ORIE4741 Project Midterm Report

Siyao Gu(sg2238), Yingying Zheng(yz949), Yiling Jiang(yj333)

I. Introduction

Our project is aimed at predicting inpatient charges based on the data from New York State hospitals in 2012 in the hope of detecting potential medical bill frauds and abuses for healthcare insurance companies.The data that we are going to examine are obtained from New York State data portal. The dataset contains 2012 New York State hospital discharge level details on facility name, hospital county, patient characteristics, diagnoses, treatments, services, and charges. After completing preliminary analyses of our data, we decided to predict the inpatient charges using classification models.

II. Data Set Cleaning

The original data contains 2.54 million rows with 37 columns. We did several cleanings and selections.

First, one important feature is APR DRG Code — The APR - DRG Classification Code. We decided to only study the data of which the diseases described by APR DRG Code are the 10 most frequent among all other data.

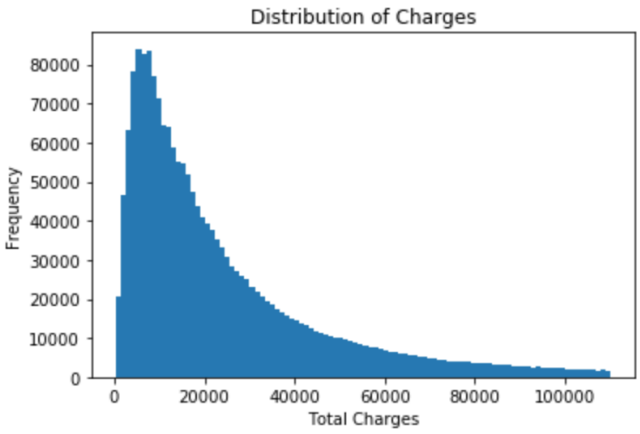
Second, we encoded three important features — Age Group, Type of Admission, and Patient Disposition — and added additional three columns for them, namely Age Group Indicator, Admission Type Indicator, and Patient Disposition Indicator for future studies. For example, the way we encoded the Age Group Indicator is 0 for Age 0 to 17, 1 for Age 18 to 29, 2 for Age 30 to 49, 3 for 50 to 69, and 4 for Age 70 or older.

Third, we found there were about 3000 rows containing missing values across many columns of the table. We then realized these observations all corresponded to abortion records. As stated on the New York State Department of Health online data portal, discharges identified as abortions have been redacted with all hospital related fields replaced with ?N/A?. Therefore, due to the limited amount of information available for the abortion records, we decided to remove them from the data set and thus our study of interest. In addition, we also removed records with the Total Charges below $500 because we think it is almost impossible to have inpatient treatment with cost less than $500.

III. Data Set Understandings

The cleaned data set contains 1.98 million rows and 40 columns with five numerical columns and 35 categorical columns. In particular, among the columns we think that could be included in the initial modeling tryouts, Hospital County, Age Group, Gender, Race, Type of Admission, Patient Disposition, APR DRG Code, Emergency Department Indicator, APR Severity of Illness Code are categorical and and Length of Stay is numerical.

The summary statistics of the Total Charges is shown in the table below followed by its distribution plotted up to the 95th percentile.



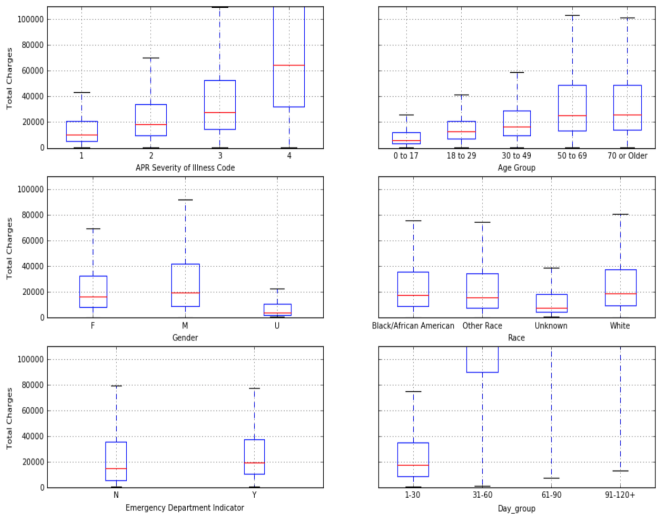
The histogram shows that the Total Charges is significantly skewed to the right as the median is about twice as much as the mean.

Since we speculate variables Hospital County, Age Group, Gender, Race, Length of Stay, Type of Admission, Patient Disposition, APR DRG Code, APR Severity of Illness Code, Emergency Department Indicator are likely to play an important role in determining hospital charges, we examined the effect of each of the features on the predicted variables.

We found the average charges across the counties range from $7.1k to $48k with a standard deviation of $11k. The average charges also vary significantly among different types of patient disposition with a standard deviation of $27k.

The following six box plots show how the Total Charges vary with different levels of the remaining

features. Based on the plots, the Total Charges generally increase as age, severity level, length of stay increase. The charges of the majority of the male patients are higher than that of female patients. Moreover, the comparison between the charges for emergent patients and those for non-emergent patients shows that being admitted as an emergent patient is usually associated with a higher payment. However, charges do not vary significantly across different races.



IV. Preliminary Regression Analyses

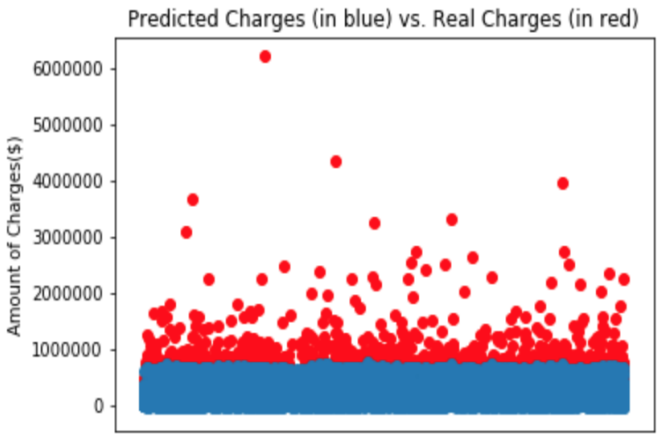
The initial model we considered was linear regression model. The dependent variable is the Total Charges. Predictor variables include Hospital County, Age Group, Gender, Race, Length of Stay, Type of Admission, Patient Disposition which explains patient's status upon discharges, APR DRG Code which records the treatment, APR Severity of Illness Description and Emergency Department Indicator which indicates whether the patient went through the emergency process. Among these 10 selected predictor variables, 9 of them are categorical variables. Specifically, 7 out of 9 categorical variables include more than 2 levels. To deal with them, we used one-hot encoding, which resulted in a 1,607,648*52 feature vector.

Following the selection of predictor variables, we randomly split the data into 80% training set and 20% test set. Then we built a simple linear regression model with mean squared error (MSE) as the loss function. The result is shown below:

	features	estimatedCoefficients
0	Length of Stay	5060.547796
1	F	-1713.676039
2	Black/African American	-4425.703619
3	Other Race	343.305375
4	139	-7945.360383
5	140	-8776.105080
6	194	-7721.611419
7	540	-10848.389763
8	560	-9145.247347
9	640	-23044.020569
10	720	-15623.524510
11	Major	-34852.922971
12	Minor	-41426.374208
13	Moderate	-42237.381544
14	Y_emer	-3693.802843
15	1_age	-7062.493717
16	2_age	-5521.492984
17	3_age	687.159569
18	4_age	-2940.651762
19	1_admin	-14479.934255
20	2_admin	-4389.823046
21	3_admin	1585.963711
22	4_admin	-5672.663903
23	5_admin	1466.307586
24	0_dispo	-33211.582577
25	10_dispo	-8842.043021
26	11_dispo	-9709.636751
27	12_dispo	672.192851
28	13_dispo	-3453.554176
29	14_dispo	24200.871658
30	15_dispo	6300.042885
31	16_dispo	3214.221942
32	17_dispo	8561.269938
33	1_dispo	-3286.256371
34	2_dispo	-12617.028440
35	3_dispo	4647.130208
36	4_dispo	72740.136422
37	5_dispo	3967.800799
38	6_dispo	-1564.886100
39	7_dispo	-42051.045203
40	8_dispo	-6878.272934
41	9_dispo	-32077.059307
42	Kings	-14059.280323
43	Queens	-12411.848902
44	Nassau	2828.275589
45	Bronx	-7949.402402
46	Suffolk	-1179.729853
47	Erie	-22865.774176
48	Westchester	-8574.283696
49	Monroe	-26875.380279
50	Onondaga	-22100.060959
51	Albany	-11136.690905
52	Richmond	-7428.885729

We have training set MSE = 2,085,046,469, test set MSE = 2,101,808,599 and R^2 = 0.564. This is not a satisfying result because of the high test set MSE.

To explore why the linear regression model has such as high MSE, we looked at the real values and prediction values on our test set:



The majority of the predicted charges lie below the $700,000 line, while the number of real charges that exceed $700,000 is 496. If we exclude those 496 numbers exceeding $700,000 and recalculate the test set MSE, we get a value of 1,202,442,540, which is almost half of the previous test set MSE. Combining the fact that our dataset is highly skewed to right, we conclude that values on the right of the distribution impact the prediction accuracy of the model. Realistically, the following situations may be reasons that cause the inaccuracy above:

- A billionaire and a man with low income both get cancer. They would be charged a high amount of money based on our model. However, the poor man may refuse to receive treatment because he cannot afford it, while the billionaire may spend more than what our model predicts since he can afford more.
- The same billionaire and the poor man get cancer. They both find a way to fund their treatments, yet they do not want to pay for them since they both get late-stage cancer. In that case, the real charges would be significantly lower than the predicted charges.

The above situations tell us that our model may have underfitting issue since there are many additional factors that should be considered in our model, such as patients' income level, patients' willingness for treatments and even patients' emotional status. However, it is almost impossible for our model to take all of these direct and indirect factors into consideration. Thus we would mainly focus on fixing the model so that an improved model with higher complexity can deal with the underfitting problem.

We then used Lasso to see if we should delete any predictor variables. We ran cross validations on the value of λ and decided to use $\lambda = 9$ because of its lowest test MSE. The final result is as follows:

	features	estimatedCoefficients
0	Length of Stay	4403.743005
1	F	-0.000000
2	Black/African American	-0.000000
3	Other Race	0.000000
4	139	-0.000000
5	140	-0.000000
6	194	-0.000000

The result from Lasso says that among all the predictor variables, only the predictor variable Length of Stay is considered powerful for prediction. It confirms the idea that our model is not a proper one in terms of the dataset.

V. Future Plan

We then thought about changing our model. The nature of our dataset, with the majority of the data being categorical, is more similar to the format of a decision tree, where each node of the tree corresponds to each level of a categorical variable. Therefore, we plan to fix our model following the steps below:

- The underfitting problem of our model will be fixed by changing the prediction model. Instead of solving a regression problem to predict the exact value of the Total Charges, we will try to transform the problem to a classification problem and predict the level of Total Charges.
- We will separate all the Total Charges into several levels and put them into different bins. For instance, we may say charges within the range $0 - 5000 belong to level 0, $5001 - 20000 belong to level 1, etc.
- We will then use decision tree or random forest to make predictions on the level of Total Charges. By doing that, we are solving a classification problem.
- Instead of using MSE to determine the performance of our model in the previous linear regression problem, we will use classification accuracy table to calculate true positive, true negative, false positive and false negative ratio, evaluating the performance of our classification model.