INTRODUCTION

In this project, have a dataset on COVID-19, which contains information on number of Deaths, Recovered, Confirmed and Active cases for dates starting from 23rd January 2020 to 16th March 2020. This data includes the information for 156 countries, which are further divided into provinces with their latitude and longitude mentioned. We have used Tableau to visualize the major states affected in United States and have also created a dashboard to visualize the trends in different covid cases and forecasted the trend for the dates from 17th March 2020 to 28th March 2020 of deaths happening across few predominant countries under limelight. Further we have used Logistic Model and Random Forests in R to predict the number of deaths. We have used k-fold validation, Forward Selection and Backward Selection during the process of analysis.

ANALYSIS IN R

Cleaning dataset and creating new variables

We have changed the date column into Date format. We have checked for outliers and correlation plot of the dataset. We did not find any significant positive and negative correlation within the variables. During the process of cleaning the dataset, we have created dummy variables for the column “Case_Type”. There are 4 new columns created for dummy variables, namely, “Active”, “Confirmed”, “Deaths” and “Recovered”. We have also created 2 subsets, “Death_rate” for the list of cases which ended up in death and another one “Survival_rate” for cases which were recovered. We have also created a Month variable from Dates.

Logistic Model:

Using Logistic Model, we ran the model to predict deaths with factors like have Province, Cases and Month. The idea here is to predict the province that is having more significance with the Death cases. Initially we modelled using Country and could not find any significance variables because the data gets diluted at a country level. Using glm function we build a model for the training

dataset. It is based on survival and death created from dummy variables of Case Type variable. Here we achieved 51.2% accuracy. After building the model using the Forward selection, we used it for prediction by applying it on testing dataset. For the built model we computed the optimal cut Off (0.4199), calculated the misclassification error (0.4886), plot the ROC graph to verify the area under the curve, which was 0.50, computed confusion matric and obtained the accuracy of 51.13%. The significant variables obtained were from month 3 (*March*) and few of the provinces from China province (*Hubei, Zheijiang, Henan, Guangdong*). Cases variable was also significant.

We also used Backward selection approach to see if we find better accuracy or get any new significant variables. Here the most significant variables were similar to Forward approach, yet this model had better accuracy. We now used the new model for prediction by applying it on testing dataset. For the built model we computed the optimal cut Off (0.4399), calculated the misclassification error (0.4873), plot the ROC graph to verify the area under the curve, which was 0.509, computed confusion matric and obtained the accuracy of 53.26%. (Appendix: LOGISTIC MODEL – ***BACKWARD*** SELECTION). Further we used Kfolds validation with 10 iterations to cross validate the performance of the model. This also showed similar accuracy.

Random Forest Model

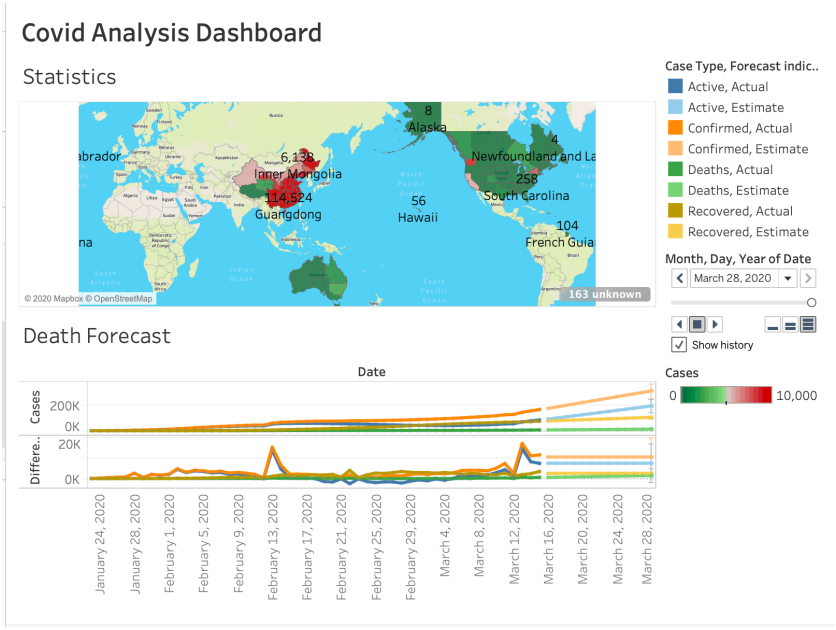
We also created the random forest model for comparison. We get the original accuracy of random forest is 53%, which is the highest accuracy among all three models. Here, we fine-tuned the algorithm using tuneRF function and from the graph (Appendix: RANDOM FOREST MODEL). we find the mtry value drops until 4 and starts to increase. So, using mtry value as 4 we ran the model with target as ‘Deaths’ with all other variables. Also, from the order of importance we found the Country, Province and cases listed at top ones with Gini index measurement. These variables are the ones clearly impacting the deaths of the COVID-19 cases.

COVID -19 Data Analysis

The prediction of patients recovering or dying from all the three models output is tabulated (Appendix: Recover or die?)

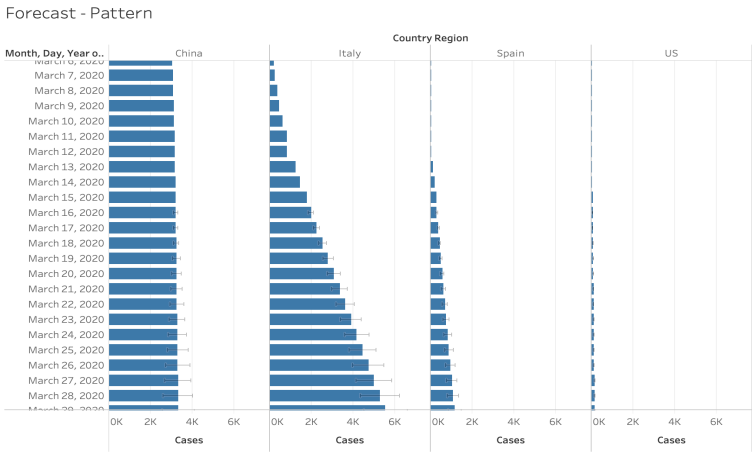
ANALYSIS IN TABLEAU

We created dashboard based on the existing data set to analysis COVID 19 cases situation.



Date

Case Type	Country Region	March 17, 2020	March 18, 2020	March 19, 2020	March 20, 2020	March 21, 2020	March 22, 2020	March 23, 2020	March 24, 2020	March 25, 2020	March 26, 2020	March 27, 2020	March 28, 2020	March 29, 2020	March 30, 2020
Deaths	China	3,225	3,235	3,245	3,255	3,264	3,274	3,284	3,294	3,304	3,313	3,323	3,333	3,343	3,353
	France	123	137	150	163	176	189	202	216	229	242	255	268	281	295
	India	3	3	4	4	4	5	5	5	6	6	7	7	7	8
	Italy	2,260	2,535	2,810	3,085	3,360	3,635	3,910	4,185	4,460	4,735	5,010	5,285	5,560	5,835
	Spain	377	442	507	572	637	701	766	831	896	961	1,026	1,091	1,156	1,221
	US	76	84	91	99	106	114	121	129	136	144	151	159	166	174



From the graph, we can see China (Red) has the most cases among the world. And in USA, we can see that Washington state, California state and New York state have the most serious status for COVID 19.

Further, we wanted to have exact number of deaths predicted against top countries. From the table, we find China is attaining a saturation whereas the cases in Italy and Spain are rapidly increasing in numbers. At this point, US has not attained a state of critical hotspot. More visual analysis using Map and Forecast is provided in the (Appendix: Analysis using Tableau)

There are two turning points appeared at February 13th and March 12th. During February 13th, the recovered rate is higher than that of deaths and from March 12th, the whole situation reverses. The global evolution of death versus recovery is also added in the appendix.

CONCLUSION

We have achieved an accuracy of 51.26% in Logistic Model and 53% in Random Forests model. We feel this because, the other factors available in the dataset have a very low correlation with the number of deaths. However, from the significant variable on March month having greater p value, we feel March could potentially be the crucial month which is affecting the COVID-19 deaths. This means appropriate social distancing, travel restrictions and awareness to use masks and gloves needs to be spread across the globe should be made in practice to save us from the pandemic effects before a vaccine is in the market.

REFERENCES

[1] COVID-19 Map. (n.d.). Retrieved from https://coronavirus.jhu.edu/map.html

[2] Tahseenahmad. (2020, April 20). COVID-19 Exploratory Data Analysis - RandomForest. Retrieved from https://www.kaggle.com/tahseenahmad/covid-19-exploratory-data-analysis-randomforest

[3] Tableau Tips and Tricks: Tableau Jedi Tricks. (2019, May 22). Retrieved from https://www.edureka.co/blog/tableau-tips-and-tricks/

APPENDIX

ANALYSIS USING R

LOGISTIC MODEL - FORWARD SELECTION:

```
> summary(model)

Call:
glm(formula = Deaths ~ Province_State + Cases + Month, family = binomial(link = "logit"),
    data = trainingData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.4254  -1.1770   0.0052   1.1758   4.2416

Coefficients:
                                         Estimate Std. Error z value Pr(>|z|)
(Intercept)                            -0.0880450  0.2338203  -0.377  0.70651
Province_StateAlaska                    0.0290949  0.3300252   0.088  0.92975
Province_StateAlberta                   0.4361790  0.3319211   1.314  0.18881
Province_StateAnhui                     0.7471181  0.3548001   2.106  0.03523 *
Province_StateArizona                  -0.0219802  0.2584198  -0.085  0.93222
Province_StateArkansas                  0.0548593  0.3267158   0.168  0.86665
Province_StateAustralian Capital Territory  0.0053644  0.3227097   0.017  0.98674
```

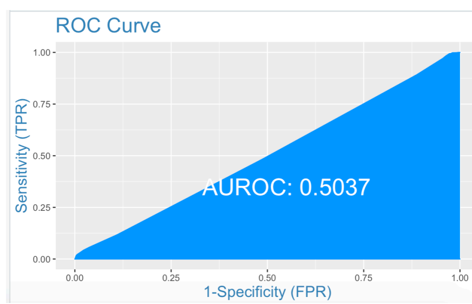
```
Province_StateGuam                      0.1431337  0.3349537   0.427  0.66914
Province_StateGuangdong                 0.8309132  0.3580599   2.321  0.02031 *
Province_StateGuangxi                   0.3875901  0.3293348   1.177  0.23924
Province_StateGuizhou                   0.1368740  0.3228447   0.424  0.67159
Province_StateHainan                    0.2955876  0.3362855   0.879  0.37941
Province_StateHawaii                    0.1736516  0.2840618   0.611  0.54099
Province_StateHebei                     0.3654923  0.3369155   1.085  0.27800
Province_StateHeilongjiang              0.4519942  0.3371414   1.341  0.18003
Province_StateHenan                     1.0909609  0.3738274   2.918  0.00352 **
Province_StateHong Kong                 0.0772966  0.3375519   0.229  0.81888
Province_StateHubei                     9.9973365  1.0024968   9.972  < 2e-16 ***
Province_StateHunan                     0.7363246  0.3562472   2.067  0.03874 *
Province_StateIdaho                     0.1310269  0.3218099   0.407  0.68389
Province_StateIllinois                  0.1465532  0.2550047   0.575  0.56549
Province_StateIndiana                  -0.0068624  0.2448081  -0.028  0.97764
Province_StateInner Mongolia            0.1849823  0.3219912   0.574  0.56563
Province_StateIowa                      0.0368918  0.2673905   0.138  0.89026
Province_StateJiangsu                   0.6479254  0.3414276   1.898  0.05774 .
Province_StateJiangxi                   0.8199377  0.3588813   2.285  0.02233 *
Province_StateJilin                     0.1747720  0.3240603   0.539  0.58967
Province_StateKansas                   -0.0389212  0.2839959  -0.137  0.89099
```

```
Province_StateWisconsin                 0.1095999  0.2704977   0.405  0.68535
Province_StateWyoming                   0.0533260  0.3246429   0.164  0.86953
Province_StateXinjiang                  0.2058201  0.3327576   0.619  0.53623
Province_StateYunnan                    0.3187471  0.3225073   0.988  0.32299
Province_StateZhejiang                  0.8000409  0.3669808   2.180  0.02925 *
Cases                                  -0.0041890  0.0003286 -12.747  < 2e-16 ***
Month2                                  0.0248206  0.0301979   0.822  0.41112
Month3                                  0.0702617  0.0331223   2.121  0.03390 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 48523  on 35001  degrees of freedom
Residual deviance: 48108  on 34876  degrees of freedom
AIC: 48360

Number of Fisher Scoring iterations: 8
```

```
> misClassError(testData$Deaths, predicted, threshold = optCutOff)
[1] 0.4888
```

```
> Concordance(testData$Deaths, predicted)
$Concordance
[1] 0.4860043

$Discordance
[1] 0.5139957

$Tied
[1] 0

$Pairs
[1] 56265001
```

```
> accuracy = ((p[1,1] + p[2,2])/sum(p))*100
> accuracy
[1] 51.11985
```



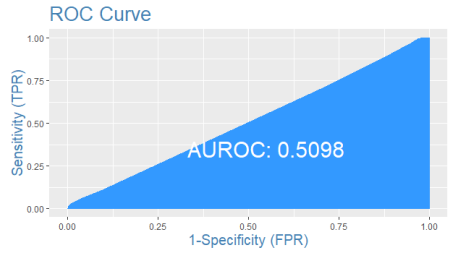
KFOLDS:

COVID -19 Data Analysis

```
> print(model)
Generalized Linear Model

35002 samples
    3 predictor
    2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 31502, 31502, 31502, 31501, 31502, 31502, ...
Resampling results:

  Accuracy   Kappa
  0.5071139  0.01423841
```

LOGISTIC MODEL - BACKWARD SELECTION

```
> model <- glm(Deaths ~ Province_State + Cases + Difference, data=trainingDat
a, family=binomial(link = "logit"))
> summary(model)

Call:
glm(formula = Deaths ~ Province_State + Cases + Difference, family = binomial(link = "logit"),
    data = trainingData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.5366  -1.1819   0.0043   1.1729   5.2071

Coefficients:
                                     Estimate Std. Error z value Pr(>|z|)
(Intercept)                        -5.407e-02  2.326e-01  -0.232  0.81618
Province_StateAlaska                2.667e-02  3.300e-01   0.081  0.93559
Province_StateAlberta               4.371e-01  3.319e-01   1.317  0.18789
Province_StateAnhui                 8.058e-01  3.582e-01   2.249  0.02448 *
Province_StateArizona              -2.167e-02  2.584e-01  -0.084  0.93318
.
.
.
.
.
Province_StateWisconsin             1.114e-01  2.705e-01   0.412  0.68037
Province_StateWyoming               5.407e-02  3.246e-01   0.167  0.86772
Province_StateXinjiang              2.119e-01  3.328e-01   0.637  0.52419
Province_StateYunnan                3.267e-01  3.226e-01   1.013  0.31114
Province_StateZhejiang              8.922e-01  3.725e-01   2.396  0.01659 *
Cases                              -3.770e-03  3.130e-04 -12.044  < 2e-16 ***
Difference                         -1.556e-02  3.403e-03  -4.574 4.79e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 48523  on 35001  degrees of freedom
Residual deviance: 48091  on 34877  degrees of freedom
AIC: 48341

Number of Fisher Scoring iterations: 8
```

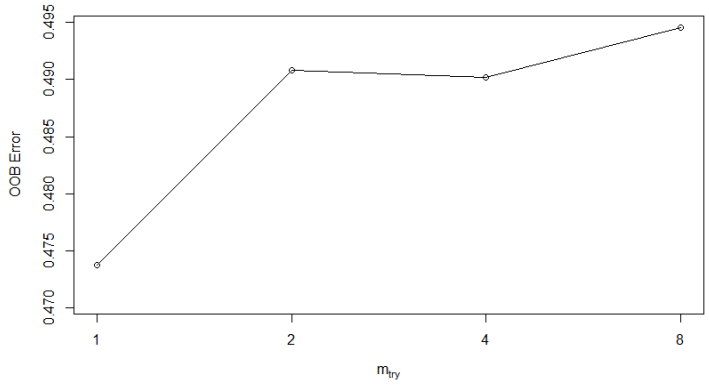



```
> optCutOff
[1] 0.43997699
```

```
> misClassError(testData$Deaths, predicted, threshold = optCutOff)
[1] 0.4873
```

```
> accuracy
[1] 53.2689
```

RANDOM FOREST MODEL



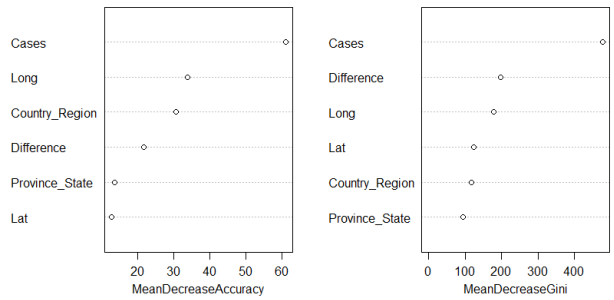

```
predTrain     0     1
        0  3711  1567
        1 13807 15917
>
```

```
predtest    0    1
       0 1347  900
       1 6137 6618
```

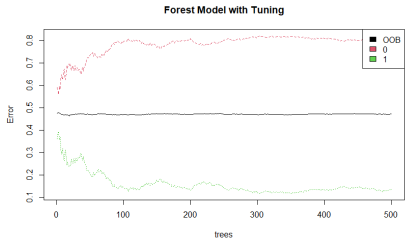
```
> accuracy
[1] 53.09292
```

```
> importance(model_r)
                        0           1 MeanDecreaseAccuracy MeanDecreaseGini
Country_Region  9.4612295 -0.6984559             30.67858         118.41691
Province_State -0.3155131  3.6130498             13.62395          94.73482
Cases          25.4636800 28.4057517             61.03682         478.51444
Difference      4.9195383 22.0028197             21.83178         197.06367
Lat             6.4326283 -2.9022901             12.76609         123.92905
Long           13.1840216 -1.6016845             33.87305         178.12362
```

model_r



COVID -19 Data Analysis



RECOVER OR DIE?

FORWARD MODEL APPROACH

	Active.Patient	Dead.or.Recovered
160	Patient_640	Recovered
161	Patient_644	Recovered
162	Patient_648	Recovered
163	Patient_652	Dead
164	Patient_656	Dead
165	Patient_660	Dead
166	Patient_664	Dead
167	Patient_668	Dead
168	Patient_672	Dead
169	Patient_676	Recovered
170	Patient_680	Dead
171	Patient_684	Dead
172	Patient_688	Dead
173	Patient_692	Recovered
174	Patient_696	Dead

	Confirmed.Patient	Dead.or.Recovered
160	Patient_639	Recovered
161	Patient_643	Recovered
162	Patient_647	Recovered
163	Patient_651	Dead
164	Patient_655	Dead
165	Patient_659	Dead
166	Patient_663	Dead
167	Patient_667	Dead
168	Patient_671	Dead
169	Patient_675	Recovered
170	Patient_679	Dead
171	Patient_683	Dead
172	Patient_687	Dead
173	Patient_691	Recovered
174	Patient_695	Dead

BACKWARD MODEL APPROACH

	Active.Patient	Dead.or.Recovered
160	Patient_640	Recovered
161	Patient_644	Recovered
162	Patient_648	Recovered
163	Patient_652	Dead
164	Patient_656	Dead
165	Patient_660	Dead
166	Patient_664	Dead
167	Patient_668	Dead
168	Patient_672	Dead
169	Patient_676	Dead
170	Patient_680	Dead
171	Patient_684	Dead
172	Patient_688	Dead
173	Patient_692	Dead
174	Patient_696	Dead

	Confirmed.Patient	Dead.or.Recovered
160	Patient_639	Recovered
161	Patient_643	Recovered
162	Patient_647	Recovered
163	Patient_651	Dead
164	Patient_655	Dead
165	Patient_659	Dead
166	Patient_663	Dead
167	Patient_667	Dead
168	Patient_671	Dead
169	Patient_675	Dead
170	Patient_679	Dead
171	Patient_683	Dead
172	Patient_687	Dead
173	Patient_691	Dead
174	Patient_695	Dead

RANDOM FOREST MODEL APPROACH

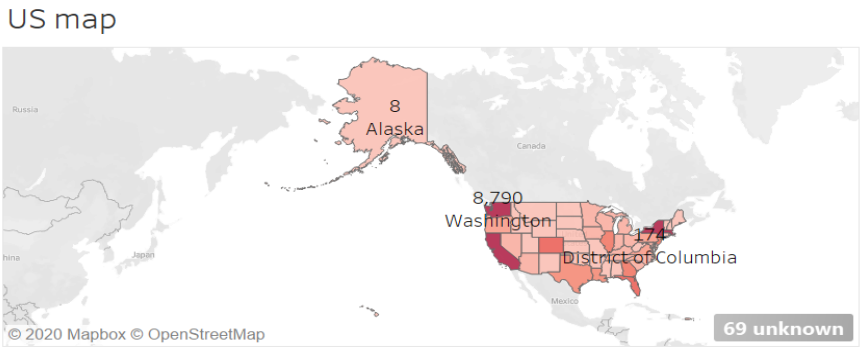
COVID -19 Data Analysis

	Active.Patient	Dead.or.Recovered
160	Patient_640	Recovered
161	Patient_644	Dead
162	Patient_648	Recovered
163	Patient_652	Dead
164	Patient_656	Dead
165	Patient_660	Dead
166	Patient_664	Recovered
167	Patient_668	Recovered
168	Patient_672	Dead
169	Patient_676	Recovered
170	Patient_680	Dead
171	Patient_684	Dead
172	Patient_688	Dead
173	Patient_692	Recovered
174	Patient_696	Dead

	Confirmed.Patient	Dead.or.Recovered
160	Patient_639	Recovered
161	Patient_643	Dead
162	Patient_647	Recovered
163	Patient_651	Dead
164	Patient_655	Dead
165	Patient_659	Dead
166	Patient_663	Recovered
167	Patient_667	Recovered
168	Patient_671	Dead
169	Patient_675	Recovered
170	Patient_679	Dead
171	Patient_683	Dead
172	Patient_687	Dead
173	Patient_691	Recovered
174	Patient_695	Dead

ANALYSIS USING TABLEAU

MAP ANALYSIS DASHBOARD



RECOVERY VS DEATH EVOLUTION