COVID19 Global Forecasting

Talha Hanif Butt
Computer Science Department
FAST-NUCES
Lahore, Pakistan
thanifbutt@gmail.com

***Abstract*—Kaggle had launched a competition [1] for global forecasting of COVID19. The goal of this competition was to provide better methods for estimates that can assist medical and governmental institutions to prepare and adjust as pandemics unfold.**

In this project, I predicted the cumulative number of confirmed COVID19 cases in various locations across the world, as well as the number of resulting fatalities for future dates.

To the best of my knowledge, this is the first attempt to solve this problem using PySpark. Along with the traditional algorithms available like Linear Regression, Decision Tree, Random Forest, Gradient Boosted Tree, I also applied High Dimensional Data Projection and was able to top the leader board in Week-4 of the competition.

I. Introduction

The 2019 novel coronavirus (COVID-19) pandemic appeared in Wuhan, China in December 2019 and has become a serious public health problem worldwide [7], [8]. The virus that caused COVID-19 pandemic disease was called as severe acute respiratory syndrome coronavirus 2, also named SARS-CoV-2 [9]. Coronaviruses (CoV) are a large family of viruses that cause diseases resulting from colds such as the Middle East Respiratory Syndrome (MERS-CoV) and Severe Acute Respiratory Syndrome (SARS-CoV). Coronavirus disease (COVID-19) is a new species that was discovered in 2019 and has not been previously identified in humans.

In this project, I aim to explore different solutions for predicting the number of confirmed cases and fatalities in different regions of the world. To the best of my knowledge, it is the first attempt to solve the problem under consideration using PySpark. Moreover, apart from traditional approaches like Linear Regression, Decision Tree, Random Forest and Gradient Boosted Tree, I have also tried Data Projection in to Higher Dimensions along with Gradient Boosted Tree obtaining reasonable results. Although, the Kaggle competition in which I participated got over by the time I got better results but I was able to top the leader board using Projection in Higher Dimensions followed by Gradient Boosted Tree.

II. Data

The data for the competition has been provided by John Hopkins CSSE [2]. Currently, the competition is in week 3, the training dataset contains 32,707 cases comprising of 72 days and 180 countries.

For training, each case has corresponding Id, Province State, Country Region, Date, ConfirmedCases and Fatalities.

For test data, each case has Id, Province State, Country Region and Date.

Out of the parameters provided, Province State had NULL values which required to be taken care of.

In PySpark, it is not possible to train a regression model with multiple ouputs, as a result of which, separate training and testing needed to be performed for Confirmed Cases and Fatalities followed by joining their individual outputs in a single file for submission.

III. Some Previous Submissions

People had tried Linear Regression using LightGBM, Bayesian Model, Log-domain lognormal as a sigmoid and SIR models. Majority had focused on data visualization so that better understanding could be attained for proposing a solution.

IV. Methodology

The challenge involved forecasting confirmed cases and fatalities between April 1 and April 30 by region, the primary goal wasn't only to produce accurate forecasts. It was also to identify factors that appeared to impact the transmission rate of COVID-19.

When I started, the competition was in Week-3. Following is what I did in the project:

Visualization, Pre-Processing, Linear Regression, Decision Tree, Random Forest, Gradient Boosted Tree, Higher Dimension Data Projection + Gradient Boosted Tree, Post-Processing.

A. Visualization

The data set provided had 32,707 training records while 13,158 test records having following schemas:

```
root
 |-- Id: integer (nullable = true)
 |-- Province_State: string (nullable = true)
 |-- Country_Region: string (nullable = true)
 |-- Date: timestamp (nullable = true)
 |-- ConfirmedCases: double (nullable = true)
 |-- Fatalities: double (nullable = true)
```

Fig. 1. Train

```
root
 |-- ForecastId: integer (nullable = true)
 |-- Province_State: string (nullable = true)
 |-- Country_Region: string (nullable = true)
 |-- Date: timestamp (nullable = true)
```

Fig. 2. Test

Id	Province_State	Country_Region	Date	ConfirmedCases	Fatalities
1	null	Afghanistan	2020-01-22 00:00:00	0.0	0.0
2	null	Afghanistan	2020-01-23 00:00:00	0.0	0.0
3	null	Afghanistan	2020-01-24 00:00:00	0.0	0.0
4	null	Afghanistan	2020-01-25 00:00:00	0.0	0.0
5	null	Afghanistan	2020-01-26 00:00:00	0.0	0.0
6	null	Afghanistan	2020-01-27 00:00:00	0.0	0.0
7	null	Afghanistan	2020-01-28 00:00:00	0.0	0.0
8	null	Afghanistan	2020-01-29 00:00:00	0.0	0.0
9	null	Afghanistan	2020-01-30 00:00:00	0.0	0.0
10	null	Afghanistan	2020-01-31 00:00:00	0.0	0.0
11	null	Afghanistan	2020-02-01 00:00:00	0.0	0.0
12	null	Afghanistan	2020-02-02 00:00:00	0.0	0.0
13	null	Afghanistan	2020-02-03 00:00:00	0.0	0.0
14	null	Afghanistan	2020-02-04 00:00:00	0.0	0.0
15	null	Afghanistan	2020-02-05 00:00:00	0.0	0.0
16	null	Afghanistan	2020-02-06 00:00:00	0.0	0.0
17	null	Afghanistan	2020-02-07 00:00:00	0.0	0.0
18	null	Afghanistan	2020-02-08 00:00:00	0.0	0.0
19	null	Afghanistan	2020-02-09 00:00:00	0.0	0.0
20	null	Afghanistan	2020-02-10 00:00:00	0.0	0.0

Fig. 3. Some Training Records

1) Confirmed Cases but no Fatalities: Fig.5-Fig.10 represent different visualizations performed for the cases which have confirmed virus but have no fatalities. From Fig.9, Top 5 countries include China, US, Australia, Canada and France. It can be seen in Fig.6 that the spike occurred after 30 days while it started going downwards after about 55 days.

2) Per Day Confirmed Cases Sorted: Fig.11-Fig.14 represent different visualizations performed for the cases which have confirmed virus having fatalities or not. It can be seen in Fig.12 that the spike occurred after 30 days but never goes down in the data provided.

3) Per Day Fatalities Sorted: Fig.15-Fig.18 represent different visualizations performed for per day fatalities. It can be seen in Fig.16 that the spike occurred after 45 days but never goes down in the data provided.

ForecastId	Province_State	Country_Region	Date
1	null	Afghanistan	2020-04-02 00:00:00
2	null	Afghanistan	2020-04-03 00:00:00
3	null	Afghanistan	2020-04-04 00:00:00
4	null	Afghanistan	2020-04-05 00:00:00
5	null	Afghanistan	2020-04-06 00:00:00
6	null	Afghanistan	2020-04-07 00:00:00
7	null	Afghanistan	2020-04-08 00:00:00
8	null	Afghanistan	2020-04-09 00:00:00
9	null	Afghanistan	2020-04-10 00:00:00
10	null	Afghanistan	2020-04-11 00:00:00
11	null	Afghanistan	2020-04-12 00:00:00
12	null	Afghanistan	2020-04-13 00:00:00
13	null	Afghanistan	2020-04-14 00:00:00
14	null	Afghanistan	2020-04-15 00:00:00
15	null	Afghanistan	2020-04-16 00:00:00
16	null	Afghanistan	2020-04-17 00:00:00
17	null	Afghanistan	2020-04-18 00:00:00
18	null	Afghanistan	2020-04-19 00:00:00
19	null	Afghanistan	2020-04-20 00:00:00
20	null	Afghanistan	2020-04-21 00:00:00

Fig. 4. Some Test Records

Date	count(ConfirmedCases)
2020-01-22 00:00:00	27
2020-01-23 00:00:00	34
2020-01-24 00:00:00	35
2020-01-25 00:00:00	38
2020-01-26 00:00:00	39
2020-01-27 00:00:00	40
2020-01-28 00:00:00	41
2020-01-29 00:00:00	43
2020-01-30 00:00:00	46
2020-01-31 00:00:00	50
2020-02-01 00:00:00	51
2020-02-02 00:00:00	50
2020-02-03 00:00:00	50
2020-02-04 00:00:00	50
2020-02-05 00:00:00	48
2020-02-06 00:00:00	48
2020-02-07 00:00:00	47
2020-02-08 00:00:00	45
2020-02-09 00:00:00	42
2020-02-10 00:00:00	41

Fig. 5. Confirmed Cases but no Fatalities Sorted

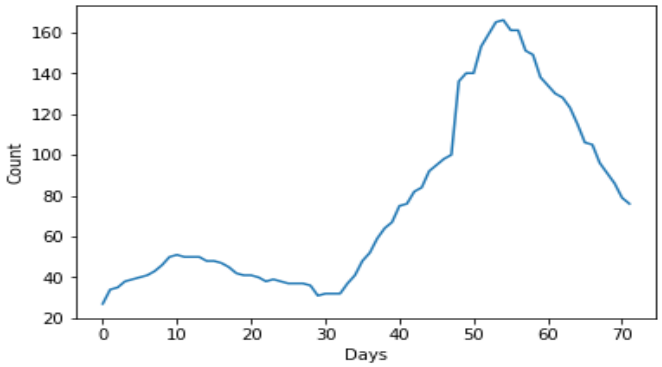

Fig. 6. Per Day Confirmed Cases but no Fatalities Counts

4) Countrywise Confirmed Cases: Fig.19-Fig.21 represent different visualizations performed for countrywise confirmed cases. From Fig.20, Top 5 countries include China, US, Australia, Canada and France

5) Countrywise Fatalities: Fig.22-Fig.24 represent different visualizations performed for countrywise confirmed cases. From Fig.23, Top 5 countries include China, US, Canada, Australia and France.

Province_State had NULL values so I tried removing records having NULL values and removing Province_State as

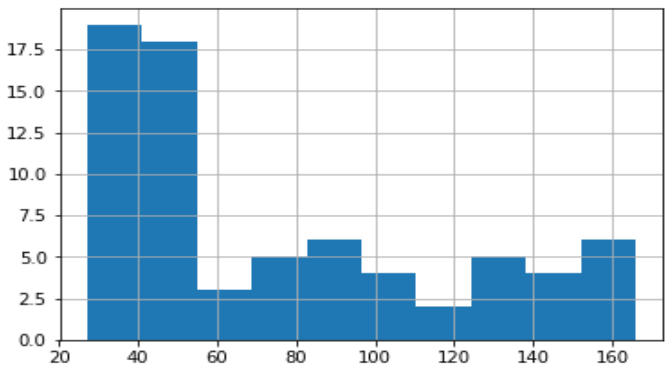

Fig. 7. Per Day Confirmed Cases but no Fatalities as a Histogram

count	72.000000
mean	76.291667
std	43.521546
min	27.000000
25%	40.000000
50%	51.500000
75%	105.250000
max	166.000000

Fig. 8. Basic Stats about Confirmed Cases but no Fatalities

Country_Region	count(ConfirmedCases)
China	830
US	505
Australia	290
Canada	248
France	193

Fig. 9. Top 5 Countries w.r.t Confirmed Cases but no Fatalities

a feature to include all the records. I got better results by removing Province_State as a whole.

B. Pre-Processing

Pre-processing tasks performed were as follows: Remove NULL values, Convert Timestamp to UnixTimestamp and Convert Categorical Attributes to Nominal.

C. Linear Regression

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable

Country_Region	count(ConfirmedCases)
Panama	1
Botswana	1
Albania	2
Guatemala	2
Sierra Leone	3

Fig. 10. Bottom 5 Countries w.r.t Confirmed Cases but no Fatalities

Date	count(ConfirmedCases)
2020-01-22 00:00:00	28
2020-01-23 00:00:00	36
2020-01-24 00:00:00	38
2020-01-25 00:00:00	41
2020-01-26 00:00:00	44
2020-01-27 00:00:00	47
2020-01-28 00:00:00	48
2020-01-29 00:00:00	51
2020-01-30 00:00:00	54
2020-01-31 00:00:00	58
2020-02-01 00:00:00	60
2020-02-02 00:00:00	60
2020-02-03 00:00:00	60
2020-02-04 00:00:00	61
2020-02-05 00:00:00	61
2020-02-06 00:00:00	61
2020-02-07 00:00:00	62
2020-02-08 00:00:00	62
2020-02-09 00:00:00	62
2020-02-10 00:00:00	62

Fig. 11. Per Day Confirmed Cases Sorted

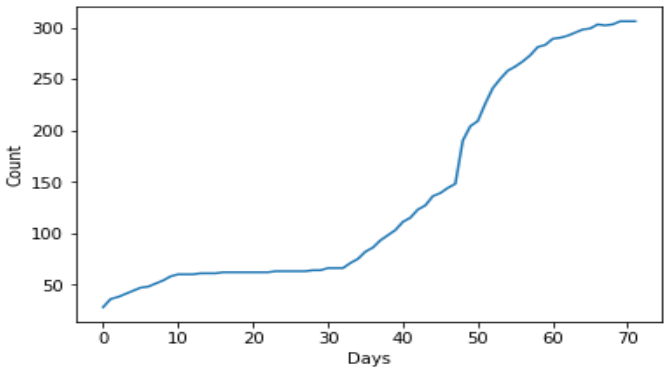

Fig. 12. Per Day Confirmed Cases Counts

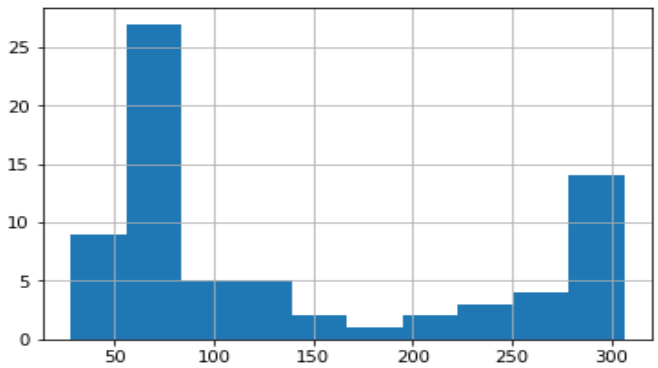

Fig. 13. Per Day Confirmed Cases as a Histogram

count	72.000000
mean	139.819444
std	99.352315
min	28.000000
25%	62.000000
50%	84.000000
75%	252.000000
max	306.000000

Fig. 14. Basic Stats about Per Day Confirmed Cases

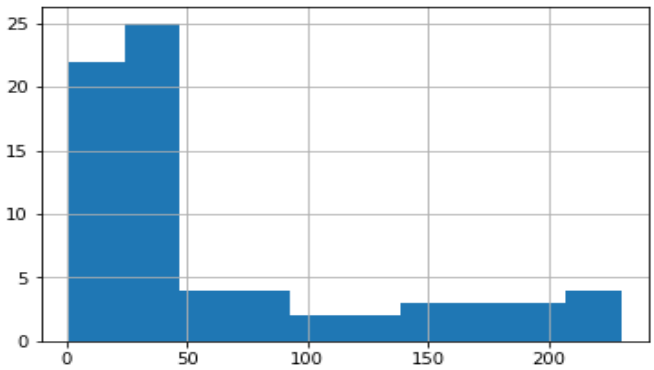

Fig. 17. Per Day Fatalities as a Histogram

Date	count(Fatalities)
2020-01-22 00:00:00	1
2020-01-23 00:00:00	2
2020-01-24 00:00:00	3
2020-01-25 00:00:00	3
2020-01-26 00:00:00	5
2020-01-27 00:00:00	7
2020-01-28 00:00:00	7
2020-01-29 00:00:00	8
2020-01-30 00:00:00	8
2020-01-31 00:00:00	8
2020-02-01 00:00:00	9
2020-02-02 00:00:00	10
2020-02-03 00:00:00	10
2020-02-04 00:00:00	11
2020-02-05 00:00:00	13
2020-02-06 00:00:00	13
2020-02-07 00:00:00	15
2020-02-08 00:00:00	17
2020-02-09 00:00:00	20
2020-02-10 00:00:00	21

Fig. 15. Per Day Fatalities Sorted

count	72.000000
mean	63.527778
std	66.643887
min	1.000000
25%	19.250000
50%	34.000000
75%	86.750000
max	230.000000

Fig. 18. Basic Stats about Per Day Fatalities Sorted

score, c = constant, b = regression coefficient, and x = score on the independent variable [3].

Linear Regression	Including Province State	Without Including Province State
Score	3.65539	3.35323

TABLE I
RESULTS USING LINEAR REGRESSION

D. Decision Tree

A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. The construction of decision tree classifier does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision

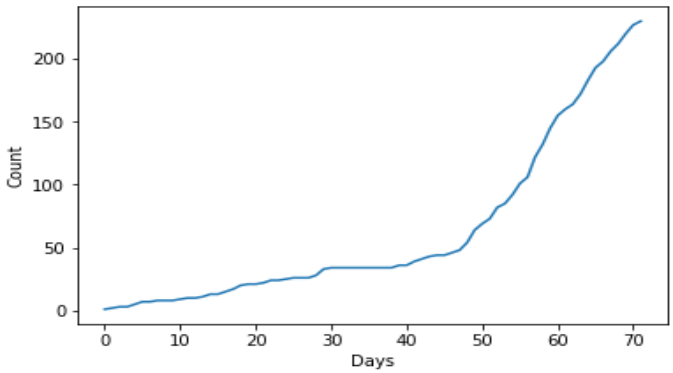

Fig. 16. Per Day Fatalities Counts

count	180.000000
mean	55.927778
std	199.527711
min	3.000000
25%	19.000000
50%	28.000000
75%	37.000000
max	2357.000000

Fig. 19. Basic Stats about Countrywise Confirmed Cases

Country_Region	count(Fatalities)
China	2357
US	1245
Australia	376
Canada	334
France	278

Fig. 20. Top 5 Countrywise Confirmed Cases

Country_Region	count(Fatalities)
Sierra Leone	3
Burundi	3
Botswana	4
MS Zaandam	6
Burma	7

Fig. 21. Bottom 5 Countrywise Confirmed Cases

trees can handle high dimensional data. In general decision tree classifier has good accuracy. Decision tree induction is a typical inductive approach to learn knowledge on classification.

Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. An instance is classified by starting at the root node of the tree,testing the attribute specified by this node,then moving down the tree branch corresponding to the value of the attribute. This process is

count	133.000000
mean	34.390977
std	145.487052
min	1.000000
25%	8.000000
50%	14.000000
75%	22.000000
max	1527.000000

Fig. 22. Basic Stats about Countrywise Fatalities

Country_Region	count(Fatalities)
China	1527
US	740
Canada	86
Australia	86
France	85

Fig. 23. Top 5 Countrywise Fatalities

Country_Region	count(Fatalities)
Libya	1
Congo (Brazzaville)	1
Zambia	1
Senegal	2
MS Zaandam	2

Fig. 24. Bottom 5 Countrywise Fatalities

then repeated for the subtree rooted at the new node [4].

Decision Tree	Depth = 3	Depth = 5
Score	2.49157	2.38298

TABLE II
RESULTS USING DECISION TREE

E. Random Forest

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction [5].

Random Forest	Trees = 2	Trees = 20	Trees = 100
Score	3.13457	3.15120	3.20820

TABLE III
RESULTS USING RANDOM FOREST

F. Gradient Boosted Tree

A gradient boosted model is an ensemble of either regression or classification tree models. Both are forward-learning ensemble methods that obtain predictive results through gradually improved estimations. Boosting is a flexible nonlinear regression procedure that helps improving the accuracy of

trees. By sequentially applying weak classification algorithms to the incrementally changed data, a series of decision trees are created that produce an ensemble of weak prediction models. While boosting trees increases their accuracy, it also decreases speed and human interpretability. The gradient boosting method generalizes tree boosting to minimize these issues [6].

Depth	Score
3	2.54647
5	2.05806
7	2.01467
9	1.98333
30	1.98171

TABLE IV
RESULTS USING GRADIENT BOOSTED TREE

G. Higher Dimension Data Projection + Gradient Boosted Tree

In this experiment, I tried projecting data to higher dimensions by first taking square and then cube of Country_RegionIndex followed by Gradient Boosted Tree.

HDDP + GBT	Square	Cube
Score	2.2225	2.22225

TABLE V
RESULTS USING HIGHER DIMENSION DATA PROJECTION + GRADIENT BOOSTED TREE

H. Post-Processing

Post-processing step included the creation of submission file by combining the outputs of multiple files including a separate file each for confirmed cases and fatalities.

V. EVALUATION

Submissions were evaluated using root mean square logarithmic error as:

$$\sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\Big(log(p_i+1)-log(a_i+1)\Big)^2}$$

where:

n is the total number of observations

p_i is the prediction

a_i is the actual value

log(x) is the natural logarithm of x

VI. CONCLUSION

I found out that Gradient Boosted Tree performed the best among the algorithms tested with a score of 1.98171 on week-3 and 2.22225 on week-4 which was the top score on the private leader board as the competition was closed.

VII. FUTURE WORK

It will be a good direction to try different values for records having NULL values and see the effect.

REFERENCES

[1] https://www.kaggle.com/c/covid19-global-forecasting-week-3/ data.

[2] https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series.

[3] https://www.statisticssolutions.com/what-is-linear-regression/

[4] https://www.geeksforgeeks.org/decision-tree/

[5] https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[6] https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/gradient_boosted_trees.html

[7] Roosa, K., Lee, Y., Luo, R., Kirpich, A., Rothenberg, R., Hyman, J. M., Yan, P., and Chowell, G. Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020. Infectious Disease Modelling, 5 : 256-263, 2020.

[8] Yan, L., Zhang, H. T., Xiao, Y., Wang, M., Guo, Y., Sun, C., Tang, X., Jing, L., Li, S., Zhang, M., Xiao, Y., Cao, H., Chen, Y., Ren, T., Jin, J., Wang, F., Xiao, Y., Huang, S., Tan, X., Huang, N., Jiao, B., Zhang, Y., Luo, A., Cao, Z., Xu, H., and Yuan, Y. Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan. medRxiv preprint doi: https://doi.org/10.1101/2020.02.27.20028027, 1-18, 2020.

[9] Stoecklin, S. B., Rolland, P., Silue, Y., Mailles, A., Campese, C., Simondon, A., Mechain, M., Meurice, L., Nguyen, M., Bassi C., Yamani, E., Behillil, S., Ismael,S., Nguyen, D., Malvy, D., Lescure, F. X., Georges, S., Lazarus, C., Tabaï, A., Stempfelet, M., Enouf, V., Coignard, B., Levy-Bruhl, D. and Team, I. First cases of coronavirus disease 2019 (COVID-19) in France: surveillance, investigations and control measures, January 2020. Eurosurveillance, 25(6) :2000094, 2020.