COVID-19 Time to Hospitalization

Team: N/A – Peter Yi / Amos (Wei-Liang) Li / Charlotte(Dai jing) Lin/ Candy (Jia Hui) Xiao

GitHub Links inside documents

Introduction

At start of 2020, novel coronavirus named "Covid-19" has stormed its way across the globe with the ability to do long term health damages to anyone infected by it. An important factor of analyzing the effect of the disease would be the time to hospitalization from the point of time of symptom onset. This is because if there is a large gap between the time of symptom onset and the hospitalization, we would be able to realize which age group or which symptoms would likely to be confirmed cases of Covid19.

This report aims to explore the ability of Random Forest, Ridge Regression and linear regression to predict time to hospitalization. At the end, the three models will be compared, and the better model will be used for submission in which it has been determined to be using linear regression with a score of 4.2568 on the public leader board. Random Forest has scored a best score of 4.3075 and ridge regression has scored a best score of 4.35.

Random Forest Approach

Data Preprocessing

For the dataset, there are 10 variables given in which "Duration" will be our dependent variable, "Age" will be counted as numerical variable and remaining variables will be considered as categorical variables. Additionally, age was further categorized by groups with ranges defined in the research "Redefining meaningful age groups in the context of disease". During the investigation, it was discovered that while some provinces has the same encoding, it has a different encoding value in the corresponding country column. This problem does not just exist in the province column but also the city as well. The cause of such issue could be "due to missing data in this column" as Dr.Elliott described in his email(*1). For the variables "confirmed" date, we have decided to investigate into the total case of confirmed Covid19 cases and realized that February 10th to February 13th was a spiking point in which we decided to group the "Confirmed" date variable into 3 categories. First category will be before Feb 10th, second category will be in between Feb 10th and February 13th and third category would be anything after February 13th. The original value of "Confirmed" variable are also converted into timestamp and scaled. From the scatterplot of sex and outcome, we did not see any predictability possible from these two variables thus it was discarded. Since the "Symptoms" variable consists of strings, it was parsed and dummy variables "throat", "fever", "cough", and "diarrhea" are derived. Additionally, the number of symptoms from each patient encountered are also tracked into another variable. Another indicator variable indicating whether the patient got other symptoms than fever and cough were created as well.

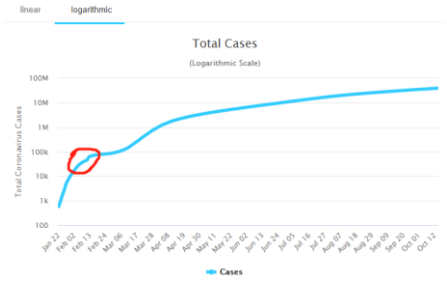
Filter

	age	sex	city	province	country
1	68	8a467	ba1b5	89dd9	59dcd
2	25	8a467	44886	fe869	59dcd
3	73	8a467	b4ff4	38fc4	16725
4	20	d516d	07f3f	8ac5e	59dcd
5	50-59	8a467	38fc4	38fc4	c263d
6	77	8a467	38fc4	d7cac	59dcd

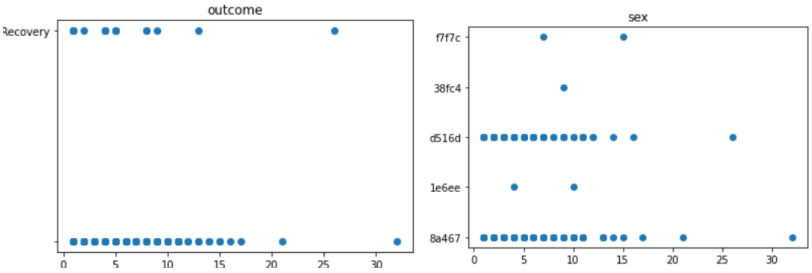
38fc4	f163b	10
38fc4	72c53	9
38fc4	ca1bd	9

Figure showing duplicated province but different country

*1) "Thanks for your email - please refer to lecture for explanation (likely due to missing data in province column, but I won't investigate). Congrats on 5 bonus points for your team."



Graph displaying the sudden increase in Covid19 cases



Methodology

Initially, Grid search is applied over all possible parameters for selected variables from preprocessing. It reached a score of 4.4 Kaggle. Through running feature importance, we discovered that level encoding of city, province and country did not contribute to the model and "fever" and "cough" are even decreasing the performance. We decided to switch to one hard encoding for the geographic variable and discard the "fever" and "cough" variable. Variable with feature importance less than 0.0001 are also discarded and investigated. The results produced from the combined 3 steps achieved a score of 4.30750 on Kaggle.

```
R^2 Training Score: 0.25
R^2 Validation Score: 0.11
```

```
R^2 Training Score: 0.30
R^2 Validation Score: 0.12
```

```
R^2 Training Score: 0.35
R^2 Validation Score: 0.13
```

First pictures show initial model, second shows effect of one hot encoding and third is removing more unimportant variables

kaggle_submit_rf.txt 4.30750

Ridge Regression

Data Preprocessing

Same preprocessing steps were used like in the Rand Forest Approach.

Methodology

Repetitive trials and errors were use for feature selection. Then Grid search was applied on the alpha parameter, and then the MSE of the results will be compared to our best submission. Upon selecting the variable that indicates the non major symptoms (Fever, Cough), confirmed, all geographic variable, V1, Count of symptoms, throat, diarrhea, age and age group, the best model from this result has achieved a score of 4.35 on Kaggle

kaggle_submit_lr.txt 4.35228

Code to above approaches can be found at https://github.com/Rai-2018/Stat440-module1

*1) "Thanks for your email - please refer to lecture for explanation (likely due to missing data in province column, but I won't investigate). Congrats on 5 bonus points for your team."

Linear Regression

Data Preprocessing

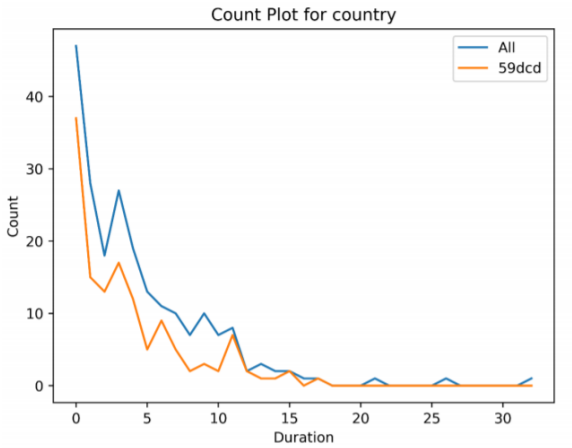
The variable age undergoes the same preprocessing strategy as in the Random Forest Approach in which means are filled to NA. Two dummy variables indicating the status of "throat" and "diarrhea" were derived from the column "Symptoms" based of fuzzy string matching. The count of symptoms for each individual are also being counted and stored into another variable. The variable "Confirmed" is also converted into timestamp and being scaled. All other string variables are treated as categorical and has dummy variable for each level

Methodology

The whole model was fed into linear regression model with the training data and it reached a score of 4.4 on Kaggle. Then we decide to specifically look at some variable in depth and plotted the frequency for each country in each 25% of data after sorting the column "duration".

Country 0-25%	frequency	Country 25-50%	frequency	Country 50-75%	frequency	Country 75-100%	frequency
59dcd	40	59dcd	34	59dcd	35	59dcd	25
fb9d7	3	c263d	7	fb9d7	8	fb9d7	12
c263d	2	fb9d7	7	c263d	6	c263d	9

Correlation of the country "59dcd" with duration reached a score of 0.97 as well



Thus, we decided to discard every other dummy variable and fit a linear regression and the RMSE decreased. Similar approach was then done for dummy variables of symptoms, province, city, V1 and country. After series of trials and errors, we configured the model of duration ~ confirmed + age + # of symptoms + symptom-throat + symptom-diarrhea + province-38fc4 + province-55fe6 + country-59dcd + city-0bd76 + V1-dd554 + V1-9a45a + V1-f9037.

Train/ Valid – Below shows the score and RMSE for retraining of the same model to confirm consistency

Score: 0.2150316427390516 0.17564627732456306
RMSE: 3.9921055173725764 4.779226107682275

Score: 0.20225003064418026 0.21460885917494765
RMSE: 4.149551278259917 4.418058582913841

*1) "Thanks for your email - please refer to lecture for explanation (likely due to missing data in province column, but I won't investigate). Congrats on 5 bonus points for your team."

Score: 0.21761513786134556 0.1651598480901143
RMSE: 4.27706424960816 4.0333521769489975

The code to approach above is at https://github.com/XYYaaa/COVID19_Pre/

Conclusion

Overall, we learnt that the dataset contains a lot of noise and may not be suitable for random forest. Additionally, random forest relies a lot on its hyperparameter and it is easily overfitted. Due to the small volume of data, random forest may also produce different results every time meaning it does not have very much stability. The effect of bootstrap in Random Forest can also effectively increase the accuracy of the prediction in Random Forest approach as submission with bootstrap parameter turned off only scores a 4.5 in Kaggle. Another observation is that while there is some predictability in Random Forest, it is hard to describe the effect from each of its parameters and selected variable.

As of linear regression, it is only a matter of effectively removing the variables such that it produces a good result. However, linear regression did provide lots of stability in the model and it is rather easy for comparison.

Attempt in Additional Bonus:

During class, Dr.Elliott mentioned that there is an additional contamination in the testing dataset in which we believed is the missing outcome variable in the test set. This is likely due to an error in the code.

*1) "Thanks for your email - please refer to lecture for explanation (likely due to missing data in province column, but I won't investigate). Congrats on 5 bonus points for your team."