Author: Alysen Casaccio

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

June 20,2024

Topic

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

Business Problem

Accurately estimating the cost of excess inpatient days is a significant challenge for hospital administrators. For this real-life example, a CFO I'm working with has provided a general estimate of $569.00 per excess day across the entire patient population. This uniform approach does not account for variations in patient characteristics or the complexity of care.

Using publicly available data as a testing ground, this project aims to first cluster patients into distinct cost categories using unsupervised learning techniques and then predict the cost of an excess day within each cluster using supervised learning. This hybrid approach will provide a more nuanced and precise estimation of excess day costs, leading to better resource allocation and cost management.

Datasets

The project will utilize publicly available discharge data from New York State for the 2017 year, including:

- Inpatient discharge and admit dates
- Calculated length of stay (LOS)
- Total charges and costs for each stay
- Primary Diagnoses
- Diagnosis-Related Groups (DRGs) assigned during the stay
- Demographic information

Methods

The project will integrate the following methods:

1. Data Cleansing and Preparation

- Data Quality: Ensuring the data is accurate, complete, and formatted correctly for analysis.
- Feature Engineering: Creation of relevant features that may influence costs and LOS.

2. Unsupervised Learning

- Clustering (e.g., K-Means Clustering): Clustering patients into groups based on features such as LOS, total charges, diagnoses, and demographic information to identify cost categories.
- Cluster Analysis: Exploring and validating the characteristics of each cluster to understand the underlying patterns.

3. Supervised Learning

- Cost Prediction Models: Development of regression models to predict the cost of an excess day within each identified cluster. Techniques may include:
  - Linear Regression
  - Random Forest Regression
  - Gradient Boosting Machines
  - Support Vector Machines (SVM)
- Model Evaluation: Using metrics such as R-squared, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) to evaluate the predictive accuracy.

4. Comparative Analysis

- Evaluation: Comparing the predicted costs with actual costs within each cluster to assess the model's performance and identify cost-saving opportunities.
- Scenario Analysis: Analyzing potential scenarios to understand the impact of reducing excess days within each cluster.

5. Visualization

- Charts and Graphs: Creation of visual representations of clusters, predicted costs, and potential savings.

Ethical Considerations

- Patient Privacy and Security: I would ordinarily ensure that data is anonymized and handled in compliance with HIPAA and other relevant regulations, but for this project, I am using a publicly available dataset where this work has already been accomplished.
- Bias and Fairness: Evaluating clustering and prediction models for biases that may affect certain patient groups unfairly.

Potential Challenges or Risks

1. Data Quality: Incomplete or inconsistent data could affect clustering and prediction accuracy. Mitigated through rigorous data cleansing and preprocessing.
2. Cluster Interpretability: The clusters identified may not be easily interpretable. Mitigation involves testing different clustering algorithms to find the most meaningful results.
3. Model Integration: Effectively combining clustering with predictive models requires careful integration and validation to ensure coherent elegance and simplicity wherever possible.

References

Abbott, D. (2014). *Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst*. Indianapolis: Wiley.

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 AI, 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