ORIE4741 Project Final 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 conduct feature engineering. And based on the transformed data, we examined the prediction accuracy of linear models and Random Forest in order to find the best prediction model.

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.

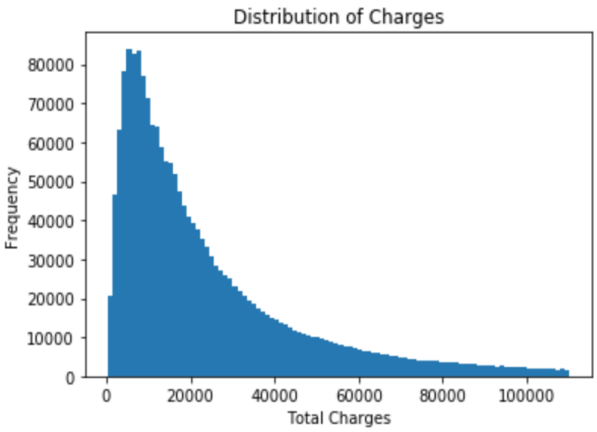
Second, we encoded three important features — Age Group, Type of Admission, and Patient Disposition — and added additional three columns for them, namely Age Group Indicator, Admission Type Indicator, and Patient Disposition Indicator for future studies. For example, the way we encoded the Age Group 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.

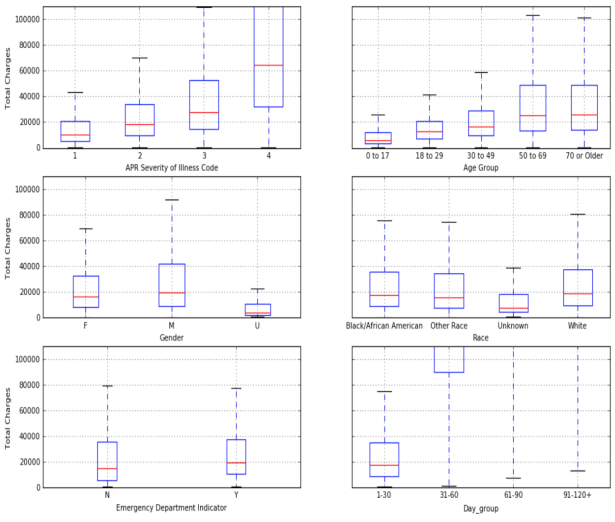


The histogram shows that the Total Charges is significantly skewed to the right as the median is about twice as much as the mean.

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.



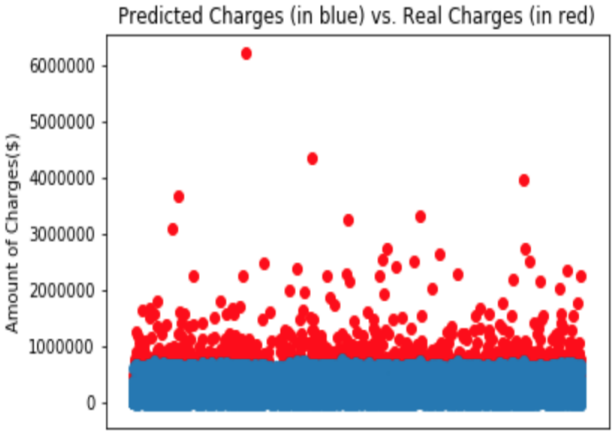
IV. Preliminary Regression Analyses

A. Initial Modeling

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. We had 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:



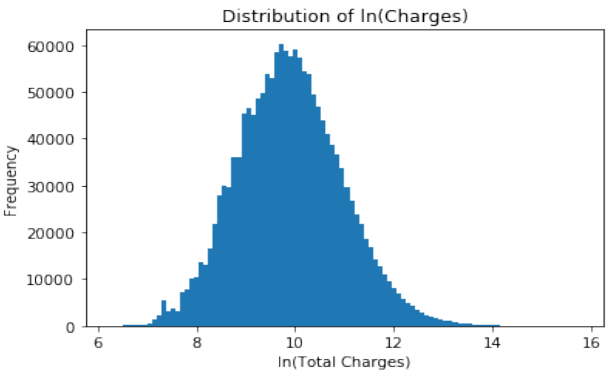
The majority of the predicted charges lie below the $700,000 line, while the number of real charges that exceed $700,000 is 496. If we excluded those 496 numbers exceeding $700,000 and recalculated the test set MSE, we got a value of 1,202,442,540, which was almost half of the previous test set MSE. Combining the fact that our dataset is highly skewed to right, we concluded that values on the right of the distribution impacted 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 could deal with the underfitting problem.

B. Feature Engineering

As illustrated above, the distribution of Total Charges is highly skewed to right, and the extremely high charges within the dataset had significantly negative impact on our model's accuracy. Therefore, we considered using feature engineering to transform the data into a format that our algorithms could work on. The method we took was taking the $\ln()$ of all the Total Charges. And the distribution of $\ln(TotalCharges)$ would be less skewed to right, as shown below:



After the transformation, the distances between the higher charges and lower charges became smaller, which could make our models more robust to those extremely high charges.

C. Model Determination

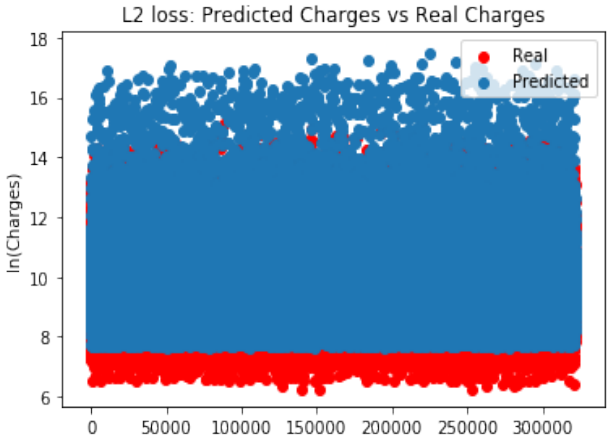
Considering that the predicted values are numeric, we still thought linear regression is still a strong candidate. In addition, we also determined to try decision tree algorithm as most of the variables of our data are categorical. Both models were applied on $\ln(TotalCharges)$ because the logarithmic transformation had already normalized the data.

V. Further Analysis

A. Linear Model

Based on the preliminary analysis result, the high test MSE was the primary challenge we faced. We decided to alter the loss function to compare the model performance of L2 regression, L1 regression, Huber regression, and quantile regression.

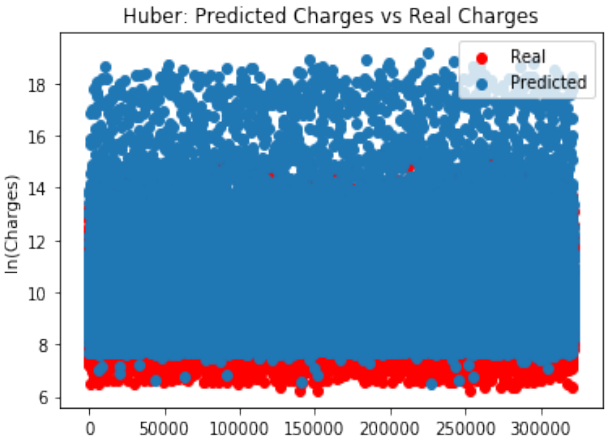
1) L2 Regression: The first model we implemented was a L2 regression model. We chose this model as a first attempt to see how the data fit in our model. The training MSE was 0.398 and the test MSE was 0.399. We also made a scatter plot of predictions and true values:



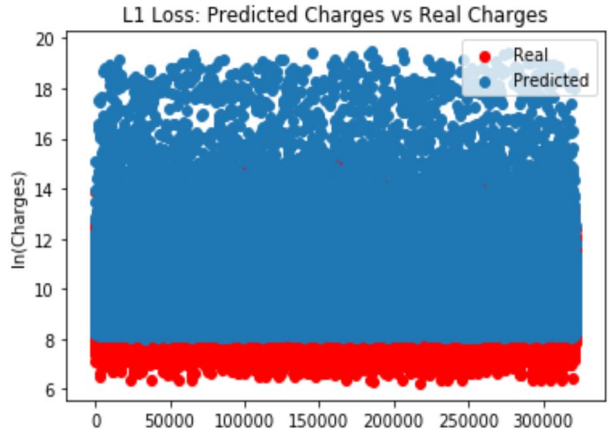
According to the plot, the predictions are shifted up. We believed this is still due to the real data with relatively high values. For instance, as shown in the Feature Engineering section, real data that exceed 14 only take part of approximately 3% of the total data,

while they can leave an impact on our model, forcing the model's predictions to be higher than it should be. Therefore, we planned to eliminate the effect of those extremely high true values in the following models.

2) Huber + L1 Regression: Two helpful ways to deal with extremely high values are to use Huber regression model and L1 regression model. For Huber regression model, the training MSE was 0.418 and the test MSE was 0.420. And the scatter plot of predictions and true values is shown below:



We also tried L1 regression to increase the model robustness with extreme values. The training MSE was 0.500 and the test MSE was 0.502. And the related scatter plot is shown below:



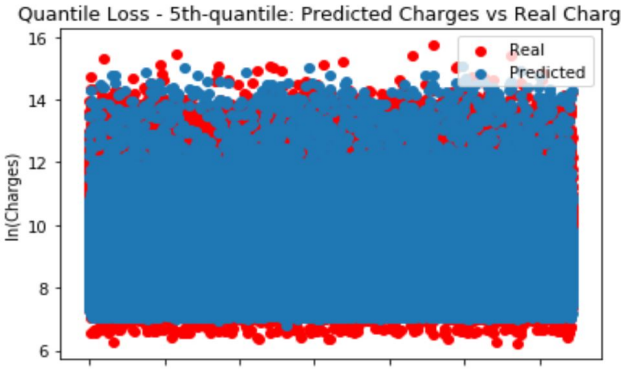
As shown from both plots, the predicted values are again generally shifted up due to the presence of large extreme values. In other words, the model prediction shifted towards the median of the hospital charges, which is bigger than the mean, given that the data are right-skewed. The issue with right-skewed distribution still remained.

3) Quantile Regression: Since all the loss functions we applied above tended to make predictions that are greater than the true data, we considered using quantile regression on the same set of features to moderate the effect of large extreme values. We determined to try quantile loss function also because L1 loss is a special case of quantile loss and we could then study the effect of the model parameters on the model performance.

$$\frac{1}{n}\sum_{i=1}^{n}\alpha(y_i - w^T x_i)_+ + (1-\alpha)(y_i - w^T x_i)_-$$

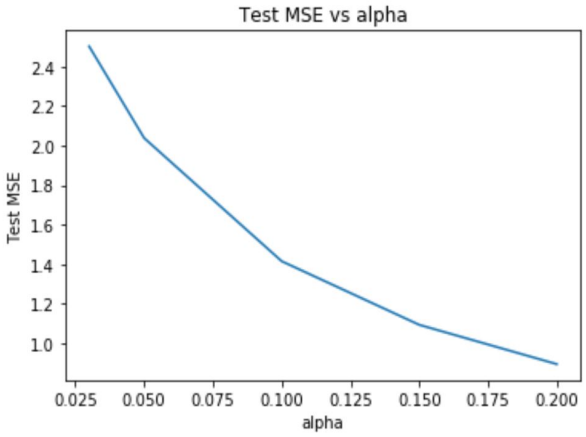
As the formula defined for quantile regression shown above, in order to prevent the model from over-predicting the values, we should decrease the penalty of making predictions that are bigger than the true values; namely, we should less emphasize the positive difference between the predictions and the true values by decreasing the value of α in the loss function. Hence, based on the distribution of the hospital charges, we first tried predicting the 5th quantile of the data by setting α to be 0.05 in the function.

After fitting the model, we again plotted the predictions and the true values together and found out that the overall predictions were indeed shifted down. However, the test MSE, estimated to be 1.85, is much higher than the ones generated from any of the previous models, we argue this was due to the large difference between the predictions and the large extreme values as we allowed the model to tolerate the positive difference more.



To study the effect of quantiles on the model performance, we also carried out a 5-fold cross-

validation with alpha ranging from 0.03 to 0.2. The plot of the estimated test MSE v.s. α values proves our argument that as we increased the value of α, we decreased the model tolerance on the positive difference and thus decreased the test MSE. However, we also noticed that as alpha increased, the model was more likely to overpredict the true values because it tended to fit the large extreme values better.



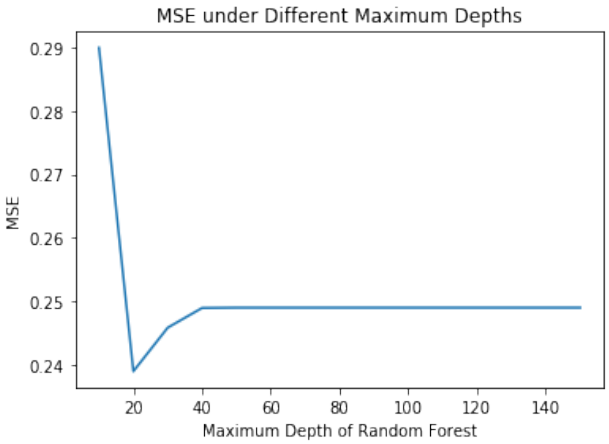
Therefore, there is a tradeoff between the test MSE and prediction range by selecting the value of α of interest. In the context of this project, since we are interested in detecting fraudulent medical bills that overcharge the inpatient care, we suggest using a small value of α to prevent the large extreme values from being over-predicted.

B. Random Forest

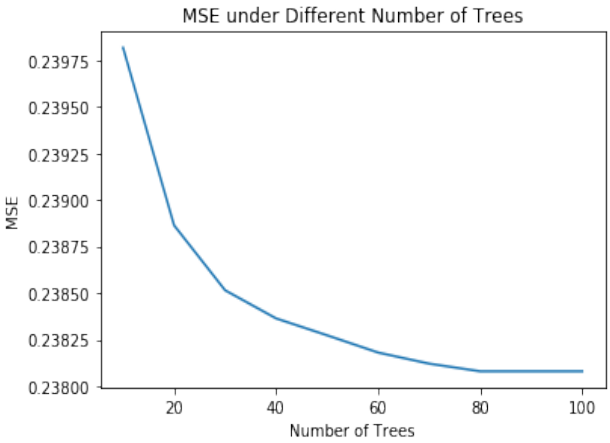
Another model that we implemented to improve the prediction accuracy was a Random Forest model. Our decision was based on the nature of our predictor variables. Out of all the predictor variables, only one of them, Length of Stay, is a numeric variable, while the rest are categorical variables. Therefore, a regression tree may be a great way to solve the problem and the value of categorical variables can be fully captured in this model. To avoid overfitting and prune the regression tree better, we decided to use Random Forest.

To optimize the performance of the Random Forest, we considered two parameters: the maximum depth of the Random Forest and the number of trees. We aimed at finding a maximum depth and a number of trees that minimize the MSE.

By setting number of trees to be 10, we calculated the MSE under different maximum depths and obtained the following graph:

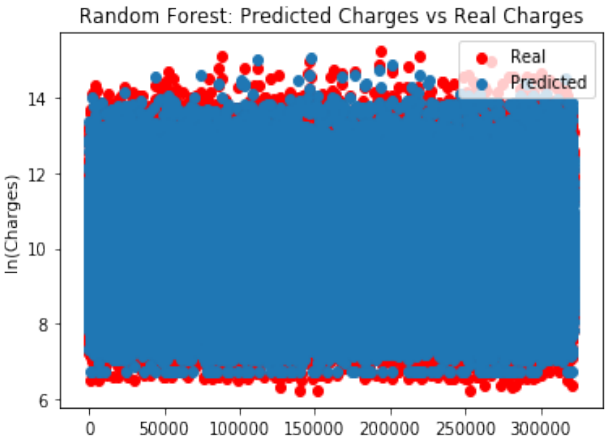


According to the graph above, when the maximum depth of the Random Forest is equal to 20, the algorithm has a lowest MSE of 0.239. To study the optimized number of trees, we picked maximum depth of 20, calculated the MSE under different number of trees and obtained the following graph:



The graph shows that the algorithm converges after the number of trees is 80. And MSE was equal to 0.238 in that case.

Our analysis above shows that the value of Random Forest MSE was significantly smaller than the MSE of all the algorithms we tried in the previous sections. We also looked at the scatter plot of predictions and true values generated by Random Forest when maximum depth equals 20 and number of trees equals 80. The training MSE was 0.213 and the test MSE was 0.238 in this case.



As seen in the graph, the prediction accuracy of Random Forest is the highest among all the algorithms we have implemented. The real data with relatively lower values that are less than 8 are more likely to be predicted, and the predictions of extremely high real values tend to result in less error.

VI. Conclusion

A. Model Comparison

Based on test MSE and prediction range, the Random Forest model performed relatively better than the linear models. Specifically, the tree-based model made better predictions on both low and high charges. However, there is not an absolute advantage of one model over another. If we had assumed there were no frauds in our data set, we think the Random Forest model would be better because it has a higher accuracy. If we assumed there were frauds in the data, we believe the quantile regression would outperform the tree-based model because it makes predictions closer to expectations. Therefore, in the context of detecting fraudulent medical charges, the model selection would be dependent on one's initial assumption on the existence of frauds.

B. Recommendation

Even though the Random Forest model yields the smallest error among all the models we tried, its training MSE and test MSE are both quite high. With a test MSE around 0.24, we calculated that the average ratio between the predicted value and the true value is either 1.88 or 0.53. The deviance between the predictions and the actual data will be magnified especially when the actual data are of large magnitude. Therefore, we suspected that the model is underfit under the existing features. We later replaced the feature 'Hospital County' with the county average income data we obtained from U.S. Census Bureau.

However, the model performance did not improve by much as the model error was approximately the same as before. We believe that in addition to the current features in our data set, there are many other factors that are highly related to hospital charges, such as patients' willingness to accept the treatments, patient's emotional status or maybe patient's education level. Therefore, in order to improve the model accuracy, it would be helpful if more information about the patients could be provided.

VII. Reference

- Hospital Inpatient Discharges (SPARCS De-Identified): 2012 . New York State Department of Health. 2017. https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/u4ud-w55t
- "SELECTED ECONOMIC CHARACTERISTICS 2006-2010 American Community Survey 5-Year Estimates". U.S. Census Bureau. http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_10_5YR_DP03&prodType=table
- "Profile of General Population and Housing Characteristics: 2010 Demographic Profile Data". U.S. Census Bureau. https://web.archive.org/web/20140305164937/http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=DEC_10_DP_DPDP1&prodType=table