HOSPITAL SELECTION

FINAL REPORT

ORIE 4741

Dayin Chen (dc652), Anne Ng (an428), Willie Xu (wx44)

2016/12/05

Introduction and Overview

Our team undertook the task of building a recommendation model for patients to help them decide which hospital to go to given their illness. In the process, we hoped to extract data insights for patients regarding hospital selection, and also inform hospital administrators on the efficacy of procedures.

However, due to the lack of public data with raw indicators of success by procedure and hospital, we were not able to asses which hospitals have the highest success rates for different illnesses. Therefore, early on we decided to analyze the question from a pricing perspective. We also narrowed the scope to heart failure. The reason being that price is highly correlated with cause of admission. You'll be charged a very different fee for heart failure compared to a sprained ankle, so it was not feasible to fit a model to the entire data set of patient discharges. We chose heart failure because it is one of the most common hospital admission reasons by volume in New York state, behind child birth and septicemia.
Given that we know both the associated costs and final charges to a patient for each visit, we want to predict how much a hospital will charge you for a visit given your personal information and the type and severity of your illness. If available, it will also take into account the quality of hospitals in terms of safety, mortality rates and preventable re-admission rates.The purpose of this analysis is to help patients and insurance providers identify whether a hospital is overcharging. At the same time, hospitals can use this information to reflect on its pricing model. Essentially, we hope to answer the following questions:

1. Patient: Given my medical illness and the desired quality of hospital, which hospital should I go to to avoid being overcharged?
2. Hospital administrator: Is it costing my hospital a lot more than other hospitals to perform the same procedure?
3. General: Does the median income and poverty rate of a county influence the cost of care at its hospitals?

The overall goal is still the same - to help patients, insurance providers and hospitals make better choices.

Data Sets and Processing

We used 5 data sets: the first data set contains all patient discharges information in New York in the year 2012. We joined the first data set with the four other data sets with information about hospital qualities measured by different indicators - Inpatient Quality Indicators (IQI) Composite and individual conditions, Patient Safety Indicators (PSI) Composite Measures, and Potentially Preventable Readmission Rate (PPR).

1. Hospital Inpatient Discharges (SPARCS De-Identified): 2012

   https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/u4ud-w55t

2. All Payer Inpatient Quality Indicators (IQI) by Hospital (SPARCS): Beginning 2009

   https://health.data.ny.gov/Health/All-Payer-Inpatient-Quality-Indicators-IQI-by-Hosp/xyfc-qbbr#column-menu

3. All Payer Inpatient Quality Indicators (IQI) Composite Measures by Hospital (SPARCS): Beginning 2009

   https://health.data.ny.gov/Health/All-Payer-Inpatient-Quality-Indicators-IQI-Composi/ba3n-bkk4

4. All Payer Patient Safety Indicators (PSI) Composite Measures by Hospital: Beginning 2009

   https://health.data.ny.gov/Health/All-Payer-Patient-Safety-Indicators-PSI-Composite-/jjir-cuty

5. Hospital Inpatient Potentially Preventable Readmission (PPR) Rates by Hospital (SPARCS): Beginning 2009

   https://health.data.ny.gov/Health/Hospital-Inpatient-Potentially-Preventable-Readmis/amqp-cz9w

6. US Census Small Area Income & Poverty Estimates (SAIPE) 2012 for NY counties

   https://www.census.gov/did/www/saipe/data/interactive/saipe.html?s_appName=saipe&map_yearSelector=2014&map_geoSelector=aa_c

The original data sets contain 2,544,731 rows of data points, corresponding to the same number of patient discharges. Before we conducted any analysis on the data sets, we cleaned and turned the data into usable form. We fixed or removed some corrupted or missing data points, extracted features we wanted to analyze, and converted some values by the following rules:

1. Shrinking data

   Because cost is naturally correlated with the type of illness you are experiencing (a hospital visit due to cardiac arrest is probably much more expensive than one for a sprained ankle), we are not able to fit one single model to the data set. Instead, we extracted visits associated with 3 of the most common hospital admission reasons (vaginal delivery, septicemia, and heart failure) and fit separate models to each.

2. We removed 4308 rows of data points with Abortion Record - Facility Name Redacted since the data set doesn't disclose which hospitals perform abortions.

3. Features transformation

   (a) Boolean Data- For data of this type, with only two categories, such as *Emergency Department Indicator* which tells us whether the patient was admitted through the emergency room, we simply assigned values of 1/0 or 1/-1 to the categories. One feature of note is *Gender*, in which we assigned 'M'/'F' values of 1/-1 and those with 'U', from which we inferred indicated transgenders or unknowns a value of 0.

   (b) Ordinal Data- For data of this type, such as with *APR Risk of Mortality* and *Age Group*, we assumed even spacing of the ordinal data values and thus assigned them numerical values corresponding to consecutive integers. Thus, 'Minor' might equate to 1, 'Moderate' to 2, 'Major' to 3, and so on forth.

   (c) Categorical Data- There are columns such as *Patient Disposition* that tell us where a patient is discharged (home, inpatient rehabilitation etc). Because there was no clear cut numerical ordering of values to this type of data, we broke these categories up into multiple boolean features using one hot encoding. Thus, each category was extracted

as its own separate new feature, giving us a number of new features equivalent to the number of categories present in the original.

(d) Miscellaneous- Some numerical data was not fully numerical in the original data-set. For instance, some were capped at a certain amount and any data-points that were over that amount were just assigned a flag such as, '120+'. These flags were simply converted to the maximum possible number for that feature (in this case, 120) and a new boolean feature was added for the presence of that flag. In addition, money values for costs and charges in the form of strings were simply converted to their float counterparts.

4. Bias - we added a new column of ones to account for bias.

After data processing, we have 3 data sets for 3 different conditions: vaginal delivery of babies, septicemia, and heart failure.

1. Vaginal Delivery: 151,610 rows
2. Septicemia: 66,790 rows
3. Heart Failure: 53,118 rows

   Note : Besides the composite IQI composite measure, we also have the risk adjusted rate (RAR) of heart failure mortality. As a result, we added an additional specific feature to our heart failure data set.

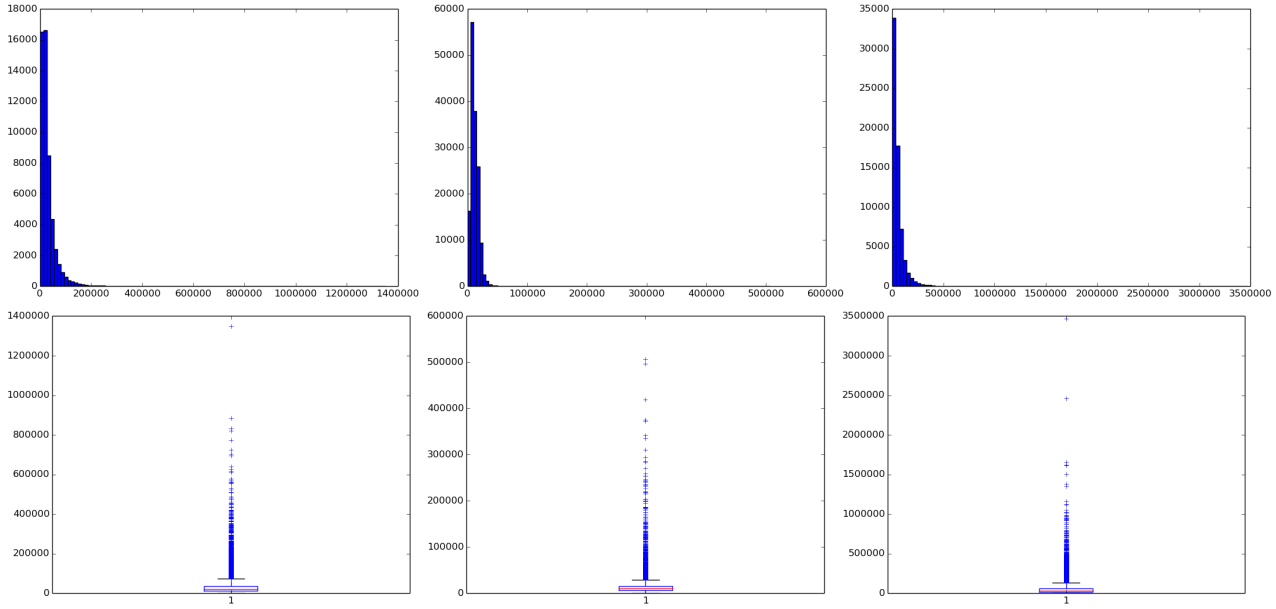
Feature Explanations

All of our data is extracted from Statewide Planning and Research Cooperative System (SPARCS). Detailed explanations of the features can be found on http://www.health.ny.gov/statistics/sparcs/. In our final data sets, there are 22 or 23 unique features, including 3 (vaginal/septicemia) or 4 (heart failure) hospital quality indicator features. Here is a brief explanation of some of the key features:

1. Hospital name and country
2. Patient age group, gender, and length of stay
3. Indicators of whether patient was admitted for urgent care, or emergency room (ER) care
4. APR Severity of illness: extent of organ system loss of function
5. APR Risk of Mortality code: likelihood of death
6. IQI (Inpatient Quality Indicator) Risk Adjusted Rate of heart failure mortality, per 1,000 discharges. (Observed Rate / Expected Rate) * Statewide Rate
7. IQI Composite Measure of selected conditions, the weighted average of the observed-to-expected ratios of its component indicators (mortality rates for selected conditions) by hospital
8. PSI (Patient Safety Indicator) Composite Measure, the weighted average of the observed-to-expected ratios of its component indicators (rates of selected hospital complications and adverse events following surgeries, procedures, and childbirth)
9. PPR (Risk Adjusted Potentially Preventable Readmission Rate), per 100 people, the observed PPR rate divided by the expected PPR rate, multiplied by the statewide PPR rate.
10. Median Household Income and Poverty Rate

Visualizations and Statistics

The following are the histograms and boxplots (from left to right: heart failure, vaginal delivery, septicemia) for the total charges. We can see that the total charges are strongly right-skewed with a lot of outliers, posing challenges to fitting a regression model, such as quadratic regression, which is sensitive to outliers.



Approach

1. Cross-Validation

   To avoid over-fitting, we split the data into a training set and a test set. 60% of the data went to the training set, and 40% went to the test set. We do cross-validation for each of the following approaches.

2. Linear Regression (Quadratic)

   We start with quadratic linear regression as it usually acts as the most popular base case for supervised learning.

3. Random Forest

   In order to identify variable importance, we implemented a random forest algorithm in R using the package "RandomForest". Random forest is a good way to avoid over-fitting. However, since the error given by random forest algorithm (shown below in results) did not improve that significantly, we are only concerned with variable importance in this report.

   Reference: https://www.stat.berkeley.edu/~breiman/RandomForests/

4. Linear Regression (Quantile Loss)

   From the previous data exploratory in "Visualizations and Statistics", we observed a large amount of outliers. We choose this approach since quantile loss function is not as sensitive to outliers. Also, we are interested in what causes the huge parity within the total charges for similar conditions. A quantile regression analysis of coefficients can help us identify the major variable that contributes to the large differences in total charges.

5. Advanced: Neural Net

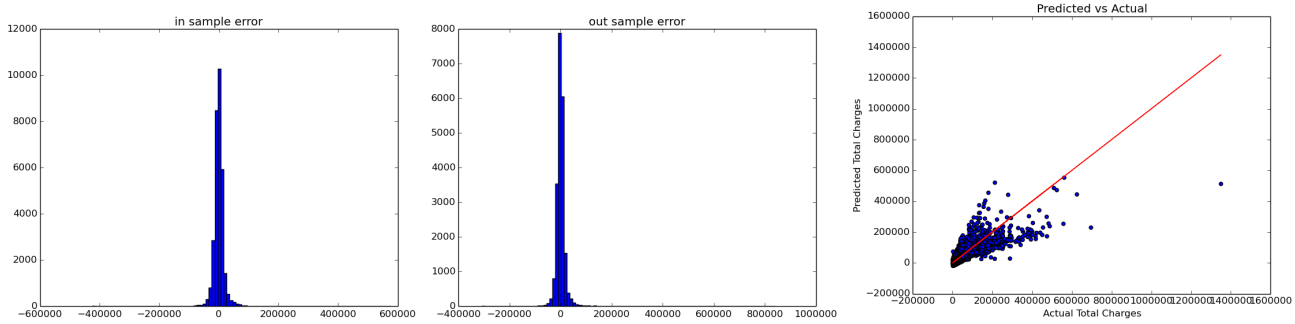
Data Analysis

Results

After running similar algorithms on all three different data sets for heart failure, vaginal delivery and septicemia, we find that the they show very similar patterns and results. Given constraints on the length of this report, we will analyze and discuss the results for heart failure only.

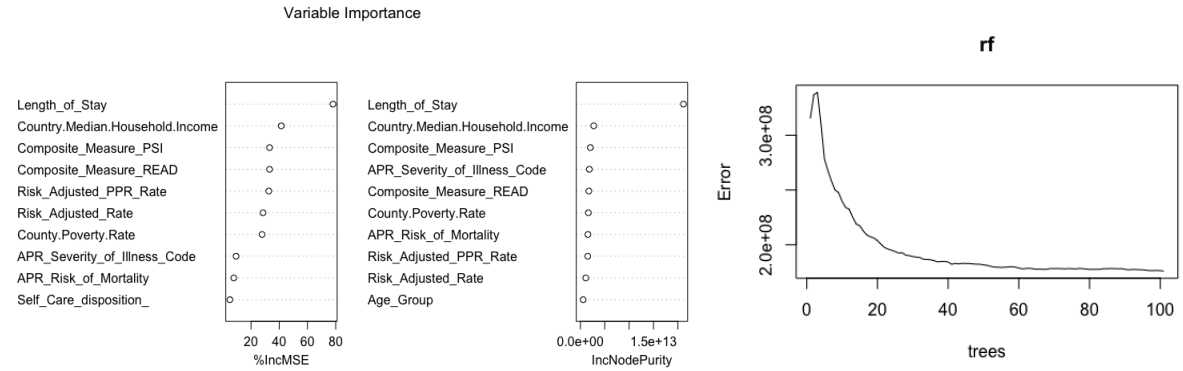
1. Quadratic

The following shows the non-absolute errors (actual - predicted) and the actual vs prediction plots. We can see that the errors have a long tail but center around 0. This means that for some data points, the errors can be so high that the prediction cannot be trusted. One reason for this observation is that the outliers in our data affect the fitting of the linear model. Also, from the predicted vs. actual plot, the quadratic regression does not give a good fit.



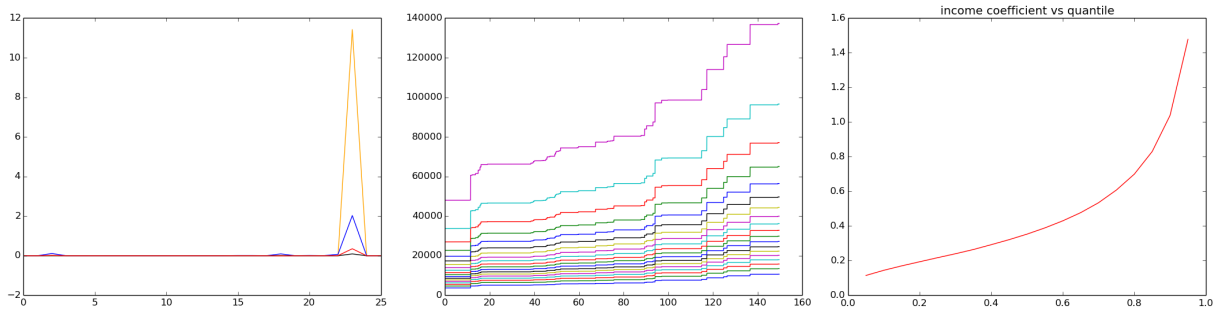
2. Random Forest

We identified some more important variables by running Random Forest Algorithms using 101 trees on R. The graph on the left indicates the significant variables and the graph on the right indicates the error against the number of trees.



3. Quantile Loss

The first plot is the plot of values of weights for each features, with offsets at last, from 2.5 percentile to 97.5 percentile (from bottom to top) The second plot is the predicted values for each quantile regression coefficients from 5 to 95 percentile (from bottom to top). The third plot shows that coefficients for the feature "Country Median Household Income" increases at an increasing rate with quantile.



Discussion of Results

From the results above, we find that the accuracy for quadratic linear regression is low and unreliable especially when the charges are at the high end due to the amount of outliers.

Looking at the results for random forest, we see that length of stay, county median household income and some hospital quality indicators are the variables that are more significant. On the other hand, patient information is not as significant. The most significant patient information is APR Severity of Illness Code, which likely correlates with length of stay.

Since quantile loss functions are not sensitive to outliers, they seem to be a better fit for the purpose of our data. We can also see how the total charges are stratified according to different variables. From the results, we can see that the total charges increase at an increasing rate. As we compare the values of weights, we see that significant differences in prices come from the variable "County Median Household Income". The weights on the county median household income increases at an increasing rate with quantiles, showing that hospitals in affluent counties charge more for high-cost stays.

Based on our interpretation of the results from quadratic, random forest and quantile regressions, we see that county median income and hospital qualities play significantly larger roles than the demographic information of patients. This is partly against our initial assumptions that charges would be dependent on patients' severity of condition. Among the variables, county median household income plays exponentially greater roles as the charges increase. Therefore, when a patient is charged more, it could be due to the neighborhood of the hospital where service has a premium.

Deep Learning

Given the low success rates with fitting regressions to the data, we decided to use deep learning to predict charges to patients and formed a neural network with input nodes corresponding to the number of features in each data-set (heart failure contained an extra RAR feature indicating mortality rate associated to it) and an output node indicating charge to the patient.

Neural networks consist of a network of nodes, or "neurons", that pass information from previous nodes to the next. Input nodes simply pass along information that they are given by the user to nodes following it in the neural network, while output nodes accumulate information from all of the messages being passed around to make a prediction for the user. Neural networks are organized in layers, and all of the nodes between the inputs and outputs belong to hidden layers that gather

information from the previous layer, apply weights to the received information according to the nodes, and pass new information to the next layer.

The learning happens in a method called *back-propagation*, in which the weights in the neural network are tuned according to new data by minimizing a loss function with respect to the weights. The optimization method can be done with methods such as gradient descent.

Many aspects of a neural network are fully customizable, including the size and shape of the network, the rate of learning, the loss function being used, etc. This allows deep learning to be applicable in a wide variety of situations, and to many different problems. Some of the main drawbacks to deep learning with neural networks include the relative ease in which the model can over-fit the data and the relatively longer run-time needed for the learning process.

For our problem, we played around with neural networks of a few different sizes, ranging from 3-5 layers and with varying amounts of neurons in each, as well as differing learning rates and learning decay. The loss function used was squared loss of the output neuron (charge to patient).

After much experimentation, we achieved the following results with the optimal corresponding neural net for each of our three patient conditions when verified on our testing data. Since the error is squared loss, I have also provided the square root of the avg and median error so that they can be compared to the average output value:

1. *Vaginal Delivery*

$$\begin{aligned}\text{Average Output: } & 11,911\\ \text{Average Error: } & 4,409,085\\ \text{Sqrt of Avg Error: } & 2,100\\ \text{Median Error: } & 986,022\\ \text{Sqrt of Median Error: } & 993\end{aligned}$$

2. *Heart Failure*

$$\begin{aligned}\text{Average Output: } & 31,057\\ \text{Average Error: } & 33,509,046\\ \text{Sqrt of Avg Error: } & 5,789\\ \text{Median Error: } & 9,465,832\\ \text{Sqrt of Median Error: } & 3,077\end{aligned}$$

3. *Septicemia*

Average Output: $53,445$
Average Error: $20,696,764$
Sqrt of Avg Error: $4,549$
Median Error: $2,780,765$
Sqrt of Median Error: $1,668$

As can be seen from the data, Neural Networks provide a fairly accurate estimation of patient charge given, the patient's information, disposition, condition, etc. This can be seen by the relatively low square root of error, in comparison with the average output for that condition. As an example, for septicemia patients, the average charge is \$53,445, while the square root of median error(squared loss) is only \$1,668. The most difficult out of the three conditions to predict was heart failure, with its higher error relative to the average patient charge, in comparison to the other two conditions.

Conclusion

We conclude with the following insights:

1. Patient: Given my medical illness and the desired quality of hospital, which hospital should I go to to avoid being overcharged?

   Our Answer: The most important factors that influence cost of care are 1)length of stay, 2)median household income, and 3) the PSI indicator that reflects the observed-to-expected ratio of complications.Thus, assuming the same quality of care, to avoid high costs, you should go to a hospital in a county where the median household income is not very high, but has a high PSI (data which is publicly available), which would reduce the chance of having a prolonged stay due to complications.

2. Hospital administrator: Is it costing my hospital a lot more than other hospitals to perform the same procedure?

   Our Answer: This can be answered directly by visualizations of cost of treatment across hospitals for the same illnesses, and doesn't require any modeling.

3. General: Does the median income and poverty rate of a county influence the cost of care at its hospitals?

   Our Answer: Based on our results, they are very much correlated. Median income plays a significant role, even more than hospital quality or patient information. However, we should be mindful that this may not be a causal relationship. Instead, method of payments, insurance coverage, and quality of medication and service might be a factor in why hospitals in higher median income counties charge more.

This data set was hard to perform predictive analytics on, because at the end of the day, patients care about success, and cost of care is not directly correlated with success. Unlike predicting credit card approvals, our model does not mimic the process for determining hospital charges. Hospitals consider cost of medication as well as number and complexity of procedure, which is not captured in this data set. This data can be more reliably used for conducting hind-sight diagnostics for hospitals.