost predictions for excess inpatient days. This nuanced estimation can significantly improve hospital administrators' resource allocation and financial planning.

Implications:

Implementing such models can enhance decision-making by providing more precise cost estimates, ultimately leading to better resource management and improved patient care.

Assumptions

Model Simplicity:

Assumes that linear and ensemble models can capture the relationship between features and costs.

Data Representativeness:

Assumes the publicly available dataset represents the broader population adequately.

Limitations

Data Scope:

The analysis is limited to a single year's data from New York State, which may not generalize to other regions or years.

Cluster Interpretability:

Clusters may not be easily interpretable or applicable to all hospitals.

Challenges

Data Quality:

Handling missing and non-numeric values presented challenges in ensuring clean and accurate data for analysis.

Model Integration:

Integrating unsupervised and supervised learning techniques required careful coordination to maintain coherence and simplicity.

Future Uses/Applications

Extended Models:

Applying similar models to other datasets or healthcare systems to validate and refine the approach. I want to attempt a similar model using the live clinical discharge data of a healthcare system I am consulting with today to see if I can provide additional insights into discovered cost clusters.

Predictive Analytics:

Expanding the model to include predictions for other healthcare costs or outcomes.

Recommendations

Model Deployment:

Implement the model in hospital financial planning systems to provide dynamic cost estimates. I also believe obtaining buy-in from key stakeholders in cost modeling is essential, such as the CFO, VP of Finance, and other cost, payor, and accounting leaders.

Ongoing Validation:

Regularly update and validate the model with new data to ensure continued accuracy and relevance.

Implementation Plan

Steps:

Data Integration: Incorporate the model into the existing hospital data system.

Model Training: Train the model with current data periodically to keep it updated.

Validation: Conduct regular validations to ensure accuracy and adaptability.

Ethical Assessment

Privacy:

Ensure all patient data is anonymized and handled in compliance with HIPAA and internal privacy and security regulations.

Bias:

Evaluate and mitigate any biases in the model to ensure fair predictions for all patient groups.

Appendix 1 - Data Dictionary

Facility Name: Name of the facility at which the discharge occurred.
Age Group: Pre-generated age groupings to include 0-17, 18-29, 30-49, 50-69, and 70+.
Gender: M or F
Race: Black/African American, White, Other Race, Multi-racial
Ethnicity: Not Span/Hispanic, Multi-ethnic, Spanish/Hispanic, Unknown
Length of Stay: Length of stay is a calculation of the number of days from admission to discharge.
Type of Admission: A selection of six different categories for admission types.
Patient Disposition: A selection of 20 or fewer categories describing a patient's care setting after discharge, including home or self-care, skilled nursing home, left against medical advice, and others.
CCS Diagnosis Code: primary discharge diagnosis code
CCS Diagnosis Description: primary discharge diagnosis description
APR DRG Code: Medicare-applied statistical system to classify inpatient stays into groups for the purposes of payment. Codes are based on factors such as illness severity, mortality risk, prognosis, treatment difficulty, need for intervention, and resource intensity.
APR DRG Description: a narrative description of the DRG code.
APR Severity of Illness Code: A rating of 1-4.
APR Severity of Illness Description: a narrative description of the four levels of severity of illness.
Payment Typology 1: Name of the patient's primary payor for that inpatient stay.
Total Charges: The dollar amount of the total charge to the patient or payor for that inpatient stay.
Total Costs: The dollar amount of the total cost incurred by the hospital or health system for that inpatient stay.

Appendix 2 - Potential Audience Questions:

1. How did you handle missing or non-numeric values in your dataset?
2. Why did you choose K-Means Clustering for this analysis?
3. How do you ensure the model generalizes well to unseen data?
4. What are the key characteristics that define each cluster?
5. How do the costs estimated by your model compare to traditional estimation methods?
6. Can this model be applied to other types of healthcare costs or outcomes?
7. What steps would you take to implement this model in a real hospital setting?
8. How do you handle potential biases in your model?
9. What are the limitations of your analysis, and how might they be addressed in future work?
10. What ethical considerations did you consider when developing your model?

References

Andy Field, J. M. (2012). *Discovering Statistics Using R.* Los Angeles: SAGE Publications.

Daniel B. McLaughlin, J. M. (2009). *Healthcare Operations Management.* New Delhi: PHI Learning Health Administration Press.

Health, N. Y. (2024, March). *Statewide Planning and Research Cooperative System (SPARCS)*. Retrieved from https://www.health.ny.gov/statistics/sparcs/

Kyle Gallatin, C. A. (2023). *Machine Learning with Python.* Sebastopol, CA: O'Reilly Media, Inc.

Leusder M, P. P. (2022 Dec). Cost Measurement in Value-Based Healthcare: A Systematic Review. *BMJ*, 7-12. doi:10.1136/bmjopen-2022-066568

Morid MA, K. K. (2018 April). Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation. *AMIA*, 1312-1321.

Taloba Al, A. E.-A.-B. (2022 March). Estimation and Prediction of Hospitalization and Medical Care Costs Using Regression in Machine Learning. *Journal of Healthcare Engineering*. doi:10.1155/2022/7969220