Cornell University

ORIE 4741/5741 Learning with Big Messy Data

Final Project Report

Ruobing Shui, Hehong Li, Ren Ruize

12/05/2021

● Abstract

This project aims to assist residents in New York State on choosing the most economical hospital to go to when they are experiencing Septicemia. We use inpatient and hospital data including length of stay in the hospital, age group, patient disposition (The patient's destination or status upon discharge), facility ID(Permanent Facility Identifier) and operating certificate number (The facility Operating Certificate Number as assigned by NYS Department of Health) to generate a model to recommend a hospital for patients. We also evaluate the fairness of our model and analyze the limitations of our project. Since there are a large number of hospitals and patients may have trouble making decisions, or may not have enough information to make best choices, this project has the possibility to help residents get the best treatment they can afford.

● Background

There are 6,090 hospitals in the US investigated by the US hospital association.[1] If we narrow down to New York state, there are still 214 hospitals.[2] The question of how to make right decisions always annoys people.

The process of making decisions of the hospitals normally costs time and money, sometimes even lives. There are several barriers for people to pick out the best option. Not only the tons of hospitals, but also information barriers stop people from making decisions properly. Besides, due to the different education level and poor health announcements, many people even have no idea what the illness they get is and how serious it is, which might result in a tragedy.

● Dataset

The dataset used is the Statewide Planning and Research Cooperative System (SPARCS) Inpatient De-Identified dataset [3] which includes details on patient characteristics, diagnoses, treatments, services and charges in 2012. The shape of this dataset is approximately 2,550,000 rows and 35 columns and is integrated based on the features of patients' ages, their races, genders, the locations of the hospitals and so on. There are different kinds of data in this dataset, such as categorical data, numerical data and miscellaneous data and also some with the values of 'NULL'. As only the numeric data could be analyzed and plotted by the mean squared error (MSE), the original dataset was reintegrated and cleaned before performing the regressions on the data.

The dataset is required to perform some preprocessing at first in order to fit the future models using the information in this dataset. At first, some features were selected by ourselves according to the assumptions on features which could have high effects on the costs of visiting hospitals for each patient. These features are 'Health_Service_Area', 'Gender', 'Age_Group', 'Race', 'Payment_method', 'CCS_Diagnosis_Description', 'Length_of_state', 'Total_Costs', 'APR_Risk_of_Mortality', 'ARP_Severity_of_Illness_Description', 'Zip_Code_3_digits', in which the 'Payment_method' feature is generated with the operations of combining three columns of payment method. After constructing the new dataset, all of the values of 'NULL' are also dropped for cleaning data and the data of zip codes which represent areas outside the NY state is also dropped.

After cleaning the data, the feature transformations are applied on two different kinds of data: categorical data and miscellaneous data. For the first one, the one-hot method is used on all the categorical data, such as the payment method, the description of the severity of illness, age group and so on. And the many-hot method was used on multiple-payment-method features to label the feature. For the Miscellaneous data, which means the combination of numerical and text data, it also requires the process of transformation which is converting the value of '120+' which represents the staying time longer than 120 days to '120'.

At first, the descriptive statistics of the data acquired were graphed using histogram and the values of mean, standard deviation, maximum and minimum were calculated by the program. These values of the total charges for various patients are shown in the following Table 1.

Table 1: Descriptive Statistics of Total Costs

Statistics	Values
mean	11,870
maximum	32,300,000
minimum	0.1
standard deviation	31,950

To visualize the relationship between the total costs and the frequencies of patients, the histogram of both total costs and charges are plotted in Figure 1.

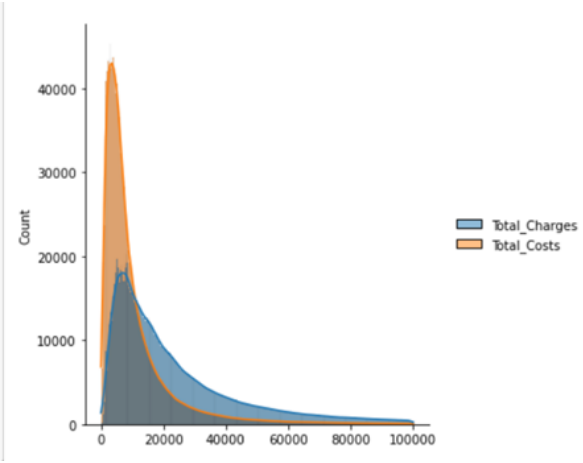

Fig. 1 Histograms of Total Costs and Total Charges

As shown in the figure above, the total costs of patients could be up to about 50,000 dollars while the charges are only lower than 20,000 dollars. This means the insurances, such as Medicaid, Blue Cross and Managed Care could take charge of most of the costs for each patient. Then some line figures are plotted to distinguish the differences of total costs by category and the relationship between costs and three different classification methods are shown in the following figure.

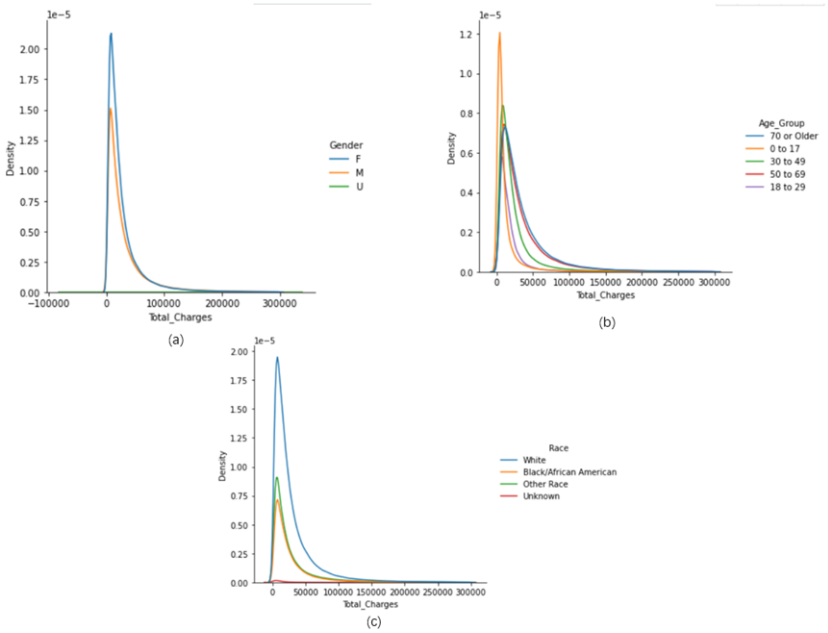

Fig. 2 Line Figures of Total Charges in three classification methods: (a) Gender; (b) Age group; (c) Race

After the pre-processing of the dataset, the linear model regression was made and the mean squared errors of it were calculated to analyze the performance of this model. Both of the results and the model graph will be represented in the next section and the availability and accuracy of the model will also be discussed.

● Linear Regression

For modeling, the data was split into training and testing sets at first, where 80% for the training set and 20% for the testing set. The dataset was separated mainly into three types: one-hot data, many-hot data and as well as numerical data. These three data sets were applied to fit the ordinary least squares (OLS) model respectively at first, then they were combined together and the mean squared errors were calculated again. The comparison shows that both the training and testing data increase when more features are added.

Table 2: Results of Training Error and Test Error

Data type	Train_MSE	Test_MSE
many_hot	2.8e8	4.4e8
one_hot	3.5e8	5.1e8
numerical	4.8e8	6.8e8
many_hot+one_hot	3.5e8	5.1e8
many_hot+one_hot+numerical	6.9e8	4.9e8

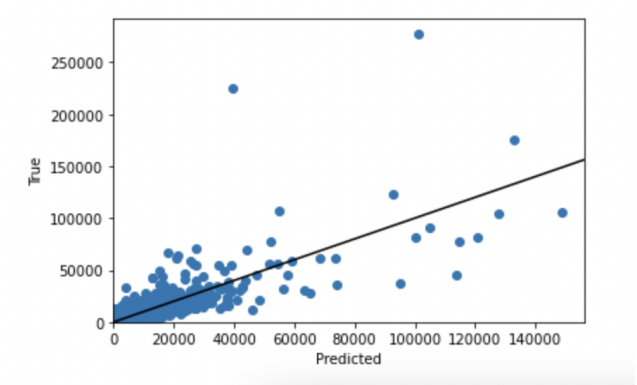

Fig. 3 Linear Regression for combined features

To visualize the errors, a line graph with data points was plotted as in Figure 3, but it is obvious that both of the errors were so large which means that the model chosen is not appropriate for this dataset. Therefore, some methods for the further analysis were raised: One is shrinking the dataset by only choosing one of the diseases which is Septicemia in this case; The other is considering establishing more complicated models, such as quantile and ridge models.

One method raised for making improvements on the model, the random forest model was generated to choose the features with higher relationship with the total costs. This process has been performed based on the tool packages obtained online and the features with the top 5 highest coefficients were used to do the future analysis, which are the length of stay, age group, patient disposition, facility ID and the operating certificate number.

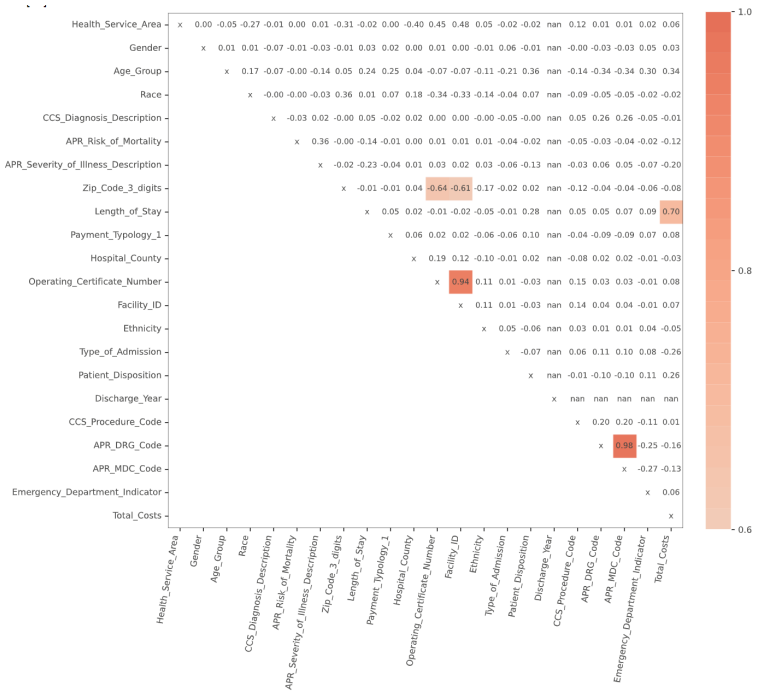

Fig. 4 Random Forest Model for Feature Selection

• Model Selection

We choose two techniques covered in class, the first technique is Quantile Regression, the second technique is Rigid Regression.

Predicting Total Charges by Quantile Regression

Broadly used in quantitative modeling, we want to see how there is a causal relationship between total cost and the average effect, so we applied quantile regression. Quantile regression is a procedure to model the relationship between independent variables and specific percentiles of a dependent variable. Compared with Ordinary Least Squares regression, it has two advantages: it doesn't make assumptions about the distribution of the target variable. It can also circumvent the influence of the outliers. Since total charges largely depend on disease, we narrow down our scope of disease only to Septicemia. Due to the capacity of the computer, we choose the top three important features from random forest, as well as randomly drop 99% of the data. We estimate the set of quantiles from 5% to 95%, compute its training and testing mean square error respectively. As shown in Fig. 5, when quantile is 42%, the model will achieve its least training and test mean square error.

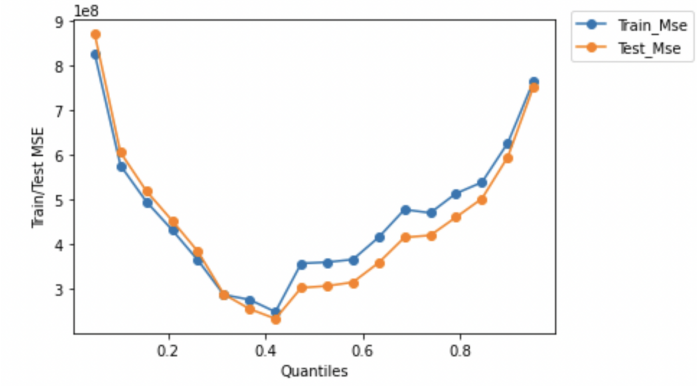

Fig. 5 Quantiles against the training and testing dataset in Regression

Predicting Total Charges by Rigid Regression

Rigid Regression applies when we have more independent variables than observations, or the current data set suffers multicollinearity. Due to multicollinearity, the least squares method from the preliminary analysis will have high variance. This will result in a large divergence between predicted values and the actual values. Same as the first model, we choose Septicemia, the top five important features, and compute corresponding training and testing mean square error. The regression strength ranges from 0.001 to 1, when alpha equals to 0.041, it has the smallest testing mean square error.

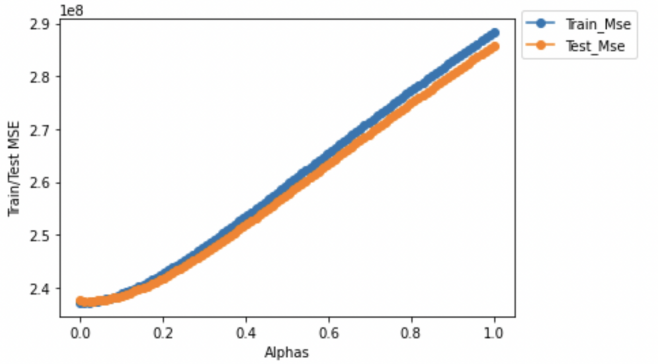

Fig. 6 Alphas against the training and testing dataset in Rigid Regression

• Result Analysis and Conclusion

Table 3 Fitting Result for 2 models

Data type	Train_MSE	Test_MSE
Quantile Regression	2.5e8	2.3e8
Rigid Regression	2.4e8	2.4e8

Model Comparison

From the table 3 above, both quantile regression and rigid regression have significantly decreased the training and test error from preliminary analysis.

Feature Consideration and Result Analysis

From the heatmap from random forest, it shows that 'Length_of_stay', 'Operating_Certificate_Number', 'Patient_Disposition', 'Facility_ID', 'CCS_Diagnosis_Description' are the features that have the most significance. In the meanwhile, individual information such as gender is not as important.

From preliminary analysis, random forest to quantile regression and rigid regressions, we found 'Length_of_stay', 'Operating_Certificate_Number' are significantly larger weights than the demographic information of a person, which is against our initial assumption. Among the variables, the accuracy value of length of stay can achieve up to 0.7 with regard to total cost. We can conclude that an individual's charge largely depends on his stay in the hospitals, which is common. As for the model, compared with testing mean square error, it is clear that Quantile Regression would be the best candidate method.

Though our model has several limitations, such as the single dimension of prediction result, the model still can provide patients with valuable insights in the finance view. In fact, after the fairness analysis in the following part, the hospital can also know the current cost situation so they can make prices more properly. Further, government staffs who manage the medical market, they can decrease the discremination and give a fair medical treatment environment to residents.

● Fairness Matrics

Fairness is a key element we need to consider when solving the real world problem. It studies how to decrease and even eliminate the bias in the dataset to avoid that the model results lead to the discrimination of the clasification groups, such as gender, age and race.

As for our project, the fairness refers to the total cost of treatment being at the same level for every group. If a hospital decides the charge based on the group types, then certain groups will find it more difficult to get good medical treatment in the future. For age groups, as the graph 7 shows, the cost distribution has little differences with ages above and below 50 years old which is around 20,000 dollars so there is no discrimination on age group. For racial groups, as the graph 8 shows, the cost of White people is around 20,000 dollars while other racial groups are around 29,000 dollars. It's 45% lower than other races. So when the government makes medical related policies, they need to consider minor race groups and give the residents a fair medical treatment environment.

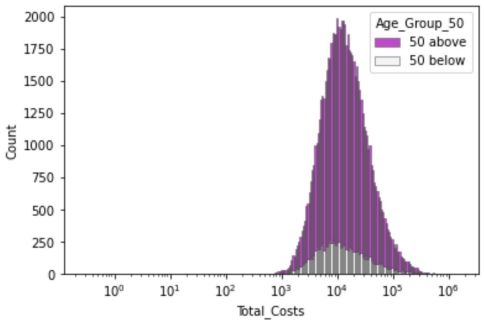

Fig.7 The total cost of Septicemia treatment by Age

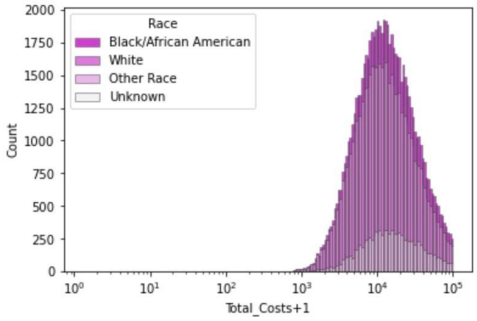

Fig.8 The total cost of Septicemia treatment by Race

● Current Limitation and Future Work

The project has several limitations. First, this project focuses on the total cost of treatment, there are other dimensions important to the cost of treatment results as well. In the future we can take the quality of treatment and the success rate into account. Second, the dataset has information bias and it's easy to lead to outliers. For example, the disease has different severity levels so that it will impact the results, as for this limitation, we can segment each disease and divide it into three kinds according to the emergency level. Third, the dataset just has the area of the hospitals instead of hospital's names. In the future,we can add another dataset with more detailed information to give patients more specific suggestions.

To help users get the suggestion easily and clearly, we can make a user friendly website or online AI chat bots to launch the project idea to real products.

- **References**

[1] Fast Facts on U.S. Hospitals, 2021, The American Hospital Association, Retrieved Dec 5th, 2021 from https://www.aha.org/statistics/fast-facts-us-hospitals

[2] NYS Health Profiles. Retrieved Dec 5th, 2021, from https://profiles.health.ny.gov/hospital/bed_type/Total+Beds

[3] Hospital Inpatient Discharges (SPARCS De-Identified): 2012 Available at: https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/u4ud-w55t