CornellEngineering
Operations Research and Information Engineering

ORIE 4741

LEARNING WITH BIG MESSY DATA

Final Report

Yijia Cui — yc2576

Jia Tang — jt854

December 10, 2019

Contents

Problem description

Our project aims at giving the patients helpful information on choosing a hospital. Given their personal information and medical history, we hope to estimate their length of stay and the cost for staying at particular hospitals. For our project, we decided to focus on patients with depression. Unlike other major diseases that have obvious symptoms, the diagnosis for depression is highly complicated and personalized. The length of stay in hospitals and the costs, just as the diagnosis itself, also depends significantly on the patient's information. Therefore, in our project, we want to develop models that predicts the length of stay and costs for patients with depression. But with similar approaches, future work can also be done for analysis focusing on patients with other diseases.

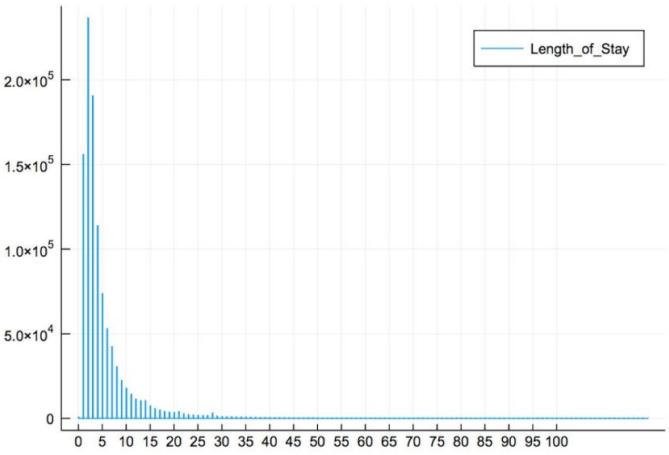
Dataset description

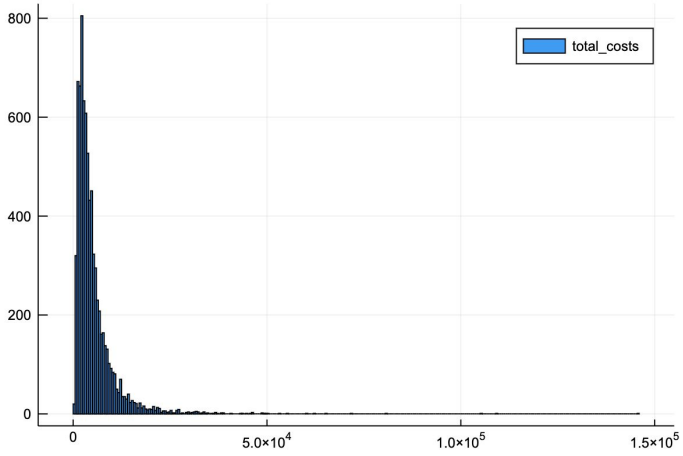
In this project, the dataset we used is the SPARCS Hospital Inpatient Discharges dataset for hospitals in New York State. Providing the basic information of all patients, this dataset includes the age, race, gender and ethnicity of the patients discharged from 2012. The data set also provides information about length of stay, type of admission, patient deposition, CCS Diagnosis, severity of illness, risk of mortality, payment method, which allows us to show the possible relationship between them clearly. Since the dataset also includes the data of total charges and total cost corresponding to each patient, we can also investigate the costs covered by insurance, which can help us to investigate which hospital saved money best for depression patient. Furthermore, the dataset also allows us to compute other features, including ratio of severity of

illness, with 1 representing minor and 4 presenting extreme, to length of stay for each patient in corresponding hospital.

The dataset is comprehensive and almost complete. This dataset contains 2,544,543 examples and 35 different features. The features of the data are broken down into the following categories:

Feature Type	Count
Boolean	**3**
Nominal	**15**
Numerical	**14**
Ordinal	**3**
Total	**35**





According to the histogram of length of stay and total costs feature shown above, the average length of stay for all patients in New York State is approximately 5.32 days with maximum days of 119 days and minimum days of 0 days. And for depression patients, the maximum length of stay is 97 days and the minimum is 1 day. Since the average length of stay in

a hospital is used to gauge the efficiency of a healthcare facility and the national average for a hospital stay is 4.5 days, less than 50% of the hospitals in New York States are able to operate in an efficient manner. Also, there are many outliers that are far from the mean length of stay days. According to the Agency for Healthcare Research and Quality, the average cost is $10400 per day, which exceeds the maximum cost, $999.98, of all patients in New York State and the maximum charge, $998.5, for depression patients in New York State.

Models

Model 1: We developed this model to predict the patient's length of stay in the hospital.

Approaches: We applied a filter to our dataset, selecting only samples whose APR_DRG_Description column have value "Depression except major depressive disorder". There are 7818 such samples. In order to decide which features are most influential and should be examined for our model, we ran preliminary regressions on all features and picked the features that are significant to prediction of length of stay. Those features are Age_Group, Race, Gender, Ethnicity, Type_of_Admission, Hospital_Service_Area, Hospital_County, and APR_Severity_of_Illness_Code. Among these features, APR_Severity_of_Illness_Cod is real value data and does not need any feature engineering. All other features are categorical data, and we used one-hot encoding to transfer these features into the following vectors:

Age_Group: ["70 or Older", "50 to 69", "30 to 49", "0 to 17", "18 to 29"];

Race: ["White", "Black/African American", "Other Race", "Unknown"];

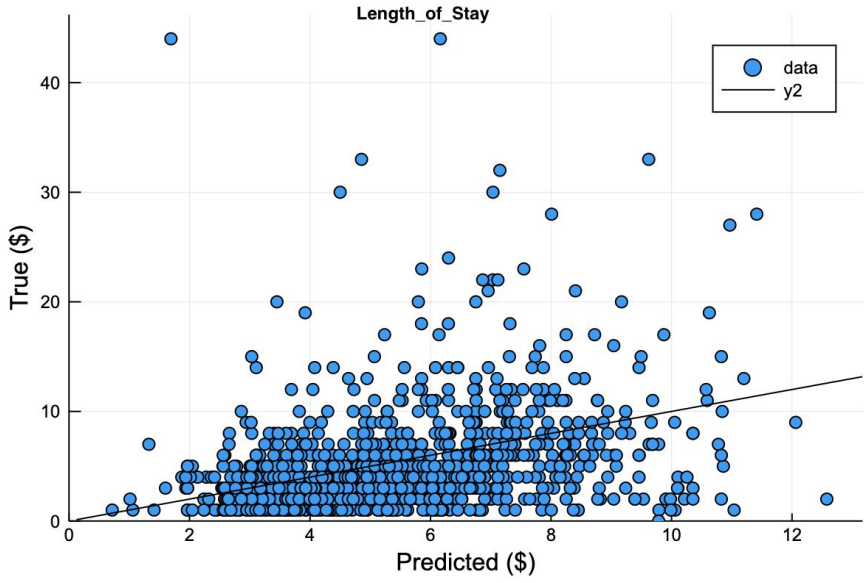
Gender: ["F", "M"];

Ethnicity: ["Not Span/Hispanic", "Unknown", "Spanish/Hispanic"];

Type_of_Admission: ["Emergency", "Urgent", "Elective", "Not Available", "Trauma"]; Hospital_Service_Area: ["Finger Lakes", "Hudson Valley", "Western NY", "Capital/Adirond", "New York City", "Central NY", "Southern Tier"]; Hospital_County: ["Monroe", "Chemung", "Westchester", "Erie", "Schenectady", "Niagara", "Rensselaer", "Ulster", "Orange", "Warren" … "Chautauqua", "Columbia", "Rockland", "Franklin", "Herkimer", "Allegany", "Broome", "Cattaraugus", "Schuyler", "Genesee"]. Then we cleaned data to eliminate rows with missing values from the columns mentioned above. We got 7802 data points. Since the data points with missing values only accounts for less than 10% of the population, we simply deleted those data points instead of estimate for the missing values.

Result:

We fit a linear model using the features mentioned above and an offset. Using the MSE function, we calculated that the training set error is 26.367 and the test set error is 32.803. On average, patients with depression stays in the hospital for 5.474 days.

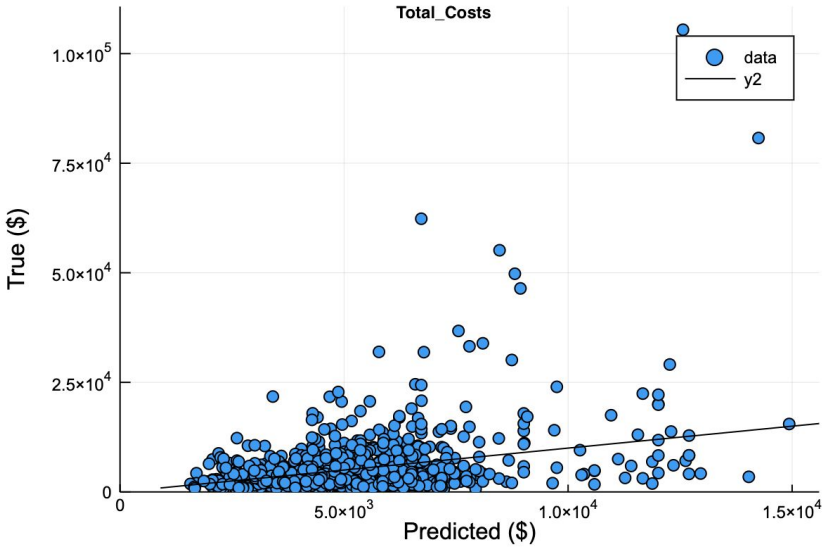


Model 2: We developed this model to predict the total costs for patients with depression.

Approaches: In a similar way as in model 1, we selected Age_Group, Race, Type_of_Admission, Hospital_Service_Area, Hospital_County, Payment_Typology_1, Payment_Typology_2, Payment_Typology_3 as the significant features. However, if we only look at data points without missing values from all these features, there only 1152 data points left, which we considered not enough. In addition, after applying feature transformations, the resulting real-valued matrix had 75 columns, which concerned us as overfitting. Therefore, as we did in our preliminary model analysis, we run preliminary regressions on different combinations of the features. Finally, we selected Age_Group, Race, Type_of_Admission, Hospital_Service_Area, and Payment_Typology_1 as features. In addition to the one-hot encoding we did for model 1, we also transformed Payment_Typology_1 into ["Medicaid", "Medicare", "Private Health Insurance", "Self-Pay", "Blue Cross/Blue Shield", "Federal/State/Local/VA", "Miscellaneous/Other", "Managed Care, Unspecified"].

Result:

We fit a linear model using the features mentioned above and an offset. Using the MSE function, we calculated that the training set error is 2.610e7 and the test set error is 3.774e7. On average, patients with depression spend 5199.055 dollars in the hospital.



Limits

First, we only have a population size of 7818. And there is only one description in the entire APR_DRG_Description column ("Depression except major depressive disorder") that indicates that the patient has depression. However, in reality, there are more layers to the diagnosis of depression. For example, how long have the patients be diagnosed with depression and whether or not the patients are taking medication can both affect the length of stay or costs. Even though we did include level of severity in the first model, our model is still not comprehensive enough due to the lack of samples and features.

Second, due to the limited time and our limited knowledge, we are unable to fully apply what we learned in class in this project. Although linear regression performs alright in our case, it might be better if we can use a more complicated model in the prediction. It might also be useful if we can use decision trees or random forest in our feature selection. We could also apply PCA in the analysis.