Cornell University
Operations Research and
Information Engineering

ORIE 4741

Learning with Big Messy Data

Final Report

Matthew Dalton
mgd67

Katelyn Glassman
kmg237

December 5, 2016

Contents

1 Problem Specification

Decisions as sensitive as healthcare planning call for careful consideration and thoughtful quantitative analysis to ensure informed decisions on the parts of patients, hospital administrators, and insurance providers. When initially pursuing medical treatment, it is often difficult to foresee patient outcome, length of stay, and costs incurred, although clarifying these factors can offer great decision-making value for patients and families.

Our project's goal is to provide patients with valuable insights and predictions regarding medical care based on their diagnosis and some basic information. Given their personal information and medical status we sought to develop a tool to provide the patient with an estimate of their expected outcome, length of stay at the hospital, and costs incurred. We believe that these three questions will be the most significant for a prospective patient in a hospital. Patient outcome, or whether he or she will survive, be released home, or be released elsewhere, is naturally the primary concern when making health care decisions. Length of stay is also an important factor when a patient is trying to plan for recovery after hospital admission. And lastly, costs are of course a large consideration for both patients and insurance providers when deciding on procedures and hospitals.

2 Description of Dataset

Our project examines the Statewide Planning and Research Cooperative System's (SPARCS) Hospital Inpatient Discharges for New York State. This dataset provides comprehensive information about all patients that have been discharged from a New York hospital in the year 2012, including age group, gender, and race. It also provides information about each patient's length of stay, reason for admission, severity of illness, condition upon leaving, method of payment and costs incurred. Specifically, the dataset quantifies severity of illness on a scale from 1 to 4, with 1 being minor and 4 being extreme, as well as an ordinal field for risk of mortality.

The dataset is comprehensive and mostly complete, with intuitive, colloquial values. Data validation was certainly employed in the collection of the data, as nominal fields like Diagnosis Description and Ethnicity are sensible and consistent. The features of the data are broken down into the following categories:

Feature Type	Count
Boolean	4
Nominal	16
Numerical	15
Ordinal	3
Total	*38*

In addition to the 38 features listed above, an age column was included with age values pre-grouped into ranges (0-17, 18-29, etc.), which we felt had qualities of both an ordinal and nominal feature.

3 Modeling Approach

3.1 Preliminary models

To see if our goal was feasible, we developed preliminary models to examine cases of a sudden onset condition that requires immediate care and predict outcome. These models were fit to data on heart attack diagnoses in New York hospitals, and they attempt to predict patient outcome. Our preliminary models formulated parameters to predict survival rate based on varying features and, through splitting the data into train and test sets, tested the accuracy of these fits on untrained data.
In order to determine which features are most influential and should be isolated and examined, we ran a preliminary regression across all of the features and picked those that were the most statistically significant in holding predictive value for survival rate. These features were `Severity of Illness, Length of Stay, Gender`, and `Age`, and so we fit several preliminary models over different combinations of these features to predict survival rate.
As shown in the file `Heart Attack Linear Models.ipynb` in our GitHub repository, we developed 3 models to be fit to the dataset that predict survival rate. Each model is fit to and cross-validated against a training set to develop a predictive fit parameter $\mathbf{w}$, then this $\mathbf{w}$ is tested on a previously unvisited test set.

3.2 Generalized Model

Overview In developing our final product, we realized that our target user is the patient, and as such, our intention was to automate as much of the prediction process as possible. Doing so would allow the patient to obtain the predictive guidance they seek solely by entering their current medical condition and personal information without having to perform the actual quantitative data analysis and prediction methods themselves. Accordingly, we developed scripts at all stages of the process which could handle most of the task for any inputted diagnosis.

Cleaning Given that each prediction our system makes is within the context of a specific diagnosis, our general cleaning script performs an initial clean on the subset of the SPARCS dataset that contains only entries with that diagnosis. The script removes rows containing missing or unknown entries for any column of the data set. While an unsupervised learning method for imputing these values may have offered an added benefit, the relatively small number of missing entries made us think this was unnecessary. Next, the script converts the entries of `Length of Stay` and `Total Charges` from strings to integers and floats, respectively. Finally, the script deletes the columns we deemed repetitive or unnecessary for analysis such as `Discharge Year`, because all data came from 2012, and `APR Severity of Illness Description`, because there is already an ordinal

category `APR Severity of Illness` that assigns a numerical value to the nominal description value. After the data was cleaned, we split the data into training and test sets following the 80% / 20% convention.

Feature Transformations To make our dataset more approachable we included a function that could convert a nominal feature into a vector of boolean features using one-hot encoding in order to allow for regression. For example, we converted the single `Race` column, which contained text values of a patient's race, into three columns: `White, Black` and `Other Race`, and for each patient we placed a 1 in the applicable column and a 0 in the others. We used the same logic for the `Gender` column, splitting it into two binary columns for `Male` and `Female`, and for the `Method of Payment` column into `Private Insurance, Medicare`, and `Medicaid`.

Model Fitting Our system uses linear models of the same type to make its predictions. We have defined these models as 5-tuples composed of a desired prediction, a set of features, a function to map a prediction from the model output space to the prediction space, a loss function, and a regularizer. We then included general functions associated with this model type to form real valued $\mathbf{X}$ and $\mathbf{y}$ for each model, and to find the models' coefficients $\mathbf{w}$ using proximal gradient descent.

4 Prediction System Results

To demonstrate our system's performance, we have focused on one archetypal diagnosis for each feature that we are aiming to predict. We will examine how our system performs while predicting the expected outcome of a heart attack, the duration of stay in a hospital for depression, and the total cost of a hip replacement.

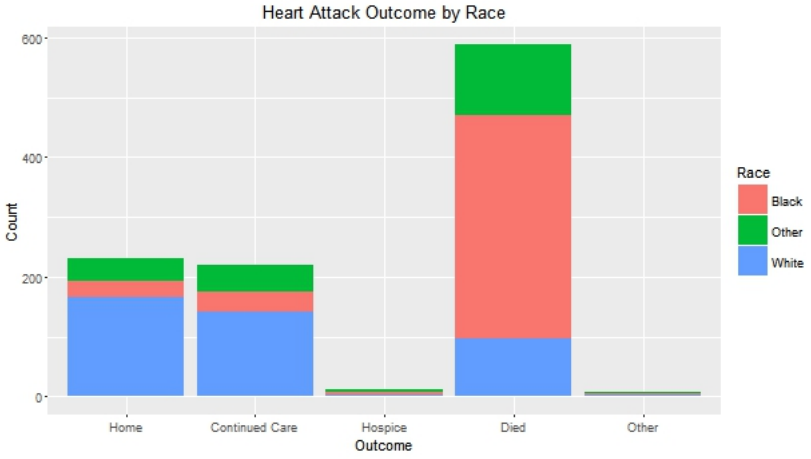
4.1 Predicting Expected Outcome Case Study: Heart Attack

Motivation There were 1072 hospital admissions diagnosed as cardiac arrest in New York State in the year 2012. Cardiac arrest is a serious, life-threatening and sudden-onset condition, and its symptoms are often consistent among cases. Because of this, we found heart attack cases to be an appropriate set for analyzing and predicting outcome.

Specific Data Cleaning In addition to the general data cleaning that our system performs, we grouped the patient outcomes into five intuitive categories to make the classification simpler. These categories are `Home, Continued Care, Hospice Care, Expired` (Died), and `Other`. We are using the `Other` category as a catch-all grouping for atypical outcomes, such as release to law enforcement or to a psychiatric hospital.

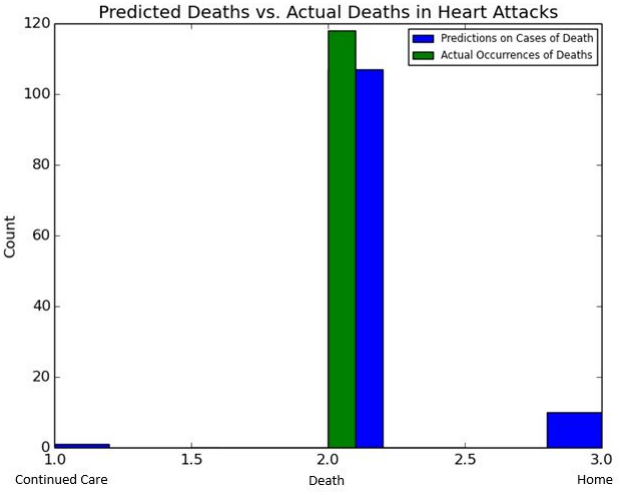
Model Fitting and Analysis To predict the expected outcome we used multinominal classification on the 5 simplified categories of outcomes using information about health service area, age group, gender, race, admission type, severity, and payment type ad features, as well as an offset.

As shown to the right, race is an extremely influential feature on survival rates of heart attacks. Black patients have much higher mortality rates after diagnoses of heart attacks than do white patients or patients of other nonspecified races. As such, race was an applicable feature to our classification.



The two methods we used to perform the classification were one-vs-all with hinge-loss and learning probabilities. In both cases, we one-hot encoded each outcome in a vector of length 5. Then with one-vs-all, we used hinge loss to predict each entry value, and with learning probabilities, we estimated the probability of each outcome. Finally, we predicted the outcome to be the vector entry with either the value closest to 1 (for one-vs-all), or the value with the highest probability (for learning probabilities). We then cross validated each model on the training set and found the learning probabilities model to be better, with an average misclassification rate that was over 5% lower than that of the one-vs-all method. We selected the learning probability multinominal classification model as our choice for heart attacks and then fit it to our validation set. This method was particularly effective in predicting outcomes of death, with only a 0.0789 misclassification rate for outcomes of death.

As shown in the histogram to the right, the learning probabilities classification model was particularly effective in predicting outcomes of death. Most occurrences of death in the test set were classified as such, with few cases being misclassified as either continued care or home release. We suppose that our outcome predictions were most successful on outcomes of death because there were far more data points in the training set with an outcome of death than, say, hospice care, which helped to avoid overfitting in death prediction models.

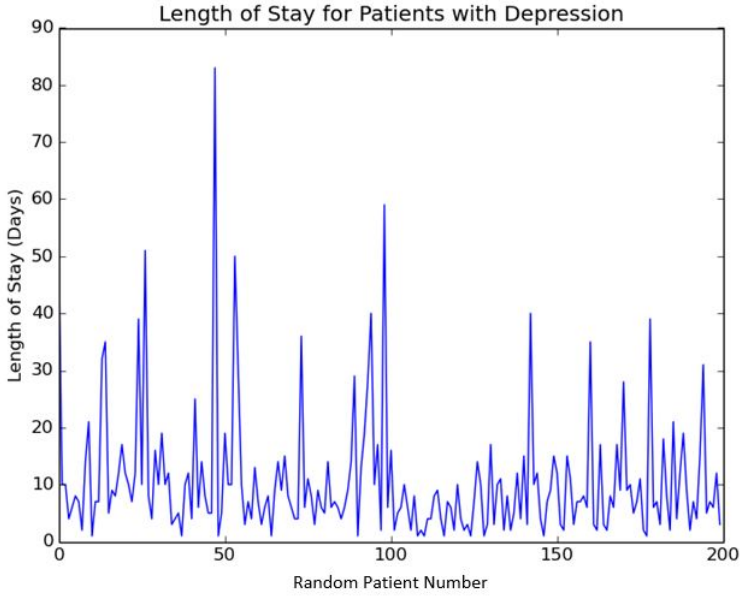


4.2 Predicting Length of Stay Case Study: Depression

Motivation Depression diagnoses require nuanced care and personalized, highly monitored treatments. The lengths of hospital stays can vary significantly for depression patients depending on severity of diagnosis and other personal information about the patient, and as such we believed depression to be a powerful subset of diagnoses for predicting a patient's length of stay.

Specific Data Cleaning In this particular case, our general cleaning function was sufficient to allow us to fit a model to the data. No special manipulation was required.

Model Fitting and Analysis To predict length of stay we used a linear model and features relating to health service area, age group, gender, race, type of admission, and severity, as well as an offset. In choosing a loss function, there were some unique characteristics about both predicting length of stay and the specific diagnosis that we had to consider. First, for a prediction of length of stay, there is an inherent bias against predictions that underestimate how long someone will be in the hospital. A patient would prefer to make plans in accordance with expecting to stay in the hospital a long time and then getting out sooner, than being told they have to stay longer than they were expecting too. Therefore, we had to define a new metric to test our models that captured the full squared error for low predictions but only a fraction (assumed to be .5) of the squared error for high predictions. This new metric allowed us to capture the patient's bias for high predictions while also allowing us to continue to penalize very inaccurate predictions as well. The second point relates to patients with depression in particular. On average, patients with depression were in the hospital for around 10 days, but in particular cases, this length was much larger. This trend is captured in the graphic below:



From this analysis, we concluded that a quantile loss function at the 1.0 level minimized the error, as shown in the visual below, and accordingly is our choice for predicting length of stay for patients with depression.

4.3 Predicting Costs Incurred Case Study: Hip Replacement

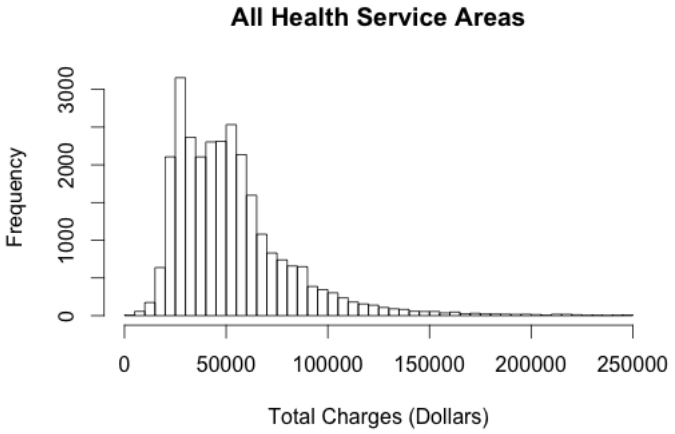
Motivation In 2012, over 28,000 people received a hip replacement in New York hospitals. Hip replacement surgery is a fairly common medical procedure and is oftentimes a major event in people's lives. Although the procedure is fairly consistent in most cases, the costs of hip replacements can vary significantly. For these reasons, we thought it to be a good procedure with which to demonstrate our system's ability to estimate cost and provide useful cost predictions for both patients and insurance providers.

Specific Data Cleaning In our 2012 SPARCS dataset there were 28,373 recorded hip replacements. From this data, we removed cases where the patient diagnosis did not intuitively seem to require a hip replacement, such as a diagnosis of "Arm Fracture" or "Tuberculosis". In these cases, we thought that the unusual diagnosis indicated that those patients were not members of the typical hip replacement population and should be excluded from our data. We also removed outlier entries where the cost exceeded $250,000, as well as entries where key demographic information was missing. We felt justified in simply deleting the data in these cases instead of imputing estimates since the removed entries only accounted for 3.2 percent of the total data.

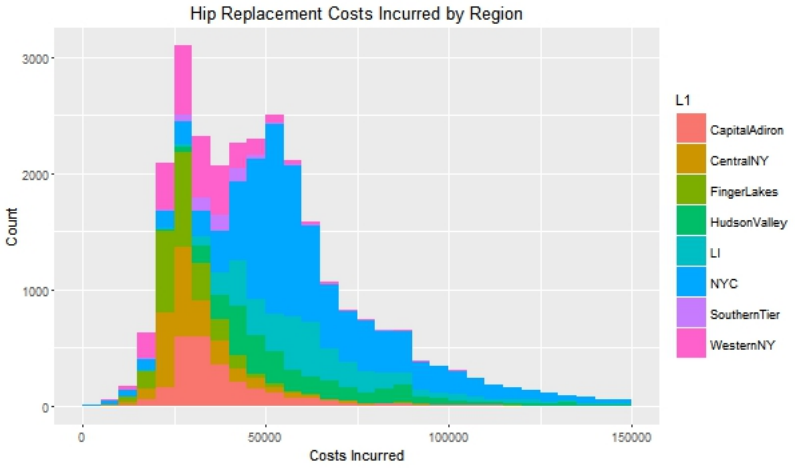
Data Analysis The next step we took was to perform some preliminary data analysis to get a sense of the features most significant to the total charge of a hip replacement. A major factor that we discovered to have an effect on cost was the health service area of the patient. By inspecting the distribution of total charges, we concluded that the overall distribution was bimodal and that the 8 complete health service areas could be divided into 2 categories of distributions centered around distinct means. The first group had a

mean total cost of around $34,000 and included the Western New York, Finger Lakes, Southern Tier, Central New York, and Capital/Adirondack regions. The second group included the New York City, Hudson Valley, and Long Island areas and had a mean total cost of around $65,000. Geographically, this division makes sense because the areas in each group are clustered together, and so we decided to include this feature in our dataset.

Displayed to the right is the distribution of cost for all hip replacement procedures. The division of regions into two distinct groups is evident by the two peaks.

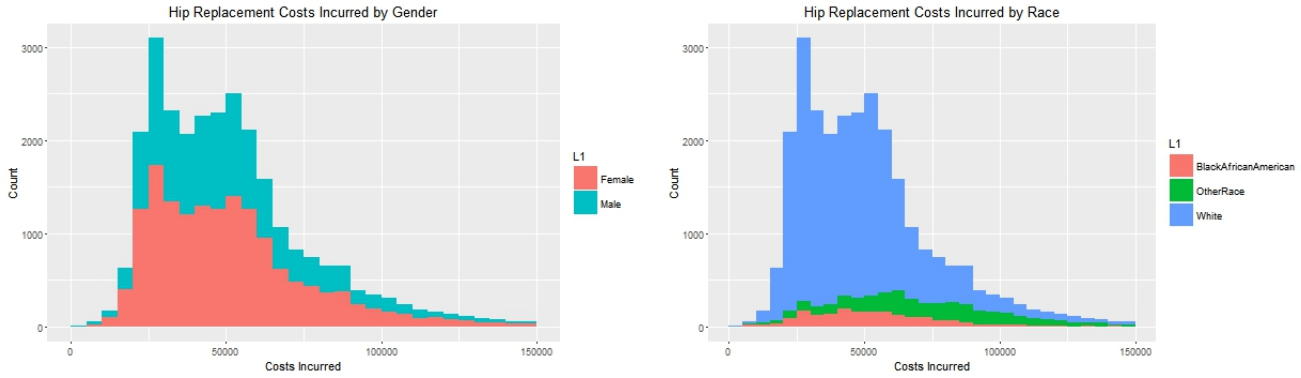


Displayed to the right is the distribution of hip replacement costs further broken down by region. As mentioned, the bimodal distribution of hip replacement costs is defined by regional divisions.



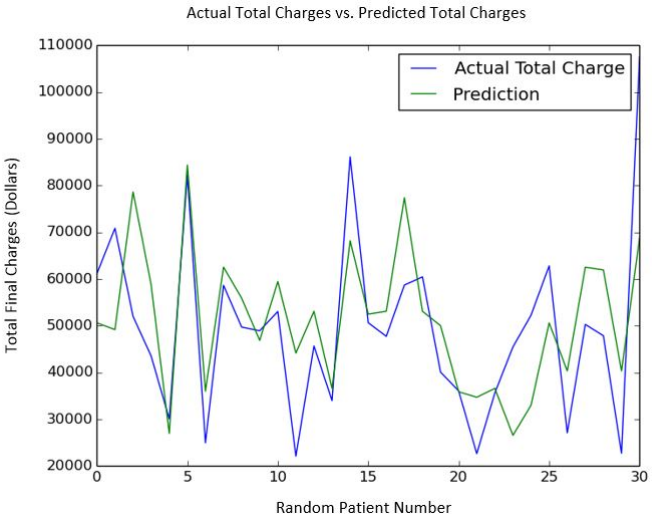
In a similar fashion, we examined other categorical features, such as gender and race, and checked for differences in distribution of total charges among the different populations. For example, we examined total charges for men compared with women, or total charges for white people against total charges for black people. Out of all of the categories, age group, race, and admission type to be the most significant.

To illustrate this point, below are the cost distributions by race and by gender. Costs seems to follow the exact same trend regardless of gender, whereas cost trends are more variable across different races.



Model Fitting and Analysis To predict the total charge of a hip replacement, our system uses a linear model with a quadratic loss function and no regularizer. Our initial model used features relating to the patient's health service area, county, age group, gender, race, admission type, method of payment, and severity rating. Converting the model to a real-valued matrix using the system's feature transformation resulted in 85 columns. We suspected that this model would overfit the training data, so we created a second model that only used the features we found to be significant in the preliminary data analysis, which were patient's health service area, age group, race, admission type, and severity rating. Feature transformation resulted in only 24 columns for this set of features. When we validated our models, we found that this second model indeed had better predictive power with a MSE of 4.99e8 compared with the previous model's MSE of 5.56e8.

Below is a plot of the final model's cost prediction compared with the real costs from a test set. Clearly, the final model offers a reasonably accurate cost prediction.



4.4 Current Limitations

Manual Restraints At this stage in development, some of the diagnoses require additional manipulation done manually to make an optimal prediction. One of these steps is additional cleaning, such as removing unforeseen outliers in the hip replacement case. More importantly, the choice of loss function for our model and which subset of features to use in prediction is still left up to the user and requires domain knowledge in data science.

Information Extraction Limitations Another limitation on our system is that we are currently not extracting information from all features of the SPARCS dataset. Moving forward, it might be useful to use perform some unsupervised learning or clustering methods on our data beforehand to divide patients of the same diagnosis into subgroupings that possess unique common characteristics.

5 Conclusion

Through thoughtful modeling and analysis, we realized the great predictive power of this given healthcare dataset through the use of careful model fitting and error evaluation. Our targeted approach for the prediction of fields like patient outcome, length of stay, and costs incurred allowed us to focus our training efforts on those subsets with the highest predictive power, helping us to refine our models. Keeping in mind that 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. Although we are limited by some requirements for specialized cleaning and manipulation, we are confident in the ability of our models to offer valuable insights regarding a patient's future. As our case studies and models have shown, there is great potential to ameliorate some difficulties with pursuing medical care through analytical and heavily-tested methods as we have employed.