rs2579, hl778, rr779

Mid-term Project Report

1 Project overview

This project aims to assist residents in New York State on choosing the best hospital to go to when they are experiencing a medical situation. We prepare to use inpatient and hospital data including intake reason, cost, and payments, and then generate a model to recommend a hospital for patients based on their individual information. 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.

2 Dataset

To understand the dataset better and generate our model, we observe our dataset generally and adopt data cleaning.

2.1 Dataset Description

We specifically choose the dataset from the Hospital Inpatient Discharges (SPARCS De-Identified): 2012. This dataset has 2544543 rows and 35 columns, including the description of information of patients and hospitals. Patients' information includes the patient's age, race, location, payment methods and so on. In qualitative features view, there are patients' demographic background data to support classification; in quantitative features view, there are patients' treatment data to support analysis.

2.2 Data Cleaning

Before starting the analysis, we cleaned the data by dropping the rows which have null values which are meaningless. Then we transformate several features to help predict the model. We pick up the ['Health_Service_Area', 'Gender', 'Age_Group', 'Race', 'Payment_method', CCS_Diagnosis_Description', 'Length_of_state', 'Total_Costs', 'APR_Risk_of_Mortality', 'APR_Severity_of_Illness_Description', 'Zip_Code_3_digits'] as chosen features. The Payment_method feature is created by combining three columns of payment method. Since our recommendation is for residents in NY state, we dropped the data out of NY state. After cleaning and transformation, the new dataset has 2434522 rows and 153 columns.

2.3 Feature Transformation

We use two kinds of methods to make feature transformations.

2.3.1 Categorical Data

We apply the one-hot method to all the categorical data such as payment method, the description of the severity of illness, age group and so on to help develop the predict model. For multiple payment method solutions, we use the many-hot method to label the feature.

2.3.2 Miscellaneous

The length of stay data is hybrid with numerical and text data. For example, the patients stay more than 120 days will be recorded as 120+. We convert all the points of '120+' to '120'.

3 Visualization

To analyze some descriptive statistics of the data we acquired, some histograms are generated based on the total costs and charges for different patients. When applying the statistics describe function on the total cost data, the mean, maximum, minimum and standard deviation values are calculated which are $1.187 * 10^4$, $3.23 * 10^7$, 0.1, $3.195 * 10^4$. These for the total charge values are also calculated to be $3.28 * 10^4$, $7.066 * 10^6$, 0.31, and $6.156 * 10^4$. The relationship between the total costs and total charges are plotted to find about how much money could be saved after visiting the hospitals.

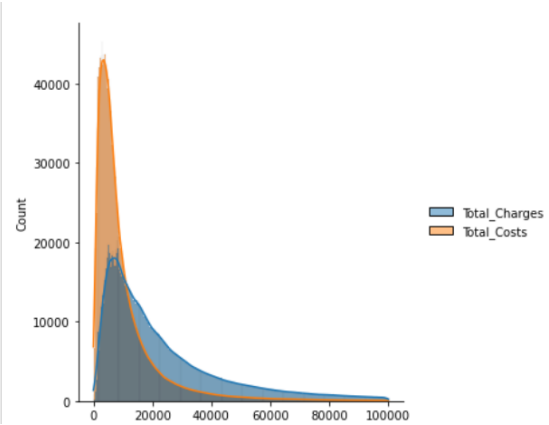

Figure 2: Histograms of Total Costs and Total Charges

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. Hence, to make the suitable decision for individuals, we will only focus on the charges in the following experiment.

After making the histogram for all the patients the total charges, some classifications are applied in the graph to make better estimation for the costs of different groups of patients. The first feature used is the gender and the histogram for the two categories are as following.

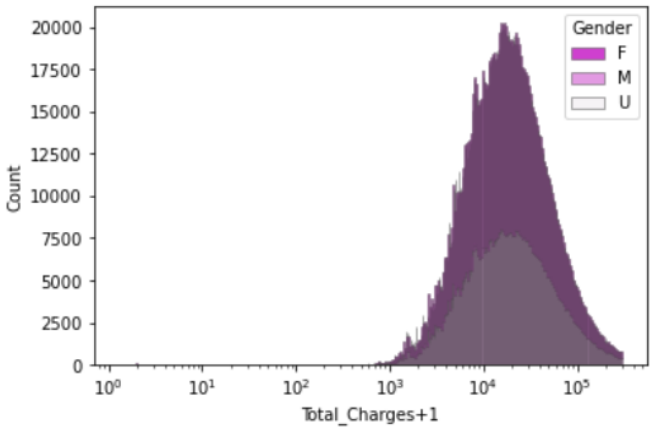

Figure 3: Histogram for Total Charges in Different Genders

However, as shown above, this histogram uses two different colors to show the charges for different genders (female, male and unknown). But this is not obvious to distinguish the differences of charges with classification. Hence, the line figures analyzing the relationship between charges and three different classification methods are generated in the following figure.

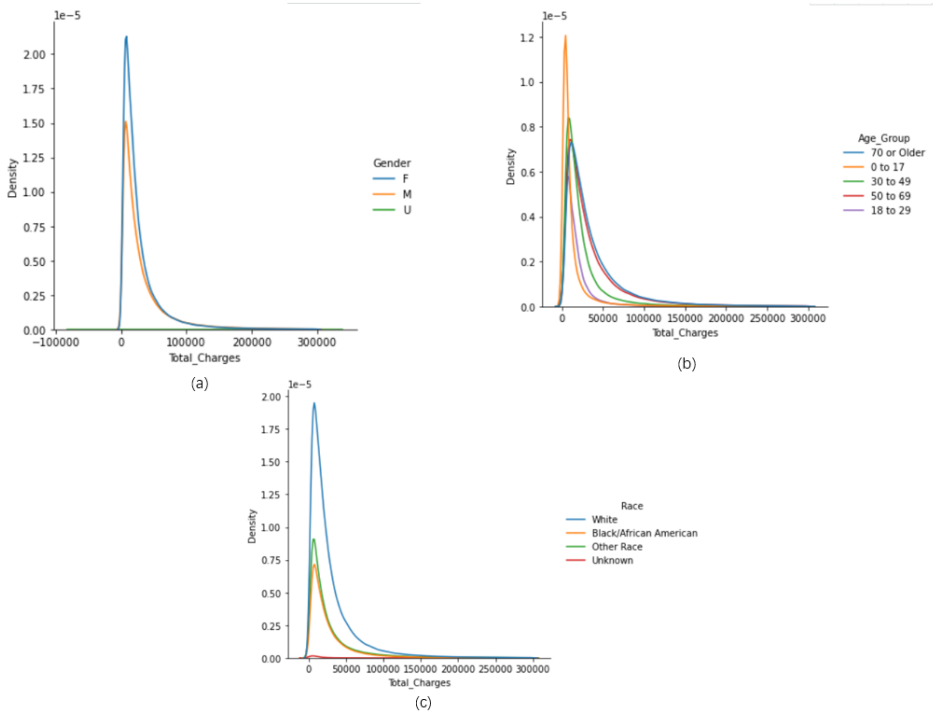

Figure 4: Line Figures of Total Charges in three classification methods: (a) Gender; (b) Age group; (c) Race

From these figures, it can be indicated that female always spend more money during visiting the hospitals than male. For the age group, younger people group shows higher total charges which leads to higher costs in hospitals. As for the race classification, white people spend the most money while the African/Black American cost the least.

4 Preliminary analysis & Effectiveness

According to the previous discussion, we divided our model into three data types, they are one-hot data, many-hot data, as well as numerical data set. We first separately applied these three kinds to fit with the ordinary least squares(OLS) method, then we combined them with each other.

Table1: The results of training error and test error

Data type	Train_MSE	Test_MSE
many_hot	284342225	444670594
one_hot	349837619	506420295
numerical	482865918	682345093
many_hot+one_hot	350278360	506869116
many_hot+one_hot+numerical	494320350	692740834

```
Size of dataset:  2434522
(19476, 158) (4869, 158)
(19476, 1) (4869, 1)
Train MSE         692740834.0516478
Test MSE          494320350.88283265
```

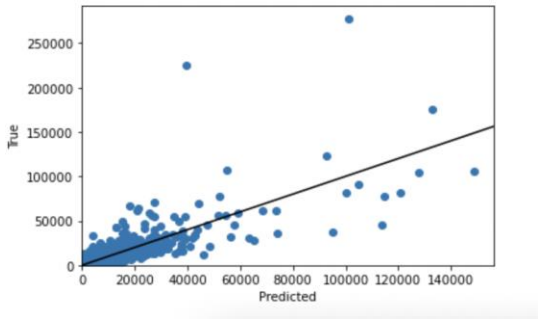

Figure 5: Result for all combined features by OLS

We can see that, as we add more features to our model, both the train and test mse would increase, this means that in our preliminary results, the result cannot be simplified trained with the linear models, which becomes less effective. Maybe some of the category data sets can be represented as linear models, but we cannot add all of them for one at one time. We will include this in our future analysis. Besides, in our model, since the original dataset was too huge, we choose to randomly drop 99% of the data set.

5 Future work

We already built a linear model currently but the train error and test error are both too large. There are several tips we need to consider in the future.

5.1 Plan of avoiding over-(under-)fitting

- To avoid overfitting, we can use regularization methods such as L1 regularization and ControlBurn to add information to our model.
- To avoid underfitting, we can compute the bias and variance of our model.

5.2 Next step

- Our next step will be more focused on the modeling for the total charges based on the features we selected.
- We will use more complex algorithms to generate our model to fit the dataset. Since this time we dropped 99% of our data, we can use cross validation to more accurately predict our features.
- Based on the results we will see if we need to add more features to generate the model and decide which feature can be helpful for the model.