CornellEngineering
Operations Research
and Information Engineering

Midterm Report

Hospital Selection

Hao Qiu, Yuting Zhao, Yuze Wang

November 6, 2019

1 Problem Statement

Our project goal is to provide the optimal hospital selection recommendation based on the patient's medical condition and other relevant features. Given a set of personal information about a certain patient, we aim to offer an estimate of their expected outcome and the total charges incurred. As a result, we will attempt to direct this patient to the hospital that would best fit their needs.

2 Dataset

2.1 Dataset Description

Our dataset is provided by the Statewide Planning and Research Cooperative System (SPARCS) Hospital Inpatient Discharges for New York State published on health.data.ny.gov. This dataset keeps track of all patients who have been discharged from New York hospitals during 2012 and records the corresponding treatments and charges.

It originally contains 2.54 million rows of data points and 34 different columns ranging from qualitative features, including Hospital Service Area and patient's demographic background, to quantitative features such as patient's Length of Stay, the Severity of Illness, Costs, and etc. The dataset scales several ordinal features into numeric levels; for example, it transforms the Severity of Illness into four levels: 1 = minor and 4 = extreme. It also groups the ages into 0-17, 18-29, etc and displayed the Length of Stay that are over 120 days as "120+."

2.2 Data Cleaning

Before conducting our analysis, we cleaned the data and performed feature transformation on several columns to prepare for model fitting. We picked one of the Major Diagnostic Categories with the most patient records: *Diseases and Disorders of the Circulatory System*, extracted features that are potentially related to our analysis, and removed rows with missing data entries.

2.3 Feature transformatiom

2.3.1 Boolean Data

For data such as Emergency Department Indicator that only falls under two categories - whether the patient was admitted through the emergency room or not, we assigned values of 1, 0 to the entries. In a similar fashion, we assigned 1, -1, 0 to the Gender which originally has values of 'M', 'F', 'U' respectively where 'U' represents an unknown gender.

2.3.2 Ordinal Data

We encoded all categorical data with ordinal levels into numerical values displayed as consecutive integers.

2.3.3 Categorical Data

Columns like Patient Disposition represents if a patient is discharged to home, inpatient rehabilitation, or etc. We grouped these types into five broader categories: Home, Continued, Hospice, Died, and Other, where Other includes all other types that do not fall under any previous category such as law enforcement, etc.

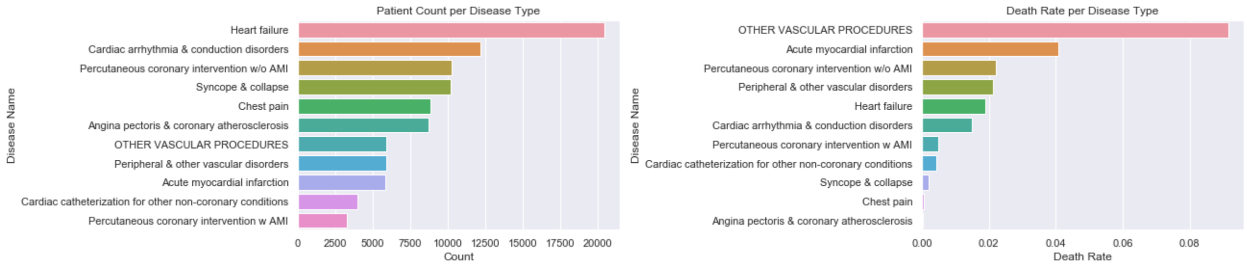
2.3.4 Miscellaneous

Some data entries under Length of Stay were not purely numerical but mixed with a text symbol. The original data set an upper limit on this feature: it assigned a value of "120+" for the data points that are over 120 days. For simplicity, we converted these values into the maximum value: 120.

3 Visualization

To gain a better understanding of our data, we first plot Patient Count by Disease Type for Diseases and Disorders of the Circulatory System, ranking the pervasiveness of diseases. As shown in the graph, heart failure is the most common disease in circulatory system disorders. We then plot the distribution of total charges for all circulatory diseases to visualize one of our output spaces.

Figure 1(a) shows the number of patients recorded for each circulatory system disease. The New York hospitals had received over 20,000 heart failure, about 12,500 cardiac arrhythmia & conduction disorders, and about 10,000 percutaneous coronary intervention w/o AMI cases which rank as top three common conditions among the circulatory diseases.



(a) Patient Count by Disease Type

(b) Death Rate by Disease Type

Figure 1: Patient Count and Death Rate by Disease Type

Figure 1(b) exhibits all circulatory diseases with decreasing order of the death rate. "Other Vascular Procedures" has the highest death rate which almost reaches 10%. However, some diseases among the top do not have a significant amount of patient records according to Figure 1(a). "Other Vascular Procedures," for example, has the highest death rate but with only about 6,000 patient count.

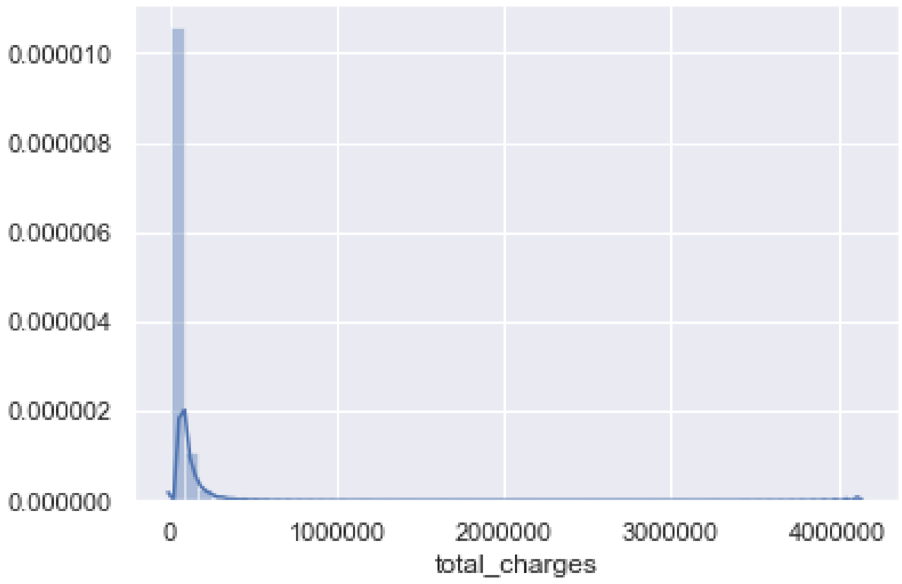


Figure 2: Total Charges

Figure 2 displays the distribution of total charges. As shown above, it is largely skewed to the right with a mean value of below 1,000,000 dollars but a maximum value of over 4,000,000 dollars.

4 Preliminary Analysis

In this study, we sought to examine diseases and disorders of the circulatory system, which is the disease with most cases in the dataset. We fitted our preliminary models to the data on circulatory diseases and attempted to predict the total charges incurred and expected outcomes. Our preliminary models formulated parameters to predict based on various features and tested the accuracy of the fits on untrained data by splitting the data into train and test sets.

We developed 1 model that predicts the total charges and 1 model that predicts the patient outcome.

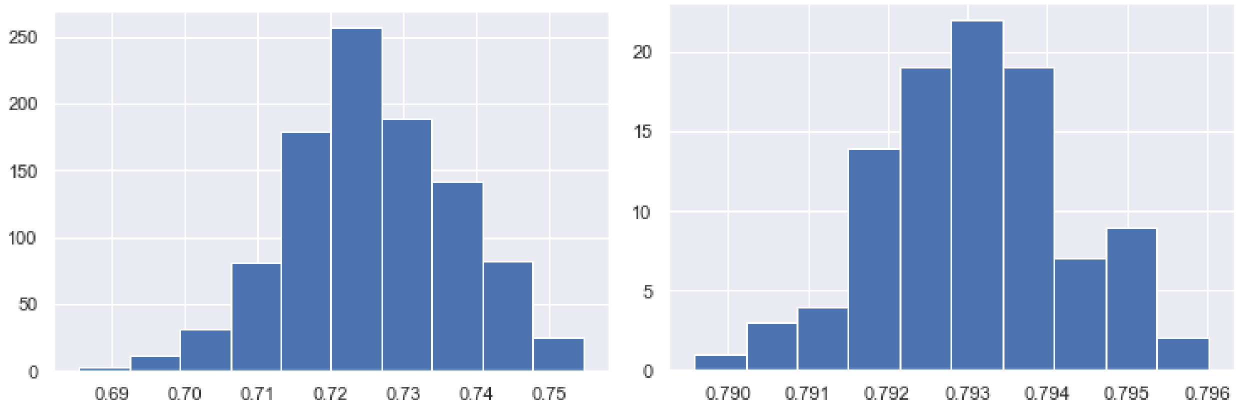
- Our first model is a linear regression that uses parameters including Severity of Illness, Length of Stay, and etc to predict the total charges.
- Our second model is a logistic regression that uses parameters including Severity of Illness, Length of Stay, and etc to classify the expected outcome of treatments.

Each model is fit to and cross-validated against a training set to develop a predictive fit parameter w, then this w is tested on a previously unused test set.

5 Validation

We developed several methods to test our models' ability to generalize and avoid overfitting. In our training stage, our models are cross-validated against different partitions of the training set. We developed a bootstrap estimator that fits the model using randomly re-samples from the training set and calculated an out-of-sample R^2 and accuracy using previously unseen test sets.

Figure 3(a) is a histogram of the R^2 of linear regression from 1000 bootstrap replications, and figure 3(b) is a histogram of the accuracy of logistic regression from 100 bootstrap replications. As we can see, the out-of-sample R^2 has a relatively high variance. We are working to develop less variant measures for our test set evaluation in order to offer a more stable evaluation measurement.



(a) Bootstrap Linear R^2

(b) Bootstrap Logistic accuracy

Figure 3: Bootstrap Logistic Regression

6 Moving Forward

Among the circulatory diseases, we will pick three severe conditions that have the lowest survival rate in practice. As illustrated in Figure 1(a) and 1(b), within the top common diseases (over 7,500 records), the ones with the highest death rates are as follows: *Percutaneous coronary intervention w/o AMI, Heart failure, and Cardiac arrhythmia & conduction disorders*. By fitting models on these diseases, we hope to provide a large population of patients who are under life-threatening conditions with an objective estimate of their expected outcome, which may serve as a reference for their future treatment.

We will improve the model that predicts the total charges. Currently, our cross-validated R^2 value is around 0.6, which is not optimal. We will seek to improve this score by incorporating more features and using alternative models.