CMS STAR RATING & PROVIDER ANALYSIS

Abhinav Sharma
abhi18av@gmail.com

Lipi Chaturvedi
clipi2013@gmail.com

Abstract

- The primary objective of the Overall Hospital Quality Star Ratings project is to develop a statistically sound methodology for summarizing information from the existing measures on Hospital Compare in a way that is useful and easy to interpret for patients and consumers.

- With the increase in the data around us nowadays, people are expressing their views or sentiments about their experience with products around them. Reviews are having a great influence on people and decisions made by them. This has led researchers and market analyzers to analyze the opinions of users in reviews and model their preferences accordingly. These ratings usually vary on a scale from one to five (stars) or very bad to excellent. In this Capstone Project we address the problem of attributing a numerical score (one to five stars) to all the hospitals in US based on multiple factors.

Stages

- **Data Retrieval :** Collected the data from the Hospital Compare Website.
- **Data Understanding :** Identified important measures for all the groups.
- **Data Preparation :** Created a grouped master dataset for all the groups, removed unrequited columns, duplicates and processed outliers. And then created a master dataframe with 72 features, merged all the groups data into one.
- **Data Modelling :** Performed supervised and unsupervised learning to analyze the best model which would give a high accuracy and be efficient in modelling.
- **Recommendations :** Provided recommendations to improve the star rating for Evanston Hospital (Provider ID : 140010).

Data Understanding : 7 Groups and 72 Measures

Under hospital compare, the data is organised into 7 groups which helps customer in making a informed decision. The groups are as follows along with their weightage :

- Mortality, Readmission, Safety of Care, Patient Experience (22% weightage groups)
- Timeliness of care, Effectiveness of care, Medical Imaging Efficiency (4% weightage)

Measures

1. Positive Measures : Communication with doctors, responsiveness of hospital staff, patient given appropriate vaccines etc.
2. Negative Measures : Readmission, complications after surgery, mortality measures, etc.

Data Understanding : Quality Issues

Data format

- The original data is in ‘wide-format’ in 55 files which was converted into master file such that each row represents a provider and columns represents measures.

Missing value

- Removed columns having more than 50% NA value in the dataframe and replaced NA values with median.
- Removed all the records having NA under the column Hospital.overall.rating, such rows were approx 24% of the entire master dataset.

Standardisation of Measure Scores

- Transformed all the measure’s score into a common scale using the StandardScaler.

Overview Of Methodology - 5 Step Approach

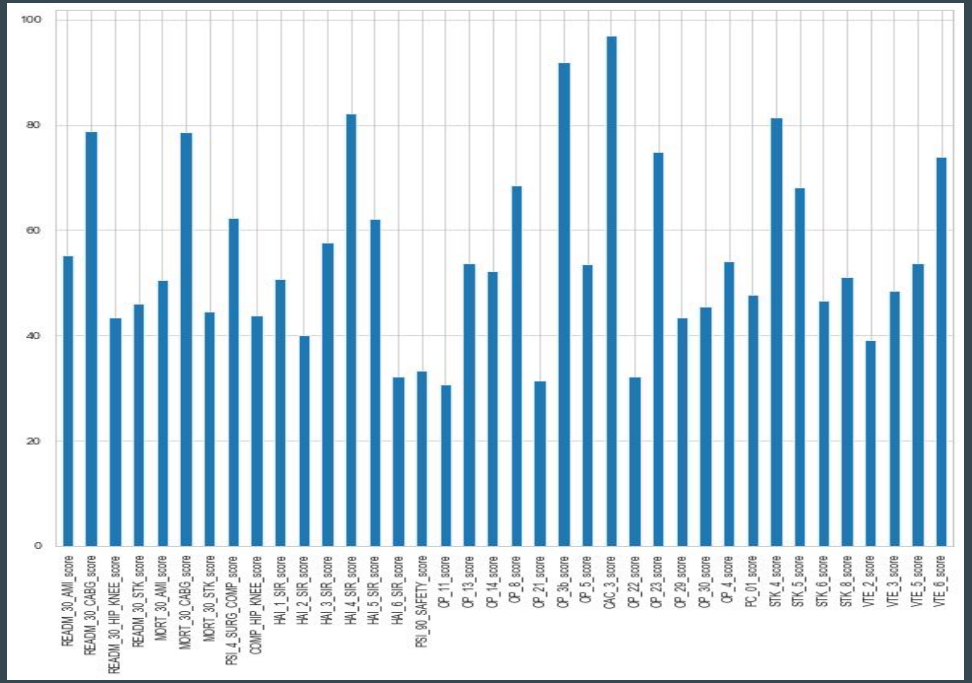
- The measures are first selected based on their relevance and importance as determined through stakeholders and expert's feedback.
- All the included measures are standardized to be consistent in terms of direction, magnitude, and then standardized measures are organized into seven groups according to measure type.
- For each group, Factor Analysis is used to estimate a group score for each hospital reporting measures in that group.
- A weight is applied to each group score, and all available groups are averaged to calculate the hospital summary score.
- Finally, to assign a Star Rating, the hospital summary scores are organized into five ordered categories using a clustering algorithm.

Data Overview

Analysing the dataset for missing

Data revealed the need for

- Imputing sparse measures
- Selecting dense measures

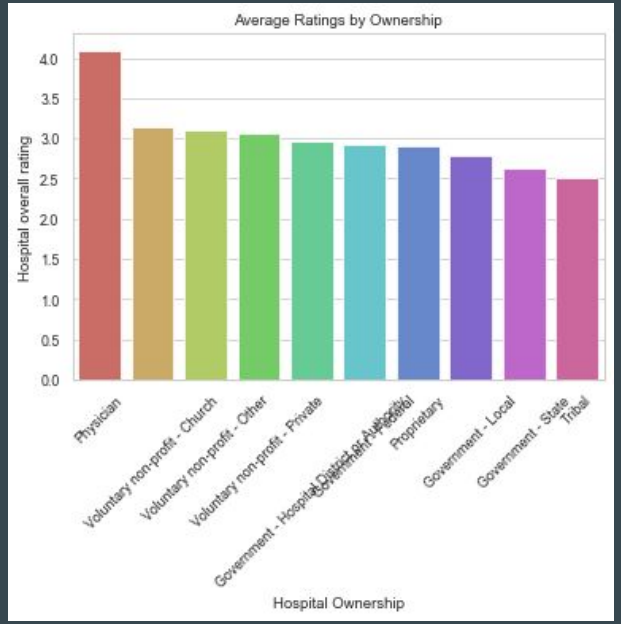


Average rating analysis

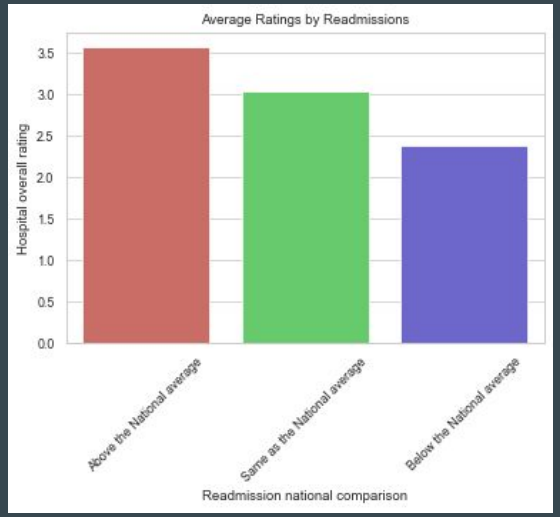
On a generic analysis we see that

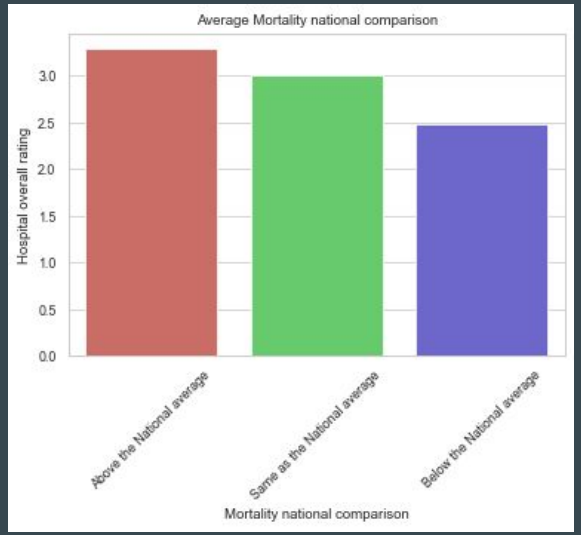
the hospitals owned by Physicians have

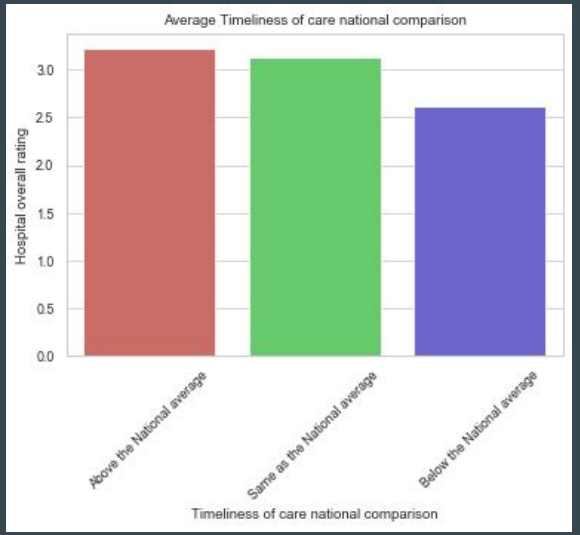
Higher rating.



Average comparative ratings

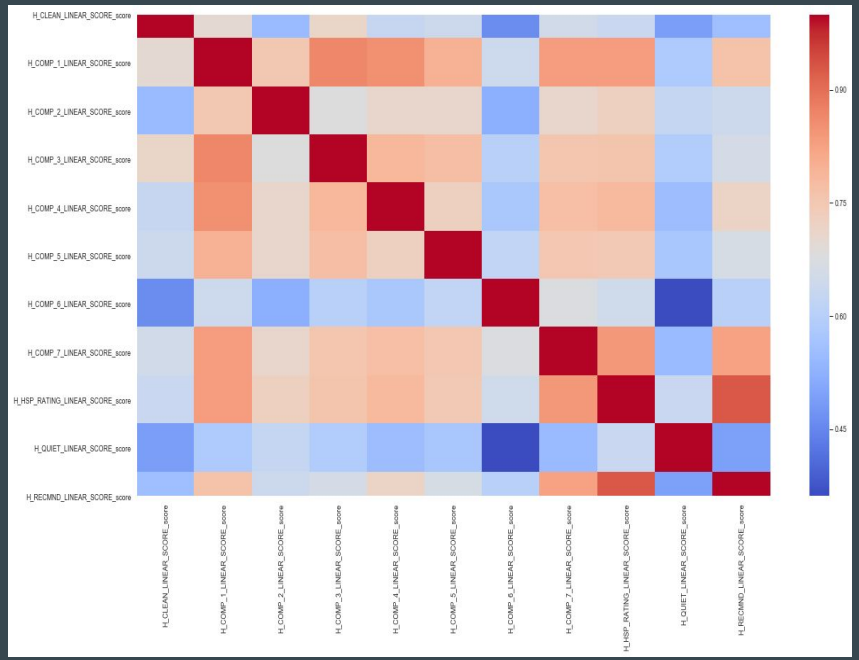




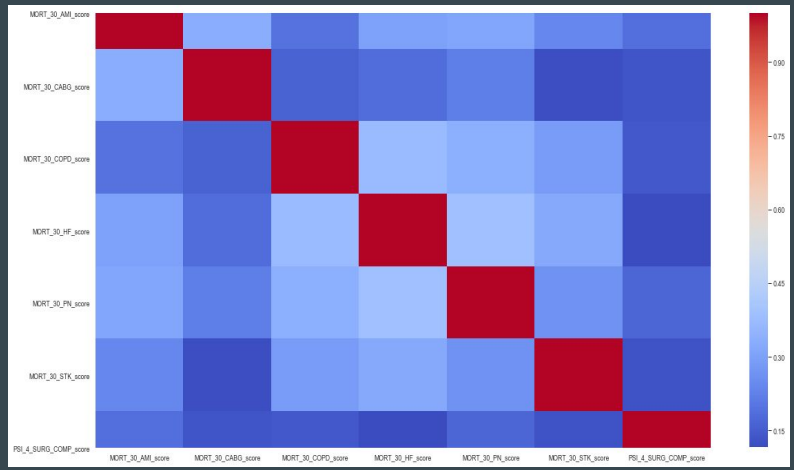


Measure Importance & Correlation Plots

Safety



Mortality



Data Modelling: Supervised Learning methods:

1. Logistic Regression

- Logistic Regression analysis is a statistical technique to evaluate the relationship between various predictor variables (either categorical or continuous) and an outcome variable which is generally binary but it can also be extended to multi class classification.-
- Challenges Faced :

i) Highly Correlated Variables : If input variables are highly correlated with one another (known as multicollinearity), then the effect of each on the regression model becomes less precise. In the data provided by the CMS, there are a lot of measures which are correlated with 1 or more measures.
ii) Assumptions regarding the relationship between input and output variables : Regression models assume that the relationship between the predictor variables and the dependent variable is uniform, i.e., follows a particular direction – this may be positive or negative, linear or nonlinear but is constant over the entire range of values. This assumption may not hold true for certain associations – for example, mortality from pneumonia may be higher at both extremes of age.

Data Modelling: Supervised Learning methods:

2. Naive Bayes

- It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.
- Challenges faced :

  i) It performs well on categorical data but poorly on continuous features if they do not have a normal distribution.
  ii) If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as "Zero Frequency". To solve this, we can use the smoothing technique.

Data Modelling: Supervised Learning methods:

3. Random Forests

- Random Forests is a versatile machine learning method capable of performing both regression and classification tasks. It also undertakes dimensional reduction methods, treats missing values, outlier values and other essential steps of data exploration, and does a fairly good job.

Data Modelling: Supervised Learning methods:

- Relative importance of top 6 features based on the model's predictive power :

Feature Name	Importance
READM_30_HOSP_WIDE_score	0.0908
PSI_90_SAFETY_score	0.7305
H_HSP_RATING_LINEAR_SCORE_score	0.04154
H_RECMND_LINEAR_SCORE_score	0.03780
MORT_30_HF_score	0.03745
H_COMP_1_LINEAR_SCORE_score	0.03238

Data Modelling: UnSupervised Learning methods:

k-means Clustering

- The k-means clustering analysis is a standard method for creating categories (or clusters) so that observations in each category are closer to their category mean than to any other category mean. The number of categories is pre-specified; CMS specifies five categories, so that k-means clustering analysis generates five categories (clusters) based on hospital summary scores in a way that minimizes the distance between summary scores (observations) and the middle value of their assigned cluster (category mean). It organizes hospitals into one of five categories such that a hospital's summary score is "more like" that of the other hospitals in the same category and "less like" the summary scores of hospitals in the other categories. The Star Rating categories are structured such that the lowest group is one star and the highest group is five stars.

Data Modelling: UnSupervised Learning methods:

- Challenges Faced :

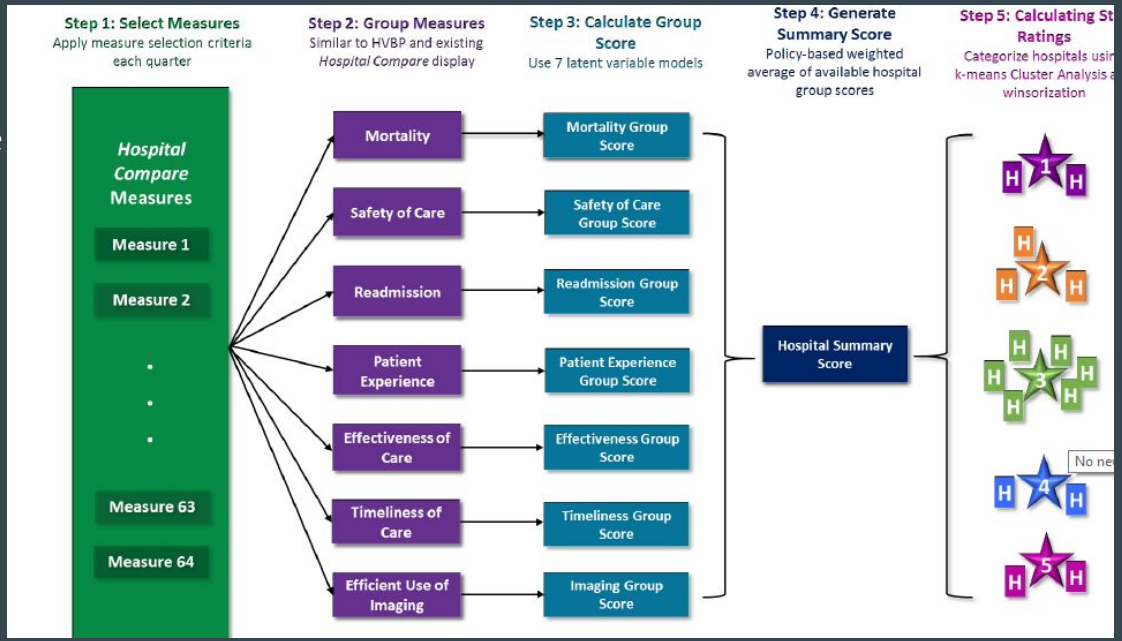
  Need to specify the number of clusters beforehand to divide the data in the initial stages only.

  Clustering model is able to predict star ratings with an overall accuracy of approx 45%.

Data Modelling: UnSupervised Learning methods:

Process

- Using the measure weights calculated using factor analysis, each group's score is calculated.
- The group scores are multiplied by the weight of group (22% or 4%) to calculate the final score which are then divided into 5 clusters.



Data Modelling: UnSupervised Learning methods:

2. Hierarchical Clustering

- It involves creating clusters that have predominant ordering from top to bottom.

  Advantage :

  In hierarchical clustering, the data is not partitioned into a particular cluster in a single step. Instead, a series of partitions/merges take place, which may run from a single cluster containing all objects to n clusters that each contain a single object or vice-versa.

  A dendrogram is formed in the end, which shows which data points group together in which cluster and at what distance.

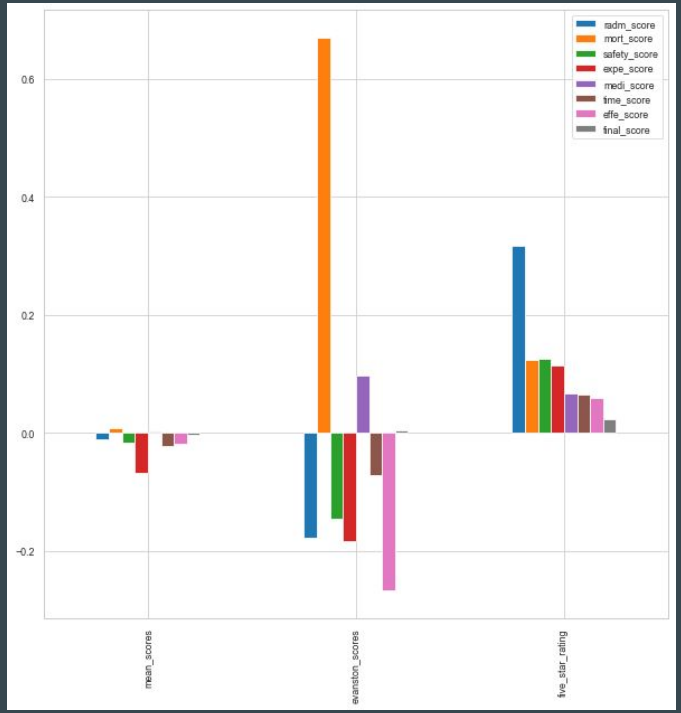
Provider Analysis for Evanston Hospital

The mean scores are negative for almost all

Groups. For 5 star rated hospitals however, the

Ratings, needs to improve in

- Readmission
- Safety
- Experience
- Effectiveness
- Timelinesss

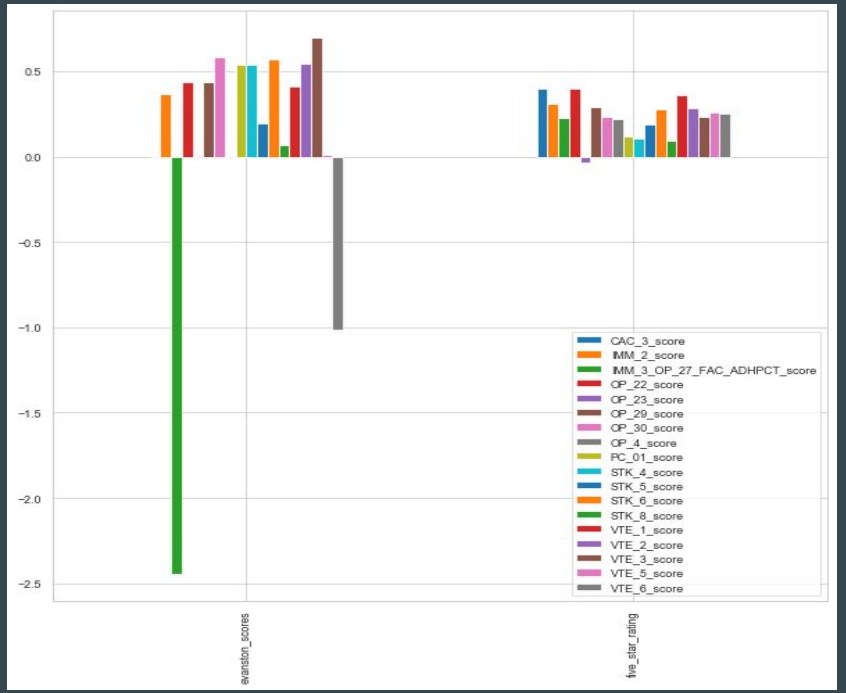


Provider Analysis for Evanston Hospital

Within effectiveness group, it's

necessary to focus on

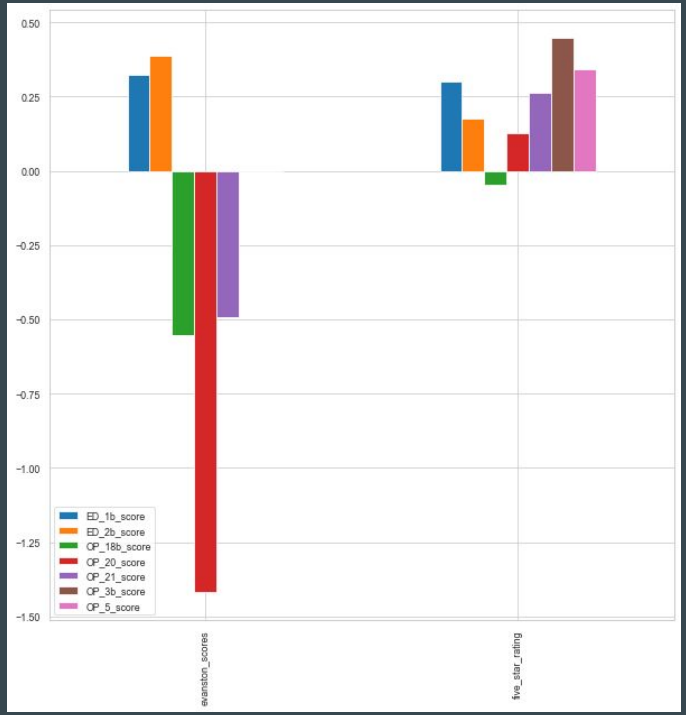
- IMM_3_OP_27_FAC_ADHPCT_score
- VTE_6_score



Provider Analysis for Evanston Hospital

In this group, Evanston hospital needs to improve

- OP_20_score
- OP_21_score
- OP_5_score



End of presentation