PREDICTING HOSPITAL PERFORMANCE IN THE U.S.

Kenneth Martin

PROBLEM STATEMENT

The U.S. has thousands of hospitals and other medical facilities that operate at various levels of effectiveness. I wanted to discover which qualities of a hospital's surrounding area are most important when predicting the hospital's levels of timeliness and quality of care. Being able to predict a hospital's effectiveness based on its surrounding area could be useful for medical professionals as well as policy makers and others in government who want to ensure high quality care for as many people as possible in the United States. With better insight into which factors are most likely to indicate an increase or decrease in a hospital's effectiveness, one can more easily identify the type of hospital that is likely to be below the national average. This would allow for a more focused approach to increasing healthcare effectiveness in the country as a whole.

For this project, I used government Medicare data that includes general information on more than 5,000 hospitals throughout the U.S., as well as different metrics to measure the effectiveness and timeliness of these hospitals. In addition, I used U.S. Census data to get a better picture of the socioeconomic situation surrounding each hospital. In addition to factors such as state, county, facility type, and ownership, the census data will allow me to consider how the race, age, and education breakdowns, as well as other statistics such as average income of the surrounding area affect the performance of a hospital. To do this, I created models using the overall hospital rating, timeliness of care national comparison, and patient experience national comparison metrics as dependent variables, and then analyzed which census features were most important for the model's ability to predict the hospitals' effectiveness.

Using census data for the surrounding zip code alone, I was able to create predictive models for the three metrics of effectiveness with ROC-AUC scores as high as .733. In addition, I was able to determine which features of the surrounding area are the most important when predicting the hospital's performance.

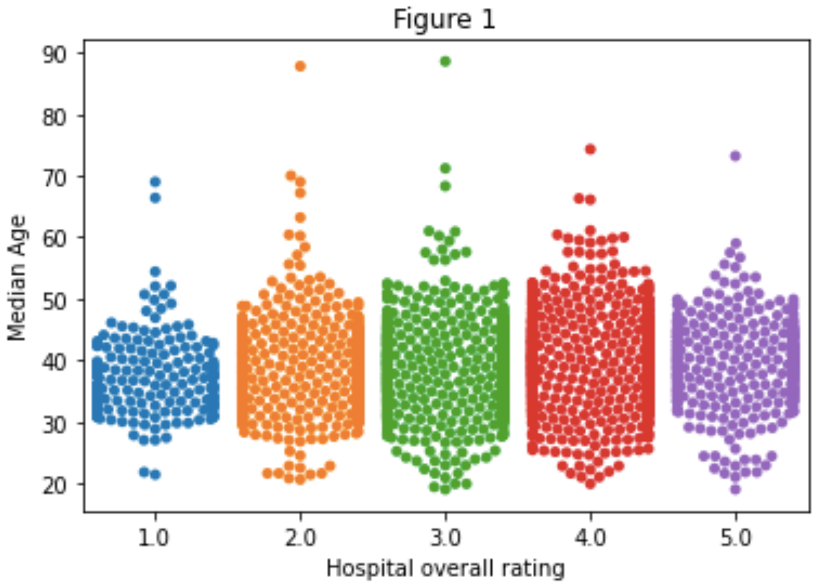
DATA WRANGLING

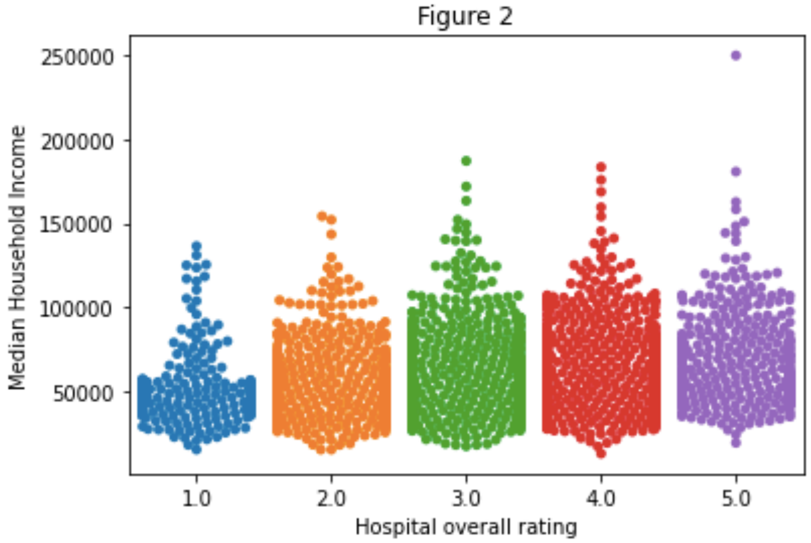
The raw data that I started with was two CSV files; one containing general hospital information, and one with the various performance scores that each hospital had received. I began by merging these two into one dataset. I then pulled the desired features from the U.S. Census, and merged these into the hospital dataset using the zip codes.

I dropped several unwanted columns and replaced the comparisons to the national average for various performance metrics (“Above”, “Below”, etc.) with integers. At this point, my dataset had 4,565 rows and 16 columns.

EXPLORATORY DATA ANALYSIS

I initially wanted to see if any census features stood out as likely to be important when predicting the performance metrics. I plotted Median Age and Median Household Income by Overall Hospital Rating to get an idea of the general spread of the data.

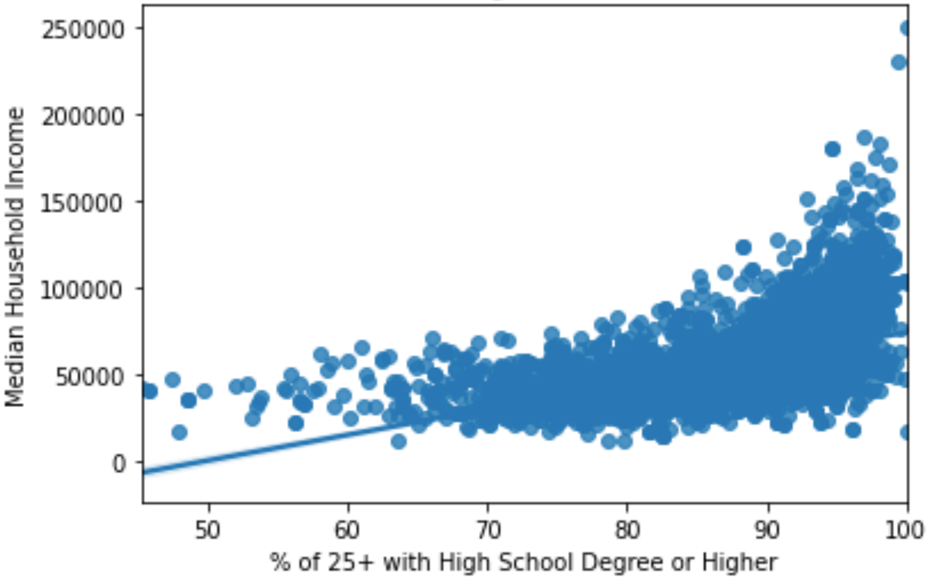




Both of these plots showed a large amount of variation of the data for each rating, but seemed to somewhat indicate a slight increase in Median Age and Median Household Income as the area's overall hospital rating increased.

I was also aware that some of my U.S. Census features could be confounding variables. To investigate this, I plotted some of the features against each other, such as % of 25+ with High School Degree or Higher and Median Household Income, to see if there was some kind of correlation.

Figure 3

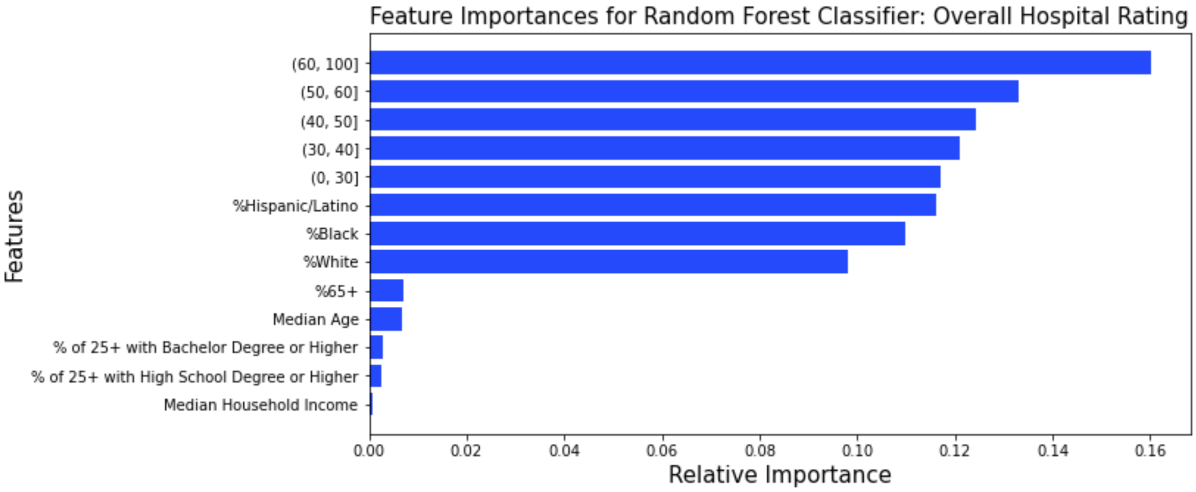


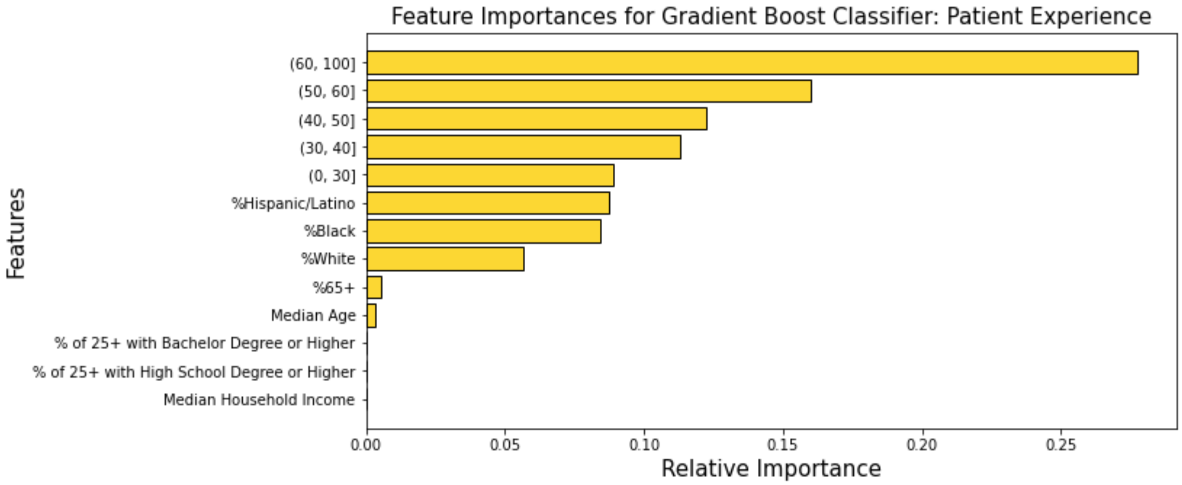
As suspected, some of the features did seem to be at least somewhat correlated, which was something for me to be mindful of as I built my models.

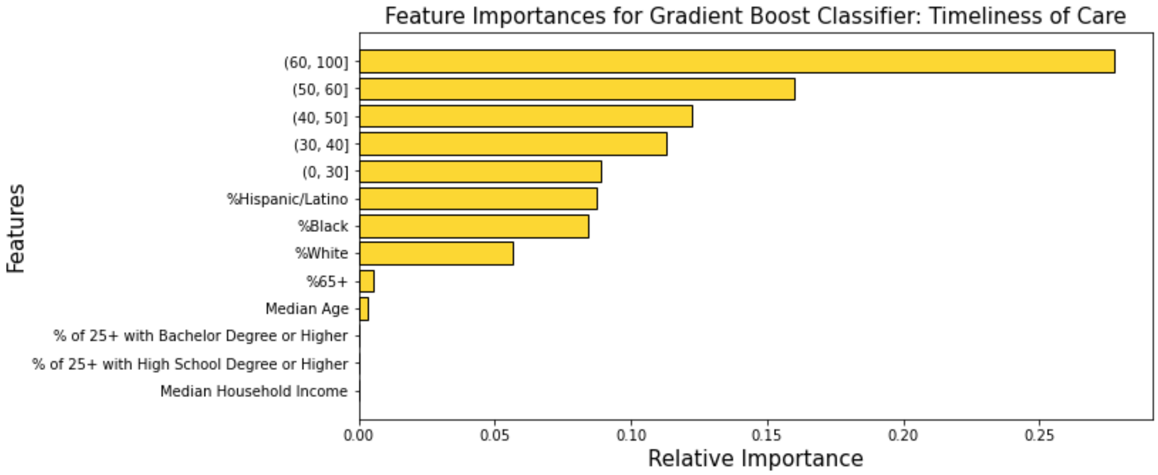
Modeling

For each of the three dependent variables (Hospital Overall Rating, Patient Experience National Comparison, and Timeliness of Care National Comparison) I binned the scores into two categories: Above Average and Equal to or Below Average. I then tested three different machine learning classification models: Logistic Regression, Random Forest Classifier, and Gradient Boosting Classifier.

For the Random Forest Classifier and Gradient Boosting Classifier, hyperparameter tuning was done using grid search cross-validation. The metric I focused on maximizing while building the models was ROC-AUC. After determining the best model for each dependent variable, I plotted the feature importances of that classifier.


Feature Importances for Random Forest Classifier: Overall Hospital Rating
Features
(60, 100]
(50, 60]
(40, 50]
(30, 40]
(0, 30]
%Hispanic/Latino
%Black
%White
%65+
Median Age
% of 25+ with Bachelor Degree or Higher
% of 25+ with High School Degree or Higher
Median Household Income
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
Relative Importance


Feature Importances for Gradient Boost Classifier: Patient Experience
Features
(60, 100]
(50, 60]
(40, 50]
(30, 40]
(0, 30]
%Hispanic/Latino
%Black
%White
%65+
Median Age
% of 25+ with Bachelor Degree or Higher
% of 25+ with High School Degree or Higher
Median Household Income
0.00
0.05
0.10
0.15
0.20
0.25
Relative Importance


Feature Importances for Gradient Boost Classifier: Timeliness of Care
Features
(60, 100]
(50, 60]
(40, 50]
(30, 40]
(0, 30]
%Hispanic/Latino
%Black
%White
%65+
Median Age
% of 25+ with Bachelor Degree or Higher
% of 25+ with High School Degree or Higher
Median Household Income
0.00
0.05
0.10
0.15
0.20
0.25
Relative Importance

After tuning hyperparameters with grid search cross-validation, my most successful model for predicting Overall Hospital Rating was the Random Forest Classifier with a ROC-AUC score of 0.691. The most successful model for predicting Patient Experience National Comparison was the Gradient Boosting Classifier with a ROC-AUC score of 0.682. Finally, the most successful model for predicting Timeliness of Care National Comparison was the Gradient Boosting Classifier with a ROC-AUC score of 0.733.

Conclusion/Future Research

Using only census data, I was able to create a decent predictor of the quality of a hospital based on its surrounding demographics. The ROC-AUC scores varied from 0.682-0.733, which is not great, but unsurprising considering the large number of factors that likely affect a hospital's success. In the future, I would like to add more features that describe the financial situation of the surrounding area. I had initially hypothesized that median income would be very important in predicting hospital performance, but clearly this is not the case. In hindsight this makes sense, as it seems that areas with an older demographic tend to have better hospitals. While older people tend to have more wealth than younger demographics, many people over the age of 60 begin to retire at some point, which would bring the median income down. Populations with people in their 40s and 50s likely have a relatively high median income, but also generally don't require a lot of really good healthcare.

I am also curious to know what causes the median age of the surrounding area to be a good predictor of above average hospital performance. I would hypothesize that as people get older and retire, they might look to move to an area with a good hospital, as healthcare is generally more important for people 60+ than it is for younger generations.

I was surprised to see such a sharp drop off in feature importance after the racial breakdown of the area. I assume that the %65+ and Median Age categories are simply confounding variables that are less helpful than the binned age groups. However, I would have thought that the education breakdown would be more important. Perhaps these are also confounding variables, either with the racial breakdowns or the age groups.

The U.S. Census provides a wealth of data for a huge variety of features. While I have used a handful of them in my project, I am sure that there are more features that could increase the success of the models that I have made. In the future, new models could be created using a different selection of features to see if the models' scores could be improved, and to hopefully get a better idea of which specific demographics are in need of an increase in hospital/healthcare quality.