Breaking the Factors Down: Predicting Medical Charges for Patients in The Bronx

By Nihar Sidhu (ns625) and Nehal Rawat (nr338)
ORIE 4741: Learning With Big Messy Data Final Project

Motivation:

There is an increasing push to collect medical data to improve quality of care given by doctors, increase survival rates for patients, decrease hospital charges, and more. However, it is especially difficult within the medical field to collect data given that a lot of work must be done to request a patient's medical data, and then also guarantee confidentiality. Medical data that is available often contains obscure terms and terminology that is not easily interpretable to individuals outside the medical field.

In this report, we examine the 2015 Statewide Planning and Research Cooperative System (SPARCS) dataset provided by the New York State Department of Health. The dataset contains information on patient discharges from New York State Hospitals, and includes information such as patient demographics, diagnosis, charges associated with the hospital visit, and other metrics as well.

Dataset and Overarching Problem:

The 2015 dataset is the most recent SPARCS dataset provided by the New York State Department of Health. Given the magnitude and unique diagnoses that individuals come to hospitals for, we decided to look into cases where patients come into hospitals because of chest pains. Chest pains is one of the most common reasons that people across the US visit hospitals, and often leads to more serious cardiovascular conditions. We considered looking into individuals going into hospital due to mental illnesses, but later learned that mental illnesses are not often treated at hospitals, but more so at psychiatrists and psychologists offices. We also considered looking into cases where individuals go to the hospital for labor, but we were interested in looking at a diagnosis that spans both genders. Given all the counties in New York State, we also focused our efforts on one county in particular: The Bronx. The Bronx has the lowest income per capita in New York State, and we were especially interested to see what features contributed the most to medical charges in this area. The main question we aimed to answer was: Can we predict hospital charges for Bronx patients admitted to the hospital because of chest pain? While we explored this question, we were also interested in looking at which factors over others play a greater role in determining hospital charges.

Our models can play an influential role. The models created using Bronx patient chest pain data can help medical officials potentially identify the main factors that cause medical charges to be high, and thus work towards lowering them. Additionally, when a patient goes to a hospital, an individual does not receive a hospital bill until sometimes as late as three or four weeks after the hospital visit. Our model can help individuals get a faster and easier sense of how much a hospital visit could cost. The model can also play an important role for insurance companies as they can derive rough estimates of hospital costs, and therefore adjust their risks accordingly.

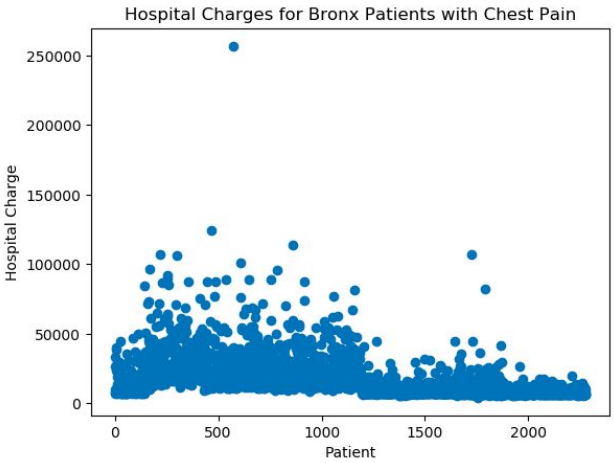
Data Cleaning:

We cleaned our data so that we were only looking at cases of chest pains in the Bronx. To select the data associated with chest pain, we looked at where "APR_DRG_Code" = 203, which was linked to chest pain. In our dataset, most details related to abortion cases were omitted. However, since we are only looking at cases related to chest pains, we were safely able to eliminate cases related to abortions. Thus, the rows and

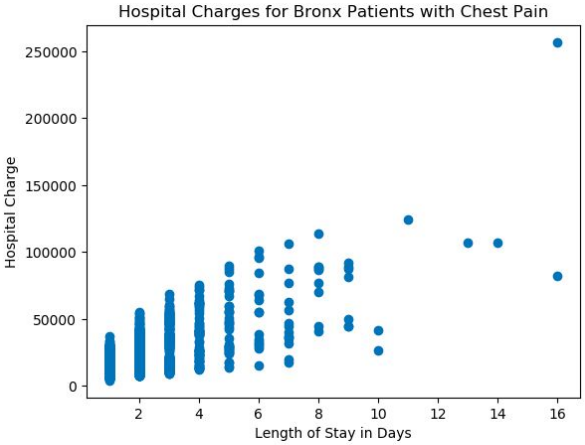
features of the dataset we were looking at actually had no NA values. After this cleaning, we were left with 2,279 rows.

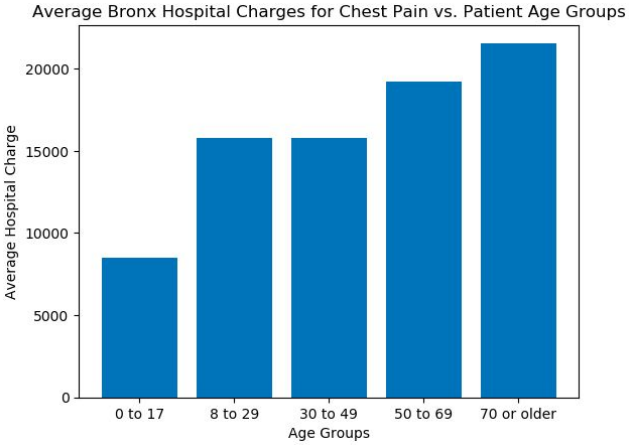
Summary Statistics:

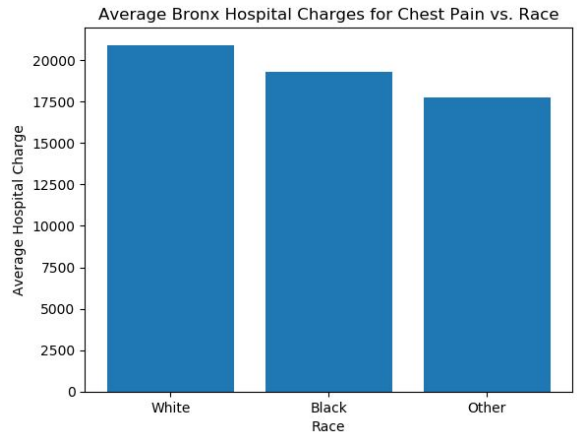
To get a sense of the range of hospital charges, we calculated some brief summary statistics. The mean of Bronx hospital charges for chest pain was $18, 618, the median was $13, 347, the maximum was $256, 762, the minimum was $4,109, and the standard deviation was $15, 650. As we plotted the hospital charges for various patients, we got the resulting plot on the right. The hospital charge above $250,000 was a clear outlier and we eliminated that from out set.

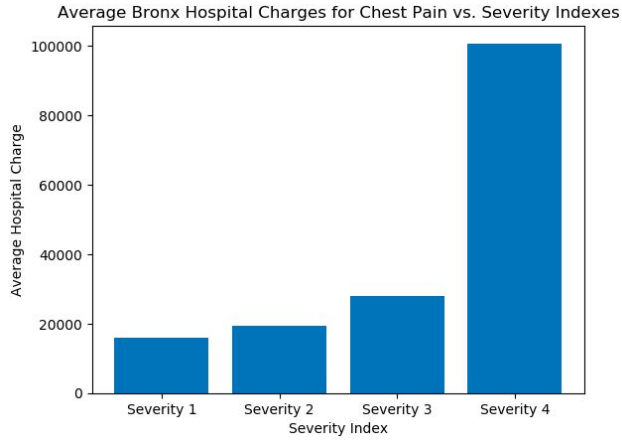


To understand our dataset better and get a sense of the features we would be interested in, we plotted some of the features, such as length of stay in days, age groups, race, and severity index of patient against the hospital charges.









Feature Selection and Feature Engineering:

From the initial scatter plots, we saw relations between some of the features and medical charges, and made sure to keep those features as part of our feature selection. In addition, we had a range of other numerical, ordinal, and nominal features:

Numerical Data:

- Length of Stay: The length of stay (in days) in the hospital for each patient was already numerical.
- Hospital Charge (dependent variable): The SPARCS dataset originally had the hospital charges formatted as Strings and with the "," delimiter, so we had to run a script to treat the dollar amounts as numerical.

Ordinal:

- Ages: The dataset already grouped ages into 0 to 17, 8 to 29, 30 to 49, 50 to 69, and 70 or older. Since this is an ordinal set, we just encoded 0 to 17 to be 1, 8 to 29 to be 2, etc.
- Type of Admission: The dataset described the type of admission for each patient into the hospital as Elective, Urgency, Emergency, and Trauma, and we encoded this as 1, 2, 3, and 4 respectively. The lower values correspond to patients that did not need help right away and higher values correspond to patients that needed attention immediately.
- Severity of Illness Code: This feature was already encoded on a scale from 1 to 4 - 1 being a mild diagnosis, to 4 being a very severe diagnosis.
- Risk of Mortality: The dataset describes the risk of mortality as either "Minor", "Moderate", "Major", or "Extreme". These were encoded as 1, 2, 3, and 4 respectively.
- Surgical Description: We encoded a non-surgical medical intervention to be 1, and a surgical intervention to be 2.

Categorical:

For each categorical feature, we used one-hot encoding to create a boolean feature for each category.

- Facility Name: The dataset had numerous hospitals that we had to one hot encode. This resulted in about 7 different hospitals we took into account.
- Gender: We used one hot encoding to create a 'Male' feature and a 'Female' feature
- Race: We used one hot encoding for each race category: 'White', 'Black', 'Other'. The complete dataset also includes 'Multi' and 'Unknown', but the dataset for chest pain patients in the Bronx does not include these categories.

Model Fitting - Regression Models:

To predict the medical charges, we started off with regression models using various loss functions and regularizers. Below are some of the models we tried.

L2 Loss Function:

We first tried to fit our model to the least squares objective function as described as below:

$$\sum_{i=1}^{n} (y_i - w^T x_i)^2$$

To test our model, we broke our data into training and test set. We put 80% of the data into the training set, and 20% in the test set. We trained our model on the training set, and then used it against the test set to see how good our model was. From this initial model, we got a resulting r^2 value of 0.792. While root mean square error (RMSE) is a common way to measure the difference between predicted and actual values, it gives a disproportional amount of emphasis to outliers and large errors. Thus, we looked at another

method of calculating error. One method to calculate the error is the mean absolute percent error. We calculated the mean absolute percent error across the n rows in the dataset as follows:

$$\frac{100}{n}\left(\sum_{i=1}^{n}\frac{|\,(actual\ hospital\ charge - predicted\ hospital\ charge)\,|}{(actual\ hospital\ charge)}\right)$$

With the L2 objective function, we got an mean percent error of 27.61% on the test set. We then were interested in trying other models to reduce the mean percent error, and to also find more sparse solutions. We wanted to find a sparse solution so our model could be easily explained to medical officials to indicate the factors that are most significant in determining hospital charges.

L1 Loss Function:

The L2 loss function fits the mean better, so we then looked into the L1 loss function, as this fits the median of the dataset better. With the L1 objective function, the model aimed to minimize the following loss function:

$$\sum_{i=1}^{n}(y_i - w^T x_i)$$

Changing to the L1 objective function, gave a huge improvement. From this model, the absolute mean percent error dropped to 21.04%.

L1 Objective Function and L1 Regularizer:

We then added the L1 regularizer to see if there were certain features we could pick out, and to prevent overfitting and reduce the test error as well. The model that we then tried to minimize was as follows:

$$\sum_{i=1}^{n}(y_i - w^T x_i) + \lambda \sum_{i=1}^{n}|w_i|$$

We were very able to slightly reduce the mean percent test error to 21.00%. While the L1 regularizer encourages sparsity, we still did not get a solution that was very sparse that could help us single out certain features.

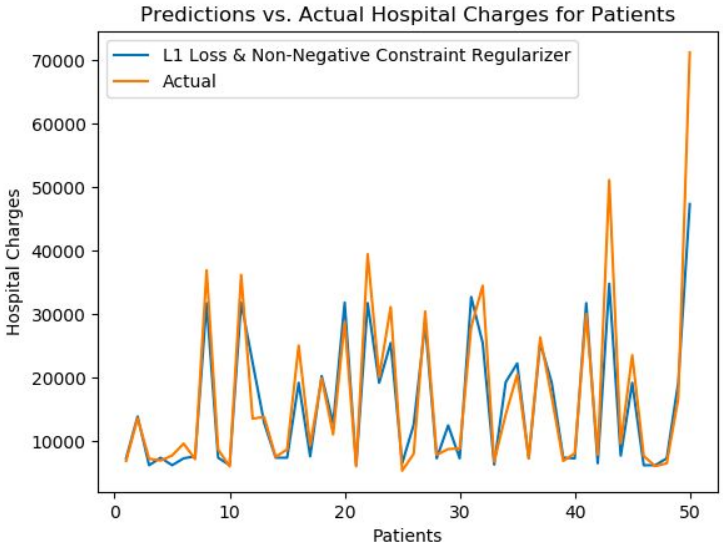
L1 Loss with Non-Negative Regularizer:

We then tried to use another regularizer such as the non-negative constraint to produce more sparse solutions. The model that we then tried to minimize was as follows:

$$\sum_{i=1}^{n}(y_i - w^T x_i) + \lambda \sum_{i=1}^{n}1(w_i \geq 0)$$

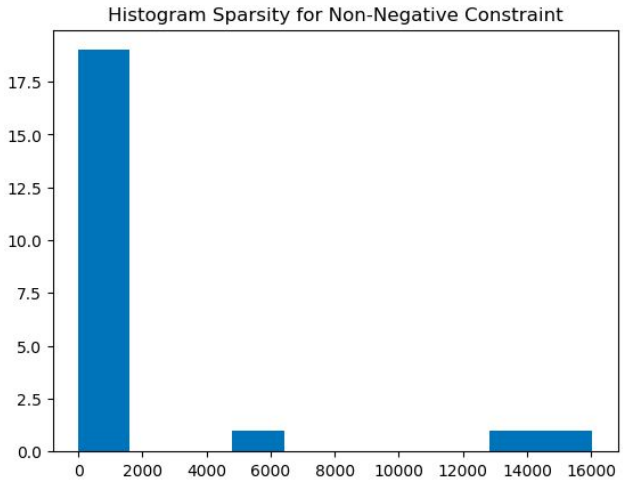
From this model, the absolute mean percent error was 20.88%, which was an improvement from the L1 objective function with the L1 regularizer model.

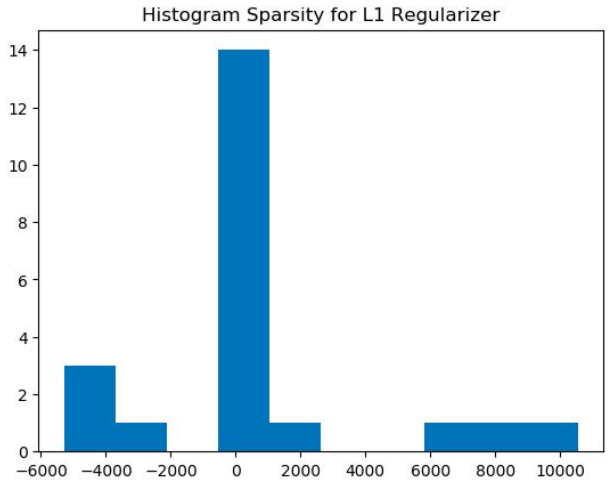
The plot below shows the predictions from the L1 loss with the non-negative regularizer model against the actual values.



We also tried a model with the Huber loss and non-negative regularizer, but this did not improve our error rate.

Sparsity:





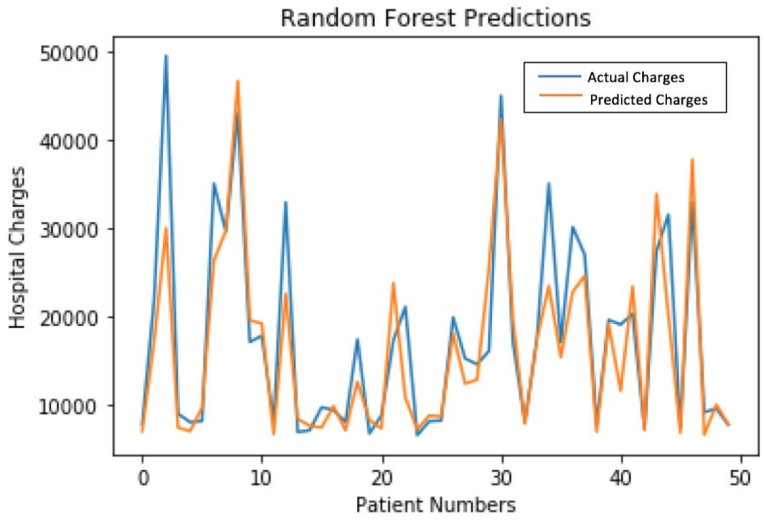
The histogram on the left highlights the sparsity of the L1 loss function with non-negative regularizer model, compared to the sparsity of the L1 loss function and L1 regularizer model on the right. The non-negative regularizer gave a more sparse solution. From the non-negative constraint, the features that had a non-zero weight on hospital charge were the length of stay, some of the hospital facilities themselves, and being male. From this, medical experts can see that length of stay factors heavily into medical charges, and thus can do further research to see what can be done in reducing the length of stay for a patient as well. Additionally, the four hospital facility features with high weightages were the Jacobi Medical Center Hospital, Montefiore - Wakefield Medical Center, Montefiore - Henry Lucy Medical Center, and the Lincoln Medical Hospital, whereas the hospitals such as as the Bronx-Lebanon and the SBH Health System Hospital had very minimal weights. At first it may seem that these four hospitals could be attracting more sick

patients that stay for a longer time, and thus lead to higher medical charges. However, our data shows that the severity of chest pain across the hospital are very close (Standard deviation for severity index across the hospitals for chest pain is 0.07). Therefore, this could be due to additional factors not included in our dataset, and more data would be needed to see why those four hospitals in particular are more significant in predicting hospital costs than the other two.

Random Forest Regression:

Another model we attempted to fit our data is random forest. Random forest for regression works by randomly selecting groups of regression trees. The decision trees are assigned certain features and are split based on minimization of errors and maximization of information gained through the split. The random forest model takes the average of outputs of many regression trees. We used random forests because this method reduces the possibility of overfitting, as is possible using only decision trees.

For this model, we also split the data into 80% training set and 20% test set. We performed feature scaling to standardize our data to have zero mean and unit variance. We chose two hyperparameters (parameters which cannot be learned by the model) to tune our model through cross validation, namely the maximum depth allowed for the decision trees in our forest and the maximum features to consider when determining the optimum split. We cross validate to ensure that our model is not overfitting by training and evaluating our model numerous times on different hyperparameters. We use GridSearchCV (part of sklearn) to perform cross validation across all possible combinations of hyperparameters. GridSearchCV also refits the model on the training set using the most ideal combination of hyperparameters. We then use our final model to predict the costs on the test set. Using this model, our r^2 value is 0.84 and our mean percent error is 19.07%. We selected 50 sample points from the test data to graph the actual costs and our predicted costs. These results can be seen below.



Quantile Regression:

Once we were able to create a model to predict the medical charges for patients admitted to the hospital for chest pains, we were further interested to see what features played a larger role in the different quantiles for hospital charges. The plots below depict the relative extent to which each feature affects hospital charges in the different quantiles.

In Figure 1 below, we see that the age group of a patient and the severity of the chest pain for a patient impact the total hospital charge in the lower and higher percent quantiles very similarly. We also see that the effect of type of admission is much more significant for hospital charges around the 60% quantile compared to hospital charges around the 90% quantile. Thus, medical officials can also take a look into what makes visits to the emergency room more expensive, compared to an elective visit, in order to reduce medical charges.

Figure 2 shows how dramatically the length of stay of a patient causes the hospital charges to increase, especially in the later quantiles. Until the 50% quantile, the feature weight rises slowly, after which it rises much faster. Again, medical officials can use this knowledge to try to minimize the length of stay a patient is in the hospital, and minimize the hospital charge.

Figure 1: Feature Weights for Bronx Hospital Quantile Charges

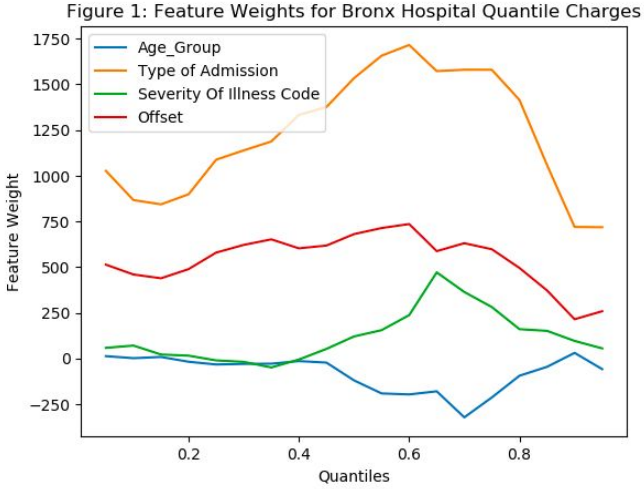


Figure 2: Feature Weights for Bronx Hospital Quantile Charges

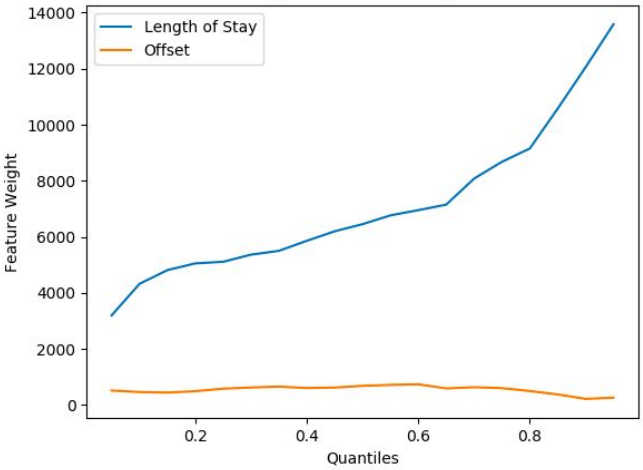
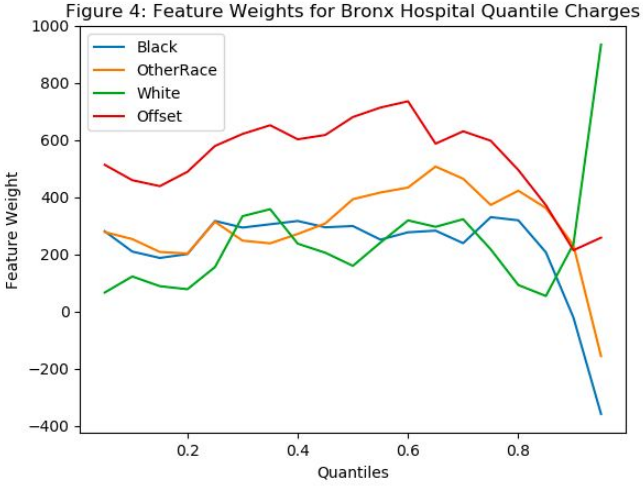


Figure 3 depicts how for all quantiles, the 'Male' feature is given a higher weight than the 'Female' feature in determining hospital charges, and the weight for both features decreases for the highest quantiles. Males on average seem to have a higher hospital charge for chest pain than females. Interestingly, the model with the nonnegative constraint also picked out the male feature as important to predicting the medical cost.

In Figure 4, for most of the quantiles up until the 90%, one race over another does not factor more into the hospital costs. However, for around hospital charges in the 90% tile, 'Black' and 'Other Race' compared to 'White' have opposite effects on the hospital charge. We cannot say why this is so, but perhaps they are certain chest pains and cardiovascular conditions that are more common within a demographic.

Figure 3: Feature Weights for Bronx Hospital Quantiles Charges

Figure 4: Feature Weights for Bronx Hospital Quantile Charges



Possible Future Extensions:

Given the numerous features, there are many other different features we could aim to predict as well. While we aimed to answer a regression question regarding our dataset, there are some potential classification questions that are also worth looking into. One interesting question is predicting where patients go after they leave the hospital.. This is a categorical feature in the original dataset consisting of the following 9 groups: 'Home or Self Care', 'Left Against Medical Advice', 'Home with Home Health Services', 'Short-term Hospital', 'Skilled Nursing Home', 'Hospice - Medical Facility', 'Inpatient Rehabilitation Facility', Psychiatric Hospital', and 'Expired'. In addition, it would be interesting to compare hospital charges for other conditions, and explore the extent to which the factors mentioned above impact the hospital charges for these conditions as well.

Conclusion:

Model	Absolute Mean Percentage Error
L2 Loss Function	27.61%
L1 Loss Function	21.04%
L1 Loss Function with L1 Regularizer	21.00%
L1 Loss Function with Non-Negative Constraint	20.88%
Random Forest	19.07%

Overall, the random forest model produced the lowest absolute mean percentage error. The L1 Loss Function with non-negative constraint also had reasonably close values for the absolute mean percentage error, and also gave sparse solutions that helped us single out some of the features that contributed the most to hospital charges.

Regarding the hospital charge predictions, we believe our models are accurate enough to give individuals a good sense of what their medical charge might be as they wait for their bill. Additionally, our models have exposed some of the factors that are of greatest importance in the different quantiles of hospital charges, and thus officials can use these factors and try to streamline and improve the healthcare system that could result in lower hospital charges for patients in the Bronx.

Citations:

1. Course material from ORIE 4741: Leaning With Big Messy Data by Professor Madeline Udell (https://people.orie.cornell.edu/mru8/orie4741/lectures.html)
2. Packages used from the Julia package "Low Rank Models"
3. *Hospital Inpatient Discharges (SPARCS De-Identified): 2015 | Health Data NY*, https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/82xm-y6g8
4. "Python Machine Learning Tutorial, Scikit-Learn: Wine Snob Edition." *EliteDataScience*, 16 Sept. 2017, https://elitedatascience.com/python-machine-learning-tutorial-scikit-learn
5. Spoon, Marianne. "10 Most Common Reasons for an ER Visit." *HowStuffWorks*, HowStuffWorks, 7 July 2011, https://health.howstuffworks.com/medicine/10-common-reasons-for-er-visit1.htm