Mercer University   
Stetson-Hatcher School of Business

**Predicting Hospital ratings**

Exploring significant factors affecting hospital ratings

Team members: Hemasree Diguva Sodum , Nilufar Nosirjonava

Advisor: Dr. Wei Xiong  
  
 **Fulfilment statement:** A capstone project for partial fulfillment of the requirements of Master of Science in the Business Analytics program offered by Stetson-Hatcher School of Business at Mercer University

**Introduction:**  
In today’s health care industry, understanding patients' hospital experiences is crucial for enhancing care standards. The Hospital Consumer Assessment of Healthcare Providers and Systems (HCAP) dataset provides valuable data and patient experience feedback regarding different aspects of their hospital journey.

In the healthcare realm, patients’ preferences and expectations have changed a lot. People are seeking not only effective treatments; however, they are also prioritizing the experience of their individual wellness and holistic health. For a real-life example, a patient receiving proper medical treatment also values a safe and welcoming atmosphere and the environment where caring staff's additional humanity.

Hospital ratings are one of the first steps that lead patients to make decisions when they are seeking healthcare services. The first impression of patients is usually built by hospital ratings. Moreover, hospital ratings are representative of hospitals quality and performance. It allows hospitals to compare their performance with other hospitals, which drives improvements in the quality of healthcare. Most importantly, ratings help identify areas requiring improvement and meet individual’s requirements and preferences. These all-important points motivated us to work on the project topic of “Predicting hospital ratings by exploring significant factors effects on hospital ratings."

The healthcare industry worldwide strives to understand and develop patient experiences. Hospital ratings are traditionally measured by patients’ satisfaction levels, and they are considered a pivotal metric in evaluating healthcare service quality. However, it is challenging to identify what factors influence hospital ratings. Patients' experiences are impacted by many factors, like the advanced level of medical areas, how well communication between hospital staff and patients is implanted, the industry’s benchmark rates in mortality, readmission, safety, and the overall atmosphere in the hospitals. Capturing all these complexities and exploring their influences on hospital ratings requires a comprehensive data-driven approach.

Our project purpose is to delve into the HCAP dataset and predict hospital ratings through various dataset details like patients’ experience, hospital operations, benchmark standards, and administrative details. Our main goal is to explore the relationship within the data environment and predict hospital ratings. Finally, when hospitals get rated, it represents everything that hospitals provide and how they perform, such as how friendly and helpful staff are, how long the waiting time is, and whether patients are satisfied with the cleanliness and safety of the facilities. Also, we want to employ this dataset to figure out why patients rate hospitals with higher and lower rates, what makes them satisfied and not satisfied, and which factors have the highest impact on patients and hospitals. Moreover, through our project, hospitals can obtain insights on where to allocate their resources, and with accurate forecasting of factors that influence them, hospitals can target their areas to improve without extensive spending. Consequently, strategically investing in specific areas will lead to increased patients’ retention and positive word-of-mouth referrals. This is one of the main steps in generating long-term business values, a positive reputation, and patient loyalty. There is no doubt that a higher rate of hospitals creates more confidence, trust, and a stronger connection with patients. Taking all the points mentioned into account, we strongly believe that our project not only predicts hospital rates but also provides tangible business value by finding targeted improvements that optimize patient satisfaction for long-term care. The current report covers our methodology, results, and significant findings in the context of the healthcare environment for patients’ satisfaction.

**Literature review:**

Hospital ratings are a pivotal metric in healthcare quality evaluation, and they are often affected by patient satisfaction. Previous research also explored versatile patient experience factors like communication effectiveness, care quality, and various environmental factors. Many different studies have used the HCAP dataset to predict hospital ratings and According to Mazurenko (2017), certain patient-level and hospital-level predictors were associated with higher patient satisfaction or lower patient satisfaction. The same dataset of 2007–2013 was implemented by Rupinder et al. (2015) to investigate changes in patient satisfaction using all available variables, and in his analysis, satisfaction scores were measured by doctor communication, which was an interesting approach. He grouped hospital scores quarterly based on satisfaction scores. In Mazurenko’s (2017) analysis, some variables had mixed relationships with patient satisfaction. The dataset Mazurenko (2017) utilized had the same 10 main questions that measure patient satisfaction across domains such as doctor and nurse communication, responsiveness of staff, pain management, hospital environment, discharge information, and communication about medication. However, our dataset did not include the demographic variables that some researchers' datasets employed. Moreover, in the HCAP dataset we employed between 2016 and 2020, we did not include any market-related variables that Mazurenko (2017) had. Several predictors, such as hospital ownership types and EHR criteria (electronic health records), were associated with hospital scores. The same result was met in our model insights by showing that private hospital ownership and hospitals with electronic health records were considered factors highly affecting higher hospital ratings and higher patient satisfaction. In comparison to our analysis, Mazurenko's (2017) methodologies implemented meta-analysis guidelines as a basic model, and as a second step goal, they focused on patient satisfaction measurement, which we did not add to our analysis because of time restrictions. However, Rupinder et al. (2015), in their research, employed a multilevel model to check the correlation between hospital observations, and they reported about 2273 hospitals in 2007. A publicly available dataset from different sources was employed for the study of Rupinder et al. (2015). Patient satisfaction scores and hospital characteristics were taken from the HCAP dataset, and the same variables we had, such as hospital ownership, states (locations), and years, were observed. Also, the author was able to obtain demographic information on the population from 2010 from the American Community Survey for the next 5 years. According to his model insights, overall patient satisfaction scores have shown high significance with a p value less than 0.001, and during the past 7 years, this index has increased. Similar to Rupinder., et al (2015) results, our model insights demonstrated that doctors and nurses communication is highly significant in hospitalsssz rating Moreover, the dataset utilized by Mazurenko (2017) covers ethnic and racial demographic group data, which led to interesting insights in Mazurenko’s research. According to their results, Asian and Native American patients demonstrated lower patient satisfaction levels compared to white Americans. However, black and Hispanic patients reported better satisfaction with nurse and doctor communication but less satisfaction with staff responsiveness and discharge information. However, we were not able to find out any demographical analysis results in our research since our dataset did not include any demographic information about patients, even gender and age. Even the pain management predictor, which was included in Mazurenko's (2017) analysis, was highly associated with high patient satisfaction. In our dataset, we had to remove this variable because only two initial years included it; later years from 2008 until 2020 did not include any information about pain management. However, we had the same results in specific variables as Mazurenko's (2017) results, such as readmission and mortality. They were negatively associated with our target variable, hospital ratings. According to Rupinder et al. (2017), they employed different multilevel models: In the first model, he aimed to examine trends and added years to the predictor list. The second model included quartiles of satisfaction, income, insurance variables, and beds. All his models were built using NLME packages. He mainly focused on seeing changes in doctors’ communication and examining the significance of this feature. His result was the same as ours, and psychiatric communication was highly correlated with patient satisfaction and the highest rates of hospitals located in the South.

**Data:**

The dataset contains high-quality data about the hospitals that are receiving payments from CMS (center of Medicare and Medicaid services). CMS is a US federal agency that works to ensure all individuals nationwide make better healthcare decisions to receive high-quality care. CMS rates hospitals throughout the United States according to the findings from patient satisfaction surveys to do this, CMS uses the HCAHPS survey, which gathers various data about the medical facility they attended. CMS assigns a numerical value to each response, and then determines the average score for each question. The process offers a consistent method for evaluating patient experiences at different facilities and assigning ratings. CMS releases this data every year publicly for the hospitals to use this information and improve their quality of services, healthcare standards and for the public to choose better healthcare facilities.

**Data Sources:**

The original source link is “https://data.cms.gov/provider-data/archived-data/hospitals”. The Centers for Medicare & Medicaid Services is the main source of data (CMS). It is available on Kaggle as well. Every year, hospitals that get funding from Medicare and Medicaid are required to submit CMS with high-quality data. Data set includes the period from 2016 to 2020. Each hospital can obtain a score of 1 to 5, where 5 stars indicate the greatest quality.

**Data Characteristics:**

The dataset comprises location and identification-related data on hospitals in the United States, such as Facility Name, Facility Id, and geographic data about the facilities, such as address, city, county, and state. Additionally, patient ratings, patient responses to questions, the number of surveys completed, and the percentage of patients who provide different types of information are all captured in the data set related to patient satisfaction. Additionally, temporal data like the start and end dates as well as the corresponding year are included. captures various hospital characteristics, such as hospital ownership, types, and whether they offer emergency services. It also includes various hospital performance metrics, such as mortality, safety, readmission, patient experience, effectiveness, and timely use of medical imaging, in comparison to national benchmark data. In general, the dataset includes patient experience, hospital performance measures, and hospital information.

**Data cleaning and Preprocessing:**

A single data frame named "hosp" was created by joining many datasets that span the years 2016 to 2020. After merging our dataset had 43 variables with 1.6M observations in them. Certain variables were eliminated from the dataset after we analyzed the data dictionary and discovered they were irrelevant for our analysis, which led to a reduction in the amount of data. Notable exclusions were variables like Start Date, End Date, and Zip Code, variables that ends with footnote etc. Some columns were renamed for improved readability.

**4.1 Data Dictionary:**

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Variable Type** | **Description** |
| Facility.Name | Character | Name of the healthcare facility |
| Facility.Id | Integer | Unique Id of each facility |
| Address | Character | Address of the healthcare facility |
| City | Character | City where the healthcare facility is located |
| State | Character | State where the healthcare facility is located |
| County.Name | Character | Name of the county where the facility is located |
| Topics | Character | Survey Questions related to patient care experiences |
| Ans.desc | Character | Answer description related to the survey topic |
| Patient.Survey.Rate | Numeric | Rating given by patients for the survey topic |
| Ans.perc | Character | Percentage related to the survey answer |
| No\_Surveys | Numeric | Number of surveys conducted |
| Response.rate.perc | Numeric | Percentage of response rate for the surveys |
| Year | Integer | Year in which the survey was conducted |
| Hospital.Type | Character | Type of hospital |
| Hospital.Ownership | Character | Ownership type of the hospital |
| Emergency.Services | Factor | Whether hospital has emergency services or not |
| EHR\_criteria | Factor | Electronic Health Record criteria indicator |
| Hosp.rating | Numeric | Overall rating of the hospital |
| Mortality | Character | Mortality rate compared to the national average |
| Safety | Character | Safety rating compared to the national average |
| Readmission | Character | Readmission rate compared to the national average |
| Patient.experience | Character | Patient experience rating compared to the national average |
| Effectiveness | Character | Effectiveness rating compared to the national average |
| Timeliness | Character | Timeliness rating compared to the national average |
| Efficient.imaging | Character | Imaging efficiency rating compared to the national average |

Two dataframes, named "star\_ratings" and "LMVs," were created by removing rows that had the values "Not Available" or "Not Applicable" for linear mean values and star ratings, respectively. “Hospitals\_data” dataframe has been created using only the required columns for the analysis from “star\_ratings” dataframe. In order to maintain data integrity, we additionally resolved the missing, "Not Available" entries in the "Hosp.rating" column. Due to the large size of the dataset, records with "Not Applicable" and "Not Available" values were considered irrelevant and were removed rather than having replaced with mean or median values to avoid bias.

**Data Extraction:**

We integrated the 'star\_ratings' and 'LMVs' datasets into the main 'hospitals\_data' dataframe after extracting specific star ratings and linear mean scores pertaining to different areas of hospital performance in order to enrich our dataset. Filtering the relevant rows based on the specified topic and merging them based on facility ID and year allowed us to extract topics like quietness, recommendation, pain management, communication about medications, staff responsiveness, cleanliness, overall hospital rating, care transition, and communication about nurses and doctors. And those columns were converted to numeric for exploratory analysis purpose. We also dropped an extracted variable called “pain\_management” because it was available only for the years 2016 and 2017.  
  
**Data dictionary after extraction and creation of new variables:**

|  |  |  |
| --- | --- | --- |
| Variable Name | Variable Type | Dictionary |
| 1. Cleanliness\_Star\_rating 2. Nurse\_communication\_Star\_rating 3. Doctor\_Communication\_Star\_rating 4. Staff\_responsiveness\_Star\_rating 5. Pain\_Management\_Star\_rating  6. Communication\_About\_Medicine\_Star\_rating  7. Discharge\_Info\_Star\_rating  8. Care\_Transition\_Star\_rating  9. Hospital\_Rating\_care\_rating  10. Quiteness\_Star\_rating  11. Recommendation\_Star\_rating | Numeric | Ratings provided by patients for the questions |
| 1. Cleanliness\_LinearMeanScore  2. Nurse\_communication\_ LinearMeanScore 3. Doctor\_Communication\_ LinearMeanScore 4. Staff\_responsiveness\_ LinearMeanScore 5. Pain\_Management\_ LinearMeanScore  6. Communication\_About\_Medicine\_LinearMeanScore  7. Discharge\_Info\_ LinearMeanScore  8. Care\_Transition\_ LinearMeanScore  9. Hospital\_Rating\_care\_ LinearMeanScore  10. Quiteness\_ LinearMeanScore  11. Recommendation\_ LinearMeanScore | Numeric | Linear Mean Score assigned for the questions |

In addition, a duplicate data frame called "hospitals\_data1" was created and has been converted to factor for comparison purposes. To facilitate effective counting and summarizing. A constant column "n" was added to the original "hospitals\_data" and assigned value "1". Also calculated mean values for different rating and survey response variables based on Facility Name and Facility ID and assigned them to 'hospitals\_data\_agg,' which can provide a summary for every hospital, enabling performance analysis and comparisons across several dimensions.  
  
**Target variable transformation:**  
Hospital rating was our target variable which was ordinal categorical variable. We decided to transform it to binary variable as it is simple and straight forward to interpret binary outcome instead of multi class. We created new variable with ratings greater than or equal to 4 as positive group (Binary “1”) and less than 4 as negative group (Binary “0”). The reason behind this segregation was driven by industry standards as well as factors pertaining to patient satisfaction, public perceptions, and the healthcare industry's reputation. A rating of four or above is typically considered positive, as it is associated with increased patient satisfaction and enhances the public's view and reputation of healthcare facilities. Ratings below 4 on the other hand indicate a possible area for improvement and are categorized as negative, indicating a lesser degree of satisfaction. In addition to simplifying the interpretation, this binary transformation aligns with industry norms and enables a targeted investigation of the factors that influence positive as well as negative ratings about the healthcare system.

|  |  |
| --- | --- |
| A graph showing the distribution of hospital ratings  Description automatically generated | After the target variable (hospital ratings) transformation to binary, we noticed a disparity in class: 70% of instances go into Class 0 (ratings less than 4), and 30% fall into Class 1 (ratings of 4 or higher). Because positive ratings are generally less common in healthcare evaluations, we purposefully decided not to balance the classes in order to maintain an ideal proportion. |

The real occurrence of positive ratings may be misinterpreted because of balancing them using any data balancing techniques like oversampling or under sampling as it may introduce bias in model results.  
  
**Relative Weight Analysis**:  
Relative weight analysis is a method used to assess the contribution of each independent variable towards target variable and R square and rank them based on their contribution to R square. We used this method in our project as part of variable selection process before running the models. As we have seen earlier from the heat map "Recommendation\_LinearMeanScore", "Hospital\_rate\_LinearMeanScore" are highly correlated. Therefore, using this method helps to choose the variable among the two to include in our models.   
  
A graph of a bar graph

Description automatically generated with medium confidenceResults of weight analysis showed us that “Recommendation\_LinearMeanScore” has more contribution towards target variable compared to “Hospital\_Rating\_LinearMeanScore” so we decided to use only recommendation linear mean score variable. RWA results showed Recommendation\_LinearMeanScore, care\_trans\_LinMeaScore, staff\_response\_LinMeaScore and nurse\_comm\_LinMeaScore are being the most weighted variable, which means these variables are more influential in predicting model outcome compared to others independent variables in the model.

Our data mining task is to predict the hospital rating whether a hospital receive positive rating or negative rating and to investigate the important variables influencing the hospital ratings Positive ratings are those that are higher than the threshold of four/ greater than 4 (>= four), while negative ratings are those that are lower than four. Our goal is to categorize hospitals according to their overall rating. This is a classification problem where the hospital rating is the target variable and the independent variables are the hospital identification factors (facility name, city, state, etc.), the patient care experience factors (cleanliness, nurse communication, quietness, etc.), and the hospital performance factors (mortality, safety, readmission, etc.). We employed logistic regression technique since it is effective for classifying hospitals as either positive or negative rated hospitals.

We want to employ logistic regression technique as it is simple to use and effective at predicting binary outcomes, we used it to categorize hospitals into positive and negative ratings. Additionally, we will be employing the random forest technique in addition to logistic regression since it can capture complex relationships in the data. Also, Support Vector Machines (SVM) as they are useful for handling dataset with high-dimensions, linear and nonlinear relationship data. We will be using different performance metrics like accuracy, sensitivity, specificity, precision, f1 score etc to assess/evaluate the performance of our models. Since we are working with class imbalance

We divided our dataset into a training set, with 70% of the data, and a testing set, with 30%. This method helps us, evaluate our classification models on a separate sample of data after training them on training data set, which guarantees accurate assessment and generalization to new/unseen data.

**Descriptive Analytics**

In the section of descriptive analytics, we explore the subject to understand it more extensively. Statistics, visual aids, and some insights aim to clarify the story behind hospital ratings and factors affecting them.

|  |  |
| --- | --- |
|  | According to hospital ratings distribution, hospitals in the U.S are rated between 1- and 5-star rates and more than 50% of hospitals are 3 star rate hospitals and rates 2 and 4 are also common rates. The least group of hospitals is 5 star and in entire U.S only 118. |

|  |  |
| --- | --- |
|  | According to insight from the histogram of hospital types, majority types of the hospitals in U.S is Voluntary non-profit-Private hospitals, followed by Proprietary, Voluntary non-profit-Church and Governmental and Governmental-Local Hospitals. |

A comparison of a graph

Description automatically generated with medium confidenceA group of graphs showing different sizes of graphs

Description automatically generated with medium confidence

A group of graphs showing different sizes of data

Description automatically generated with medium confidence

The histograms above are distribution patterns of various aspects related to specific dimension, such as cleanliness, nurse communication, staff responsiveness, communication about medicine, discharge information and others, examining these histograms all together indicates frequency of ratings falling between specific ranges and according to visual aids can be noticed that all features are normally distributed.

A group of graphs showing different types of medical rating

Description automatically generated with medium confidence

A group of numbers and symbols

Description automatically generated with medium confidence

The series of plots demonstrate the relationship between different aspects of patient care Linear Mean Values, and overall hospital ratings. By examining the plots, we can observe which hospital services scores are significant in different levels of hospitals. Also, we can see if the service is higher ranges, hospitals rate is also getting higher.

Further analysis of Correlation plot is proving that factors Recommendation and Hospital Rating are strongly correlated that further will lead to multicollinearity. To avoid this issue, we will not be adding both the variables at a time while training the model.  
A graph with blue circles and red text

Description automatically generated

**Methodology:**

**Logistic Regression models:**We used logistic regression model as it is a classification algorithm which can be used for predicting binary outcome. Advantages are It is easy to understand and implement logistic regression models. It works well when the variables in the dataset have linear relationship. Gives details about contribution of each independent variables towards target variable. Disadvantages include logistic regression might not be accurate if the sample size is small, it assumes the linearity between dependent and independent variables, dependent variable must be dichotomous.

Initially, logistic regression model was implemented on entire dataset instead of on train dataset to understand the general factors influencing hospital ratings. From the results we found that cleanliness, Nurse Communication, Staff Response, Recommendation, Mortality, Timeliness, Safety, Readmission, Efficient Imaging, and different categories of Hospital Ownership were significant factors based on p-value associated with each coefficient. Coefficient indicates influence of each independent variable on the log-odds of dependent variable. Coefficient of cleanliness is 0.107 indicating for each unit increase in cleanliness score, the log-odds of hospital rating being positive increase by 0.107. which means hospitals with better cleanliness tend to receive higher ratings. The coefficient of Mortality (Below the national average) is -5.845, indicating that hospitals with mortality rates below the national average are associated with a substantial decrease in the log-odds of higher ratings. which means lower mortality rates strongly predict lower hospital ratings.

**Features Selection:**

We used Forward and Backward Feature Selection method to select the most important factors for predicting hospital ratings since our dataset has many variables in order to find the most relevant variables that influence the hospital ratings. From the results of forward and backward selection method we found that cleanliness, nurse communication, staff communication, discharge information, care transitions, recommendation scores, mortality rates, and effectiveness, timeliness, safety, and readmission metrics are key factors influencing the overall hospital ratings.   
  
We trained our model on training dataset with hospital rating as dependent and the variables chosen from using feature selection technique as independent variables. The coefficients and significance levels of these variables are evaluated in order to comprehend their influence on hospital rating. The model performance can be accessed using error metrics like RMSE, MAE, MSE and other performance metrics like precision, F1 Score, sensitivity, specificity and accuracy as they help in evaluating the model performance more accurately.

**Random forest:**Random Forest is a machine learning algorithm that can be used for both regression and classification problems. It works by randomly creating many decision trees and combines them to make accurate predictions. Advantages of using ransom forest technique includes it reduces overfitting and improves accuracy, works well with large dimension dataset, handles different data types. When it comes to hospital rating prediction, the Random Forest model performs good. It is fairly successful at identifying high rated/positive rated hospitals but not very successful at identifying low rated hospitals. Additionally, the variable importance plot shows which variables are most important for predicting hospital rates in a random forest model. Important variables include Readmission, Safety, and Mortality significantly impacting the model’s accuracy. Variables like Staff Response, Recommendations, and Nurse Communication also play important roles, while variables like Effectiveness and Timeliness have comparatively less impact. The model performance can be accessed using performance metrics like precision, F1 Score, sensitivity, specificity and accuracy as they help in evaluating the model performance more accurately.  
  
**Support Vector Machine (SVM):**

Support Vector Machines (SVM) are set of supervised learning algorithms used for classification, regression tasks. We used in our project to predict hospital ratings (target variable), variables related to patient care experience, hospital ownership, hospital performance related factors as independent variables. SVM's advantages is that it performs better when processing high-dimensional data. It is also robust to noise in the data and can classify previously unseen data with great accuracy. Drawbacks include high computing costs and unsuitability for datasets with many missing values. Utilized a linear SVM to classify hospitals into positive and negative ratings based on the given independent variables since the variables in our dataset have linear relationship. The model performance can be accessed using performance metrics like precision, F1 Score, sensitivity, specificity and accuracy as they help in evaluating the model performance more accurately.

**Empirical Results:  
Accuracy:** Measures the ratio of correctly predicted instances to total instances.

**Sensitivity/Recall:** Measures the proportion of actual positive instances that were correctly identified by the model.

**Specificity:** Measures the proportion of actual negative instances that were correctly identified by the model.

**Precision**: Proportion of predicted positives that were actually positive.

**F1 Score:** Indicates the balance between precision and Recall.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Sensitivity | Specificity | Precision | Recall | F1 Score |
| Logistic | 0.8792 | 0.9080 | 0.8011 | 0.9254 | 0.9080 | 0.9166 |
| Random | 0.8764 | 0.9315 | 0.7366 | 0.8997 | 0.9315 | 0.9166 |
| SVM | 0.8749 | 0.9183 | 0.7647 | 0.9083 | 0.9183 | 0.9133 |

Logistic regression performed well across different metrics. With an accuracy of 87.82%, the model has 88% of the time correctly anticipated the outcome. Sensitivity, which measures the model's capacity to accurately forecast positive occurrences, indicates that the model successfully recognized positive/highly rated hospitals 90.80% of the time. The specificity rating of 80.11% signifies the model's capacity to detect negative occurrences. A precision of 92.54% indicates a low rate of false positives, meaning that hospitals predicted by the model expected to do well and are more likely to receive good ratings. The F1 score which balances precision and recall, was 91.66% demonstrated a good trade-off between false positives and false negatives.

Random forest also performed well across different metrics. With an accuracy of 87.64%, the model has 88% of the time correctly anticipated the outcome. A greater sensitivity score indicates the model's stronger ability to detect positive/high-rated hospitals (93.15%), as measured by the model's ability to properly anticipate positive instances. The specificity number, which was lower at 73.66%, shows how well the model could recognize negative situations. A precision of 89.97% implies a low false positive rate, meaning that hospitals that the model projected to be high performing are more likely to receive good ratings. With an accuracy and recall balance of 91.66%, the F1 score demonstrated a balanced performance in terms of false positives and false negatives.  
  
SVM also demonstrated a balanced performance across several metrics. With an accuracy rate of 87.49%, the model has accurately predicted the outcome 87.49% of the time. A greater sensitivity score indicates the model's superior ability to detect positive/high-rated hospitals (91.83%), as measured by the model's ability to correctly forecast positive cases. The specificity number, which was 76.47%, shows how well the model could recognize negative situations. A precision of 90.83% suggests a low false positive rate, meaning that hospitals that the model expected to do well are more likely to receive positive/high ratings. With an F1 score of 91.33%—a balance between recall and precision—the performance was well-balanced in terms of false positives and false negatives.  
  
After comparing the three models we chose logistic regression as better choice because Logistic Regression showed a strong performance by maintaining a good ratio of specificity to sensitivity. It demonstrated exceptional performance in preventing false positives and false negatives, with a precision of 92.54% and an F1 score of 91.66%, indicating it is better choice. Although there was a trade-off in specificity, Random Forest's superior sensitivity of 93.15% demonstrated a significant capacity to identify strong performing hospitals. Sensitivity (91.83%) and specificity (76.47%) were well-balanced, making SVM a competitive option too.  
  
**Conclusion:**After comparing and analyzing the results of three models we chose logistic regression as best model to predict the hospital ratings. since our objective was to develop a prediction model for hospital ratings; therefore, it is critical to have a well-balanced model, as incorrect classification may lead to patients losing faith in hospitals and negatively impacting their reputation. With a precision of 92.54% and an F1 score of 91.66%, Logistic Regression is the best option when it comes to balancing sensitivity and specificity. It also performed well overall by reducing the number of false positives and false negatives.

**Recommendations:**According to our model insight we were able to give recommendations to healthcare industries to concentrate areas mentioned below.  
Hospitals must improve patient services in the realms such as Cleanliness, staff responsiveness, doctors, and nurse communication since these areas were found one of the most important in hospitals rate. Investing training programs for various patient services would be very helpful for hospitals to develop staff skills for better connection and comprehension with patients.

Hospital performance variables such as mortality, readmission and safety are crucial metrics that determine hospital ratings. Hospitals always should strive to reduce performance metrics above mentioned by clinical protocol enhancements, comprehensive discharge planning, transitional care programs and patient safety and education by encouraging patients to be active in their care and safety.

Optimizing patient flow, managing patient time to visit with doctors and different procedures always should be in continuous improvement to decrease patients’ waiting time. Also doctors and nurses should always have transparent and honest communication with patients.

**References:**

1. Rupinder K, Mann MD “Effect of HCAHPS reporting on patient satisfaction with physician communication”, Journal of Hospital medicine, DOI: 10.1002/jhm.2490.
2. KelechiE Okonta, DaprimS Ogaji, “Predictors of patient satisfaction with surgical care in a low-middle-income country”, International Journal of Academic Medicine, 10.4103/IJAM.IJAM\_132\_20, 7, 4, (233), (2021).
3. Beata Gavurova, Jan Dvorsky, Boris Popesko, “Patient Satisfaction Determinants of Inpatient Healthcare”, International Journal of Environmental Research and Public Health, 10.3390/ijerph182111337, 18, 21, (11337), (2021).
4. KelechiE Okonta, DaprimS Ogaji, “Predictors of patient satisfaction with surgical care in a low-middle-income country”, International Journal of Academic Medicine, 10.4103/IJAM.IJAM\_132\_20, 7, 4, (233), (2021).
5. Olena Mazurenko, PhD, MD, “Predictors of Hospital Patient Satisfaction as Measured by HCAHPS: A Systematic Review”, Health Policy and Management Department, DOI: 10.1097/JHM-D-15-00050.
6. Kareem J. Kebaish, Michael R. Mercier "Spine Surgery HCAHPS Patient Satisfaction Survey Results Inversely Correlate with Survey Response Time, Spine", 10.1097/BRS.0000000000003974, 46, 18, (1264-1270), (2021).

**Appendix:**

|  |  |
| --- | --- |
| **A group of graphs showing different sizes of graphs  Description automatically generated with medium confidence** | **A group of graphs showing different sizes of data  Description automatically generated with medium confidence** |

|  |  |
| --- | --- |
| **A graph of different sizes and levels  Description automatically generated with medium confidence** |  |

**A graph of different sizes and colors

Description automatically generated with medium confidence**

**Logistic Regression:**

**A screenshot of a computer code

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A white rectangular object with white lines

Description automatically generated with medium confidence**

**Cross validation results  
A white background with black text

Description automatically generated**

**Random Forest:**

|  |  |
| --- | --- |
| **A graph with numbers and a line  Description automatically generated with medium confidence** |  |

**A screenshot of a computer

Description automatically generated**

**A white rectangular object with white lines

Description automatically generated with medium confidence**

**A screenshot of a computer

Description automatically generated**

**Support Vector Machine:**

**A screenshot of a computer program

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A white rectangular object with white lines

Description automatically generated with medium confidence**

**A screenshot of a computer

Description automatically generated**