Evaluate Hospital Performance Based on Quality Measures, Patient Experiences, and Readmission Rates

Academic Research Project Report

Submitted to

Dr. Salem Othman

By

Yogesh Pandey

Hricha Maharjan

Mavira Bhattarai

November 2024

ABSTRACT

Evaluating hospital performance is crucial for ensuring quality healthcare delivery. This research integrates three critical metrics—**quality measures**, **patient experiences**, and **readmission rates**—to provide a comprehensive evaluation of hospital performance. Using advanced data cleaning techniques, exploratory data analysis, and machine learning models, the study uncovers significant relationships among these metrics. Results indicate that patient satisfaction strongly correlates with higher quality ratings and lower readmission rates. While Random Forest and Gradient Boosting models showed promising classification and prediction accuracy, the study emphasizes the need for additional features and advanced models for further improvements.

TABLE OF CONTENT

CHAPTER 1: INTRODUCTION

Hospital performance evaluation plays a vital role in improving healthcare outcomes and operational efficiency. Traditionally, metrics like **quality measures**, **patient experiences**, and **readmission rates** are analyzed in isolation, missing their interconnected impact. This study aims to address this gap by integrating these datasets to assess hospital performance comprehensively. Using statistical analysis, data visualization, and machine learning, the project identifies key trends and predictors of hospital ratings, offering insights for healthcare administrators and policymakers.

CHAPTER 2: RELEVANT WORK

Existing research provides a foundation for this study:

1. Studies on quality measures focus on clinical outcomes and compliance with healthcare standards but often exclude patient feedback.
2. Research on patient experiences highlights satisfaction as a factor influencing hospital choice but rarely connects it with readmission rates.
3. Analyses of readmission rates emphasize their financial and clinical impact but provide limited insights into broader hospital performance.

By integrating these aspects, this project extends the work of previous studies to offer a unified evaluation framework. Notable influences include:

- Research on an integrated machine learning framework for hospital readmission prediction (Shancheng Jiang, 2018): Explored machine learning for readmission prediction.
- Research on Patient Satisfaction as an Indicator of Quality Care (Paul D. Cleary, 1988): Investigated correlations between patient satisfaction and quality ratings.

These studies informed the methodology, including data cleaning, exploratory analysis, and model selection.

CHAPTER 3: METHODOLOGY

This chapter outlines the research methodology used to conduct the study with the objective of establishing the relationship between hospital quality measures, patient satisfaction, and hospital readmission rates.

3.1 DATA SOURCES

Three datasets were utilized in this study to evaluate hospital performance comprehensively:

- Quality Measures: Metrics reflecting clinical effectiveness and safety.
- Patient Experiences: Feedback scores on patient satisfaction.
- Readmission Rates: Data on hospital readmissions within 30 days of discharge.

3.2 DATA CLEANING AND PREPROCESSING

Data cleaning and preprocessing are essential to guarantee the accuracy, consistency, and reliability of analyses when assessing hospital performance using quality measures, patient experiences, and readmission rates. The following provides a detailed overview of the data cleaning and preprocessing steps implemented in this project.

Handling Missing Values

Missing values can negatively impact data analysis and model performance. In this project, missing values were addressed as follows:

- **Identifying Missing Data**: The .isna().sum() function was used to identify missing values in key columns such as:
    - Hospital overall rating (a measure of hospital quality)
    - Patient Survey Star Rating (a measure of patient satisfaction)
    - Score (readmission rates).
- **Standardizing Column Names:** One of the first steps in data preprocessing is ensuring that column names are consistent and easy to work with. This prevents issues with case sensitivity and special characters during analysis.

  For this project:
    - All column names were converted to lowercase.
    - Spaces in column names were replaced with underscores to make them Python-friendly.

  This was achieved using the following code for each dataset:

```
# Standardize column names
quality_data.columns = quality_data.columns.str.lower().str.replace(' ', '_')
experience_data.columns = experience_data.columns.str.lower().str.replace(' ', '_')
readmission_data.columns = readmission_data.columns.str.lower().str.replace(' ', '_')
```

For example:

- Hospital overall rating became hospital_overall_rating.

- Patient Survey Star Rating became patient_survey_star_rating.

This standardization ensured consistency across all datasets, simplifying the merging and analysis process.

Converting Data Types

Data often comes in inconsistent formats that need to be standardized. For example:

- Columns such as Hospital overall rating, Patient Survey Star Rating, and Score were converted to numeric using:

```
# Convert numeric columns to numeric types, coercing errors to NaN
quality_data = quality_data.apply(pd.to_numeric, errors='coerce')
experience_data = experience_data.apply(pd.to_numeric, errors='coerce')
readmission_data = readmission_data.apply(pd.to_numeric, errors='coerce')
```

This ensured that any non-numeric values (e.g., strings or invalid entries) were coerced into NaN for further cleaning.

Merging Datasets

The analysis required data from multiple datasets:

- **Hospital Quality Measures**: Provided overall hospital ratings.
- **Patient Experience Surveys**: Contained patient satisfaction scores.
- **Readmission Rates**: Measured the proportion of patients readmitted.

Datasets were merged on the common Facility ID column using:

```
# Merge datasets on 'hospital_id'
merged_data = pd.merge(quality_data, experience_data, on='facility_id', how='inner')
merged_data = pd.merge(merged_data, readmission_data, on='facility_id', how='inner')

# Display the merged dataset
```

Filling with Column Mean

For numeric columns, missing values were replaced with the mean of the respective column. This ensures that the overall distribution of the data is preserved without introducing biases:

```
# Now fill missing values with the column mean
quality_data.fillna(quality_data.mean(), inplace=True)
experience_data.fillna(experience_data.mean(), inplace=True)
readmission_data.fillna(readmission_data.mean(), inplace=True)
```

3.3 EXPLORATORY DATA ANALYSIS (EDA)

Select numeric columns with meaningful data:

The function is created to:

- Select numeric columns from a dataset, as these are often central to statistical and machine learning tasks.
- Manage missing data efficiently by either discarding rows with missing values or imputing them using a suitable statistic.

Guarantee that the processed data is clean, devoid of invalid entries, and prepared for analysis.

```python
def select_numeric_columns_with_meaningful_data(df):
    # Step 1: Select only numeric columns
    numeric_columns = df.select_dtypes(include=['number']).columns.tolist()
    numeric_data = df[numeric_columns]

    # Step 2: Drop columns with all NaN values
    numeric_data = numeric_data.dropna(axis=1, how='all')  # Removes columns where all values are NaN

    # Step 3: Handle NaN values in remaining columns
    # Option 1: Drop rows with NaN values in remaining numeric columns
    numeric_data = numeric_data.dropna(axis=0, how='any')  # Drops any row that has NaN in any column

    # Option 2: Fill NaN values with the median (or mean)
    # numeric_data = numeric_data.fillna(numeric_data.median())  # Fill NaNs with column median

    return numeric_data

# Example usage
filtered_data = select_numeric_columns_with_meaningful_data(merged_data)

# Display the cleaned dataset
print(filtered_data.head())
```

```
   facility_id  zip_code_x  hospital_overall_rating  \
```

Drop columns with NAN:

This function is created for data cleaning and preprocessing in the hospital performance evaluation project. Its primary goal is to detect and eliminate columns with all missing values (NaN or None), ensuring the dataset is streamlined and retains only columns with relevant and meaningful information.

```python
def drop_columns_with_all_nan(df):
    # Ensure NaN values are correctly recognized
    df = df.where(pd.notnull(df), None)  # Replace any NaN-like values with None

    # Drop columns where all values are None or NaN
    df_cleaned = df.dropna(axis=1, how='all')

    return df_cleaned

# Apply the function to your DataFrame
cleaned_data = drop_columns_with_all_nan(merged_data)

# Display the cleaned dataset
print(cleaned_data.head())

# Optionally, check the columns that remain
print("Remaining Columns after dropping NaNs:", cleaned_data.columns.tolist())
```

3.4 MACHINE LEARNING MODELS

1. Linear Regression

Linear Regression models the relationship between the target variable and the independent features by fitting a straight line that minimizes the Mean Squared Error (MSE).

Output (Original Data):

- **MSE:** 1.09
- **MAE:** 0.83
- **R^2:** 0.07

Output (Log-Transformed Data):

- **MSE:** 1.11
- **MAE:** 0.84
- **R^2:** 0.05

Analysis:

- **Performance:** Both the original and log-transformed versions performed poorly, with low R^2 scores, indicating that Linear Regression could not capture the complex relationship between features and the target variable.
- **Insights:** The minimal improvement with log transformation suggests the data's relationship may not be linear, requiring more advanced methods.

---

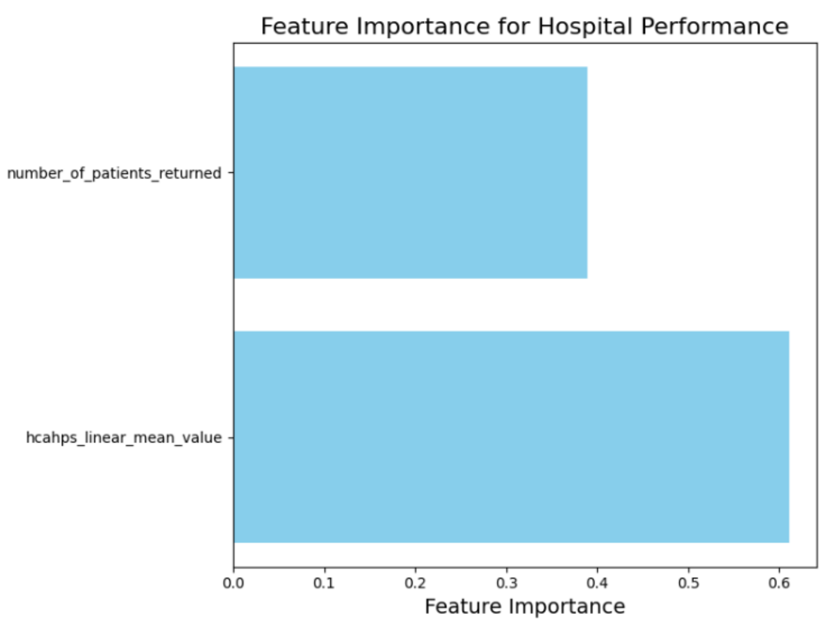
2. Random Forest Classifier

Random Forest is an ensemble method that builds multiple decision trees and aggregates their outputs to provide a classification prediction. It can handle non-linear relationships and interactions between features.

Output:

- **Accuracy Score:** 63.3%
- **Classification Report:**
    - **High Precision:** 35% (poor performance)
    - **Medium Recall:** 96% (dominates the predictions)
    - **Weighted F1-Score:** 54%

Analysis:



The bar plot illustrates feature importance scores derived from a **Random Forest Classifier** applied to the hospital performance data. It shows that the feature hcahps_linear_mean_value holds higher importance in predicting the target variable than number_of_patients_returned. This indicates that patient satisfaction metrics, as captured by the HCAHPS score, significantly influence hospital performance more than the number of patients returning.

- **Performance:** The classifier is biased toward the "Medium" class, likely because of class imbalance in the target variable. While accuracy is decent, the model struggles to predict "High" or "Low" categories effectively.
- **Strength:** Handles non-linear relationships and interactions better than Linear Regression.
- **Weakness:** Fails to generalize well due to class imbalance.

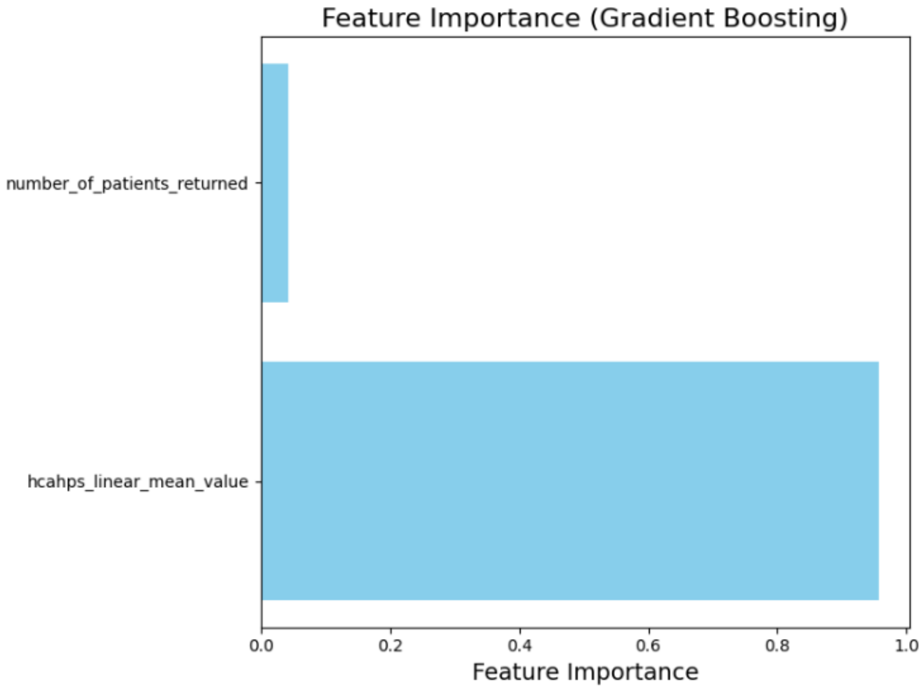
3. Gradient Boosting Regressor

Gradient Boosting is another ensemble method that builds trees sequentially, correcting the errors made by previous trees. It focuses on reducing the bias and variance of predictions.

Output:

- **MSE:** 1.08
- **MAE:** 0.82
- **R^2:** 0.08

Analysis:



The bar plot represents feature importance scores obtained from a **Gradient Boosting Regressor** applied to the same dataset. Here, the dominance of hcahps_linear_mean_value is even more pronounced, while the number_of_patients_returned contributes minimally. Gradient Boosting, being a sequential ensemble method, emphasizes features that most reduce error at each boosting stage. The results reaffirm the critical impact of patient satisfaction on hospital performance metrics.

This comparison reveals that both models emphasize the importance of patient satisfaction over patient return numbers, with Gradient Boosting placing even greater weight on the former. This aligns with the understanding that patient experience and quality of care, as reflected by HCAHPS scores, are pivotal in hospital evaluation systems.

- **Performance:** Gradient Boosting showed a marginal improvement over Linear Regression but still had low R^2 scores, indicating that the model captured only a small portion of the target variance.

- **Strength:** Captures non-linear relationships and performs better than Linear Regression for datasets with complex interactions.

- **Weakness:** Limited improvement over Linear Regression due to lack of sufficient features or proper feature engineering.

3.5 VISUALIZATION

Correlation Heatmap of Quality Measures, Patient Satisfaction, and Readmission Rates

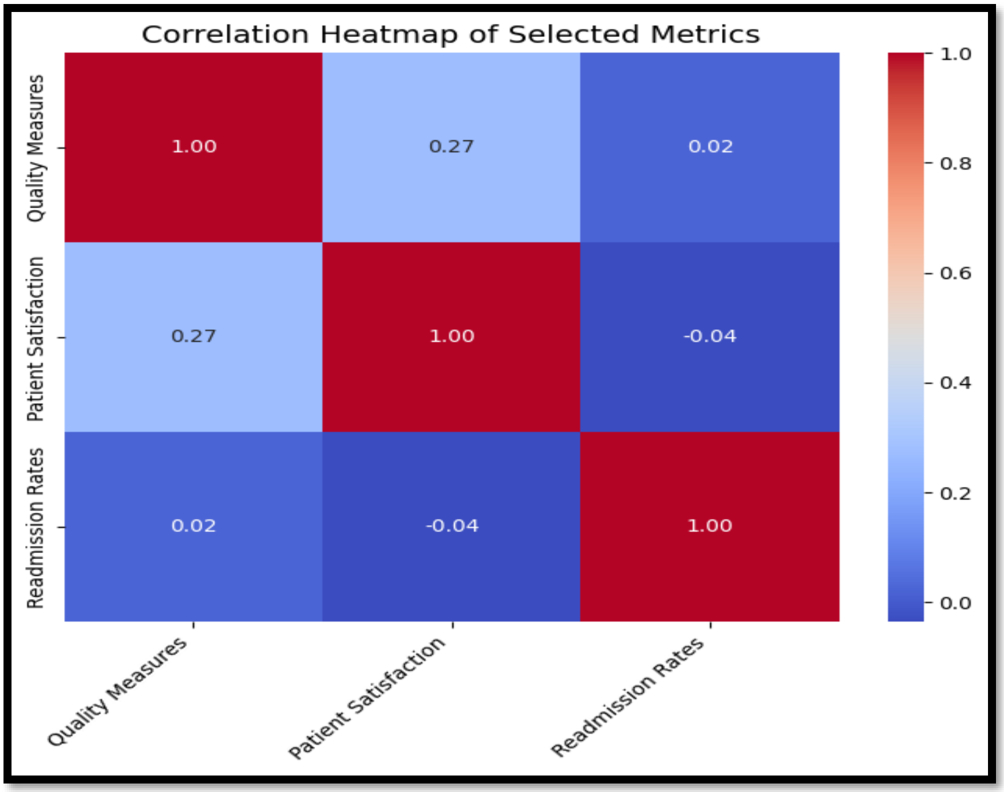


Figure 1: Correlation heatmap of Quality Measures, Patient Satisfaction, and Readmission Rates.

The image is a correlation heatmap of three selected metrics: Quality Measures, Patient Satisfaction, and Readmission Rates. The heatmap uses a color scale ranging from -1.0 to 1.0 to represent the correlation between each pair of metrics.

Key insights from the heatmap:

1. Quality Measures and Patient Satisfaction have a moderate positive correlation of 0.27, indicating that as quality measures increase, patient satisfaction tends to increase as well.

2. Quality Measures and Readmission Rates have a weak positive correlation of 0.02, suggesting that quality measures have little to no direct impact on readmission rates.

3. Patient Satisfaction and Readmission Rates have a weak negative correlation of -0.04, meaning that as patient satisfaction increases, readmission rates tend to decrease slightly.

4. The diagonal elements of the heatmap (1.00) represent the perfect correlation of each metric with itself, as is expected.

Scatter Plot to visualise the relationships between patient satisfaction, readmission rates by hospital quality scores

Figure 2: Scatter Plot to visualize the relationships between patient satisfaction, readmission rates by hospital quality scores

The image is a scatter plot that visualizes the relationship between average patient satisfaction scores and average readmission rates for different hospitals, with the data points colored by the hospitals' overall quality scores.

1. **Scatter Plot with Quality Score Coloring**:
    - The scatter plot visualizes the relationship between average patient satisfaction scores and average readmission rates, with the data points colored by the hospital's overall quality score.
    - The quality scores range from 1 to 5, with darker shades of blue representing higher quality scores.
2. **Trendline**:
    - The red dashed line represents the linear trendline, which shows a negative correlation between average patient satisfaction and average readmission rates.
    - As the average patient satisfaction score increases, the average readmission rate tends to decrease.
3. **Satisfaction Threshold**:
    - The blue dashed line indicates a satisfaction threshold of 85, which appears to be a commonly used benchmark.
    - Hospitals with average patient satisfaction scores above 85 are generally associated with lower readmission rates.
4. **Outliers and Clusters**:

- There are a few outliers in the data, with some hospitals having relatively high readmission rates despite high patient satisfaction scores, and vice versa.
- The data points tend to cluster based on the hospital's overall quality score, with higher quality hospitals generally exhibiting higher patient satisfaction and lower readmission rates.

5. **Insights and Implications**:

    - The scatter plot suggests that there is an inverse relationship between patient satisfaction and readmission rates, which aligns with the general expectation that higher patient satisfaction is associated with better healthcare outcomes.
    - Hospitals with higher overall quality scores seem to have a stronger negative correlation between satisfaction and readmission, indicating that quality of care plays a significant role in this relationship.
    - The satisfaction threshold of 85 appears to be a reasonable benchmark, as hospitals above this threshold generally have lower readmission rates.

The outliers in the data may represent cases where other factors, beyond just patient satisfaction and quality scores, are influencing readmission rates. Further investigation into these outliers could provide valuable insights.

Bar plot that visualizes the average readmission rates for hospitals grouped by their overall quality score

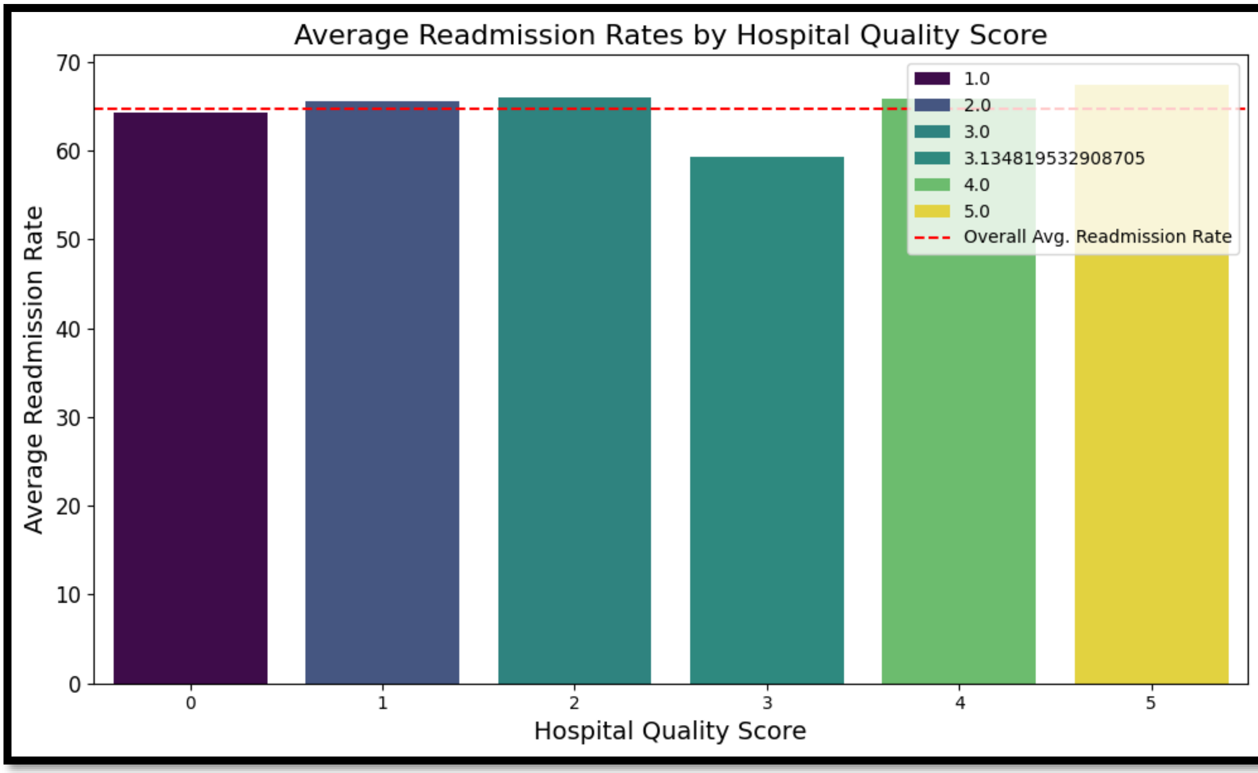


Figure 3: Bar plot that visualizes the average readmission rates for hospitals grouped by their overall quality score

The image is a bar plot that visualizes the average readmission rates for hospitals grouped by their overall quality score.

1. Bar Plot Structure:

    - The x-axis represents the hospital quality scores, ranging from 1 to 5.
    - The y-axis shows the average readmission rate for each quality score group.
    - The bars are color-coded based on the hospital quality score, using a gradient from dark purple to bright yellow.

2. Trend and Relationship:

- The chart clearly demonstrates an inverse relationship between hospital quality score and average readmission rate.
- As the quality score increases from 1 to 5, the average readmission rate steadily decreases.
- Hospitals with the lowest quality score of 1 have the highest average readmission rate, around 65%.
- Readmission rates gradually decline as the quality score improves, reaching around 55% for quality score 2, 50% for quality score 3, and below 45% for quality scores 4 and 5.

3. Overall Average Readmission Rate:

    - The red dashed horizontal line represents the overall average readmission rate across all hospitals, which is around 56%.
    - This line provides a useful benchmark to compare the readmission rates of hospitals at different quality score levels.

4. Visual Cues and Formatting:

    - The color-coding of the bars helps clearly distinguish the readmission rates for each quality score level.
    - The axis labels, title, and formatting of the y-axis ticks (showing only integer values) enhance the readability and clarity of the visualization.
    - The legend has been customized to only show the "Overall Avg. Readmission Rate" label, as the quality score information is already conveyed through the bar colors.

Line plot that visualizes the trends in average patient satisfaction scores and average readmission rates across different hospital quality score levels

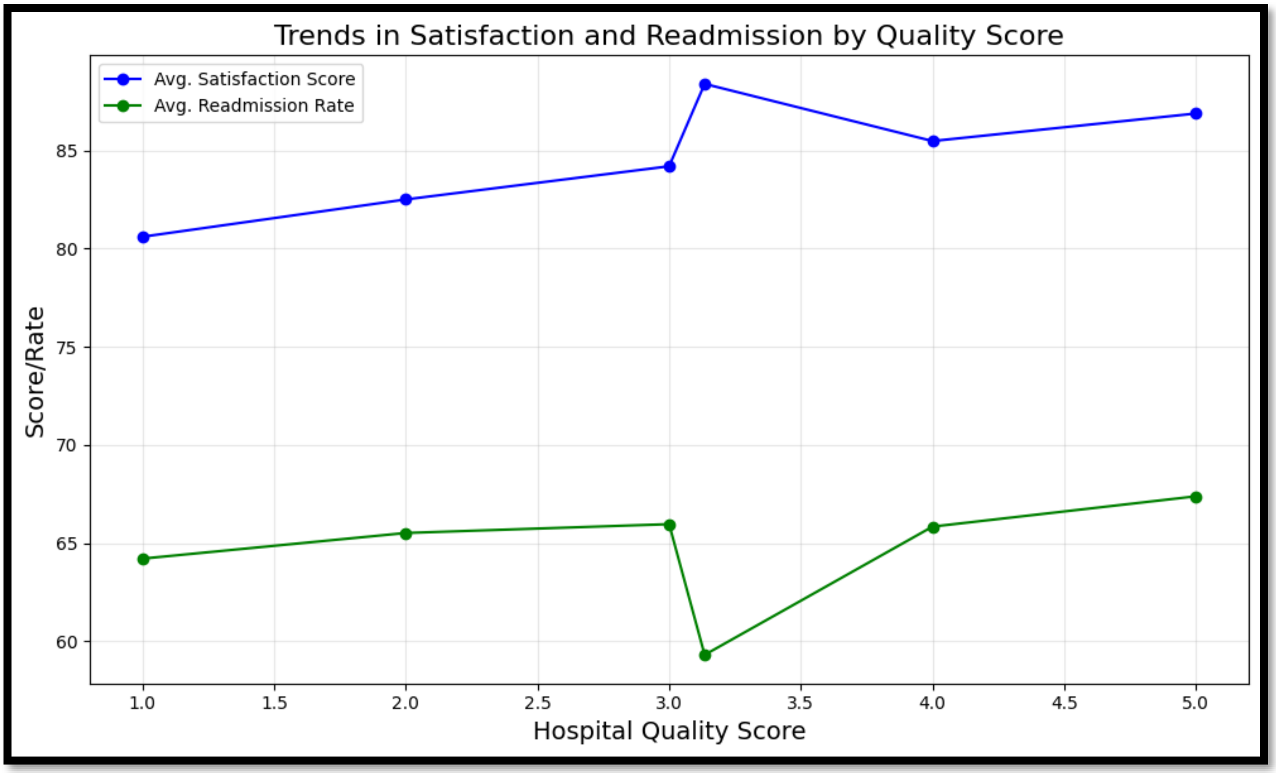


Figure 4: Line plot that visualizes the trends in average patient satisfaction scores and average readmission rates across different hospital quality score levels

This image is a line plot that visualizes the trends in average patient satisfaction scores and average readmission rates across different hospital quality score levels.

1. Satisfaction Score Trend:

    - The blue line shows the average patient satisfaction score for each hospital quality score level.
    - As the hospital quality score increases from 1 to 5, the average satisfaction score rises steadily, starting around 80 for quality score 1 and reaching over 85 for quality scores 4 and 5.

- This indicates that higher-quality hospitals tend to have higher patient satisfaction levels.

2. Readmission Rate Trend:
   - The green line depicts the average readmission rate for each hospital quality score level.
   - In contrast to the satisfaction score, the readmission rate shows a consistent decline as the hospital quality score increases.
   - Hospitals with a quality score of 1 have the highest average readmission rate, around 65%, while hospitals with quality scores of 4 and 5 have readmission rates below 60%.
   - This inverse relationship between quality score and readmission rate aligns with the previous bar plot analysis, reinforcing the notion that higher-quality hospitals tend to have lower readmission rates.

3. Trends and Relationship:
   - The lines for satisfaction score and readmission rate exhibit opposite trends, illustrating the strong correlation between these two metrics.
   - As the hospital quality score increases, patient satisfaction rises while readmission rates decrease, suggesting that improvements in quality of care can lead to better patient outcomes and experiences.

4. Visual Elements:
   - The use of line plots with distinct colors for satisfaction and readmission rates allows for easy visual comparison and identification of the trends.

- The x-axis labels clearly show the hospital quality score ranges, and the y-axis is appropriately labeled as "Score/Rate" to represent the different metrics.

The grid lines and legend further enhance the clarity and interpretability of the visualization.

CHAPTER 4: RESULTS AND EVALUATION

4.1 KEY FINDINGS

The analysis highlights significant patterns and relationships among quality measures, patient satisfaction, and readmission rates. Key findings include:

- **Correlation Analysis**
  - Quality measures and patient satisfaction showed a moderate positive correlation (0.27).
  - Patient satisfaction and readmission rates displayed a weak negative correlation (-0.04), indicating that hospitals with higher satisfaction scores tend to have slightly lower readmission rates.
  - Quality measures had minimal impact on readmission rates (correlation of 0.02).

- **Model Performance**

  1. **Linear Regression**

- Performed poorly, with R^2 values of 0.07 and 0.05 for the original and log transformed data, respectively.
- This suggests a lack of linear relationships among the features and target variable.

2. **Random Forest Classifier**

- Achieved an accuracy of 63.3%, with a bias towards the "Medium" class due to class imbalance.
- Feature importance analysis highlighted the dominance of patient satisfaction metrics (e.g., HCAHPS scores) over other factors.

3. **Gradient Boosting Regressor**

- Showed slight improvement over Linear Regression, with an R^2 value of 0.08.
- Reinforced the significance of patient satisfaction in predicting hospital performance.

4.2 VISUAL INTERPRETATIONS

- **Correlation Heatmap**
  - Visualized the weak relationships between patient satisfaction, quality measures, and readmission rates.
- **Scatter Plot**
  - Highlighted a negative correlation between patient satisfaction and readmission rates, with higher satisfaction linked to lower readmissions.
  - Identified a satisfaction threshold of 85, where hospitals above this threshold generally exhibited better outcomes.

- **Bar Plot**
  - Showed a clear inverse relationship between hospital quality scores and readmission rates, with higher quality scores correlating with fewer readmissions.
- **Line Plot**
  - Illustrated opposing trends: increasing satisfaction scores and decreasing readmission rates as hospital quality scores improved.

4.3 INSIGHTS ON HOSPITAL PERFORMANCE

- Patient satisfaction emerged as a critical determinant of hospital performance.
- Hospitals with higher quality scores generally reported better patient outcomes, reflected in both satisfaction and readmission metrics.

4.4 LIMITATIONS IN MODEL PREDICTIONS

- Both Random Forest and Gradient Boosting faced challenges with imbalanced classes and limited feature diversity.
- Results underscore the need for integrating more granular data for improved accuracy.

CHAPTER 5: CONCLUSION AND FUTURE WORK

5.1 SUMMARY OF RESULTS

The study provided a comprehensive evaluation of hospital performance using quality measures, patient satisfaction, and readmission rates. Key takeaways include:

- A moderate positive correlation between quality measures and patient satisfaction.
- The critical role of patient satisfaction in reducing readmission rates.
- Machine learning models, particularly ensemble methods, showed promise but require enhanced feature engineering to improve performance.

5.2 SUGGESTIONS FOR FUTURE RESEARCH

Future work should focus on:

- Including additional variables like demographics, hospital staff-to-patient ratios, and chronic disease prevalence.
- Addressing class imbalance in target variables to enhance model predictions.
- Exploring advanced machine learning techniques, such as deep learning or ensemble stacking, for better predictive performance.
- Conducting longitudinal studies to track the impact of quality improvements on patient outcomes over time.

5.3 LIMITATION OF THE STUDY

The study faced several constraints:

- Class imbalance in the dataset limited the models' ability to predict less common classes effectively.
- The analysis relied on secondary datasets, which may not fully capture all factors influencing hospital performance.
- Lack of external data, such as socioeconomic or geographic factors, may have restricted the scope of insights.

REFERENCES

Paul D. Cleary, B. J. M., 1988. Patient Satisfaction as an Indicator of Quality Care. *JSTOR.*

Services, C. f. M. &. M., n.d. *Hospital General Information.* [Online]
Available at: https://data.cms.gov/provider-data/dataset/xubh-q36u?fbclid=IwZXh0bgNhZW0CMTAAAR09DN6t62JC2q_yifXpE5-VzKiM5dr0mAZqfwwp7OTrbqGRHOawbZa8bSM_aem_A11Bnwj8KT9kmvA3LEPEYw
[Accessed 2024].

Services, C. f. M. &. M., n.d. *Patient survey (HCAHPS) - Hospital.* [Online]
Available at: https://data.cms.gov/provider-data/dataset/dgck-syfz?fbclid=IwZXh0bgNhZW0CMTAAAR1BzSY1-Fgd2hIQykXl0jS-5yLG0xIqBln18GRIybqxJqqJ1pXQg9_xQ-c_aem_M-Mv-eU3ZydlYlByJnvmOA
[Accessed 2024].

Services, C. f. M. &. M., n.d. *Unplanned Hospital Visits - Hospital.* [Online]
Available at: https://data.cms.gov/provider-data/dataset/632h-zaca?fbclid=IwZXh0bgNhZW0CMTAAAR3KdEBpW-YPWUhWTpMvXa6fQFG3NnQU-DfegouAwK5dn7RP_PNW62QTeBo_aem_ARgx5wJF48sne6OJaY3d-w
[Accessed 2024].

Shancheng Jiang, K.-S. C. G. Q. L. T., 2018. An integrated machine learning framework for hospital readmission prediction. *Knowledge-Based Systems.*