**factors affecting covid-19 in South Korea and forecast how covid-19 spread in South Korea in future**

**Introduction**

Covid-19 has affected the world a lot. In this project, we will discuss how covid-19 affect South Korea first and then we will apply different statistical learning methods to forecast how covid-19 spread in the future.

**Factor analysis  
*Relationship between different factors and confirmed case***

Apart from the autocorrelation from the dataset, it would be much interesting to talk about other factors that will affect the confirmed cases.

*For all figures, please refer to appendix. Press Ctrl + right click will direct you to the location.*

[***Figure 1.1***](#f1) shows the correlation between the keywords search in Korea’s most popular search engine, government policy and confirmed case.

Surprisingly, the correlation between government policy and confirmed is less 0.1, which shows that there is almost no relationship of what government has done and the number of confirmed cases. [***Figure 1.2***](#f2) shows the date government applied their policy and the confirmed case.

However, there is actually a weak correlation between confirmed case and people searching keyword for covid-19, the correlation is 0.35. [***Figure 1.3***](#f3) shows the relationship between people searching coronavirus and confirmed case, we can see the trend of people searching for keyword coronavirus and confirmed case are pretty similar after mid-February.

***Relationship between age group, gender, confirmed case and death case***

It is also interesting to condition on age group to see whether it affect the confirmed cases and mortality rate.

Since the correlations between confirmed cases of age group are over 0.95, we should plot the confirmed case against time to figure out if there is any further finding. [Figure 2.1](#f4) shows that all age group has the same trend, although we should note that 20s are more likely to be confirmed they have coronavirus.

Gender is also one possible factor that affect the confirmed case. [Figure 2.2](#f5) shows that female has more confirmed cases in the first peak of explosion, but after the peak female has the similar number of confirmed cases against male. This may cause by a single event, given that we know the first peak was due to old women who are confirmed infected by coronavirus went to church.

A much higher death rate from coronavirus compare to flu is one reason that draw the attention from world. With the similar approach to confirmed case, we will first look at the number of deaths in each age group, then look at the number of deaths against gender.

[Figure 2.3](#f6) shows that age group 50s, 60s, 70s, 80s dominate death cases. Which suggest that older people may have a higher mortality rate. For a clearer picture, please refer to [Figure 2.4](#f7).

After looking at age group, it is time to investigate on number of deaths against gender. [Figure 2.5](#f8) suggests there is not much difference between death case and gender. Thus, no further investigation is needed. We may consider there is no effect on gender against mortality rate.

***Relationship between province and confirmed case***

Condition on province, we can observe the fact that there are some providences dominate the confirmed cases. [Figure 3.1](#f9) may provide us a picture of how severe coronavirus are in those particular providence.

Since the dataset is too unbalance, we scaled it by and compare it again. [Figure 3.2](#f10) provides us the cumulative distribution of each province. As we can see, some provinces are increasing like a log function, whilst the other province are increasing like a combination of log functions, or we called it a ‘second peak’.

However, it may be too much noise to analyze provinces with only a few confirmed cases, it would be better to focus on provinces with most confirmed cases. Thus, four provinces with most confirmed case will be selected. (Daegu, Gyeonggi, Gyeongsangbuk, Seoul)

We plot the graph first and [Figure 3.3](#f11) gives us a clearer look for our further analysis. As we can see, a dominating effect from Daegu sill exist even though we are comparing with the most severe provinces. However, we can still observe that Daegu and Gyeonggi does not have a ‘second peak’, whilst Gyeonsanbuk and Seoul have a ‘second peak’. [Figure 3.4](#f12) is a reference for us if a comparison of distribution is needed.

We can also look at the correlation of the confirmed cases between different province. In [Figure 3.5](#f13), we can see there are strong correlation among most of the countries. However, in some provinces, the correlations are lower than 0.8 or even 0.7, which is not very usual. This may cause by the policies done by the local government, or the density within the province is not as high as another province, or the residence in those provinces has a higher awareness of the virus.

***Autocorrelation within the dataset***

When dealing with time series type of problem, ARIMA is obviously one of the best approaches.

Identifying if there is any seasonal pattern is our initial approach. Whereas *decompose()* function in R and by observation told us there is no seasonal pattern. For error message, please refer to [Figure 4.1.](#f14)

We then look at ACF ([Figure 4.2](#f15)) and PACF ([Figure 4.3](#f16)). ACF and PACF suggest us that is a AR(1) model. After fitting the data into the AR(1), we perform a hybrid box test on the residual from the model ,which is a combination of Ljung-box test and Box-Pierce test. The result is represented in [Figure 4.4](#f17).

As a result, we can conclude that the residual is white noise, the model fits the data well. Thus, it suggests us we can use other models which can make use of time to predict the spread of covid-19.

Summary

Forecasting

Data cleansing

From the above analysis, we figured out which factors are useful and which doesn’t tell us enough information. The idea of the dataset is to use the information a day before to predict the next day result. For the data description, please refer to table 1.1.

Data splitting

We will first train a linear regression model to see if shuffling the data will benefit the result of forecasting, due to the observation of the imbalance and limited data. If we don’t shuffle the data, the RMSE is 40.23015. Whereas shuffling the data gives us RMSE of 57.03669. This may due to model overfits the data, and in this project, we will *not* shuffle the data.

Model selection

Random forest

Although linear regression already provides us a decent result, we will always seek for a better result. Random forest is always one of the best candidates to start with if we are going for a more advance and complicate model.

Let start with ‘What is random forest’, random forest is basically a brunch of individual decision trees and work as an ensemble algorithm. 1) It selects n samples from the training dataset, and then 2) train a decision tree, 3) repeat 1-2 for k times, 4) voting from k trees to get the optimal result. Please refer to [figure 5.4](#rfpic) for a more intuitive explanation.

Advantages

One reason of using random forest is it do not require so much tuning to provide a decent result. Another reason is it usually provide a internal validation set, which means more data can be fit in to the model. Furthermore, it can deal with both continuous data and categorical data, thus no pre-processing are needed. Last but not least, it excels at dealing with outlier.

Disadvantages

One obvious disadvantage from random forest is it requires quite a lot of computation power, which means it takes a long time to deal with a large dataset. Although random forest already performs so well, there are still models perform much better than random forest. We will discuss it later after this chapter.

Parameter tuning

We first train a random forest without any tuning, and the RMSE is 16.40618, which is already a huge gain compared to linear regression.

From [Figure 5.1](#f18) we can observe that the error converges when number of trees equals to 100. Thus, the remaining parameters which influence the model most is number of variables available for splitting at each tree node.

[Figure 5.2](#f19) suggest us when the number of variables available for splitting at each tree node equals to 7, the Out-of-bag (OOB) error is the least. Thus, a new random forest is trained and the RMSE is 15.73418, a slight improvement compares to the original model.

Result

For the result of the true result and the prediction result, please refer to [Figure 5.3](#f20).

As we can see, the model basically did well. However, we can see the is a time lag between the prediction result and the true result. That may cause by some factors in the dataset which we could not figure it out easily.

Extreme Gradient Boosting (XGboost)

As mentioned before, there are some models perform even better than random forest. Whereas XGboost is one of them. In short, XGboost is a faster version of Gradient Boosting Machine (GBM).

Similar to random forest, XGboost is a combination of decision trees. However, those decision trees are corelated. 1) It trains a decision tree base on the training dataset. 2) Train a new decision tree base on the residual from the previous tree. 3) Repeat 1-2 for n times and give weights to the current tree. 4) Base on the weight to vote for the result. Please refer to [Figure 6.1](#xgpic) for a more intuitive explanation.

Advantage

One advantage of XGboost is unlike other GBM, it supports parallel computing. Thus, it reduces the time needed for training. Furthermore, a normal GBM only utilize the first term of Taylor series, XGboost utilize the second term of Taylor series, resulting a higher accuracy. In addition, it can deal with missing value, thus we don’t need to fill in the missing values by ourself.

Disadvantage

One disadvantage of XGboost is it cannot deal with category data directly. However, we are happy to know that the dataset doesn’t contain any categorical data. Furthermore, the algorithm requires memory a lot, so it demand the computer specification quite a lot.

Appendix

[**Relationship between different factors and confirmed case**](#correlation)

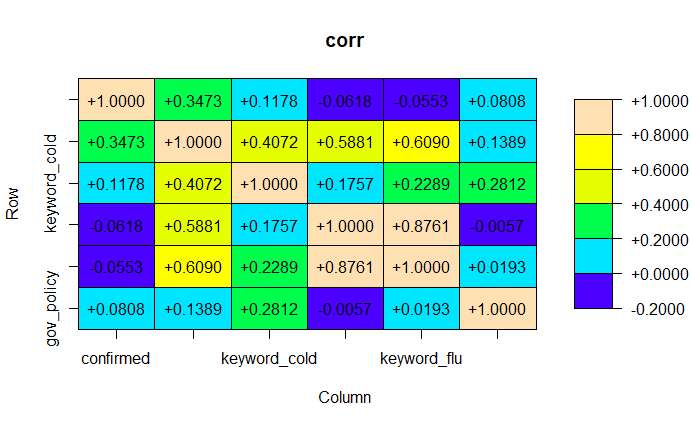
[](#relationfirmed)

Figure .1

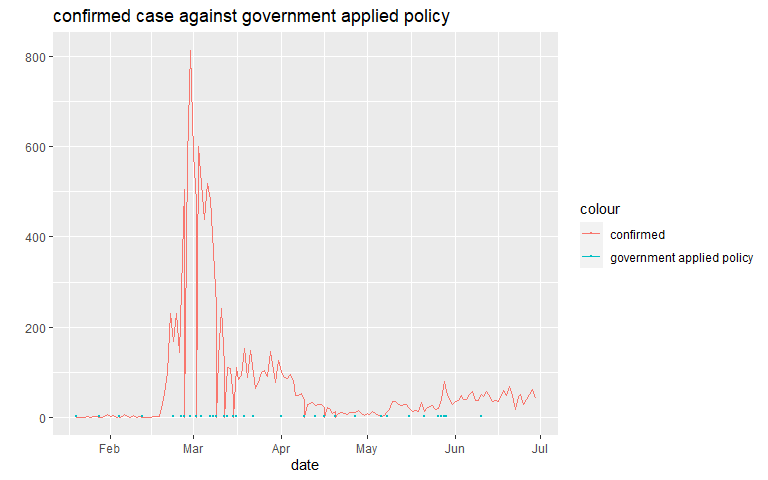
[](#relationfirmed)

Figure 1.

[**Relationship between different factors and confirmed case**](#correlation)

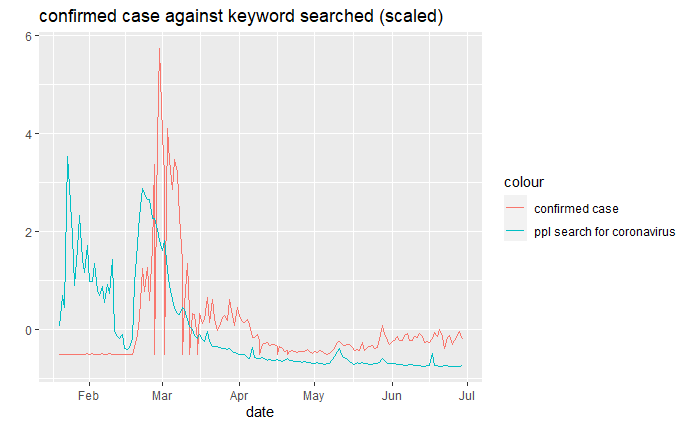


Figure 1.3

[***Relationship between age group, gender, confirmed case and death case***](#rela_age)

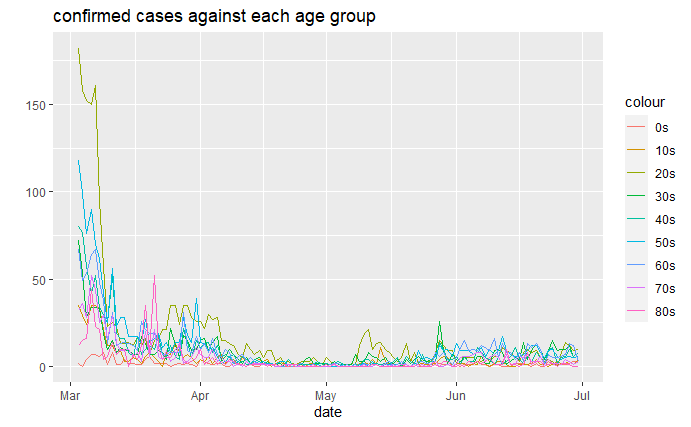


Figure 2.1

[***Relationship between age group, gender, confirmed case and death case***](#rela_age)

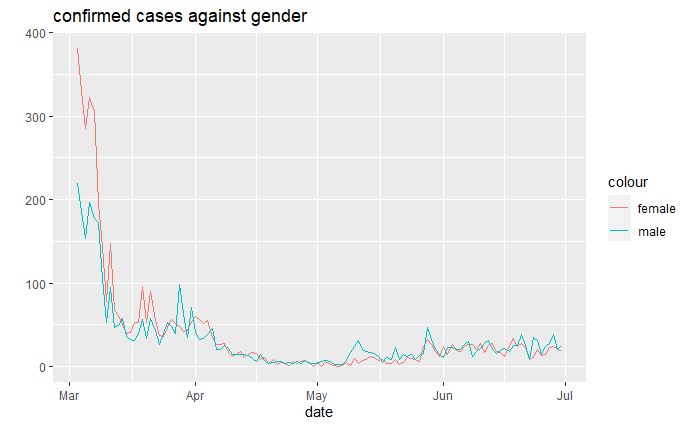


Figure 2.2

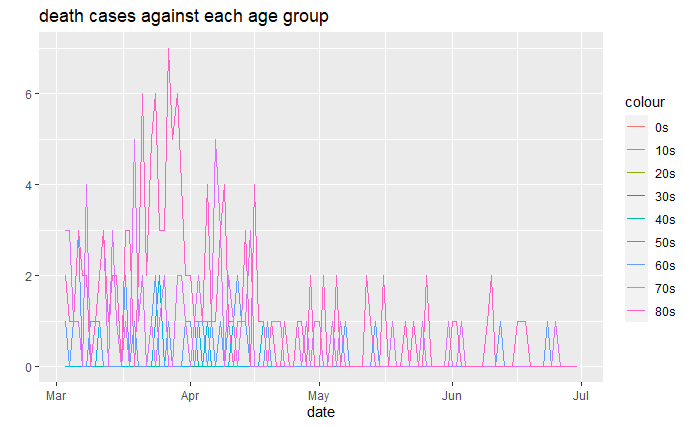


Figure 2.3

[***Relationship between age group, gender, confirmed case and death case***](#rela_age)

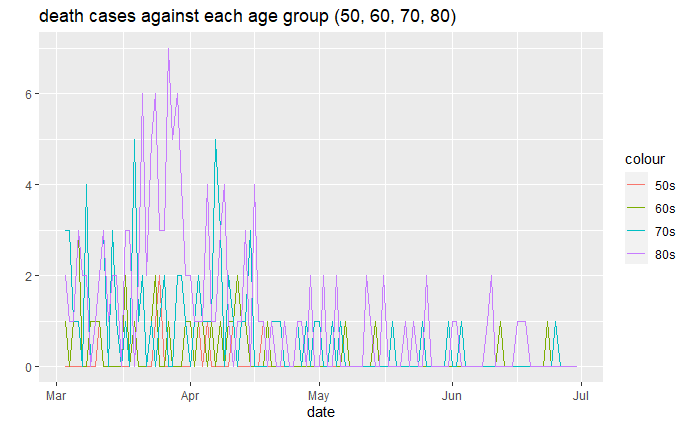


Figure 2.4

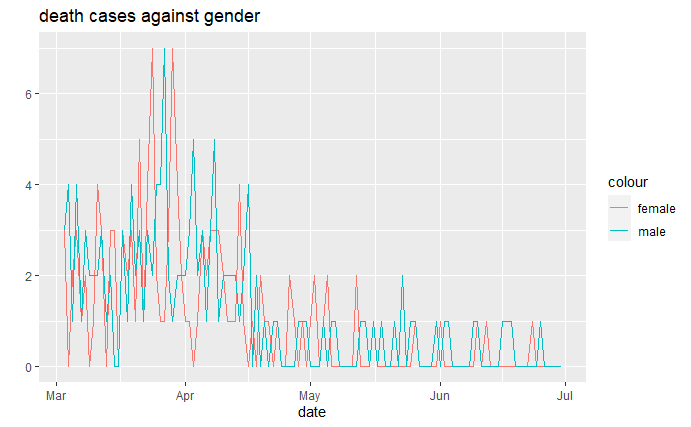


Figure 2.5

[***Relationship between province and confirmed case***](#prov)

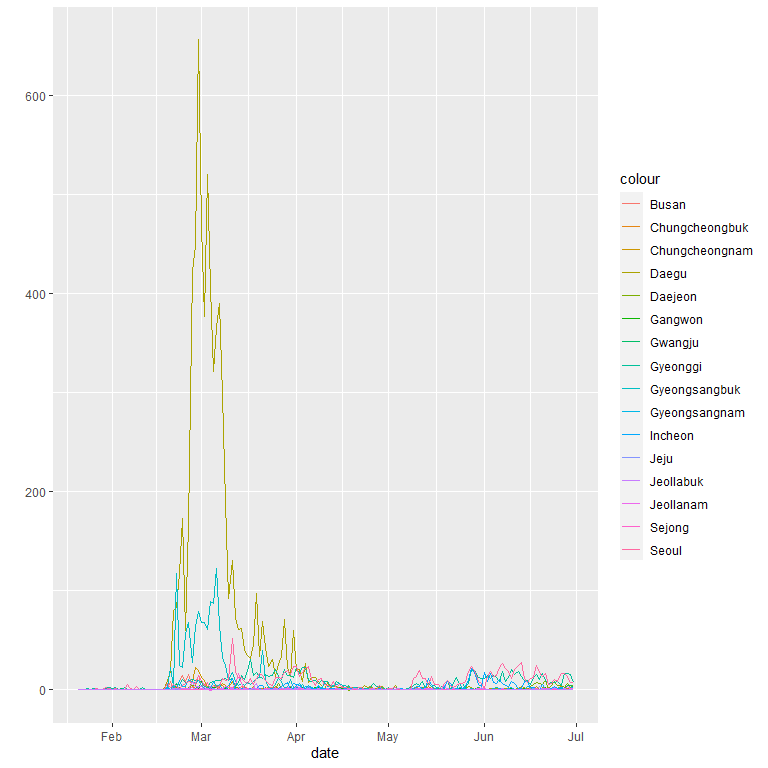


Figure .1

[***Relationship between province and confirmed case***](#prov)

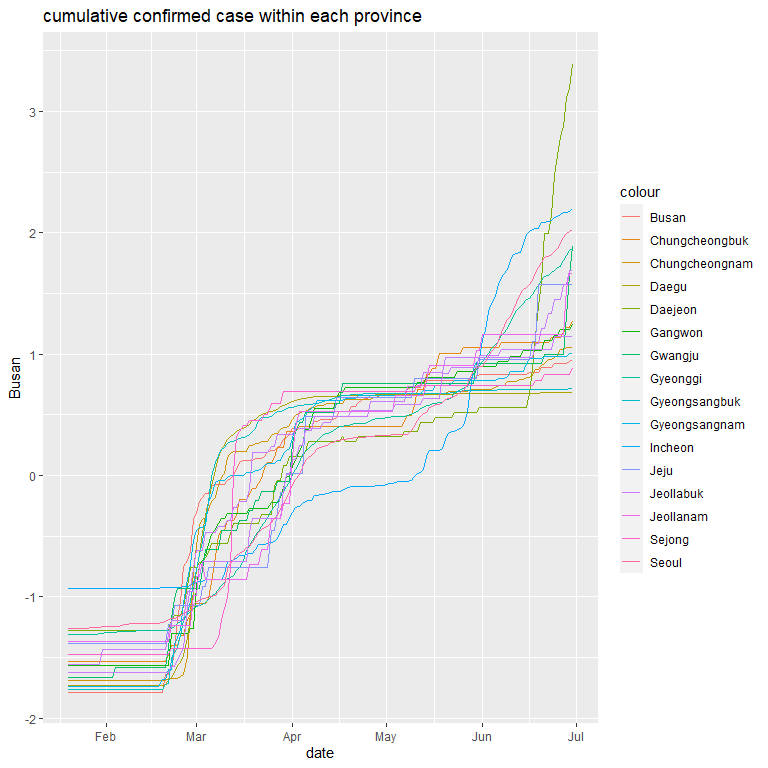


Figure 3.2

[***Relationship between province and confirmed case***](#prov)

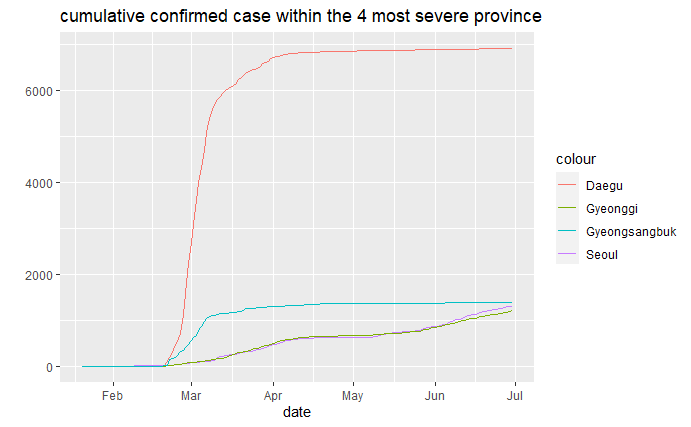


Figure 3.3

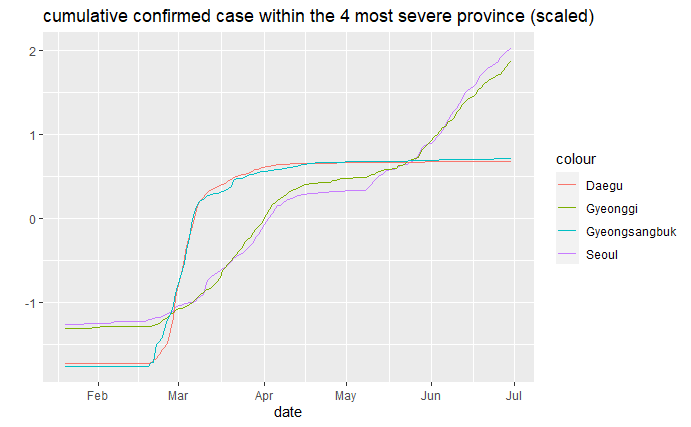


Figure 3.4

[***Relationship between province and confirmed case***](#prov)

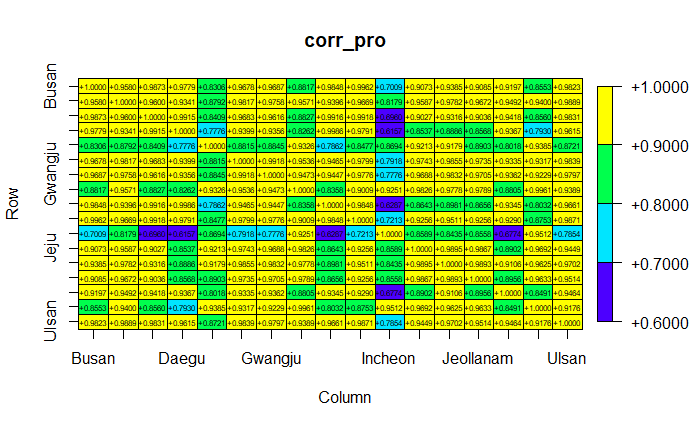
******

Figure 3.5

[***Autocorrelation within the dataset***](#autocorr)



Figure .1

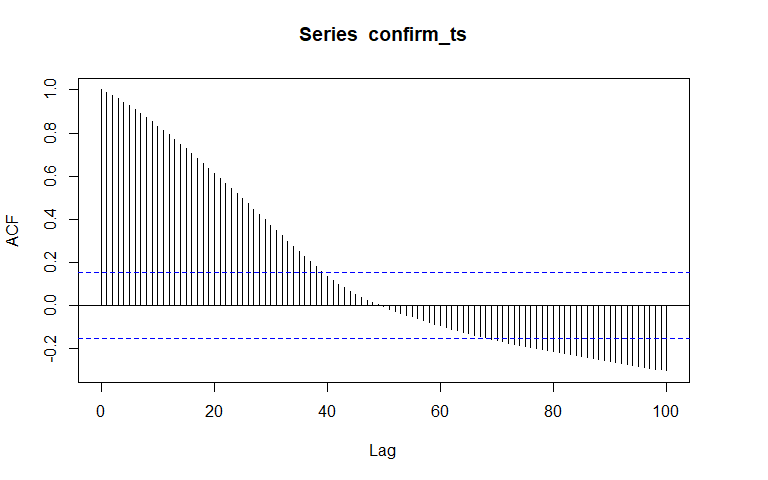


Figure 4.2

[***Autocorrelation within the dataset***](#autocorr)

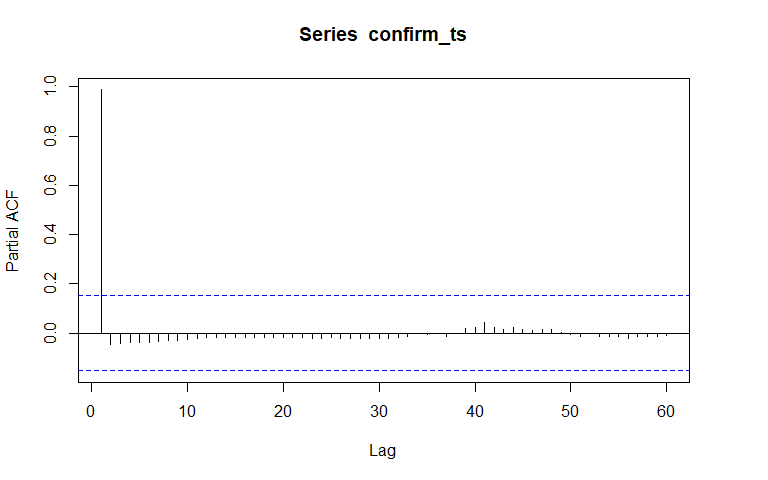


Figure 4.3

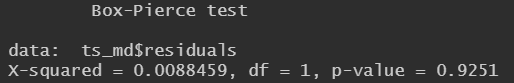


Figure 4.4

Table

[Parameter tuning (Random Forest)](#rfparam)

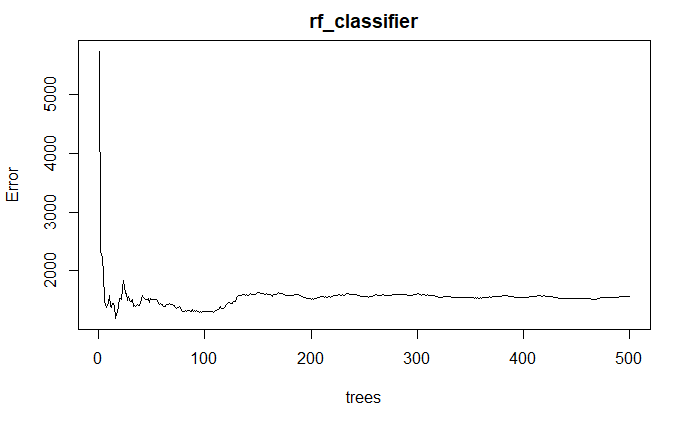


Figure .1

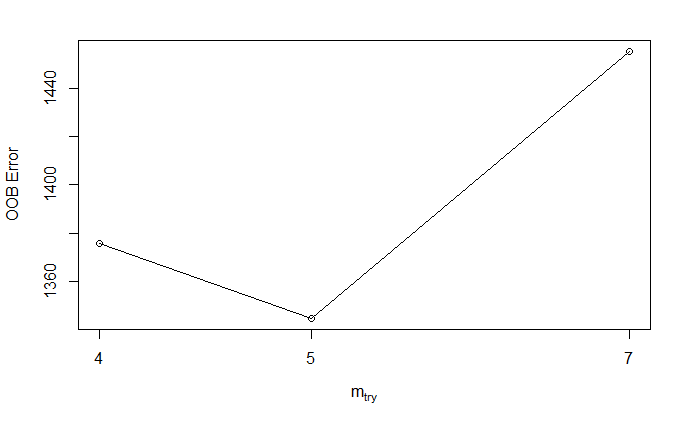


Figure 5.2

[Parameter tuning (Random Forest)](#rfparam)

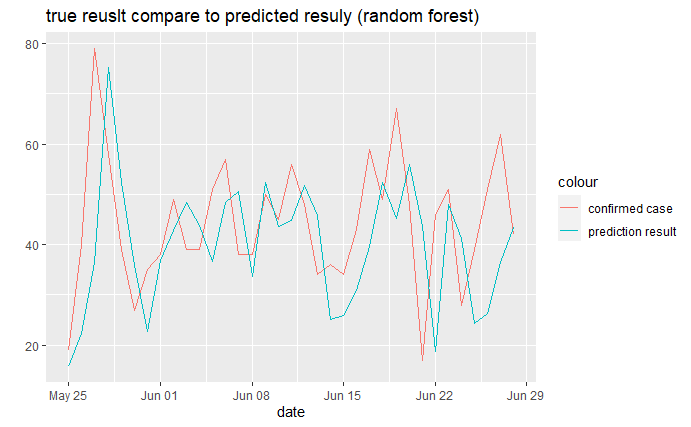


Figure 5.3

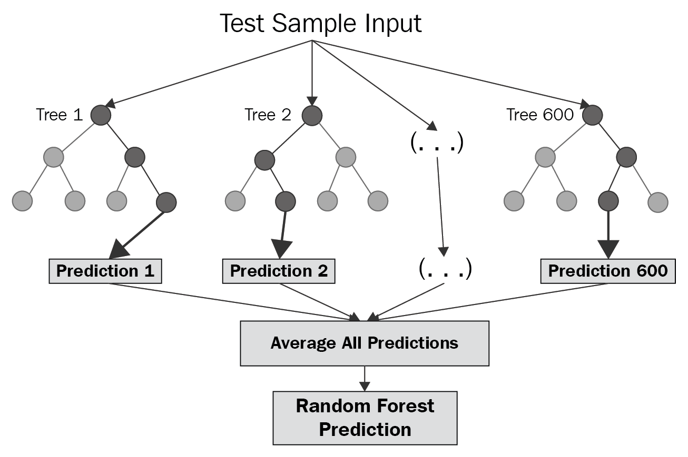


Figure 5.4

[Extreme Gradient Boosting (XGboost)](#xgboost)

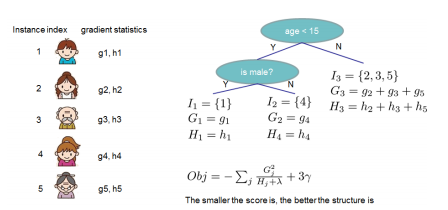


Figure .1

Reference

<https://corporatefinanceinstitute.com/resources/knowledge/other/random-forest/>

https://www.kdd.org/kdd2016/papers/files/rfp0697-chenAemb.pdf