National Hospital Price & Quality Optimization Study

Applied Analytics Practicum Project - Midterm Spring 2021

Yang, Hsuan-Han

hyang11@gatech.edu

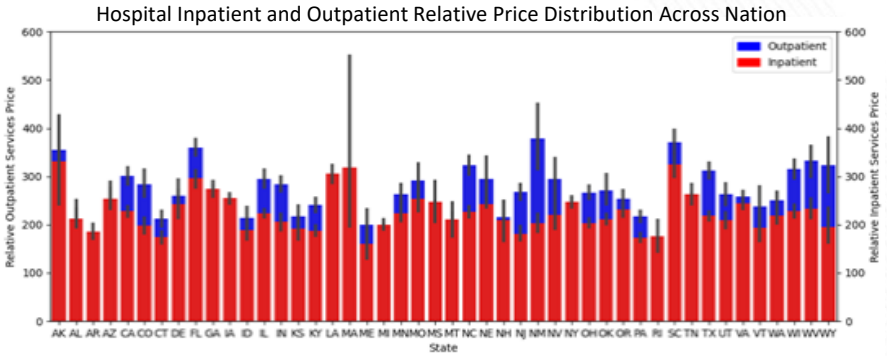
Project Background & Objective

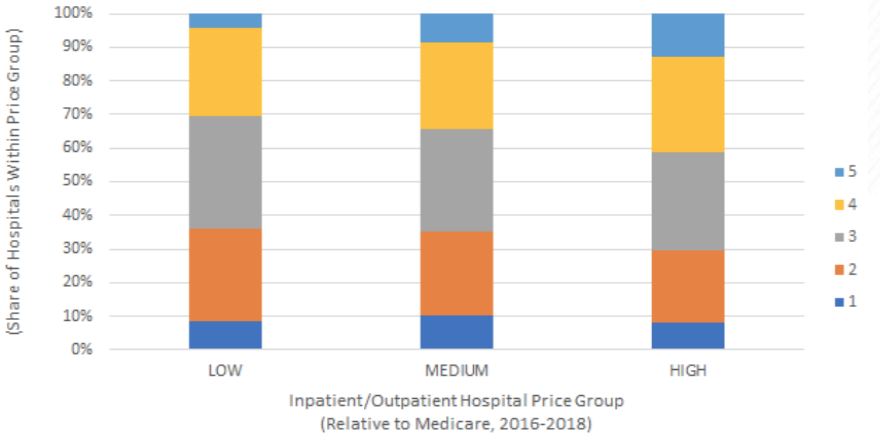
- Problem: Private payers generally pay more than what Medicare would've paid for the same healthcare service
  - Large portion of quality hospitals come with a hefty price tag for private payers
  - Customers lack tools to identify quality and low-priced hospitals

- Objective: Predict or Rank the expected inpatient or outpatient private payer costs given a hospital
  - Recommend customers low-priced hospitals with good quality services

EDA Results: High relative price for private payers, 30% low priced hospitals w/ good rating

- Mean relative outpatient and inpatient price: 269% & 227%
- Around 30% low priced hospitals have good quality rating (4-5)





Note: Relative price are derived using the ratio of amounts paid to the amount that would've been paid w/ respect to Medicare rates

Data Profile

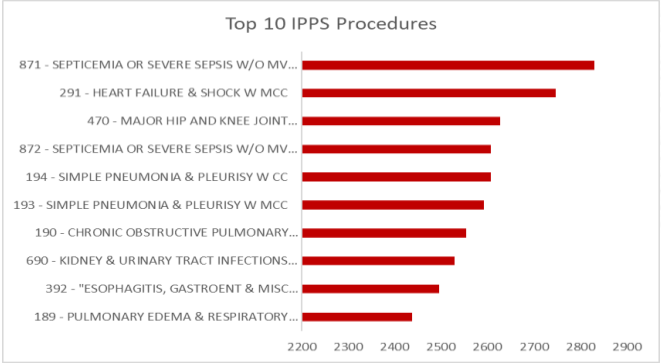
- Data Sources
  - Data_with_Labels_*.xlsx : These are the files containing the “labels” for a subset of hospitals from all over the US.
  - Rand_hcris_cy_hosp_a_2020_11_01.csv: This is the zipped full hospital cost report information system (HCRIS) data file having all the financial metrics for all hospitals across several years
  - Medicare IPPS charge data : This dataset contains the utilization and charge data for providers participating in IPPS (from hereby referred to as ‘IPPS hospitals’) for 2011-2018
  - Medicare OPPS charge data : This dataset contains the utilization and charge data for providers participating in OPPS (from hereby referred to as ‘OPPS hospitals’) for 2011-2018
- These data sources had 1500+ variables available for analysis

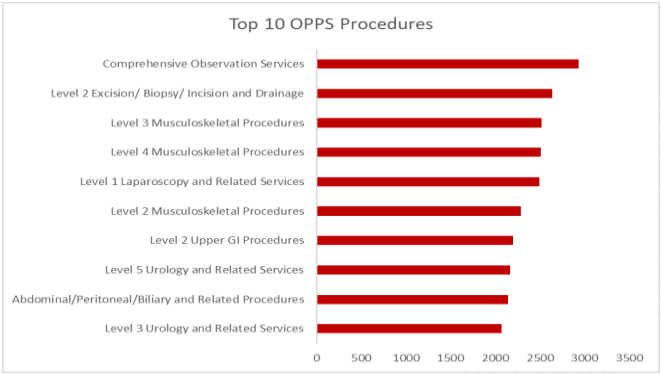
Data Ingestion

- All available data sources listed in slide 1 were merged together on a granularity of medical provider number that is a unique identifier for hospitals
- Relative price for outpatient services was selected as dependent variable. Other price/cost variables including inpatient relative price were not considered as target due to high percentage of nulls.
- All other relative/standard price variables were removed from independent variables to avoid any target leakage.
- Derived variable scores were created from IPPS and OPPS data sources using the following steps:
  - Find and filter for the most common procedure 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

OPPS/IPPS Most Common Procedures by Hospitals – For Derived Scores

- IPPS
  - 291 - HEART FAILURE & SHOCK W MCC
  - 871 - SEPTICEMIA OR SEVERE SEPSIS W/O MV >96 HOURS W MCC
- OPPS
  - Comprehensive Observation Services
- Calculation Example for OPPS normalized on State Procedure:
  - Comprehensive Observation Services Score
  - (Average_Estimated_Total_Submitted_Charges_FOR_PROVIDER - Average_Estimated_Total_Submitted_Charges_FOR_STATE)/ Average_Estimated_Total_Submitted_Charges_FOR_STATE





Data Preparation

1. Variable with **redundant** information were discarded reducing the variable list by 40 for example Hospital Name, wt_cy_1 etc

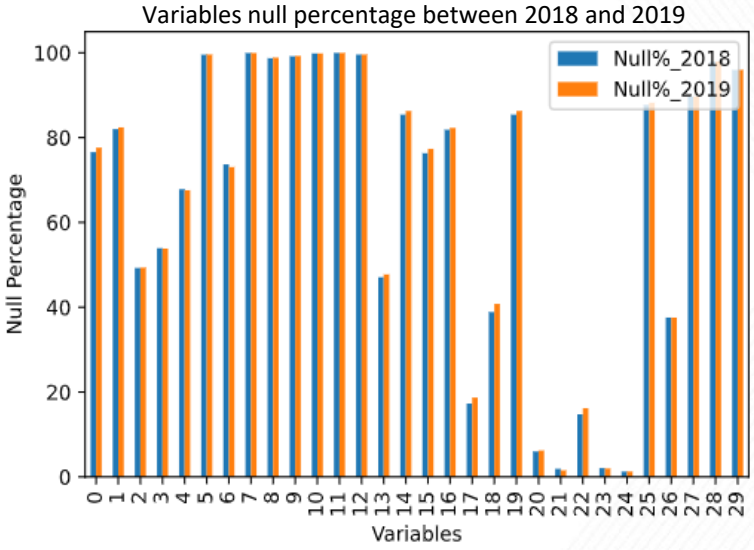
1. Variables with **categorical** values were dummy encoded for example: teach_hosp_hcr, rural_urban etc

1. Percentage of missing information was calculated for each variable as illustrated in right figure by year. It shows that both years have almost same null percentage. Variables were further shortlisted with **Nulls > 10%** leaving behind 300+ variables

1. Derived variables from IPPS/OPPS had around 20% nulls when joined with labeled data and rand_hcris. These variables were excluded from step 3 filter

1. Nulls were replaced by average of state for each variable. National average was used if the state average is null (example expns_nursing_admin_salary for state NY) for all providers

1. Rand_HCRIS 2018 was used for **training**/validation and 2019 for **scoring**

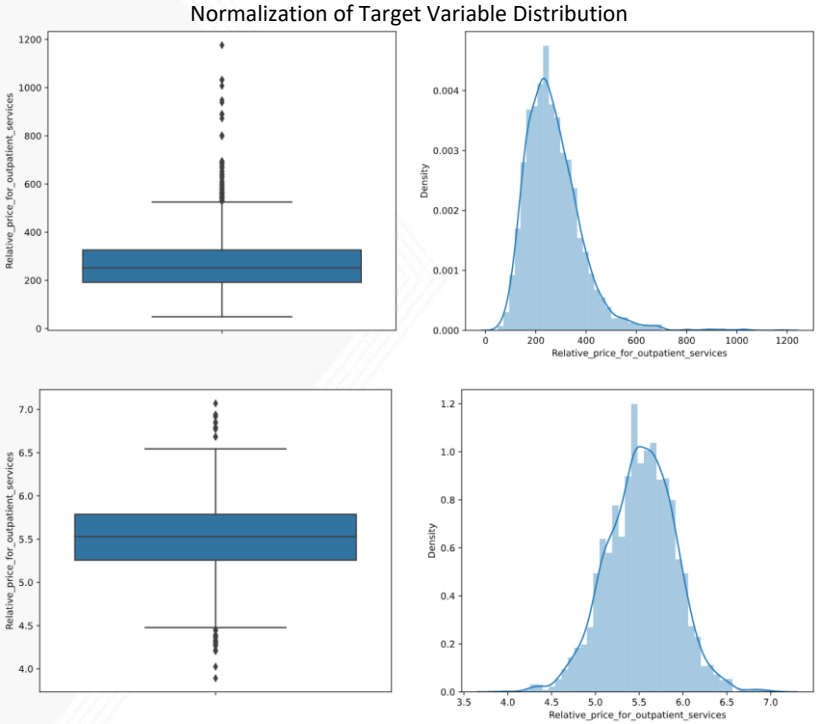


*Randomly selected 30 variables for display purposes

Modeling Approach

- Regression – Modeling Approach 1
  - Set relative price as target and use a transformation to make it normally distributed
  - Reduce variables by LASSO
  - Input the reduced set of variables to Regression model to predict relative price of hospitals
  - Repeat the regression modeling after creating cluster and evaluate if clustering would improve the results
- Classification – Modeling Approach 2
  - Transform the relative price target variable to categorical variable based on the condition relative_price < 150 as low price hospitals
  - Reduced variables by LASSO
  - Input the reduced set of variables to ensemble model (Gradient Boosting Trees) to predict the probability of a hospital to be low price
  - Repeat the modeling after creating clusters and evaluate if clustering would improve the results

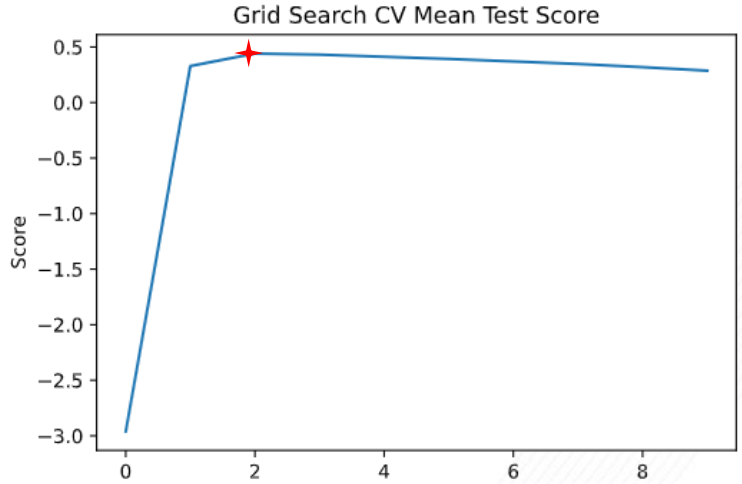
Modeling Approach 1: Regression Target Normalization



- Target variable relative price for outpatient services was normalized as illustrated in box plot and histogram

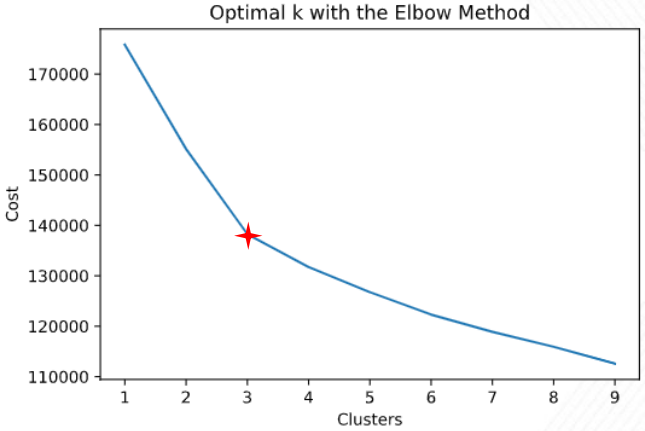
Modeling Approach 1: Lasso Regularization

- Grid search cross-validation yielded an alpha of 0.0051
- Pre-training variable count is 290
- Post-training variable count is 57
  - 57 variables with 3,084 records were chosen for clustering

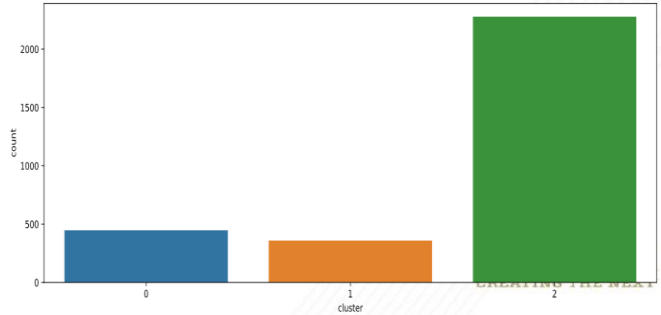


Modeling Approach 1: K-Means Clustering

- Three clusters were selected based on elbow method
- Initial summary from three clusters:
  - Cluster 0: Low price, high rating, 447 counts
  - Cluster 1: High price , low rating, 359 counts
  - Cluster 2: Medium price, medium rating, 2278 counts



cluster	Hospital_Compare_5-star_rating_(October_2018,_NA=Not_Available)	Relative_price_for_outpatient_services	Relative_price_for_inpatient_services
0	3.342282	187.237136	170.722595
1	2.908078	291.718663	259.086351
2	3.021949	282.034241	207.474100

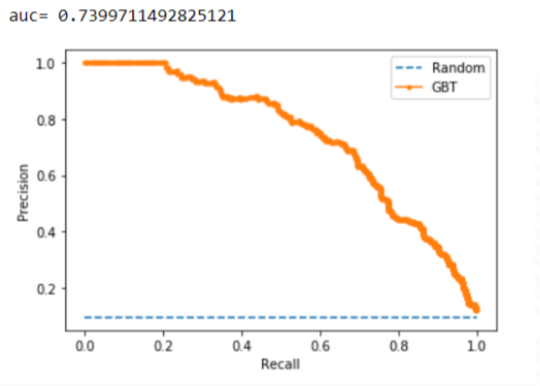


Modeling Approach 2: Classification

- Transform the relative price target variable to categorical variable based on the condition relative_price < 150 as low price hospitals
- This condition on target marks 10% of hospitals as low price hospitals
- Execute LASSO after transforming independent variables to Z-Scores using StandardScaler.
- LASSO reduced the variables to 28 that had a relationship with low price hospitals
  - These included 3 derived variables that were calculated from IPPS/OPPS
    - COVERED_CHARGES_STATE_SCORE_DRG871', 'OPPS_Comprehensive_APC_Services_Perc', 'OPPS_Total_Submitted_Charges_NATIONAL_SCORE'
- Input these 28 variables to ensemble model Gradient Boosting classifier to assign probability of being a low priced hospital. Quality star rating was not used as independent variable since we do not want quality to determine relative price. Quality star rating will be used in clustering.
- This would be the baseline model

Modeling Approach 2: Evaluation

- Scoring on 2019 rand_hcris data and evaluating on relative price with optimal F1 score, we get the following results:
  - AUC: 73.9
  - Recall: 58.2
  - Precision: 76.9
  - Accuracy: 94.4



Medicare_provide	State	Hospital_Compare_5-star_rating	Relative_price_flag	Relative_price_for_outpatient_service	Prediction_Probability	Low_price_predicted_flag
441300	TN	3	1	49	0.987676755	1
271333	MT	4	1	118	0.984414769	1
390324	PA	3	1	120	0.984034086	1
131321	ID	4	1	110	0.978017353	1
171310	KS	3	1	114	0.972672132	1

Medicare_provide	State	Hospital_Compare_5-star_rating	Relative_price_flag	Relative_price_for_outpatient_service	Prediction_Probability	Low_price_predicted_flag
110029	GA	3	0	364	0.001246511	0
450358	TX	5	0	337	0.001252978	0
100008	FL	2	0	548	0.001265959	0
310001	NJ	3	0	296	0.001335564	0
150162	IN	2	0	304	0.001360554	0

Final Report Plan

- For both modeling approaches, we plan to run respective models by Clusters
- Compare the results and accuracy of models after clustering with baseline models without clusters
- Finalize the approach (Classification/Regression, with or without cluster) to predict or rank the relative price of hospitals
- Map the results with Quality star rating so hospitals with low price and high quality can be recommended to customers.

References

To understand the problem, healthcare industry and analytic technique, we read the following papers and went through available data/documentation:

- The Utility of Clustering in Prediction Tasks[1]
- Prices Paid to Hospitals by Private Health Plans Are High Relative to Medicare and Vary Widely[2]

1. Trivedi, S., Pardos, Z.A., Heffernan, N.T. (2015). The Utility Of Clustering in Prediction Tasks. https://arxiv.org/abs/1509.06163
2. 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