HOSPITAL SELECTION
MID-TERM REPORT

ORIE 4741

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

2016/10/28

Introduction and Overview

Our team has done an initial canvas of the main data set and extracted the most relevant features for analysis. Our initial goal was to build 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, insurance providers and hospital administrators as they make decisions regarding hospital selection or hospital policies.

However, we have not been able to find raw indicators of success by procedure and hospital, so we have not been able to asses which hospitals have the highest success rates for different illnesses. Therefore, we decided to analyze the question from a pricing perspective. Given that we know the associated costs and final charges to patients 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.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. The overall goal is still the same - to help patients, insurance providers and hospitals make better choices.

Data Sets and Processing

We used three data sets: the first data set contains all patient discharges information in New York; we joined it with the second data set which contains information on number of beds available at each facility. The third data set contains information of total hospital discharges.

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

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

2. Health Facility Certification Information

   https://health.data.ny.gov/Health/Health-Facility-Certification-Information/2g9y-7kqm

3. All Payer Hospital Inpatient Discharges by Facility (SPARCS De-Identified): Beginning 2009

   https://www.health.data.ny.gov/Health/All-Payer-Hospital-Inpatient-Discharges-by-Facilit/ivw2-k

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. The values of number of certified facilities for two hospitals are missing because they closed or merged with other hospitals. Those values are replaced by past information found on the internet or mean values.

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. This is subject to change.

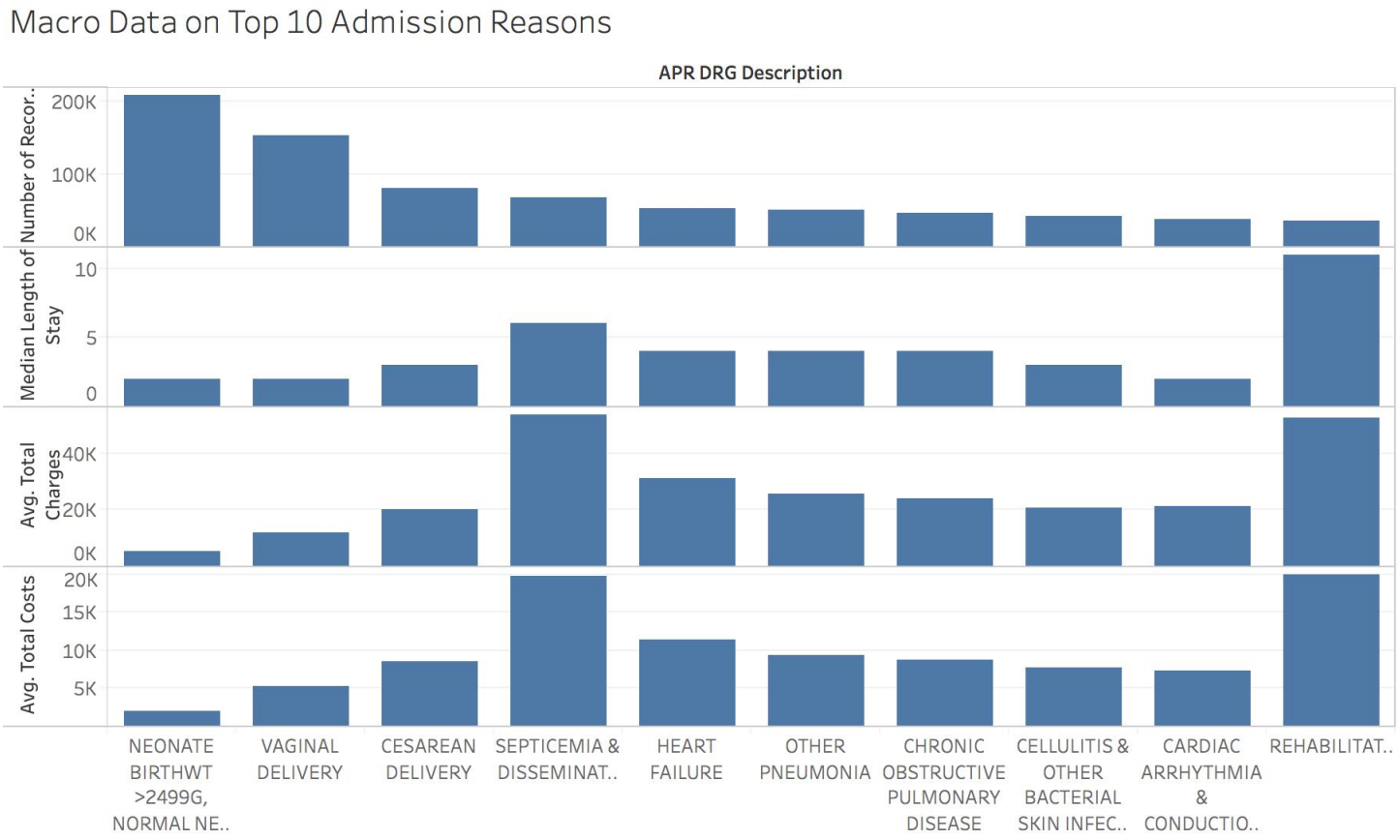
(c) Categorical Data- There are columns such as *Patient Disposition* that tell us where a patient is discharged (home, inpatient rehabilitation etc). Because there was no clear cut numerical ordering of values to this type of data, we broke these categories up into multiple features. Thus, each category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

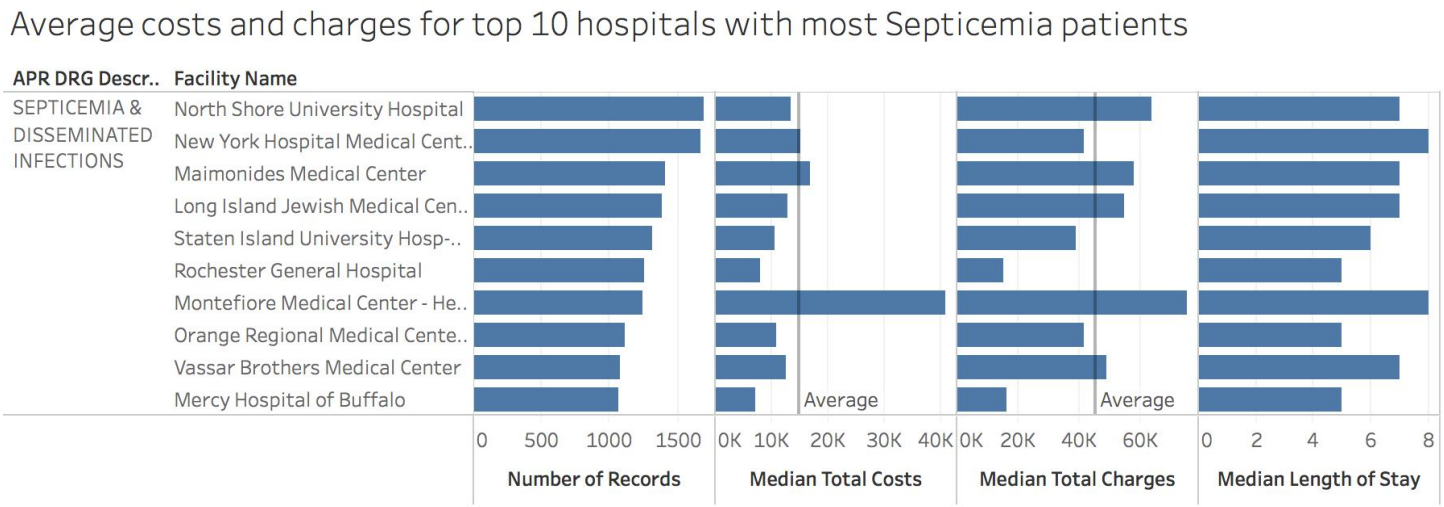
4. Combining Data sets

   We extracted information of facilities/hospitals from second and third data sets and joined them with each patient record in the first data set by facility name.

Visualizations and Statistics



Sum of Number of Records, median of Length of Stay, average of Total Charges and average of Total Costs for each APR DRG Description. The view is filtered on APR DRG Description, which keeps 10 of 316 members.



Sum of Number of Records, median of Total Costs, median of Total Charges and median of Length of Stay for each Facility Name broken down by APR DRG Description. The view is filtered on Facility Name, APR DRG Description and sum of Number of Records. The Facility Name filter keeps 227 of 227 members. The APR DRG Description filter keeps SEPTICEMIA & DISSEMINATED INFECTIONS. The sum of Number of Records filter includes values greater than or equal to 1,000.

We see that the top three patient categories are related to childbirth and infant care, followed by illnesses and diseases. From the second chart, we used a median statistic to compare patient charges and costs relative to length of stay. We see that charges are not immediately correlated with costs for different hospitals. It may be due to how some hospitals specialize in certain areas, so they may charge more for the same procedure.

How to avoid over-fitting or under-fitting

To avoid over-fitting, we can use regularization method to shrink our hypothesis space. To check whether under-fitting occurs, we can use the bias-variance approach. In general, we can use cross validation to check the strength of our model. We split our data set into three parts - 60% training set, 20% validation set and 20% test set.

Preliminary Analysis

We tried a simple linear regression model on our training set. The feature space X contains 20 numerical values and the dependent variable y is the total charges for each patient. w is found by $\mathrm{w} = X \backslash y$ on Julia.

X	w	X	w	X	w
Zip Code	422.139	Gender	-185.983	Length of Stay	700.069
Elective Ad	3579.02	Emergency	27089.4	Newborn	31329.3
Trauma Ad	29005.1	Urgent Ad	49799.5	Self-Care Dispo	26844.8
Home Service Dispo	-650.054	Nursing Home Dispo	-313.915	Short-term Hospital Dispo	-3362.42
Expired?	3303.37	APR Severity Code	6761.31	APR Risk of Mortality	1966.57
APR Medical/Surgical	4444.26	Emergency	11921.1	Total Beds	-2046.2
Total Services	18.6201				

Next Steps

From the analysis, we see that pricing models might be vastly different for different admittance reasons, so we will try fitting separate models for each of the top ten admittance reasons. Moreover, there might be some correlations between the independent variables that we need to look into in further details, to make a better choice for our feature space. In addition, we will add regularization and suitable feature transformations that better reflect differences among hospitals.