National Hospital Price & Quality Optimization Study – Analytics Practicum Project Report

Hsuan-Han Yang

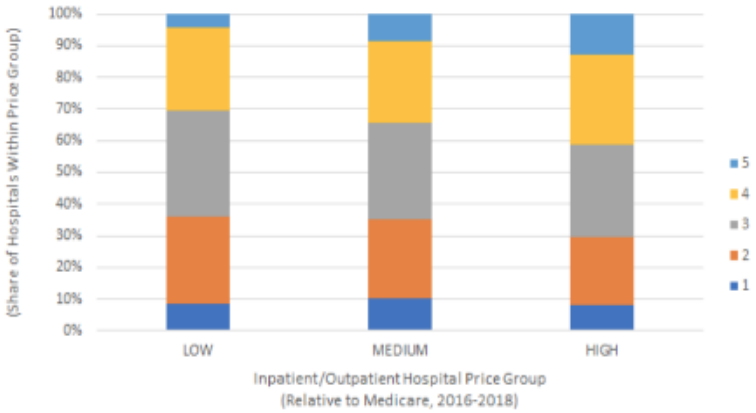
Introduction

Going to hospital can be stressful with the fact that private payers generally pay more than what Medicare would've paid for the same healthcare service (White & Whaley, 2019). Large portions of quality hospitals come with a hefty price tag for private payers. While there is an overwhelming amount of information on the internet with countless search results for each hospital, absorbing all the latest information to find the hospital with the lowest price and the highest quality can be very challenging. Currently, customers lack tools and methodologies to query all the relevant data, process them and use algorithms to identify a list of qualified hospitals.

The objective of this project is to provide analytical modeling methods that can be automated to recommend low-priced and high-quality hospitals to customers by predicting hospitals into lowest to highest price groups along with low to high quality groups based on 1000's of financial and quality metrics publicly available.

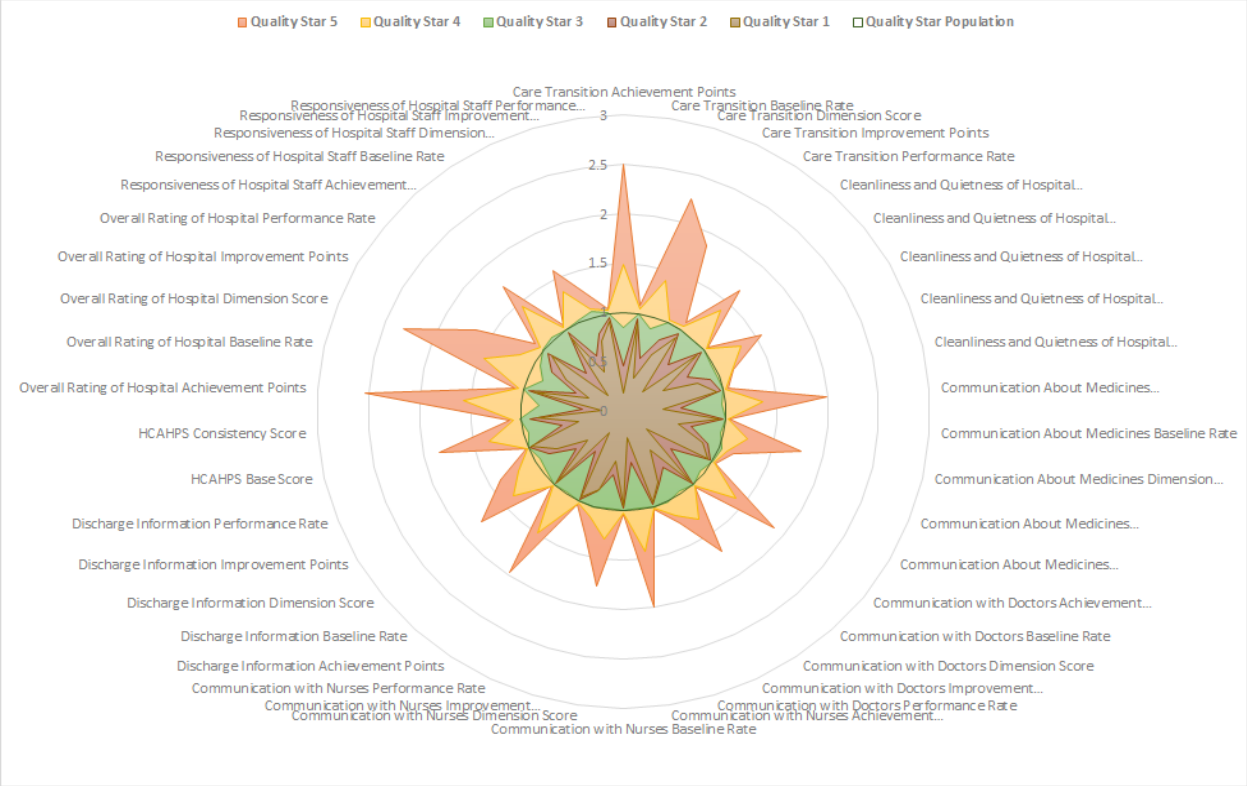
Methodology

Our proposed method is to utilize a cluster-then-predict approach which should improve prediction accuracy in most datasets (Trivedi et al., 2015). Our approach is to cluster hospitals to different quality clusters, and perform hospital price classification for each quality cluster. We performed this analysis by utilizing Python, Excel and Jupyter notebooks. For clustering, we've explored how to cluster the data to improve predictions and make an easily interpretable model. We started off with understanding the healthcare industry and exploring the data. The chart below shows that around 30% low (<150 relative price) priced hospitals have good quality rating (4-5) which illustrates the potential that there exist a significant number of hospitals that have high quality and are low priced.



After some experiments mentioned under experimentation, we utilized already done clustering of quality star rating. To better understand the quality metrics and see if these can be grouped further, we did data exploration on the quality data set. Below is a guide for the spider chart we created after data normalization and excluding variables that did not make a difference across clusters.

For the first step, we performed exploratory analysis by calculating attributes average across quality star rating, then we normalized quality star group by average of population. Afterwards, we obtain the transpose of normalized features and visualize them as the spider chart below.

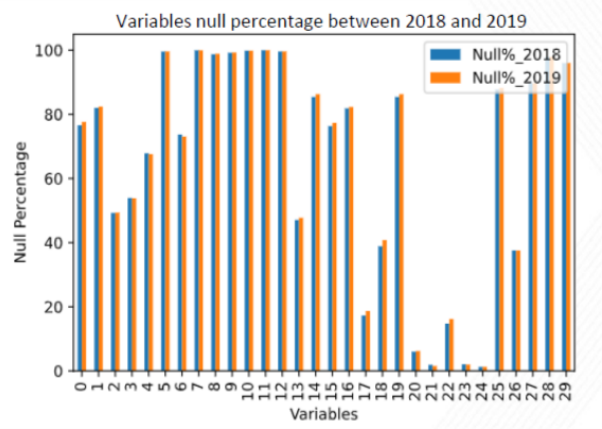


Looking at this, we can observe that hospitals are roughly distributed into three quality clusters, with normalized values within 1, around 1 (population average), and those greater than 1. As a result, we grouped quality star 1-2 into low quality, quality star 3 into medium quality and quality star 4-5 as high quality. This analysis is available in the analysis workbooks folder.
For price classification, we experimented with 3, 5, 6 and 10 price classes with and without partition of quality clusters for performance and interpretation-wise comparison. To perform price classification, we implemented feature engineering on available data sources as follows:

1. All available data sources listed in README.rtf were merged together on a granularity of medical provider number that is a unique identifier for hospitals.
2. Variables with redundant information were discarded from modeling inputs, reducing

the variable list by 40 for example Hospital Name, wt_cy_1 etc.

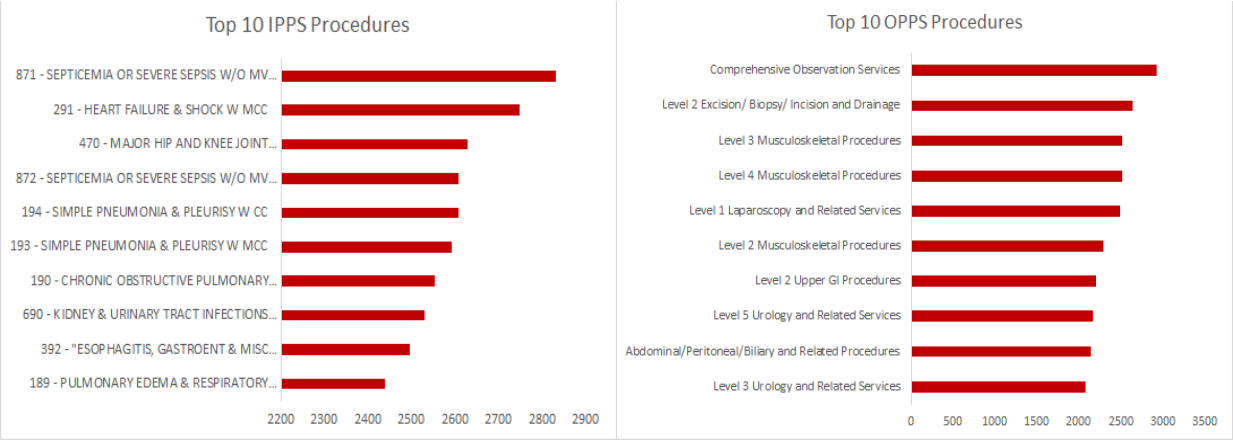
3. Variables with categorical values were dummy encoded for example: teach_hosp_hcr , rural_urban.
4. Percentage of missing information was calculated for each variable as illustrated in the figure below by year. These are randomly selected 30 variables for illustration and it shows that both years have almost the same null percentage.



5. Variables were further shortlisted that had Nulls > 10% leaving behind 300+ variables.
6. Null values were replaced by the average of state for each variable. National average was used if the state average was null (example expns_nursing_admin_salary for state NY)
7. Rand HCRIS 2018 was used for training/validation and 2019 for scoring
8. Relative price for outpatient services was selected as a dependent variable. Other price/cost variables including inpatient relative price were not considered as a target due to the high percentage of nulls
9. All other relative/standard price variables were removed from independent variables to avoid any target leakage
10. Derived variable scores were created from IPPS and OPPS data sources using the following steps:
    - Find and filter for the most common procedures in all medical providers for both IPPS and OPPS.
    - For the most common procedures, calculate percentage scores by normalizing the variables by state and national.
    - Merge these scores based on medical provider number.
    - Derived variables from IPPS/OPPS had around 20% nulls when joined with labeled and rand data. These variables were excluded from step 5 filter.

For IPPS and OPPS, most common procedures in hospitals were selected to create derived scores as follows (This analysis is available in analysis workbooks.):

- IPPS
    - 291 HEART FAILURE & SHOCK W MCC
    - 871 SEPTICEMIA OR SEVERE SEPSIS W/O MV >96 HOURS W MCC
- OPPS
    - Comprehensive Observation Services



Score calculation example for OPPS normalized on State Procedure:

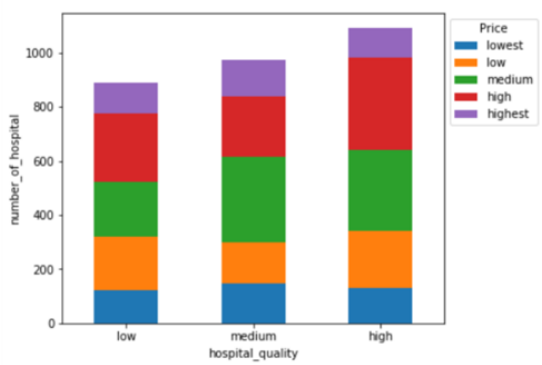
$$\text{Comprehensive Observation Services Score} = \frac{(\text{Average\_Estimated\_Total\_Submitted\_Charges\_FOR\_PROVIDER} - \text{Average\_Estimated\_Total\_Submitted\_Charges\_FOR\_STATE})}{\text{Average\_Estimated\_Total\_Submitted\_Charges\_FOR\_STATE}}$$

Comprehensive observation services score is calculated by using the difference between average estimated total submitted charges for provider and that for state divided by average estimated total submitted charges for state. Similarly, scores were calculated for each hospital by the most common procedure of IPPS and OPPS normalized on state and national level.

The recommended optimal number of classes were defined by having the greatest performance while maintaining high interpretability. Price range for each class was defined by the distribution of prices for the dataset. From the end user's standpoint, having high interpretability is useful for them to separate medium-low priced from high priced hospitals.

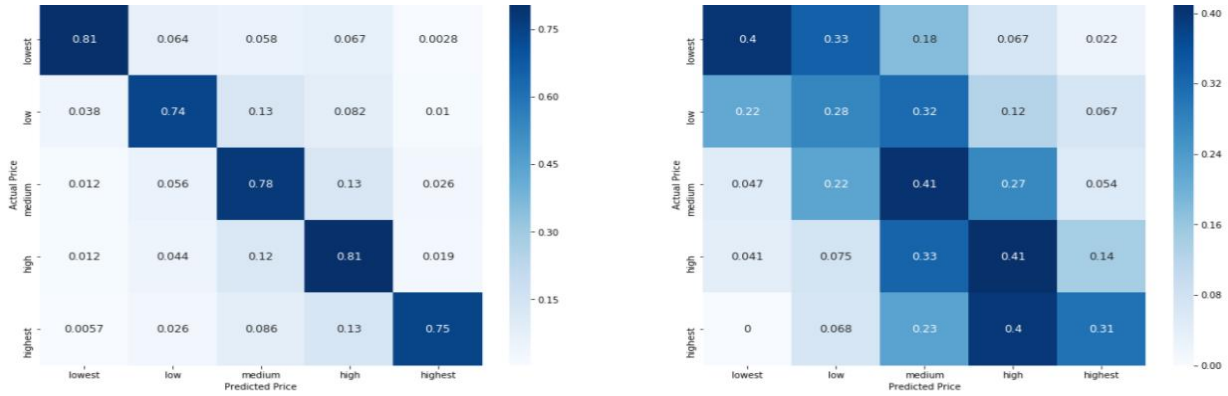
After setting the number of classes for classification, we performed modeling on data with and without partition clusters to observe performance difference and the impact of clustering on modeling performance. For each modeling iteration, variables were reduced using multinomial logistic regression with LASSO penalty and then those reduced set of variables were used in Gradient Boosting Classifier. We analyzed the performance of a model using a confusion matrix by cross tabulating predicted class against actual results for each data set that includes training, validation, test and holdout. Focus for the best model was to not only have high

interpretability, but it should also have the lowest misclassification rate and a smaller number of variables for a simpler model to avoid overfitting. Models trained on 5 price classes (lowest, low, medium, high, highest) produced the best results. These models were trained separately on each quality cluster (low,medium,high). Following chart shows distribution of hospitals by quality and price:

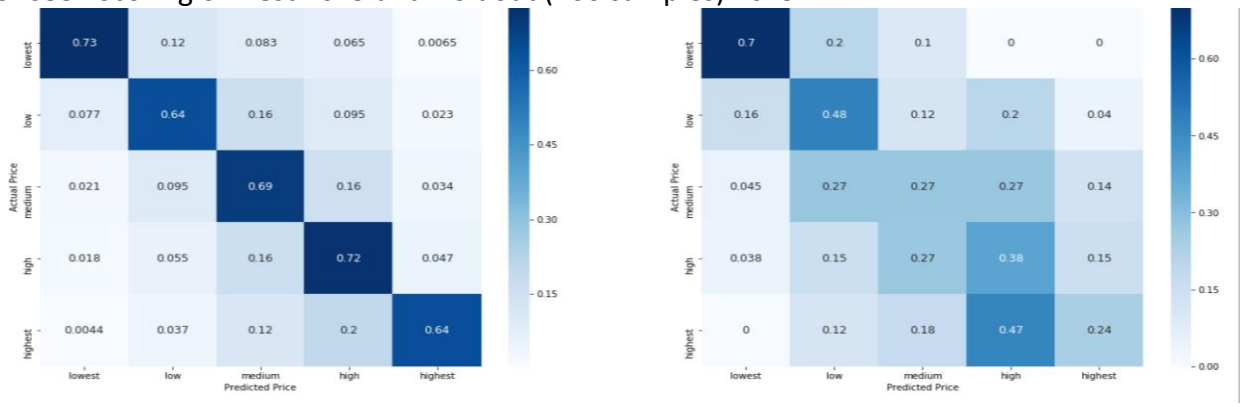


LASSO reduced variables to 34, 49 and 39 respectively by quality cluster partition. Gradient Boosting Classifier model with max depth of 1 was used to fit data with the following evaluation results:

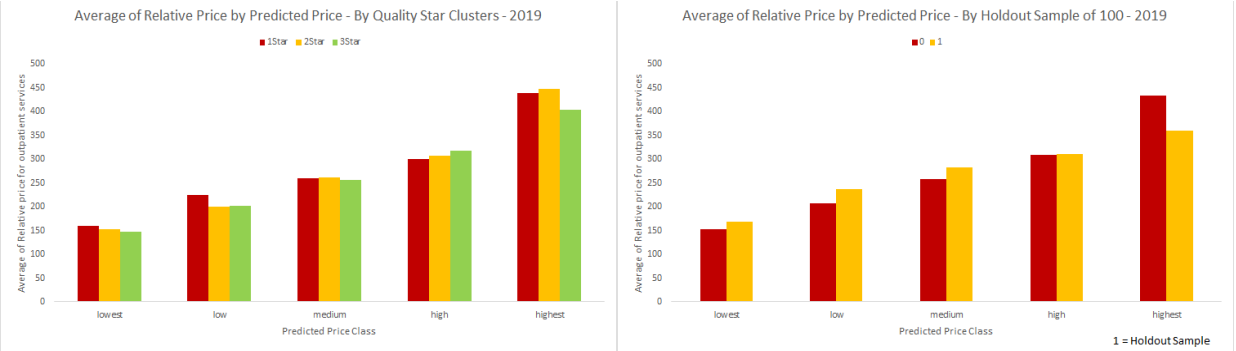
Train and Validation



Unseen Scoring on Test 2019 and Holdout (100 samples) 2019



Following graphs are the distribution of actual average relative price for outpatient services by model predicted price classes segregated on quality clusters and holdout sample:



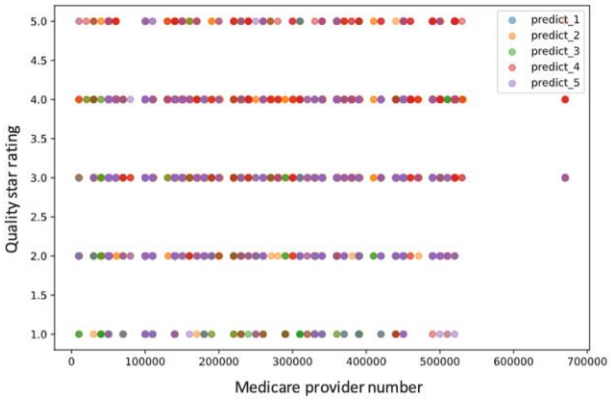
To further analyze the insights from experiment 4, we conducted analysis on feature importance from model outputs. These results are stored in analysis workbooks folder and are as follows:

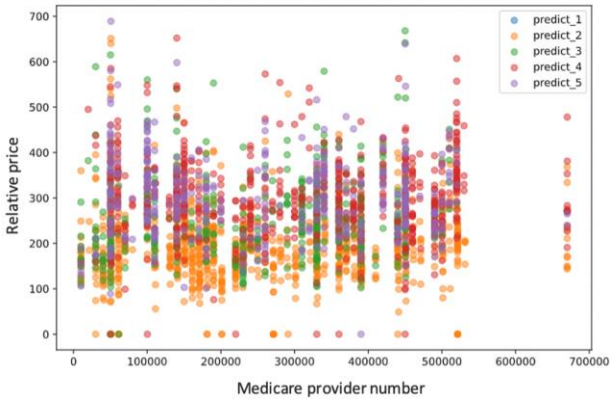
The graphs above show the insight into the top 10 features that are weighed as the most important within each cluster from gradient boosted trees and their respective relationship(positive/negative) with price can be seen in LASSO weights. Ratio of estimated

commercial revenue-to-charge ratio to Medicare revenue-to-charge ratio (commercial_to_mdcr_est) has the strongest (negative) correlation with relative price in medium and high quality hospitals whereas cost to charge ratio (used in calculation of Uncompensated care, Medicaid, and SCHIP costs) has a strong (positive) correlation with relative price in low quality hospitals.

Experimentation

The first challenge in the analysis was to reduce the number of features that had a relationship with price since there were more than 1500 variables available in the dataset. As a first step in modeling, we ran feature selection using lasso regularization with target variable of relative price to find significant variables that contribute to price. We then ran k-means on the data set using the selected 49 variables and output mean by cluster for relative price and quality. Looking at the 5-class classification results below, we concluded that these results are not distinguishing quality and price as we intended.



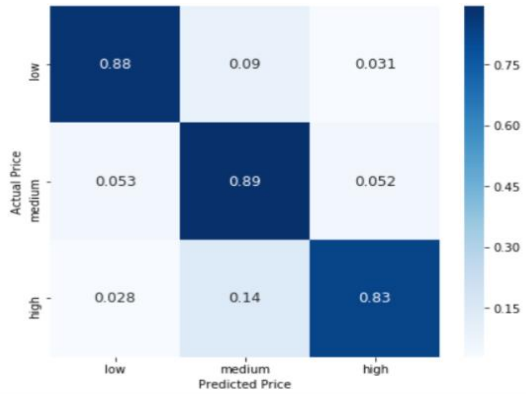


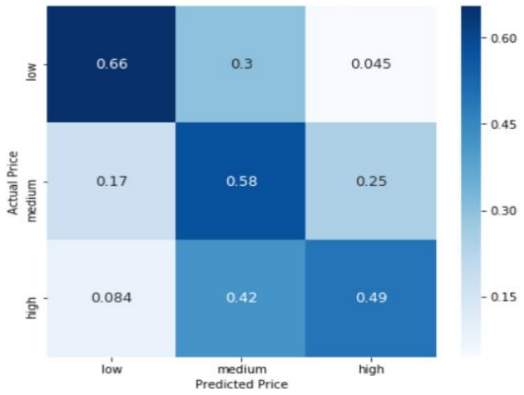
Then a total of 8 experiments were performed on 3, 5, 6 and 10 price classes both with and without partition clusters. Partitions were created by clustering quality star ratings through normalization, exploratory data analysis and spider chart visualization. Variables were reduced using LASSO in multinomial logistic regression and predictive models were trained using Gradient Boosting Classifier. Confusion matrix for each experiment is shown to display modeling performance on training, validation, unseen scoring and holdout. Training set (rand 2018) is the data used to fit the parameters. Validation set provides an unbiased evaluation of the model fit on the training set, and its confusion matrix is used to compare against the training confusion matrix to check for possible overfitting. Unseen scoring set (rand 2019) is not used in training, and it provides unbiased evaluation of model fit. Holdout set is not part of model training/validation. Both unseen and holdout confusion matrices serve as an indication of final model performance on independent data. Hyperparameters for all experiments are set to minimize misclassification and reduce overfitting.

Following are the list of experiments:

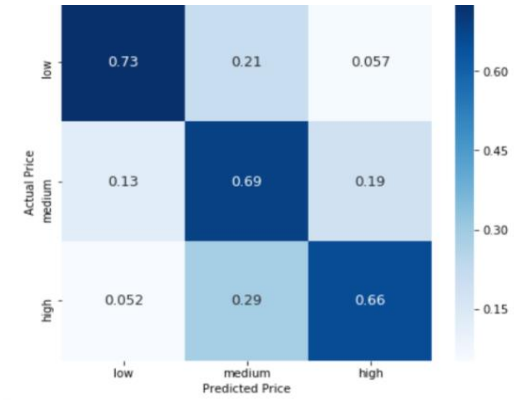
1. 3 price classes unpartitioned with 44 reduced variables

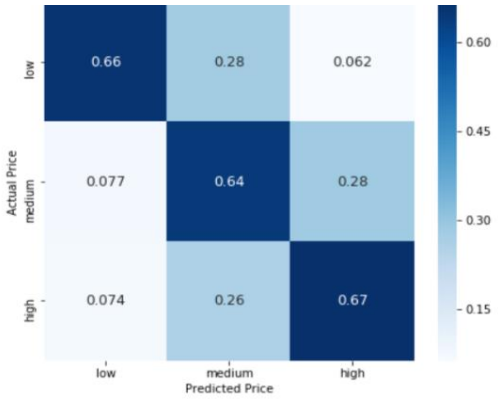
    a. Train and Validation





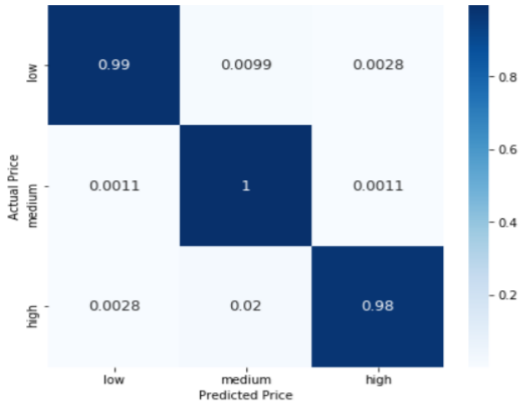
    b. Unseen Scoring on Test 2019 and holdout 2019

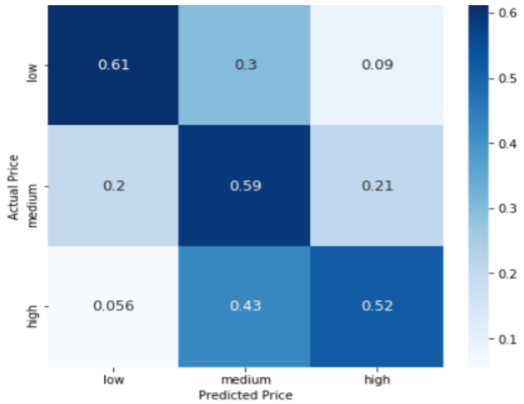




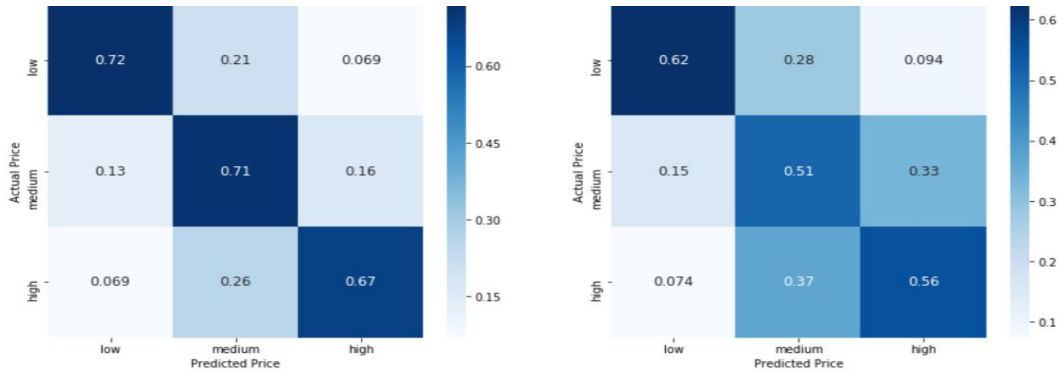
2. 3 price classes partitioned with 18, 34, 25 reduced variables by quality cluster partition

    a. Train and Validation



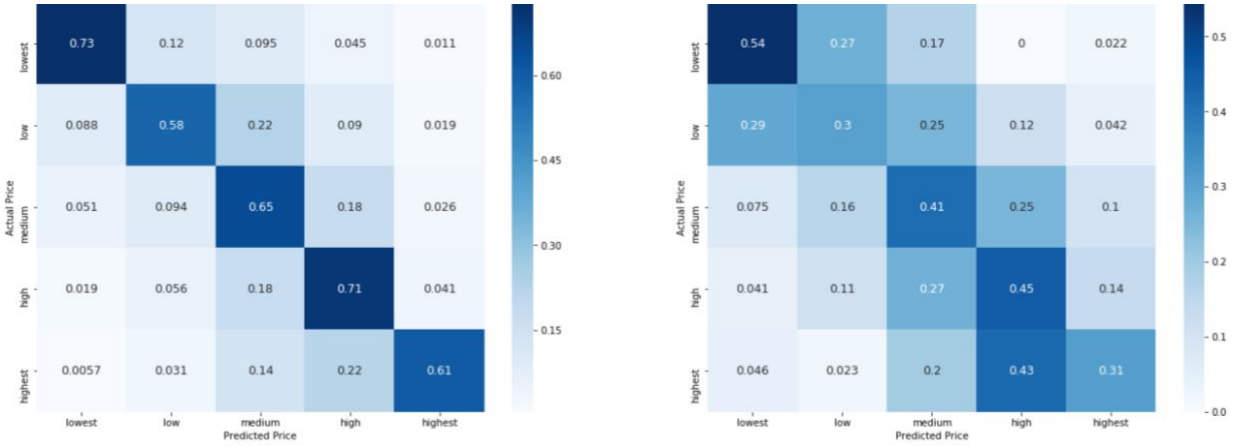


b. Unseen Scoring on Test 2019 and holdout 2019

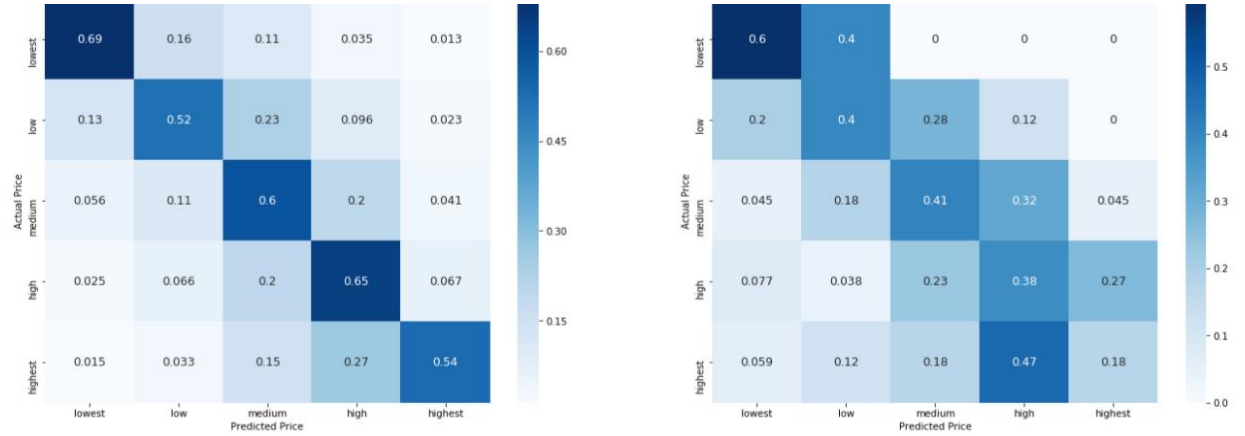


3. 5 classes unpartitioned with 110 reduced variables

a. Train and Validation

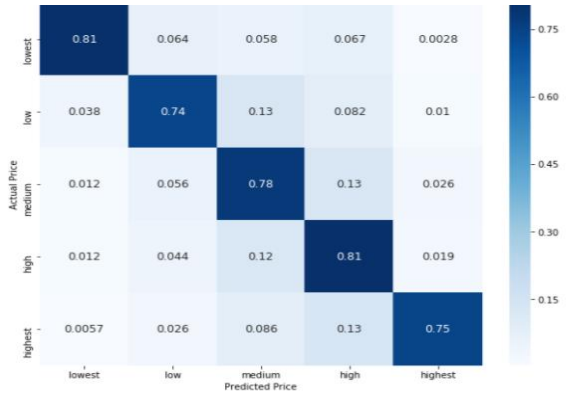


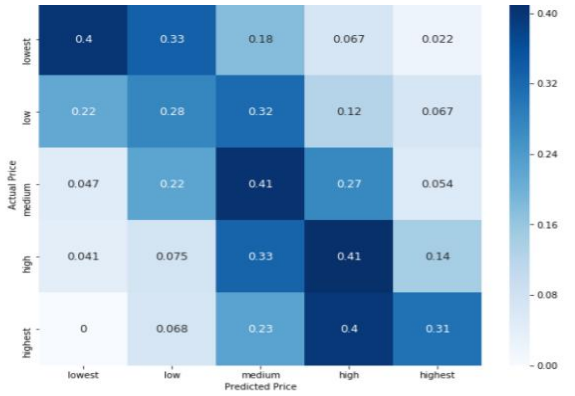
b. Unseen Scoring on Test 2019 and holdout 2019



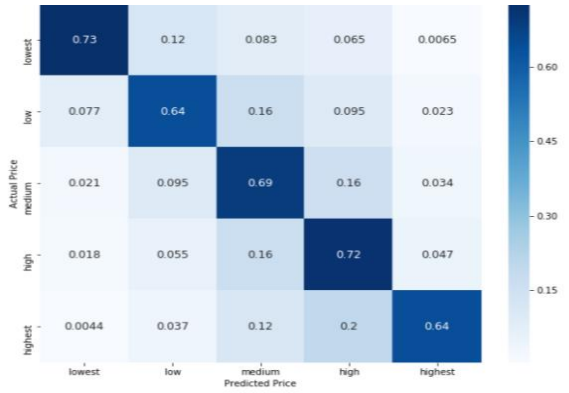
4. 5 classes partitioned with 34, 49, 39 reduced variables by quality cluster partition. **This is the model we selected for final predictions**

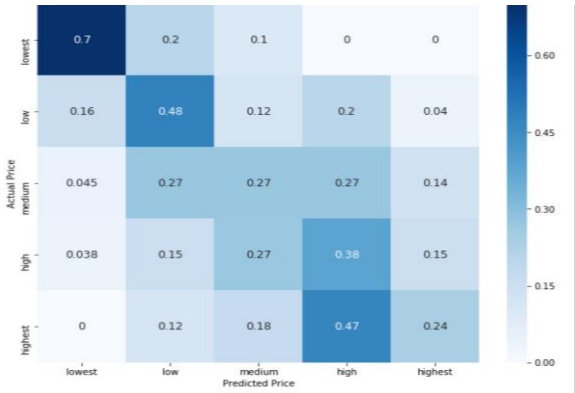
    a. Train and Validation





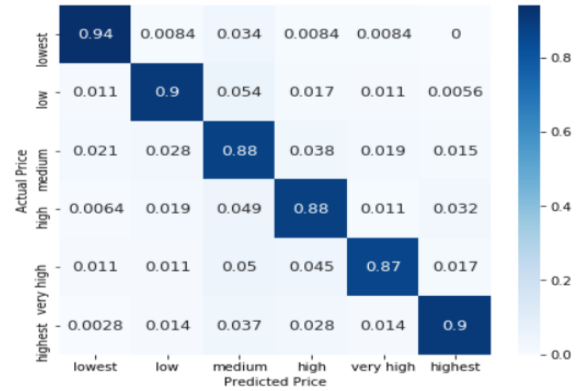
    b. Unseen Scoring on Test 2019 and holdout 2019

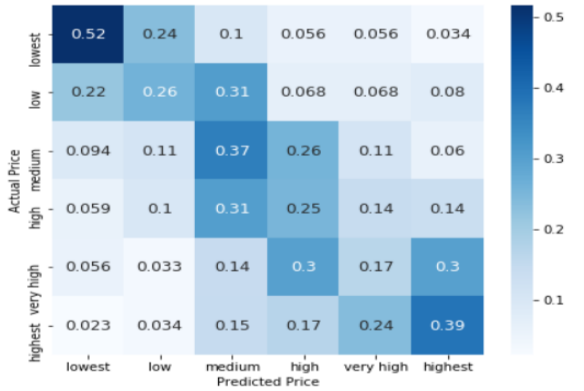




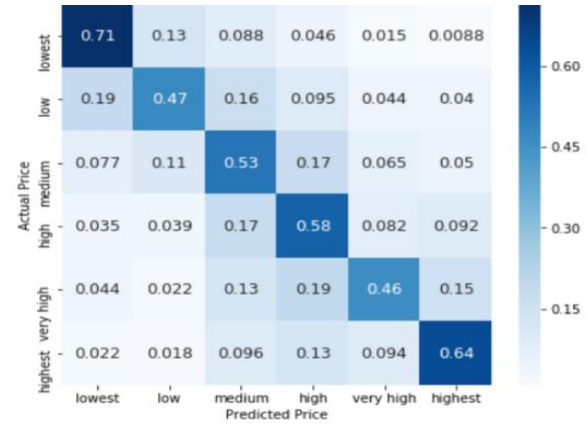
5. 6 classes unpartitioned with 60 reduced variables

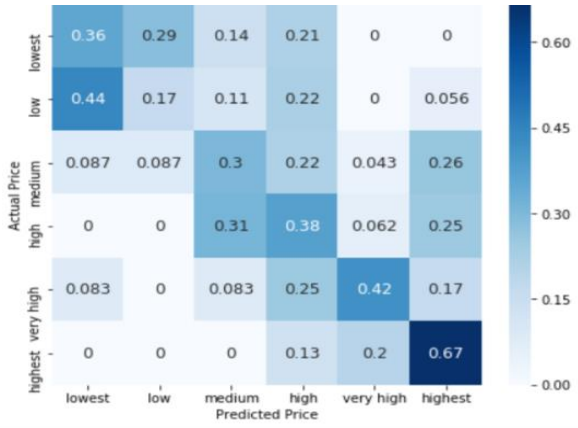
    a. Train and Validation





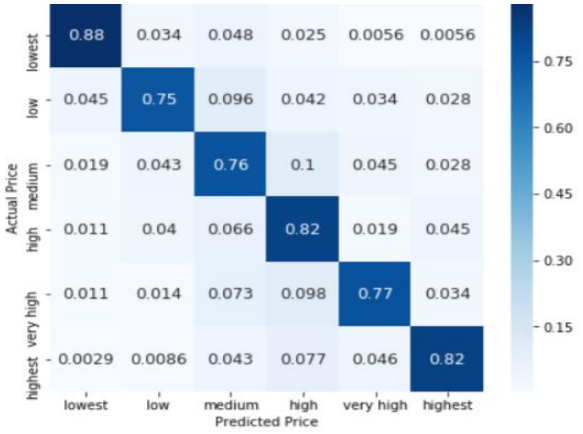
    b. Unseen Scoring on Test 2019 and holdout 2019

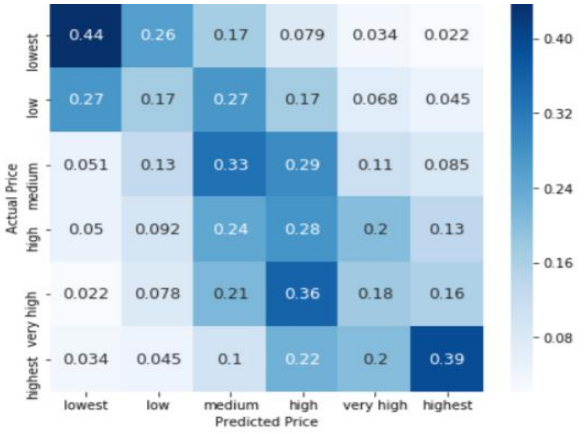




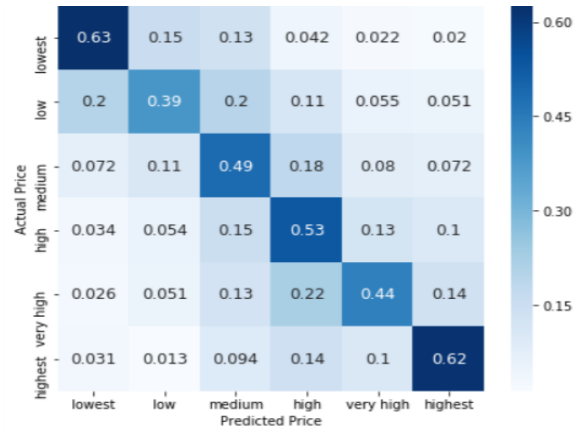
6. 6 classes partitioned with 48, 48, 36 reduced variables by quality cluster partition

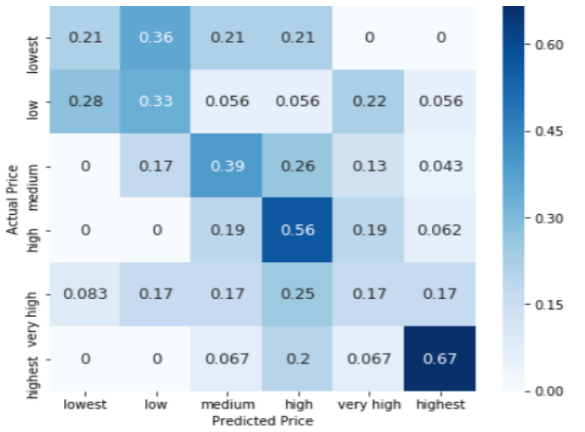
    a. Train and Validation





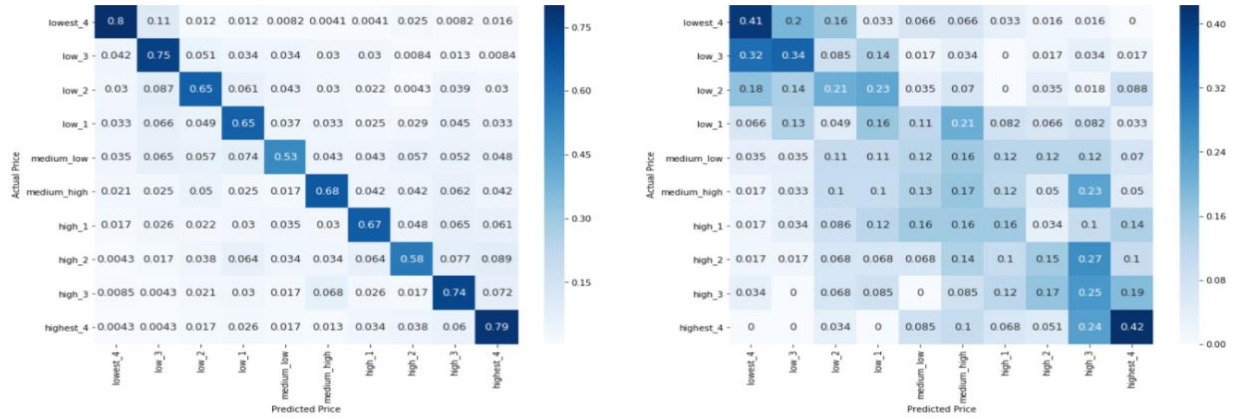
    b. Unseen Scoring on Test 2019 and holdout 2019



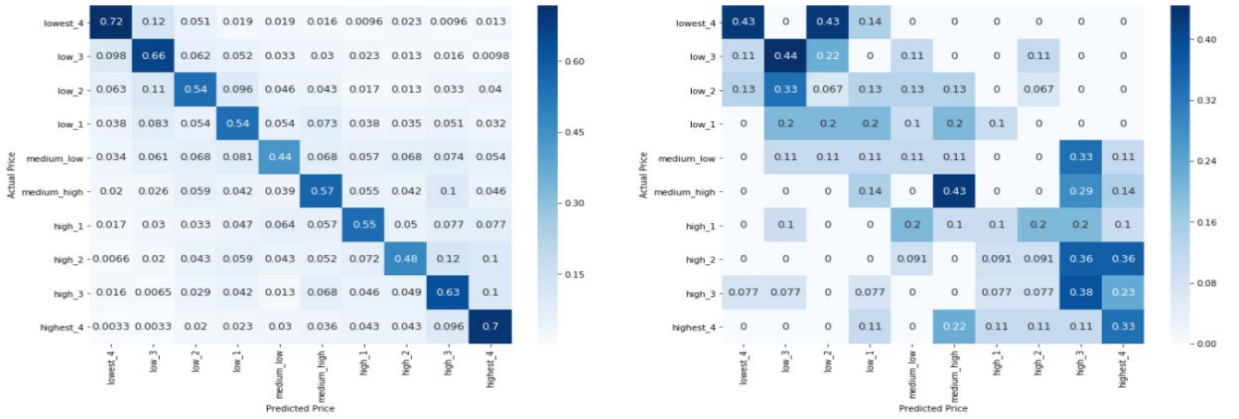


7. 10 classes unpartitioned with 294 reduced variables

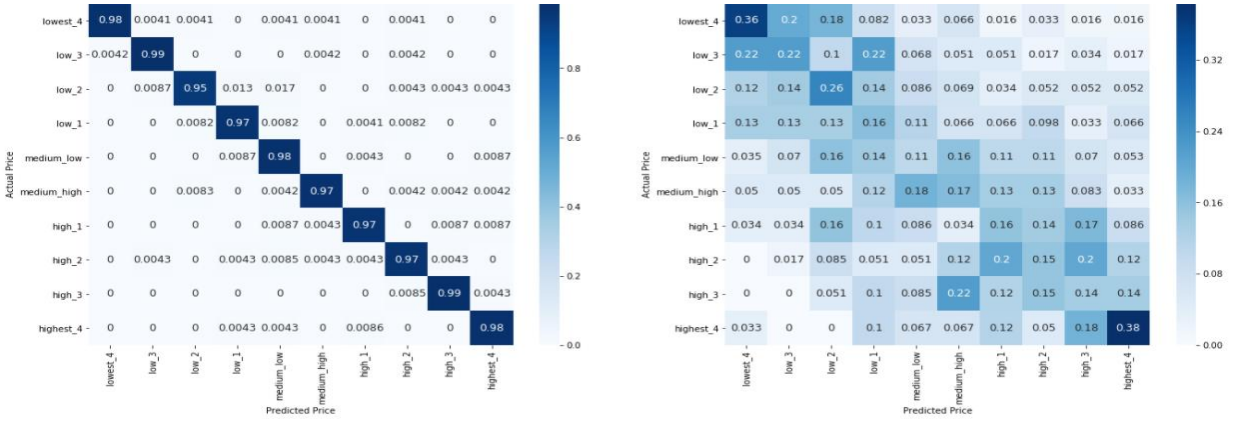
    a. Train and Validation



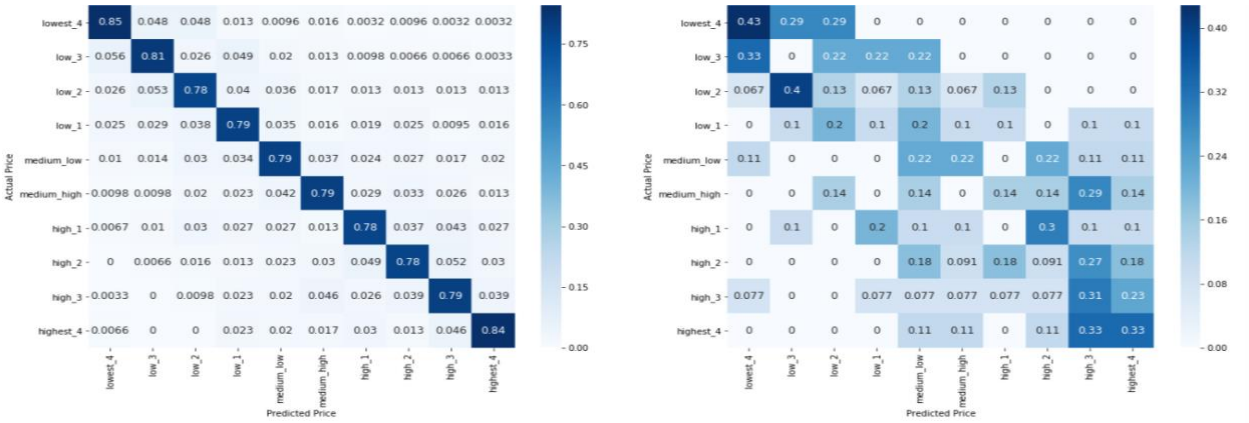
b. Unseen Scoring on Test 2019 and holdout 2019



8. 10 classes partitioned with 236, 242, 218 reduced variables
   a. Train and Validation



b. Unseen Scoring on Test 2019 and holdout 2019



Conclusion

The purpose of this project is to provide modeling methods to recommend low-priced and high-quality hospitals to customers by predicting hospitals into lowest to highest price classes. We used a cluster-then-predict approach to structure the data preparation and modeling. To evaluate the results, we tested the predicted results against actual results and visualized them using normalized confusion matrices.

The results for our classification approach showed that while our model is able to correctly identify the highest and the lowest priced hospitals for each quality cluster with relatively high accuracy, it has relatively lower accuracy when separating medium-priced hospitals from their neighboring classes. The results also show that less number of classes have higher accuracy, and higher number of classes have lower accuracy. Taking interpretability into account, it's harder for an end user to identify the target hospital with less number of classes, leading to a trade-off situation between accuracy and interpretability.

Our recommended approach is the 5-class with quality cluster partition approach in experiment 4 which has the greatest performance while maintaining high interpretability. Comparing experiment 4 against experiment 3, while both experiments utilize 5-class approach, partitioned results by quality clusters generally outperform unpartitioned results in both unseen scoring and holdout sample, suggesting that partitioning has provided a useful source of variance in the prediction process to reduce overfitting and minimize misclassification. Hence, the resulting model is simple (34, 49 and 39 variables respectively) with better predictive performance on unseen data as shown in unseen scoring and holdout confusion matrix in experiment 4.

Following files have the predictions and feature importance for the chosen model:

- Training/validation predictions with TrainTest flag:
  QoS_Output_Results\Partitioned_Models\predictions_table.csv

- Unseen/Holdout with holdout_sample flag:
  QoS_Output_Results\Partitioned_Models\unseen_scored_predictions\predictions_table.csv

- Feature Importance:
  QoS_Output_Results\Partitioned_Models\featureimportance_table.csv

Reference

1. White, C., Whaley, C. (2019). Prices Paid to Hospitals by Private Health Plans Are High Relative to Medicare and Vary Widely. https://www.rand.org/pubs/research_reports/RR3033.html
2. Trivedi, S., Pardos , Z.A., Heffernan, N.T. (2015). The Utility Of Clustering in Prediction Tasks. https://arxiv.org/abs/1509.06163