STAT 440 Module 1 Report

Rahim Jutha (301342859)
Preeya Mahantheran (301372317)
Sandy Wu (301273729)
Harry Cho (301263022)
Chris Sobczak (301335820)

October 16, 2020

1 Introduction

In this module, our objective was to predict the time between symptom onset and hospitalization *(the duration of the time until hospitalization, in days)* for confirmed hospitalized cases of COVID-19 in January and February, 2020.

The data set provided is split into a training set of 219 patients and a testing set of 200 patients. The training set contains various features such as the age of the patient *(age)*, sex of the patient*(sex)*, city in which the hospital the patient presented resides*(city)*, the province of presentation*(province)*, country of presentation, column relating to exposure*(V1)*, the date at which a positive test for COVID-19 was recorded*(confirmed)*, reported symptoms of the patient*(symptoms)*, indication of death or recovery of the patient*(outcome)* and the target variable, duration.

In order to minimize the RMSE of the model, we tried different techniques such as Multiple Linear Regression, Stepwise Regression, Gradient Boosting, Generalized Linear Model, and an Ensemble Model, Neural Networks and Random Forest. In the end, we concluded that the Random Forest was the best model as it gave us the lowest RMSE.

2 Methodology

This part was broken down into two parts: (1) Exploratory Data Analysis (EDA), (2) Feature Analysis, and (3) Model Building.

2.1 Exploratory data analysis (EDA)

The first thing we did with the datasets were to look at the values in each column and check for any NA or any strange occurrences. With this information we can make decisions on how to clean the data and what approach to take when building our model.

The age column in the train set had 2 NA values as well as values that were in were intervals.

The sex, city, province, country, and V1 columns were anonymized and contained several levels. Since these columns were categorical we thought it would be best to use them as factors. The confirmed column in the train set contained 1 NA value and contained dates from January and February. The symptoms column contained unique text strings describing patient symptoms which. The three major problems with the data was how to decide which columns to use, how to deal with the missing values, and how to clean the symptoms column.

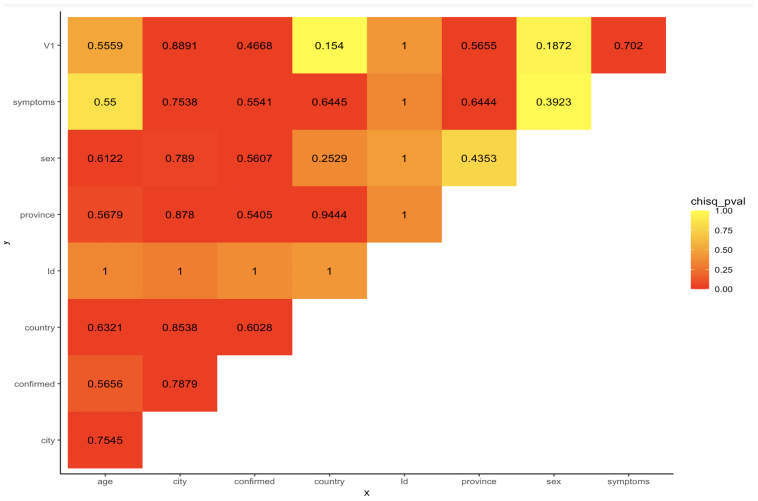

Figure 1: Figure 1: Screeplot

We also found out that there are some features in our data set that are highly correlated with one another using a correlation plot, particularly the variable *city* vs. the other remaining features present in the dataset.

The majority of the data we have is categorical except for the age column which is numerical. Due to this, we thought to try models that are effective for ordinal data.

2.2 Feature Analysis

We cleaned the age column by taking the midpoint of all the interval values and replacing the 2 NA values with the mean of the column. By doing this all the values are numerical and continuous. The factor columns had values that were only present in the training set and some only in the test set. To deal with this problem we made a level called "Other" and assigned those values into this level. The confirmed column was used as a factor column and converted into a date class. Using the date class we took the average date and assigned it to the NA value. For the symptoms column we made 2 binary columns called fever and cough which contained a 1 if a string containing the column name was present and 0 if not.

2.3 Model Building

We started with using a multiple linear regression (MLR) with all the variables and used a stepwise regression model to do our variable selection. This variable selection chose variables cough, confirmed, country, and province. This model performed well, scoring 4.48547, but we wanted to keep trying to get a more predictive model.

Next, we started experimenting with other models using the h20 package in R. Using this package, we tried Random Forest(RF), Gradient Boosting (GBM), Generalized Linear Model, and an Ensemble Model. The GBM and RF models performed the best. Then we tried to create an ensemble using these models, but it resulted in a worse model, so we decided to stick with our Random Forest model.

3 Result

The value of RMSE we have obtained is 4.28257. The value itself is an adequate value in various ways. The Q-Q plot of the studentized residual of the studied model indicates that the residuals tend to spread a

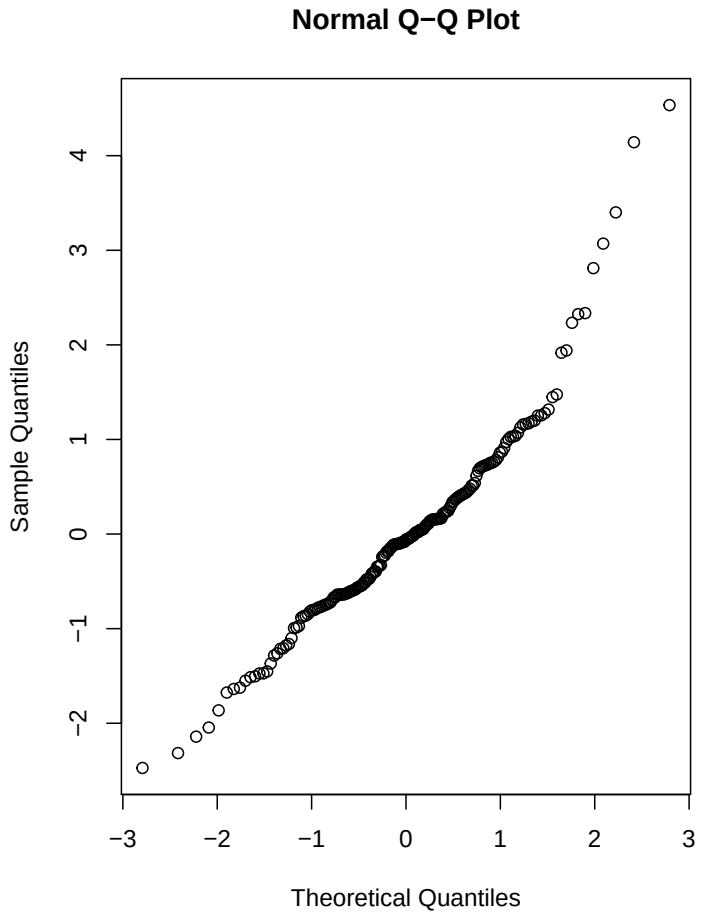

Figure 2: Q-Q plot

little towards on both ends of the plot. However, the implemented method "Random Forests" has no formal distributional assumptions unlike many other statistical methods. This allows us to neglect a very typical strong model assumptions which also allows us to handle any skewed data and categorical data that are ordinal or non-ordinal.

We can also note that the value of coefficient of variation for the model is smaller than 1. As a statistical measure of the dispersion of data points in a data series around the mean, we are able to additionally verify that the model is adequate and that the variation is fairly low and that it is a better risk

return trade off.

Lastly, an imputation with mean values has been performed as part of the study. Substituting mean values over many missing values has reduced our variance and noting the purpose of the study to reduce the RMSE value, the process is reasonable.

4 Conclusion and Future Improvement

As above, we have explained the methods that have led us to obtain a final RMSE of 4.28257 in predicting the of the time until hospitalization *(duration)* from the test set provided using Random Forest.

As for improvements, it would have been beneficial to further clean the observations in the *symptoms* feature. The *symptoms* column contains unordered categorical factors, therefore we could convert it into separate columns with multiple variables, each with a value of 1 or 0. However, due to the time constraints imposed, we were unable to do so. Other than that, we could have also utilized dimension reduction techniques like Principal Component Analysis (PCA) or Partial Least Squares (PLS) as we can see that quite a number of explanatory variables are fairly strongly correlated.

Overall, this module was pretty challenging but interesting to work on given that it is in sync with the current state of the world.