CornellEngineering
Operations Research
and Information Engineering

ORIE 4741 Final Report

Hospital Selection

Hao Qiu, Yuting Zhao, Yuze Wang

December 1, 2019

Contents

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

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.

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

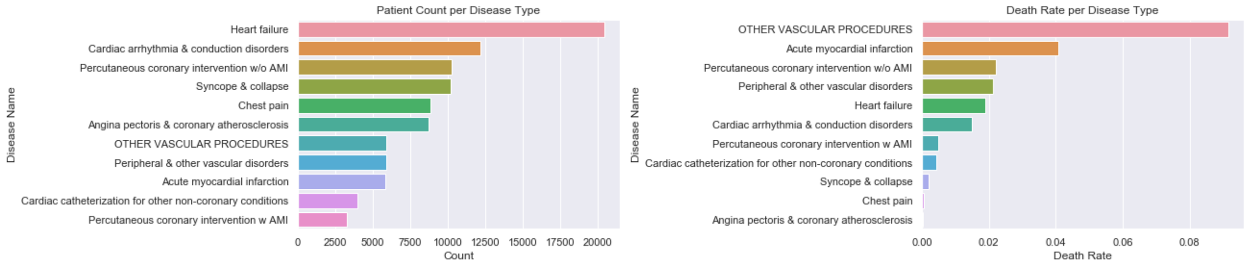
Categorical Data Columns like Patient Disposition represents if a patient is discharged to home, inpatient rehabilitation, or etc. We grouped these types into five broader categories: Home, Continued, Hospice, Died, and Other, where Other includes all other types that do not fall under any previous category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

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 for Preliminary Analysis

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.

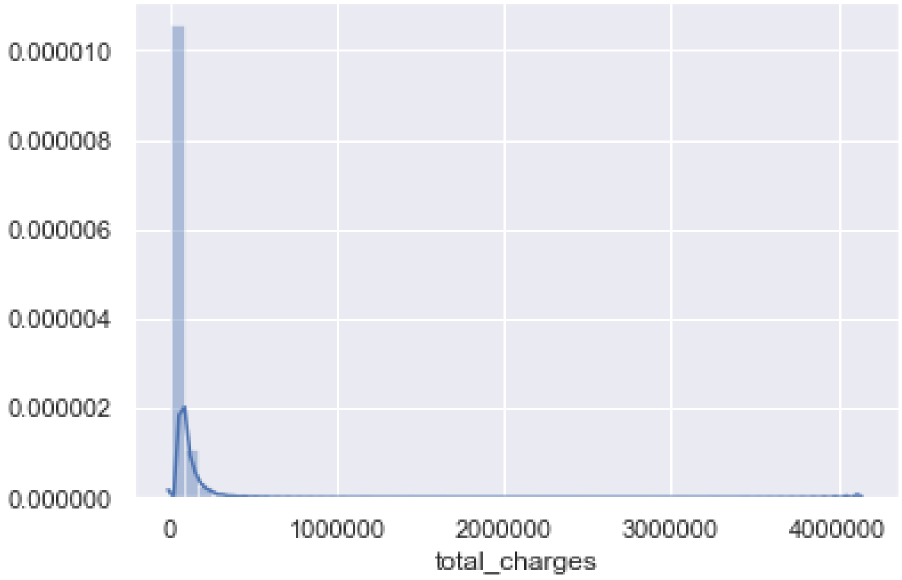


Figure 2: Total Charges

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 Modeling Results and Discussion

To examine our models’performances, we carried out predictions based on two circulatory diseases, one musculoskeletal disease and one delivery procedure. Specifically, we predicted the expected outcomes of Heart Failures and Cardiac Arrhythmia & Conduction Disorders and the total charges of a Knee Joint Replacement and Vaginal Delivery.

4.1 Predicting Expected Outcome: Heart Failures

Motivation The New York hospitals had received in total of 20,437 Heart Failures cases in 2012. Since heart failure is a common medical condition that is serious, life-threatening, and sudden-onset, it may serve as a significant field of study while predicting the expected outcome of disease treatments.

Data Cleaning Besides the general data cleaning, we created a new column of Expect

Outcome by grouping the patient disposition into five categories, namely, Home, Continued, Hospice, Died, and Other, to simplify the classification process. ”Other” contains all atypical dispositions, including release to court/law enforcement or to a psychiatric hospital.

Model Fitting and Analysis To predict the expected outcome, we used multi-nominal classification on the 5 simplified categories using information such as race, age group, gender, length of stay, risk of mortality, and etc.

The two methods we used are one-vs-all classification using logistic regression and random forest using XGBoost. In both cases, we one-hot encoded each outcome in a vector of length 5. The regularization we used in logistic regression is ridge and we did not use regularization in the random forest model. We performed 10 fold cross-validation for both models and found the random forest model to be better with 72.1% accuracy. This method was particularly effective in predicting outcomes of death, with only a 0.04172 misclassification rate for outcomes of death.

Metric	Logistic Regression	Random Forest
Accuracy	0.7098	0.7151
Precision	0.5175	0.6497
Recall	0.7192	0.7216
F_1	0.6019	0.6145

Table 1: Performance of Models Predicting Outcome of Heart Failure

We then developed a bootstrap estimator that fits the random forest using randomly resamples from the training set and calculated an out-of-sample accuracy using previously unseen test sets. Figure 3 is a histogram the accuracy of random forest from 100 bootstrap replications. As we can see, the out-of-sample accuracy is relatively stable when predicting expected outcome of heart failure.

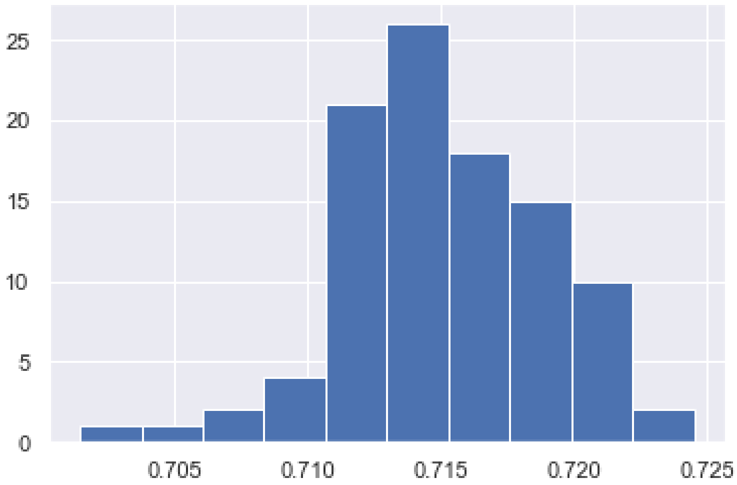


Figure 3: Bootstrap Random Forest Accuracy

4.2 Predicting Expected Outcome: Cardiac Arrhythmia & Conduction Disorders

Motivation As the second most common circulatory disease, the hospitals in the New York State had received 12,210 patients who were diagnosed as Cardiac Arrhythmia & Conduction Disorders in year 2012. Since this cardiac disease also has a relatively high death rate as shown in Figure 1(b), we considered it as another significant prediction object because the patients with this condition might also be eager to know the expected result of the upcoming treatment.

Data Cleaning We did not perform special data cleaning since the general cleaning function is sufficient for us to fit a model.

Model Fitting and Analysis To predict the expected outcome of patients with cardiac arrhythmia conduction Disorders, we used multi-nominal classification on the 5 simplified categories using information such as race, age group, gender, length of stay, risk of mortality, and etc

Similarly, we used are one-vs-all classification using logistic regression and random forest using XGBoost. In both cases, we one-hot encoded each outcome in a vector of length 5. The regularization we used in logistic regression is ridge and we did not use regularization in the random forest model. We performed 10 fold cross-validation for both models and found the random forest model to be better with 81.7% accuracy. This method was particularly effective in predicting outcomes of death, with only a 0.015 misclassification rate for outcomes of death.

Metric	Logistic Regression	Random Forest
Accuracy	0.8089	0.8120
Precision	0.6681	0.6690
Recall	0.8173	0.8170
F_1	0.7352	0.7354

Table 2: Performance of Models Predicting Outcome of Cardiac Arrhythmia

We then developed a bootstrap estimator that fits the random forest using randomly resamples from the training set and calculated an out-of-sample accuracy using previously unseen test sets. As we can see from Figure 4, the out-of-sample accuracy when predicting expected outcome of cardiac arrhythmia conduction Disorders is relatively stable.

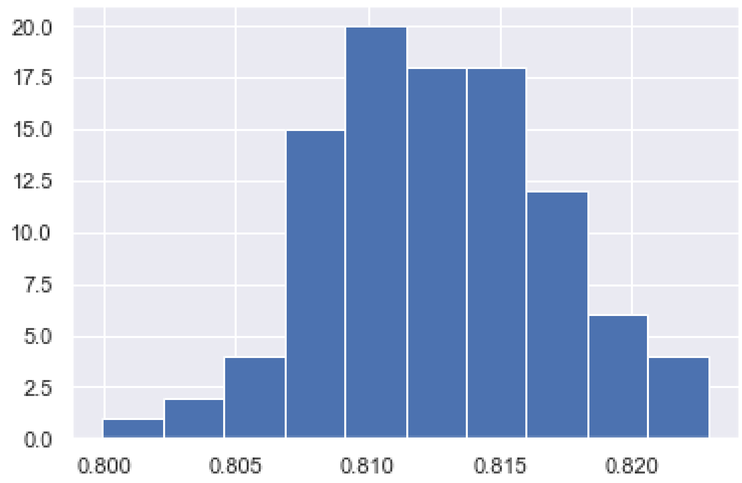


Figure 4: Bootstrap Random Forest Accuracy

4.3 Predicting Total Charges Incurred: Knee Joint Replacement

Motivation In 2012, there had been 34,230 people admitted to the New York hospitals receiving a Knee Joint Replacement. Knee replacement is a common medical procedure with an extremely low death rate, thus predicting the expected outcome may no longer be an interesting approach. However, we recognized that the total charges of replacing a knee joint vary significantly, with a majority of cases ranging from 1,000 to 200,000 dollars. Therefore, we will focus on this medical procedure to examine our models' ability to estimate the total charges incurred.

Data Cleaning A missing value in source_of_payment2 and/or source_of_payment3 indicates that the total charge is fully covered by previous source of payment(s). Thus we filled the missing values in column source_of_payment2 and source_of_payment3 by "Covered" in order to perform categorical encoding of these columns.

Model Fitting and Analysis We used two models to predict the total charge of a knee replacement surgery. The first model is a linear model with quadratic loss function and ridge regularizer, and the second is a linear model with quadratic loss function and lasso regularizer. Both models use identical features such as age race, group, gender, hospital county, method of payment and etc. When we validated out models, we found that the

Metric	Ridge Regularization	Lasso Regularization
Mean Squared Error	5.88e8	5.89e8
Mean Absolute Error	15925	15930
R^2	0.2227	0.2217

Table 3: Performance of Models Predicting Charges of Knee Replacement

linear regression that uses ridge regularizer has a lower MSE and thus has better predictive power.

Figure 5 below shows the comparison between the ridge regression model's total charges prediction and the actual total charges of knee replacement from a test set. As we can see, this model offers a reasonably accurate prediction of total charges.

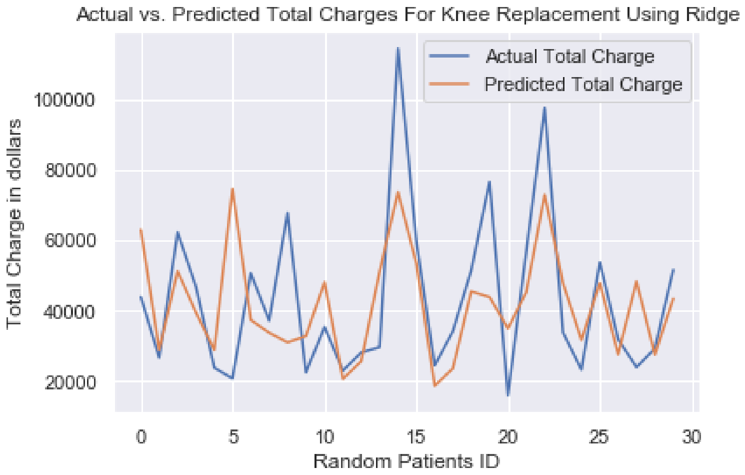


Figure 5: Actual vs. Predicted Total Charges of Knee Replacement

4.4 Predicting Total Charges Incurred: Vaginal Delivery

Motivation With a total of 151,650 cases, Vaginal Delivery is the most common medical service provided by the hospitals in the New York State in 2012. Since it might happen to almost every family probably for more than once, we believe that it would be helpful to predict the total charges of delivering a baby so that, for example, families with different income levels or ages will be able to compare the prices before making decisions.

Data Cleaning A missing value in source_of_payment2 and/or source_of_payment3 indicates that the total charge is fully covered by previous source of payment(s). Thus we filled the missing values in column source_of_payment2 and source_of_payment3 by "Covered" in order to perform categorical encoding of these columns.

Model Fitting and Analysis We used two models to predict the total charge of a vaginal delivery The first model is a linear model with quadratic loss function and ridge regularizer, and the second is a linear model with lasso regularizer. Both models use identical features such as age race, group, gender, hospital county, method of payment and etc. When we validated out models, we found that the linear regression that uses lasso regularizer has a lower MSE and thus has better predictive power.

Metric	Ridge Regularization	Lasso Regularization
Mean Squared Error	69359381	69359011
Mean Absolute Error	4426.619	4426.613
R^2	0.18320	0.18321

Table 4: Performance of Models Predicting Charges of Vaginal Delivery

Figure 6 below shows the comparison between the lasso regression model's total charges prediction and the actual total charges of vaginal delivery. The results shows that our model offers a reasonably accurate prediction of total charges.

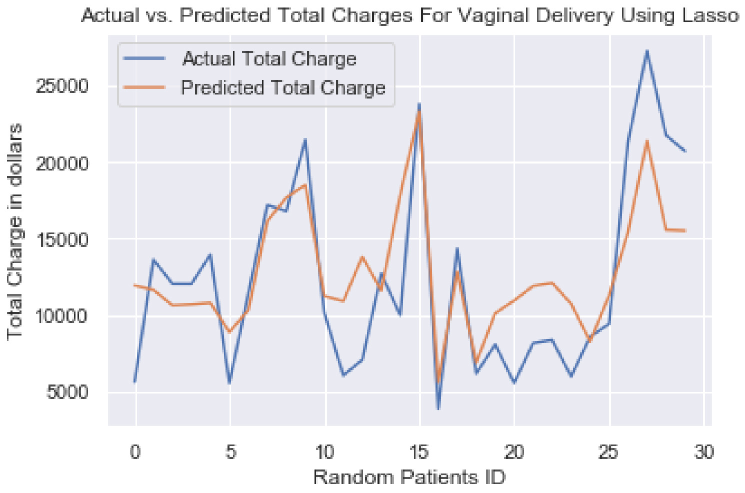


Figure 6: Actual vs. Predicted Total Charges of Vaginal Delivery

4.5 Limitations

Manual Restraints Currently, we performed some additional manual manipulations to achieve a more optimal prediction. One of these steps is additional cleaning, such as removing or imputing missing data in some of the columns. Moreover, the choice of loss function and the subset of features used in models require domain knowledge in machine learning and are still left up to the user at this stage.

Information Extraction Limitations The other limitation in our system is that we did not extract information from all features in the SPARCS dataset. Moving forward, it might be useful to apply unsupervised learning or deep learning techniques on the data to explore the underlying characteristics of patients with same diagnosis.

5 Visualization of Total Charge Levels by County

To offer useful hospital recommendations, we will link a certain prediction result to, for example, the total charge levels of a certain medical service in different counties to find the best match. We will use the information in the figures below to help the patients choose the optimal hospital service county, taking into consideration of the costs or distances.

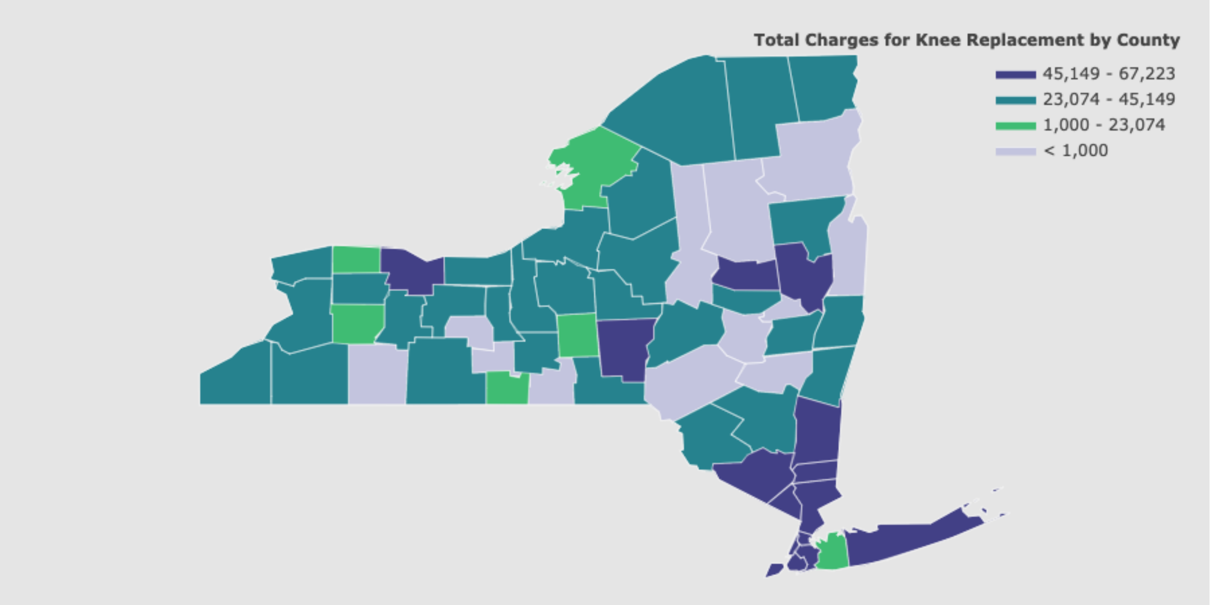


Figure 7: Total Charges of Knee Replacement by County

In Figure 7 above, we noticed that hospital service counties in light purple such as Hamilton, Essex, Allegany, and Delaware have the lowest total charges for knee replacement, which might be suitable places for patients that also have low predicted total charges or low income level. Service counties in dark blue such as Monroe, Chenango, and the New York City area had charged the highest for a knee replacement, which might indicate that those hospitals offer a better service quality or are able to carry out more complicated procedures. Thus, patients who are predicted to be more highly charged or seeking for better medical service may consider these areas.

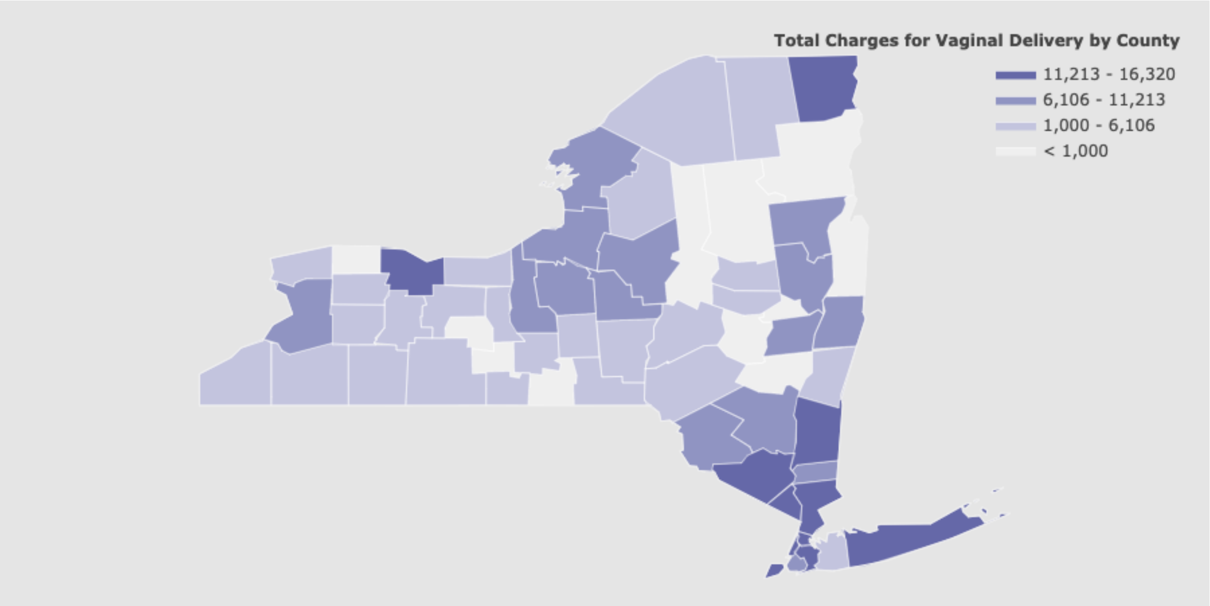


Figure 8: Total Charges of Vaginal Delivery by County

In Figure 8 above, similarly, we recognized that hospital service counties in white such as Hamilton, Essex, Yates, and Washington have the lowest total charges for vaginal delivery, which might be suitable places for patients that also have low predicted total charges or low income level. Service counties in dark blue such as Monroe, Clinton, and the New York City area had charged the highest for vaginal delivery, which might indicate that those hospitals offer a better service quality or are able to deal with more emergent, life-threatening conditions during delivery. Thus, patients who are predicted to be more highly charged or seeking for better, safer medical service may consider these areas.

6 Conclusion

Our targeted approach for the prediction of patient outcome and total charges incurred allowed us to focus our training efforts on those subsets with the highest predictive power, helping us to refine our models. Since our target consumers are healthcare patients, we worked to generalize as much of the data cleaning and model building processes as possible, allowing for robust predictions given only basic information with little manipulation. We are confident that our models would offer valuable insights regarding patients' future outcome and total charges incurred in some diagnosis despite the fact that we are limited by some requirements for specialized cleaning and manipulation.

Since patients with different conditions care about different aspects of the outcomes of medical procedures, including medical outcome or financial outcome. For life-threatening conditions, patients will seek for the success rate, while for general, common diseases, they would like to know how much it costs. Therefore, we carried out several case studies to differentiate between various diseases, offering predictions that laid emphasis on different types of outcomes. Our case studies and models have shown a great potential to alleviate some difficulties with pursuing medical care through analytical and heavily-tested methods as we have employed.

7 Bibliography

- SPARCS. https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/u4ud-w55t
- XGBoost. https://xgboost.readthedocs.io/en/latest/tutorials/index.html
- Scikit-Learn https://scikit-learn.org/stable/index.html