Cornell University
Operations Research and
Information Engineering

ORIE 4741

Learning with Big Messy Data

Final Report

Anusha Avyukt	Fei Xia	Siyang Liu
aa2686	fx43	sl2687

December 10, 2019

1.Problem Statement

When seeking medical treatment, there are constraints posed by the type of disease, emergency level, location of the patient, health insurance and other patient characteristics like income, age etc, to visit certain hospitals. The overall purpose of our analysis is to help the patient select a hospital based on their medical condition and personal information. Our project considers factors critical to hospital selection in emergency situations of septicemia, mood disorders and congestive heart failure, as these are the most prevalent diseases in our dataset. It would be ideal to have a model which can help in hospital selection for any disease but apart from prediction, we also want to understand the factors which increase expenses in the emergency situation. For this purpose, we narrow the scope of the study to these three diseases, to analyze the difference in features which contribute to hospital charges for the treatment.

Given a patient's personal information, hospital indicators, demographic indicators, location, we want to give an estimate of total average charges by a hospital, for each of the diseases in the scope of our study. We also study the difference in hospital charges for mood disorders based on the protected attribute of age and gender any disparate treatment. We then discuss the potential of our model to be a weapon of math destruction.

2.Data

2.1 Data Description

The main dataset we are using in this analysis, is the Statewide Planning and Research Cooperative System (SPARCS) inpatient de-Identified dataset, which provides details on patient characteristics, diagnoses, treatments, services, and charges for the year 2012. After the exploratory data analysis in the mid term report, we have augmented the dataset to avoid the problem of overfitting and include data on Inpatient Quality Indicators (IQI), Patient Safety Indicators (PSI), Hospital Inpatient Potentially Preventable Readmission (PPR) Rates and US Census Small Area Income & Poverty Estimates (SAIPE) 2012 for NYC counties.

The dataset provides information on service area, age group, gender, race, ethnicity of patients, type of admission, diagnosis, risk of mortality, and total charges/costs and an expanded feature set by looking at the number of beds available, insurance coverage of the patient, income and economic information at the county level, to be factored into the recommendation for hospital in an emergency situation.

Thus, our big messy dataset has data of different types: numerical data such as length of stay, ordinal data such as severity of illness, boolean data such as Emergency Department Indicator, and categorical data such as the payment type. The total number of columns in the augmented dataset after cleaning the messy data is 28. Our feature space thus considers length of stay, severity of illness along with other indicators for patient, hospital and the financial situation.

The target variable is the total hospital charges and is the only way to evaluate the quality of care available at a hospital with the current dataset. Because of this limitation, it will be difficult to

rank the hospitals which would require data on the quality of doctors, equipment, and the overall care. We thus limit our analysis to prediction of hospital charges in emergency situations for the selected diseases.

2.2 Data Cleaning

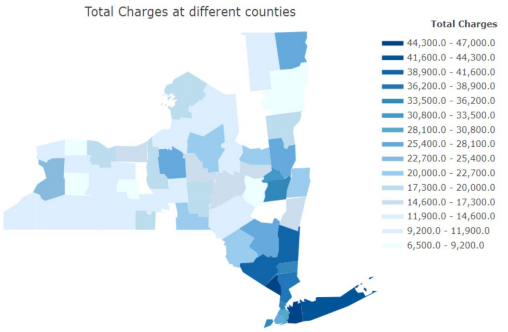
The first step was to delete the duplicated columns where they provide the identical information, for example, the column 'APR MDC Description' is a detailed explanation of 'APR MDC Code' where APR MDC stands for All Patient Refined Major Diagnostic Category. We also dropped the birth weight column because it is not considered useful in our analysis. We deleted payment typology 2 and 3 because they have too much NaN data and their information overlapped with payment typology 1. As emergency is the most frequent type of admission, we then delete the rows where they show non-emergency types. We also dropped the rows where the abortion edit indicator shows yes as those rows correspond to entries redacted to confidentiality.

We then transform the string values such as '120 +' in Length of Stay columns to numerical 120 and apply one-hot encoding to all the categorical data variables. During analysis, we also found that there are some entries corresponding to 120 in Gender, Race, Ethnicity, Severity of Illness variables and the rows corresponding to this were also dropped. Also, there is a string 'OOS' in zip code which corresponds to "Out of State". Since our data only contains information on hospitals within New York and that is also the location of our focus, we have dropped the corresponding rows for this zip code entry. For the variables which have an unknown level for a categorical variable, like for Gender, Race, Ethnicity, it has been encoded as a separate category.

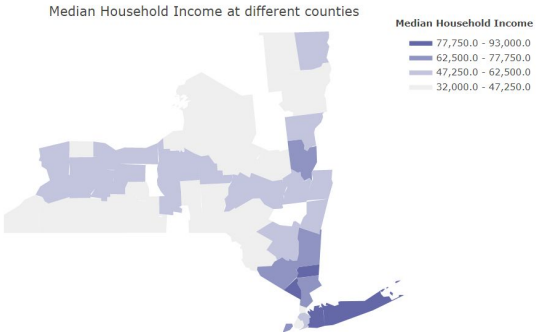
3.Exploratory Data Analysis

To better understand how the emergency charges are affected by the conditions of the patients and hospitals, we plotted the total charges of emergency versus health service area, counties, type of disease, payment types, medical procedures (non-surgical or surgical). We also explore the spatial map of hospital charges for mood disorders and median income for NYC. Below we only show four of the plots due to the space limit of the report.

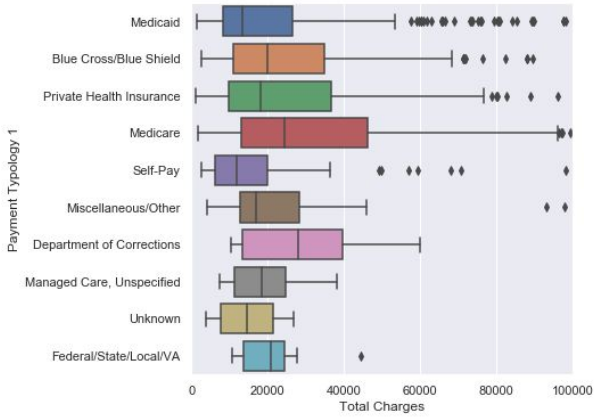
For the given dataset, we found that most of the medical charges for emergency lie between between 8,000 - 23,000 dollars, with a mean of 16,600 dollars. In Figure (a), we showed the total charges differ in different counties. We found Hudson Valley and Long Island have the highest mean charges. In Figure (b), we showed the median household income in different counties and found they seem to overlap with total charges distribution across counties. In Figure (c), we showed the charges across different payment types and found that self-pay produces the lowest charges. Figure (d) shows that certain hospitals attract more 0-17 age group patients, from which we may infer a more child related medical expertise. The data exploration analysis shows the features that can be considered for the modeling step.



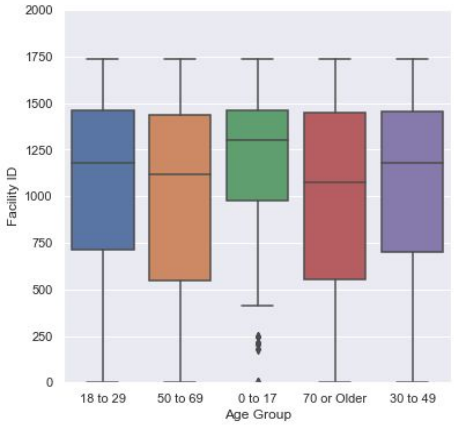
(a) Total charges for emergency at each county



(b) Median household income at each county



(c) Payment type versus total charges



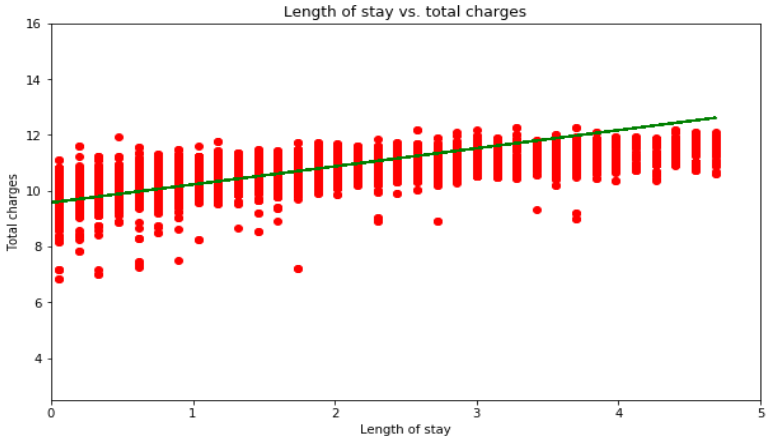
(d) Difference in hospital choices across age

4. Model Selection

We first considered feature selection more quantitatively using the Random Forests model which can naturally select the most important features for a specific regression target. The Random Forest is an algorithm consisting of many decisions trees. It uses bagging and features randomness when building each individual tree to try to create an uncorrelated forest of trees that run in parallel whose prediction by committee is more accurate than that of any individual tree [2,3]. This model returns the top 7 important factors that influence the total charges in emergency: length of stay, medical diagnosis category, medical surgery types, hospital, service area, severity of illness and hospital county.

We started the analysis with the baseline model of a simple regression which has an R squared score = 0.8. With a limited number of features, we might be overfitting and thus we consider regularized models with ridge and lasso regularizers while augmenting our dataset. As expected, the length of stay is an important explanatory variable as seen from the graph below.

Length of Stay vs Total Charges: Ridge Regression

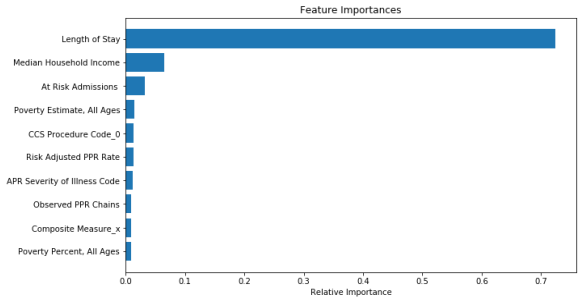


Our intuition was that location will be an important criteria in selection of hospitals in emergency situations and thus zip code would be an important variable to consider. Note that the zip code data is this dataset is trimmed from 5 digits to 3 to protect patient's privacy. Comparison of training and test errors of the model with and without zip code shows that there is minimal difference between the two and thus we select the more parsimonious model after cross-validating our results.

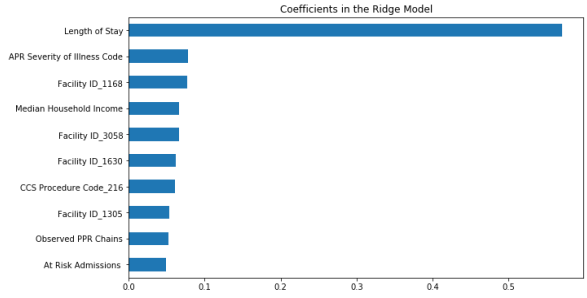
Here we are presenting results for the model fitted without the zip code, for each of the diseases using Ridge regression and Random Forest Regression which gives the feature importance.

Septicemia

Random Forest

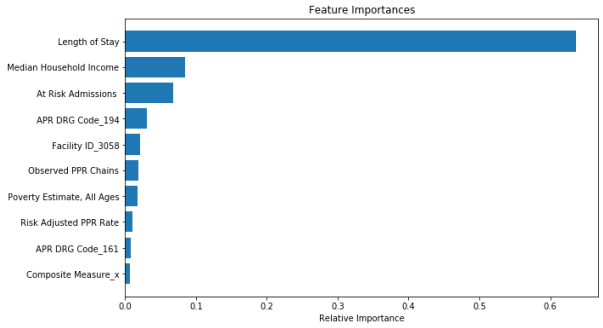


Ridge Regression

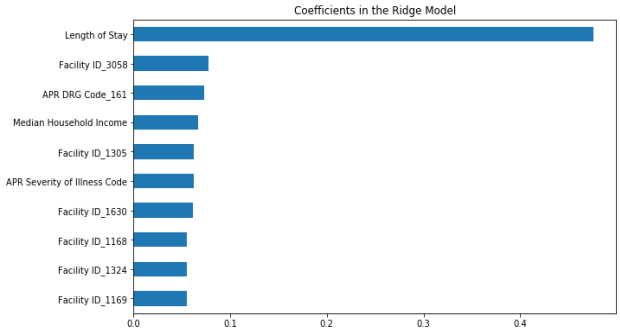


Congestive Heart Failure

Random Forest

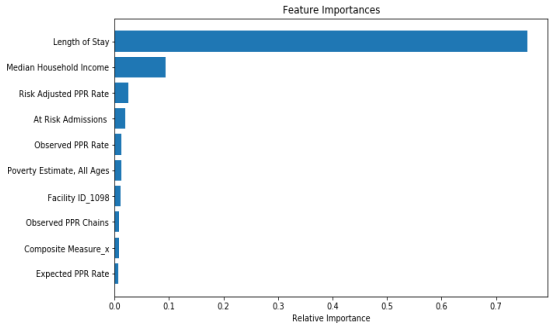


Ridge Regression

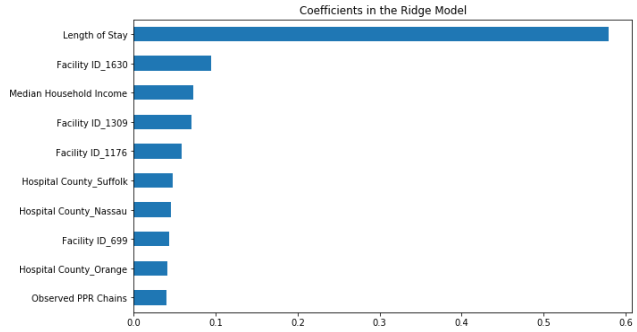


Mood Disorders

Random Forest



Ridge Regression



Models Comparison		**Septicemia**	**Congestive Heart Failure**	***Mood Disorders***
Linear regression	Train MSE	0.238	0.154	0.0858
	Test MSE	0.241	0.159	0.0853
Lasso regression	Train MSE	0.240	0.153	0.085
	Test MSE	0.244	0.157	0.085
Ridge regression	Train MSE	0.117	0.156	0.0858
	Test MSE	0.117	0.160	0.0850
Random Forest	Train MSE	0.0174	0.0155	0.0052
	Test MSE	0.0398	0.0344	0.0144

5. Analysis of Results

Based on the graphs and table above, length of stay is clearly the most significant factor for total charges. Total charges are also found to be related to the median income level, the location of the hospital, the severity of illness, at risk admission, the drug being used and the facility ID (the quality indicators of the hospitals). Notably, there is a contrast in features selected by Random Forest and Ridge regression for congestive heart failure cases. Apart from length of stay which is common in between both the models, Random Forest model picks on more indicators related to disease and finance related information which can be interpreted as patient information predicting the total charges whereas the Ridge regression is selecting dummies for hospitals, which implies that certain hospitals are more expensive controlling for patient characteristics. Both the models find the Facility_ID 3058 to be important, and that corresponds to Montefiore Medical center Jack D Weiler hospital, which implies that there might be some expensive hospitals in the dataset.

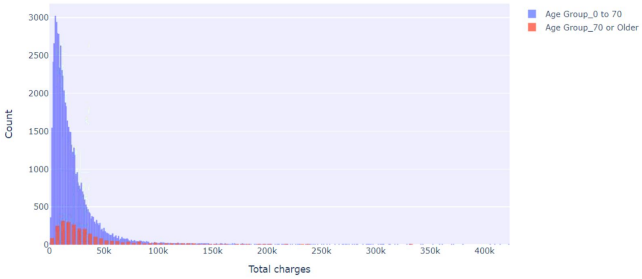
Demographic data was also found to have predictive information for total charges. Median income level, poverty estimates and the location of the hospital were found to have significant coefficients.

Other methods we experimented with was, linear regression with lasso regularizer. The comparison of training and test errors of all the methods shows that there is a difference in how each algorithm performs for that disease. For mood disorders, all the models were found to work well while for Septicemia, Random Forest model had the lowest error. For heart failure subgroup, the error for linear model with ridge and lasso gave similar results but again, the error is lowest with the Random forest. The difference in models found to work for each of the diseases implies that we need to take disease characteristics into account, when modeling hospital charges.

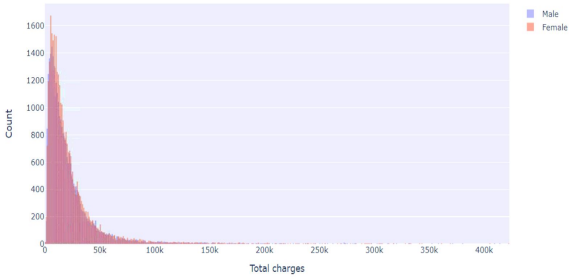
6. Fairness Metric

It is crucial to consider whether the algorithm we use for classification is fair, as discriminatory treatment can change the health outcomes for that population. If the hospital charges more from certain groups, then the lack of parity will change the data distribution in the future and make that disease more problematic for certain groups, affecting the quality of life. From the hospital regulation perspective, fairness implies that expenses for treating the same disease type must be similar, regardless of the age group, adjusting for the base rate or biological differences between the groups. In the context of our study, statistical parity makes sense as a fairness metric to check for disparate treatment of groups.

Mood Disorders Hospital Charges by Age



Mood disorders Hospital Charges by Gender



We look at the subset of data for mood disorder and classify hospital charges as more or less than $40,000 using Random Forest classifier and Logistic regression for two protected attributes, age and gender. Fairness is this context implies "disparate treatment", i.e. treatment must not explicitly depend on the group. Using Random Forest classifier, this fairness metric shows that the average charges for old people, above the age of 70 is 24.4 %, which is much larger than other group at 10.8% while there is no difference in the average charges for males vs females, with metric at 10.8% for males and 11.5% for females. For Logistic classification, this difference is starker for old people, with average hospital charges at 25% and younger people at 8.3%. The averages for males and females are similar at 6.8% and 7.8% respectively. As mood disorder can call for more treatment and care for older patients compared to younger patients, this metric is capturing the differential charges. We can conclude whether the classifier is fair or not, based only on the values of this metric after accounting for the base rate differences in treatment for age groups. For the attribute sex, this classifier is fair in terms of statistical parity i.e. there is no difference between males and females, for hospital cost of treating mood disorder.

7. Weapons of Math Destruction

Our model predictions and selection of hospitals to visit, can produce a weapon of math destruction, in both positive and negative directions. If certain hospitals are deemed to be more expensive for treatment, it could reduce the number of patients visiting that hospital, thus driving up the expenses of that hospital, making it even more expensive, and in some cases, driving the hospital out of business. If certain less known hospitals are recommended, then it can be a self fulfilling prophecy by bringing more patients and improving the scale of operations and medical expertise for that disease or if this selection was a false positive, the patient visiting that treatment may not get adequate care and treatment, which will affect quality of life of patients visiting that hospital. Either ways, the model predictions have altered the data distribution and can affect the metrics by which hospitals are evaluated, thus changing the allocation of resources within a hospital with consequences for patients relying on quality care at affordable prices.

8. Conclusion

This modeling exercise for predicting hospital charges in an emergency situation brought to fore, the importance of taking disease characteristics, income of the location and median house as factors which affect the medical expenses in emergency situations, apart from the severity of illness and type of admission etc. The visualizations and analysis also emphasizes the role of

location in treatment cost for mood disorders, with costs higher in poor counties of NYC. Although there are many improvements possible in building the predictive model, the analysis so far offers valuable insights to consider from not only the patient perspective but also from policy perspective of providing health care in a spatially equitable way. We are confident in the overall broad conclusions that we can obtain from this analysis but we would be hesitant to deploy this model for policy decisions and into production. A more detailed model which captures the complexity of location, demography and business aspect would be necessary.

9. Limitations and Future Work

Although we have a large dataset, we did not have variables to measure the quality of care or variables to explain the hospital charges and thus give a ranking of hospitals. An ordinal regression to rank the hospitals would be informative in recommendations. An unsupervised learning algorithm could also be used to cluster the hospitals based on their specialities and the patient characteristics. What we care about the most in life threatening situations is the timeliness and quality of care, which cannot be adequately captured by the hospital charges. Measures to capture the quality of care in a hospital would thus help build a better predictive system and from a broader perspective, in choosing the location and medical expertise based on the requirements of the demographic population.

For including the zip codes, instead of one hot encoding, we would include the additional census information and fit a generalized low rank model that condenses that information into a small vector with archetypes of demographic indicators. In future work, we would like to explore more spatial variables for assessing hospital charges and quality of care.

References

1. Chen X, Wang Y, Schoenfeld E, Saltz M, Saltz J, Wang F. Spatio-temporal Analysis for New York State SPARCS Data. *AMIA Jt Summits Transl Sci Proc*. 2017:483–492.
2. Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." *R news* 2.3 (2002): 18-22.
3. https://towardsdatascience.com/understanding-random-forest-58381e0602d2