**Kaggle Challenge Report**

COVID19 Global Forecasting (Week 4)



Kaggle Notebook link: [SSID-1940984](https://www.kaggle.com/code/manindersingh97/covid-19-weekk-4/notebook)

**Table of Contents**

Introduction and Background-------------------------------------------------------------------03

Dataset Description---------------------------------------------------------------------03

Literature Survey----------------------------------------------------------------------------------05

Related Work-----------------------------------------------------------------------------05

Different Models used for the challenge-------------------------------------------05

Data Processing Approach---------------------------------------------------------------------07

Machine Learning Approaches----------------------------------------------------------------16

XGBoost Regression model----------------------------------------------------------16

Linear Regression model--------------------------------------------------------------17

Legal and Ethics details--------------------------------------------------------------------------18

Kaggle submission proof-------------------------------------------------------------------------19

Result and Future work---------------------------------------------------------------------------21

References------------------------------------------------------------------------------------------22

**Figures**

Figure 1------------------ 03 Figure 17----------------14

Figure 2------------------ 04 Figure 18----------------14

Figure 3------------------ 04 Figure 19----------------15

Figure 4------------------ 04 Figure 20----------------15

Figure 5------------------ 04 Figure 21----------------20

Figure 6------------------ 7 Figure 22----------------20

Figure 7------------------ 8 Figure 23----------------21

Figure 8------------------ 8

Figure 9------------------ 9

Figure 10----------------9

Figure 11----------------10

Figure 12----------------10

Figure 13----------------11

Figure 14----------------12

Figure 15----------------13

Figure 16----------------13

1. **Introduction and Background**

If we talk about the problem here which is related to Covid-19 which is a short representation of “Coronavirus disease 2019” is a contagious disease caused by a large family of virus, which caused a Pandemic on the global scale. It has not only created an economic but also healthcare crises.

According to the reports of World Health Organization(WHO), if we look at the population, about 80% which is the majority of the population who developed symptoms have recovered from the disease without even getting treated in the hospitals, about 15% got seriously ill and 5% became critically ill and needing intensive care.

So, in this report we will talk about the challenge presented on “Kaggle platform” which is a well-known community platform for Data Science and Machine Learning admirers. Week 4 is the 4th competition launched in this series of Covid-19, where the challenge involves forecasting ‘Confirmed Cases’ and ‘Fatalities’ from 15th April to 14th May in different ‘Country Regions’. The main aim of building a model which can accurately forecast is to identify which columns from the data can be considered to find the rate of increase in the cases on daily basis. This challenge was introduced and the data was presented by the Johns Hopkins University Centre for System Science and Engineering.

Dataset Description

The dataset provided for the challenge consist of 3 main files train.csv, test.csv and submission.csv and since the available data is structured, we just have to import them and start working on the files.

The features which are available in the files are as follows:

In train.csv file we have 6 columns: Id, Province\_State where we have some null values, Country\_Region, Date, Confirmed Cases and Fatalities. Where Confirmed Cases and Fatalities are the point of interest which are unavailable in the test.csv file which we have to predict with our model which we store in a submission.csv file.

Figure 1

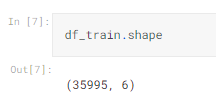


Figure 2

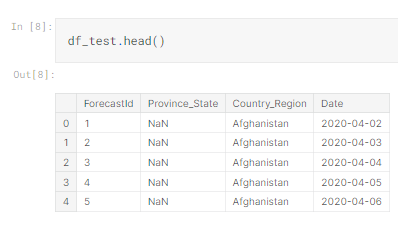


Figure 3

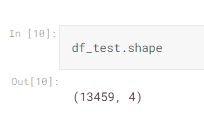


Figure 4

1. **Literature Survey**

In an interactive web-based dashboard to track Covid-19 in real time, According to E Dong, H DU, 2020([An interactive web-based dashboard to track COVID-19 in real time](https://www.semanticscholar.org/paper/An-interactive-web-based-dashboard-to-track-in-real-Dong-Du/ca019e1e38edf9d2112ea987362da454f909ac1b)), Covid-19 as discussed has been both global health and economic crisis, and if we talk about the numbers the count was 16 million people infected with more than 666000 deaths as per the reports until 29th July 2020 worldwide. Which further resulted in high degree of variances have evolved.

A study implies considering the data from different health organizations can be used for disease predictions where different Artificial Intelligence models can help from monitoring the data to visualizing and predicting Confirmed cases and Fatalities.

Related Work

Tracks Covid-19 in Real Time, ([An interactive web-based dashboard to track COVID-19 in real time ...](https://europepmc.org/article/MED/32087114)), where the researchers hosted by Center for Systems Science and Engineering(CSSE), at John Hopkins University the one which have introduced the dataset for the Kaggle challenge as well, have developed an interactive dashboard where we can track and visualize Covid-19 cases which are reported and also visualize them. The dashboard is so well organized with multiple features with the latest information. Here we can visualize the areas on the map as well where there are rise in cases and fatalities as well.

Sujatha and Chatterjee, (2020) ([A machine learning forecasting model for COVID-19 pandemic in India](https://pubmed.ncbi.nlm.nih.gov/32837309/)) explored different data analysis and visualization techniques with experimenting results using Linear Regression(LR) model which is considered to be an easy and a simple model for predicting independent variables with the help of dependent variables which are features from the training data.

According to the paper Non-contact screening system based for COVID-19 on XGBoost and logistic regression(Dong et al., 2022) XGBoost and Logistic Regression were used for the similar data and have displayed outperforming results than any other model. Inspired by which I have decided to use these two models for the current dataset to predict the ‘Confirmed Cases’ and ‘Fatalities’.

Considering all the papers I read related to the current dataset type these are few Models other challengers tried using these models, since I used one simple and one complex model in the current notebook and it is taking more than 105 minutes which is almost 2 hours to run the code, so I decided not to expand and complicate the executions, but since I tried using the following models in another notebook, I will include weblink in the end of the report.

Different Models Used for the challenge:

1. **Regression Model**

Regression models are a type of mathematical model used to explain and predict the relationship between one or more independent variables and a dependent variable. They are used to produce a quantitative measure of the relationship between the variables, which can be used for prediction and forecasting.

Regression models are commonly used in fields such as economics, finance, and marketing, as they allow researchers to quantify the relationship between different variables. They can also be used to identify relationships between variables that are not immediately apparent.

1. **Random forest**

Random forest regression is a powerful machine learning technique used for both supervised and unsupervised learning. It is a type of ensemble learning method which combines multiple decision trees in order to obtain better predictive results. It works by randomly selecting a subset of features and building multiple decision trees based on those features. The trees are built using a bagging technique where each tree is trained on a different set of randomly selected data points.

The disadvantages of random forest regression include its complexity, its slow training time, and its lack of interpretability. In addition, it is not suitable for very small datasets and can be prone to overfitting. So for this notebook, I haven’t included this model.

1. **Decision Tree Algorithm**

Decision tree algorithm is a supervised learning algorithm which can be used for classification and regression problems. It is a type of machine learning algorithm that uses a decision tree to go from observations about an item to conclusions about the item's target value. It is a graphical representation of possible solutions to a decision based on certain conditions. It works by breaking down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. Each time a subset is created, the most informative feature is selected and a decision is made. The decision tree is built with an algorithm that identifies the feature with the most information gain at each step. At the end of the process, a decision tree is generated that can be used to make predictions about the target value of a given item.

1. **Data Processing Approach**

Initially, the first thing to check once we load the files in notebook is to check if there are any null values in the dataset and if found any either we fill them with the data, or if the features are not important for our model we can just drop them.

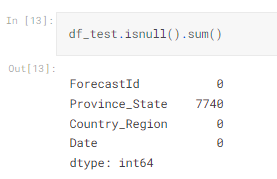


Figure 5

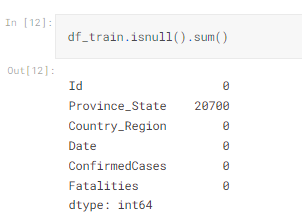


Figure 6

Since in the current dataset, ‘Province\_State’ is not that of an important feature for building or training our model I have dropped the feature. Initially I thought of using the feature by filling the null values with ‘Country\_Region’ and I tried doing it which was only helpful for data visualization purposes.

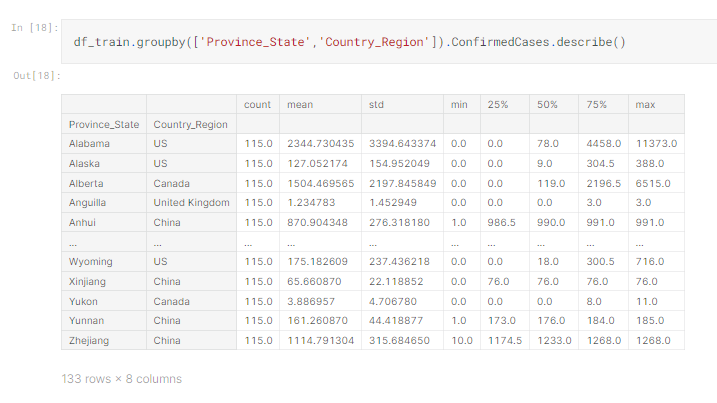


Figure 7

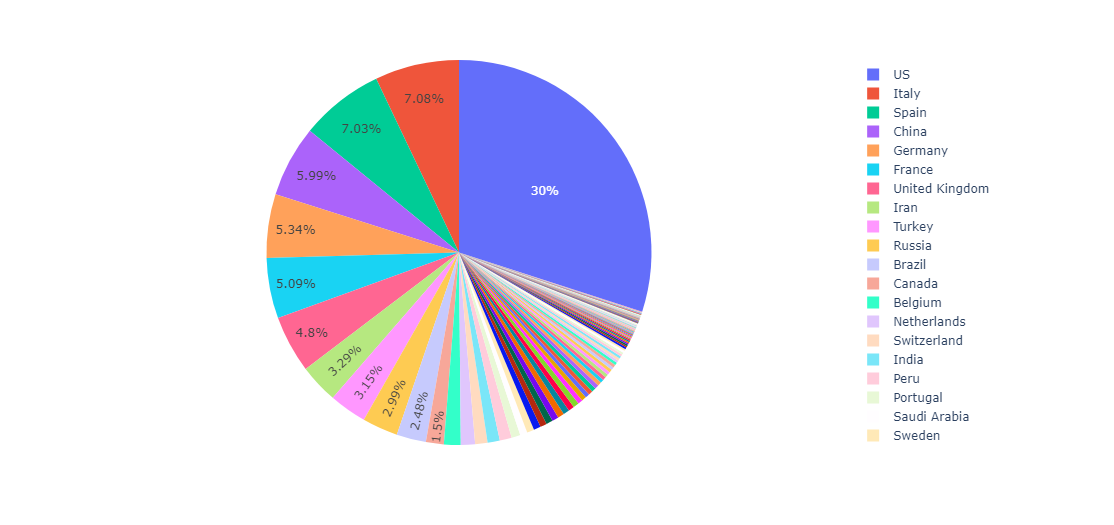


Figure 8

Plotting a pie chart, to visualize the percentage of the Confirmed cases in each country region where I can hover on the pie chart to check for the count.

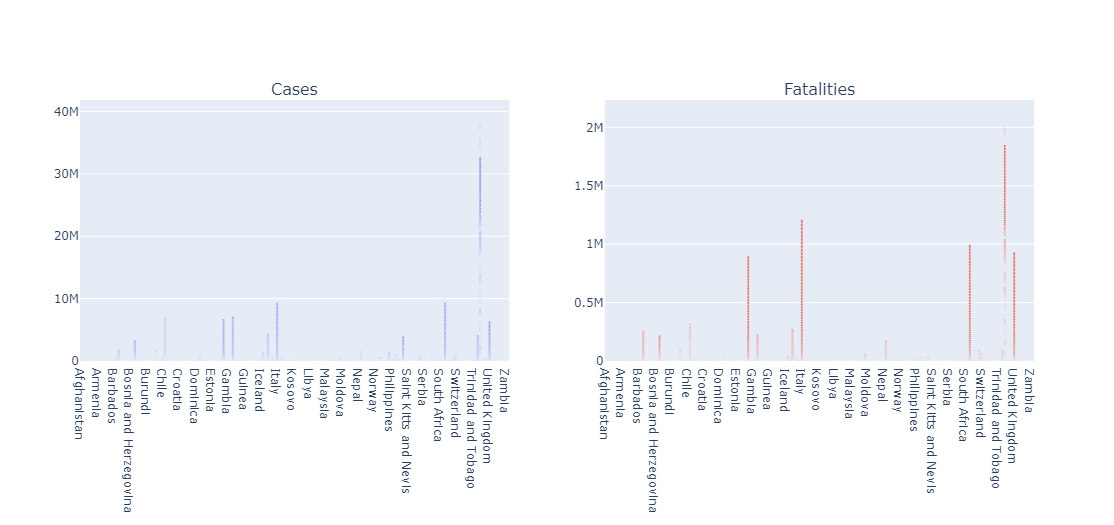


Figure 9

In this bar graph we can differentiate the count between confirmed cases and fatalities plotting the county region on the x axis. To compare them side by side, subplots were used.

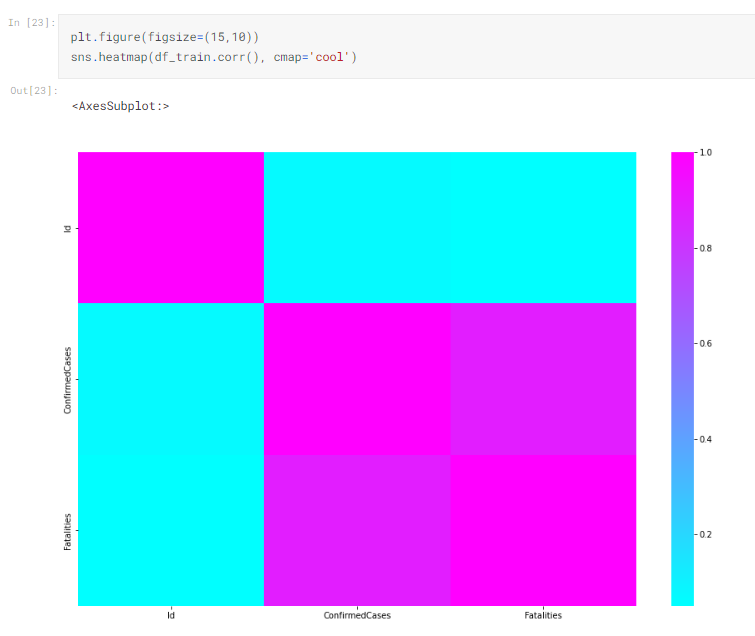


Figure 10

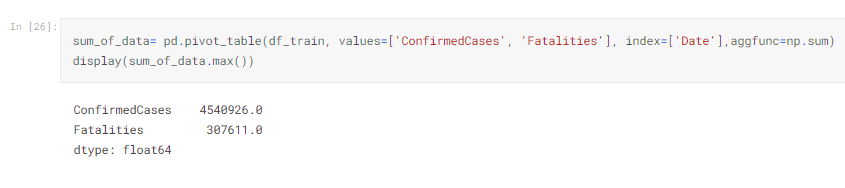


Figure 11

Here, to know the total number of cases from all the country regions combined, all we did is use sum function.

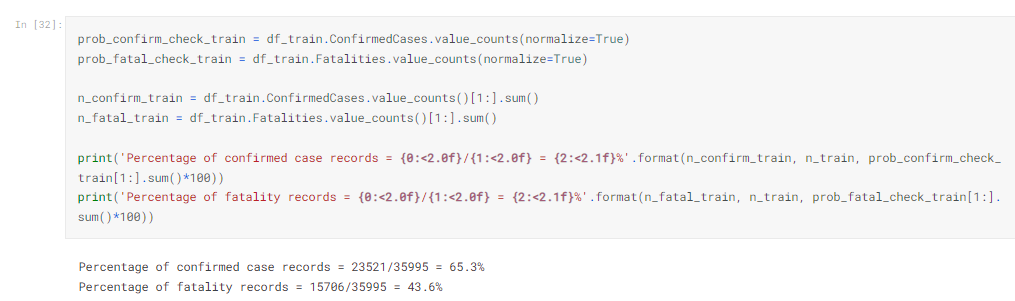


Figure 12

In the below figure is one of the most critical analysis in the notebook which shows the GrowthRate, We can see the growth rate of confirmed cases in the regions on everyday basis.

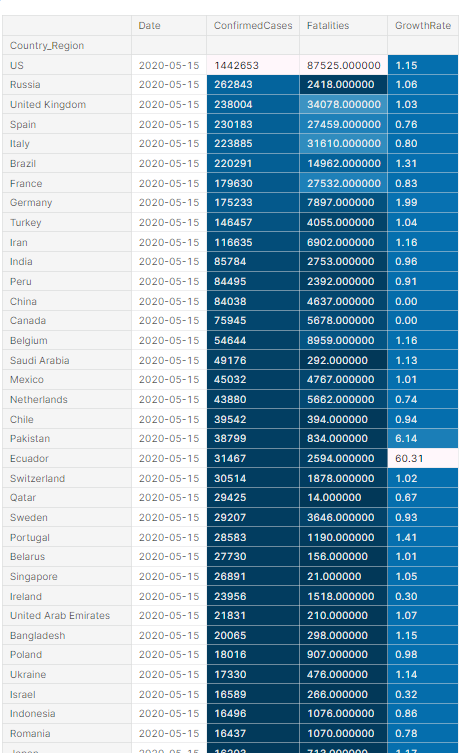


Figure 13

In the below figure, we can see the top 20 countries with highest number of confirmed cases and comparing them with Fatalities on the same graph. This is also an important observation since we get an idea of the desired output.

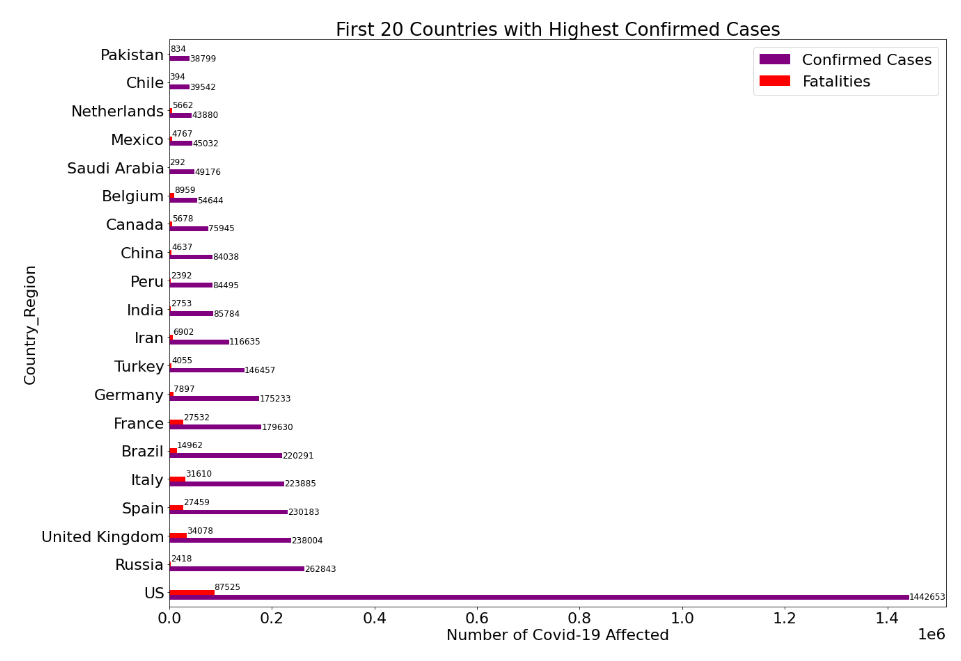


Figure 14

From the below figure,

We can see in the 1st subplot, which shows confirmed cases globally I have used line graph to plot on increase in everyday basis between confirmed cases and new confirmed cases which is a new feature we added using shift() function.

In the 2nd subplot, which shows confirmed cases Nationally, I have used line graph to plot all the country regions with increase in cases on everyday basis.

In the 3rd subplot, which shows Fatalities globally we have plotted a line graph to understand the features, fatalities and new fatalities which is again a new feature created using shift() function.

In the 4th subplot, which shows Fatalities Nationally, we have plotted all the country regions on the graph with increase in fatalities on everyday basis.

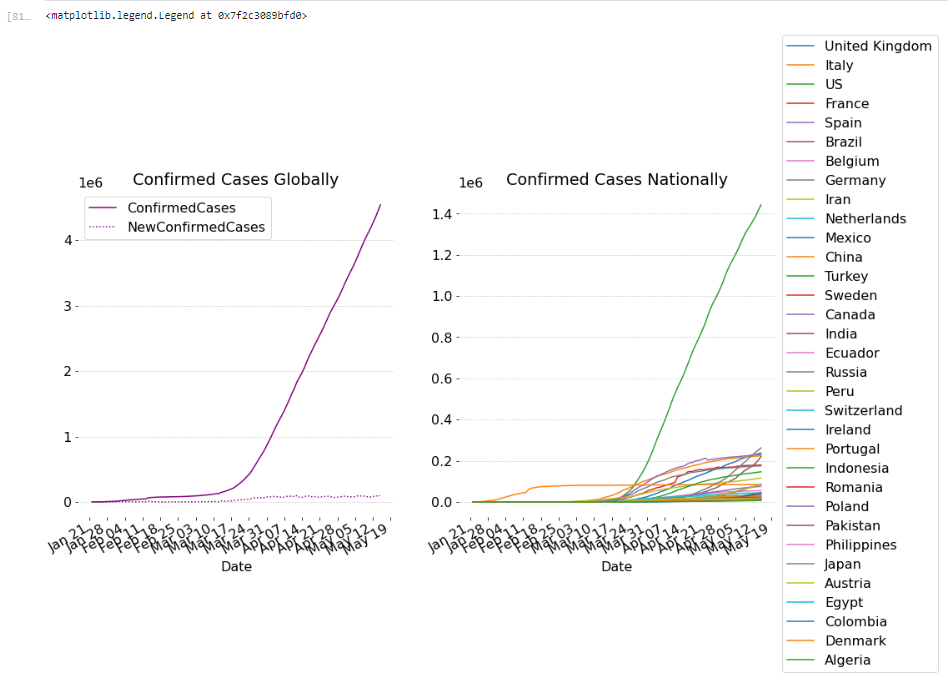


Figure 15

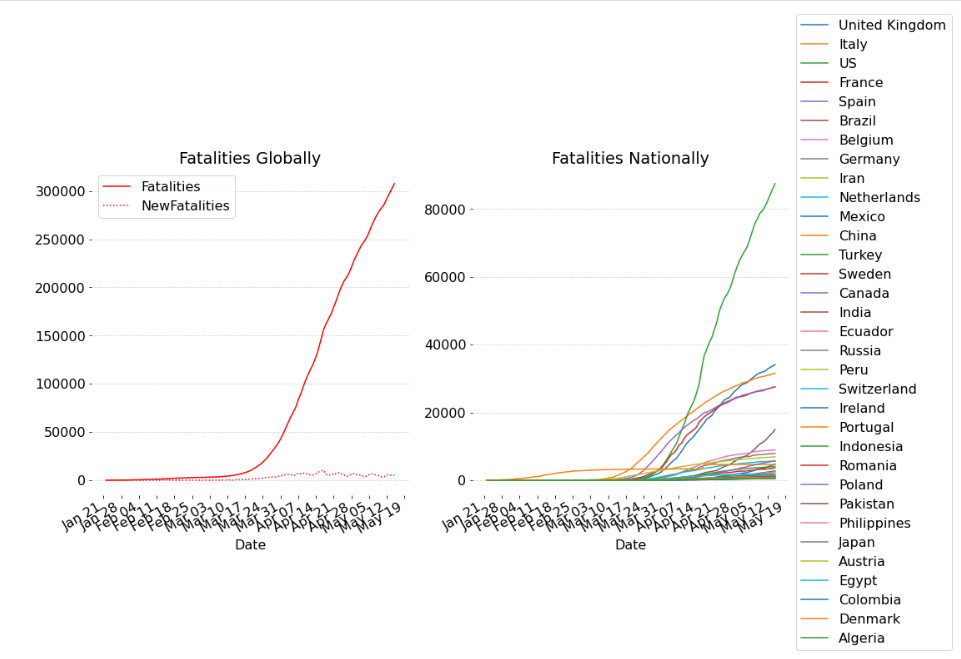


Figure 16

From the below figure, we can plot the mortality rate globally and on the 2nd subplot we have plotted with all the country regions average mortality rate on daily basis. This is one of the key findings during this data analysis.

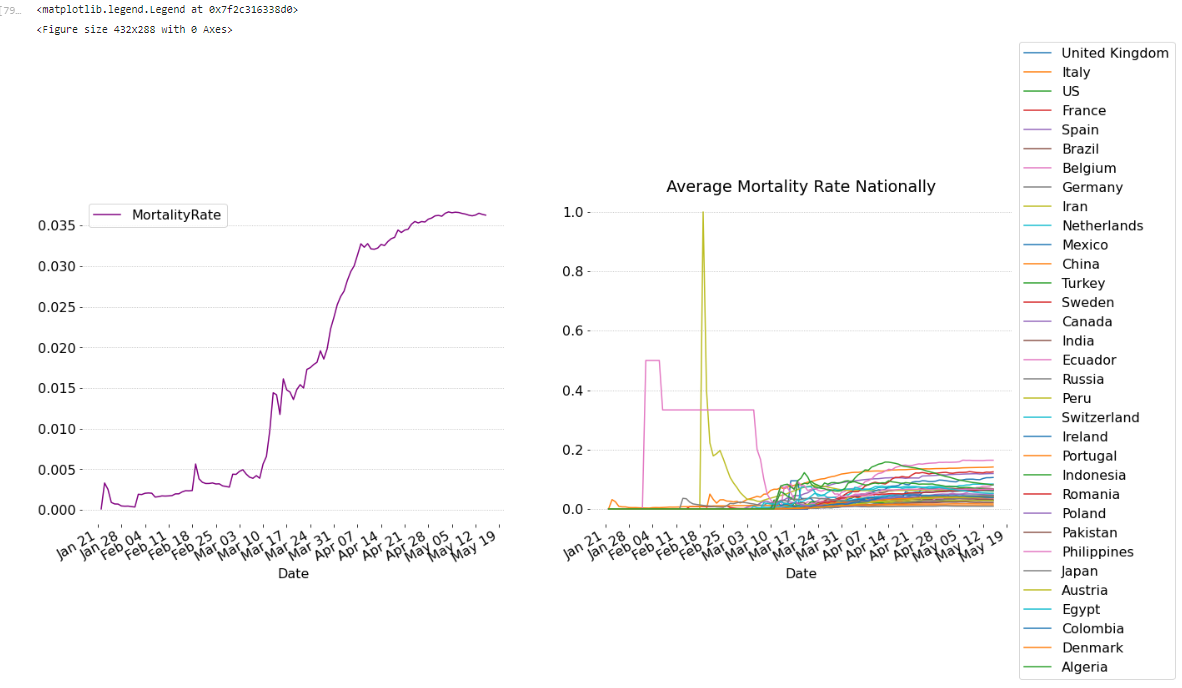


Figure 17

Taking about the key findings, in the below plot we have the top 20 countries with highest Mortality rate where country region is on Y-axis and mortality rate is on X-axis.

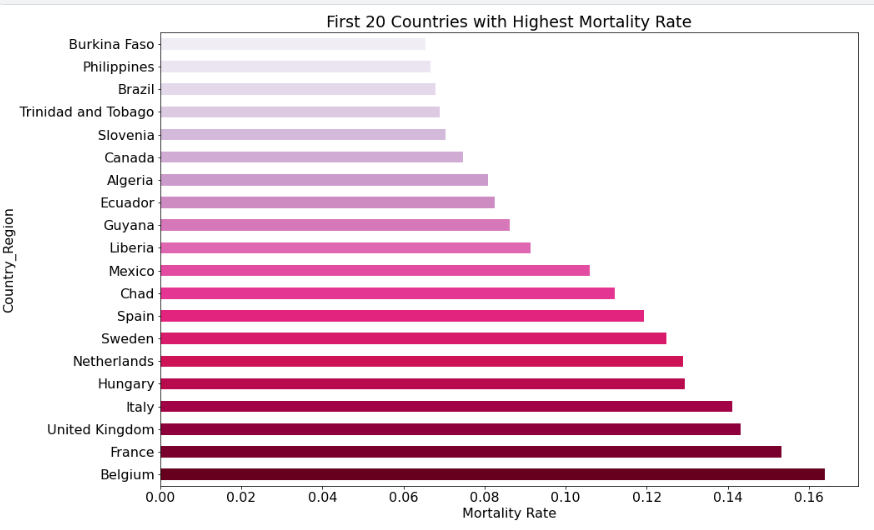


Figure 18

After all the data visualizations we can drop unwanted columns/features which are not helpful in building a model and train the data. So from train dataframe we dropped ‘Id’, ‘Province\_State’ columns.

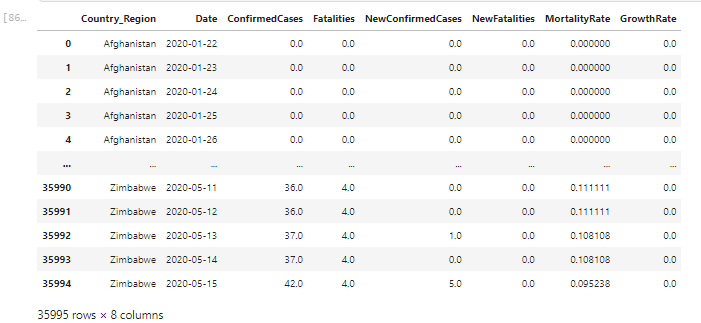


Figure 19

Below figure is the final submission file we got after prediction. This is the submission file we got with the XGBoost model.

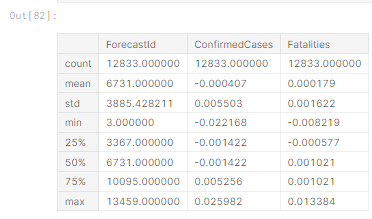


Figure 20

1. **Machine Learning Approaches**

Since there are many regression models which can be used for prediction, and I tried 4 machine learning approaches,

Linear Regression, Random Forest, Decision tree and XGboost.

Since they are 2 simple and 2 complex models which takes a lot of time to execute the code, I split 1 complex and 1simple model for each notebook. Decision Tree and Random Forest models in one notebook and Linear Regression and XGBoost in one.

Even after that it was taking 105 minutes almost 2 hours to execute each of them.

**XGBoost Model**

XGBoost is one of the well know ensemble learning algorithm which is based on Gradient Descent. It is an effective tool for regression and classification problems, particularly those with tabular data, where it outperforms other machine learning methods. For integration purposes the model uses decision tree for learning where the data is continuously split in batches to extend the tree such that each of the trees calculate threshold considering the features and comparing them with the best branch effect until the prediction results are harmonious with confirmed results.

We have used XGBoost to construct combination of features which makes every other leaf node of the decision tree is a new feature.

It is designed to be computationally efficient, so it can handle large datasets with ease. It also has the ability to handle missing values, which was very helpful for this dataset as one of the columns had null values.

With a variety of regularization techniques, which help reduce overfitting and also to improve accuracy, Xgboost works by sequentially applying and updating a set of weak learners on a given dataset.

We have also used for extracting features from training data such that the features can be used as input data for Linear Regression model

Label Encoder was used during this model to convert object data to numeric as XGBoost cannot handel categorical data.

The main advantages of XGBoost for regression include:

* High accuracy: XGBoost has shown superior performance on a variety of regression tasks.
* Fast training speed: XGBoost is an optimized algorithm that is highly optimized for speed, which allows it to train models faster than other algorithms.
* Robust and interpretable models: XGBoost models are very robust to noise and outliers and are also easy to interpret and explain.
* Scalable: XGBoost can be easily scaled to handle large datasets. It is also compatible with a wide range of hardware, including CPUs and GPUs.

**Linear Regression Model**

The Kaggle challenge I am attempting to solve is likely to be a supervised machine learning problem, which means I will need a linear regression model to solve it. Linear regression is a supervised learning algorithm used to predict a continuous target variable from a given set of feature variables. To build a linear regression model for the challenge, I will first need to identify the target variable and feature variables.

Once you have identified the target and feature variables, I will then begin to build my model. First, I will have to split the dataset into training and testing datasets, and use the training dataset to fit the linear regression model. I can then use the testing dataset to evaluate the accuracy of the model.

I have also used regularization techniques such as L1 and L2 regularization to reduce overfitting and improve the accuracy of the model. Additionally, I have used cross-validation to further improve the accuracy of the model.

Finally, one should be sure to evaluate the model using metrics such as mean absolute error, root mean square error, and r-squared. These metrics will help you understand how well your model is performing and whether it needs further improvement.

1. **Legal and ethical details**

* LEGAL- Open-source data is often publicly available, and as such, it is important to understand the applicable laws and regulations of the jurisdiction in which the data is collected and shared. The laws governing the use of open-source data can vary by jurisdiction and may include intellectual property, copyright, and data protection laws.
* ETHICAL: Open-source data may contain sensitive information that could be used to exploit or harm individuals, businesses, or organizations. Therefore, it is important to consider the ethical implications of using open-source data, such as the potential for privacy violations or the misuse of the data.
* PRIVACY: Open-source data may contain personal information about individuals, such as names, addresses, and other identifying information. As such, it is important to ensure that appropriate measures are taken to protect the privacy of individuals, such as anonymizing the data or using encryption. Additionally, data collected from open-source sources should be collected in a lawful manner and with the explicit consent of the individuals whose data is being collected.

GDPR

The UK has implemented the General Data Protection Regulation (GDPR) since 25th May 2018. This legislation was designed to give individuals more control over how their personal data is used, stored and processed by companies and organisations.

The UK GDPR applies to any organisation that processes personal data of individuals located in the UK, regardless of the organisation’s location. This means that any company that processes personal data of UK residents must comply with the GDPR.

In order to comply with GDPR, organisations must:

• Obtain informed consent from individuals before collecting any personal data.

• Inform individuals what personal data is being collected and how it will be used.

• Allow individuals to access and request the deletion of their personal data.

• Have appropriate security measures in place to protect the data.

• Notify the Information Commissioner's Office (ICO) of any data breaches that occur.

• Comply with the rights of individuals, including the right to rectification, the right to data portability, the right to object and the right to be forgotten.

1. **Kaggle Submission Proof**

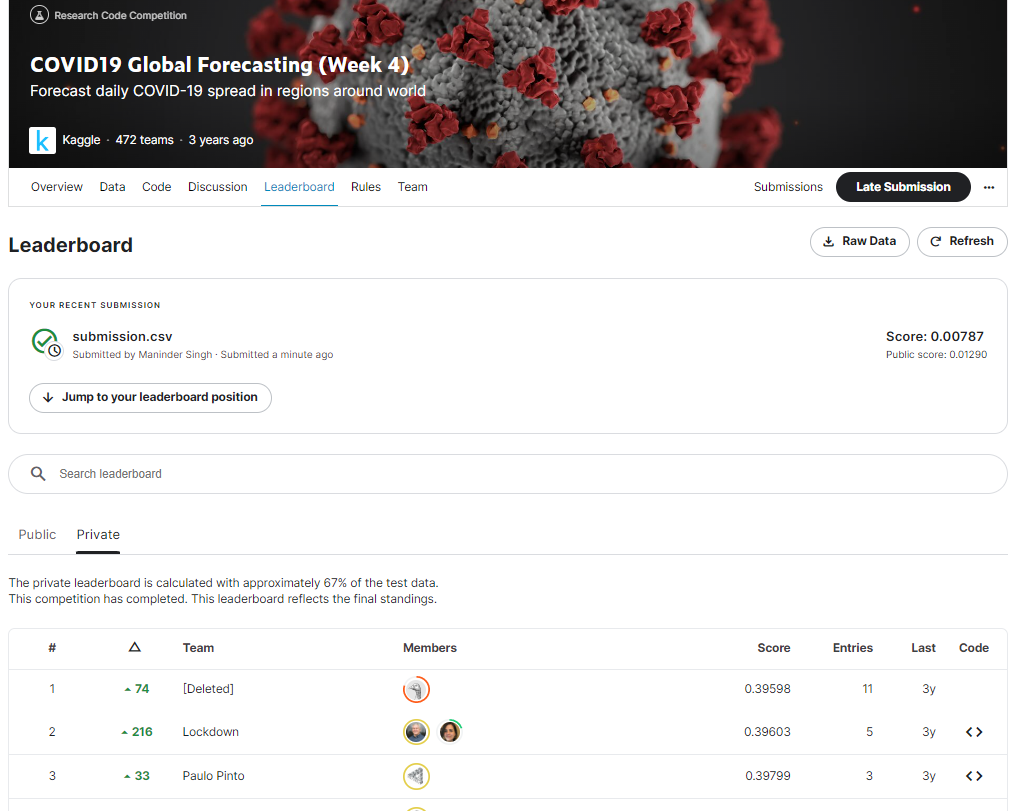


Figure 21

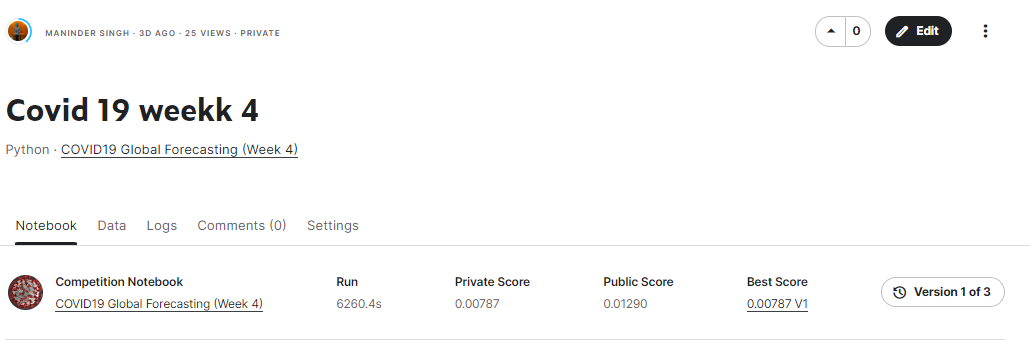


Figure 22

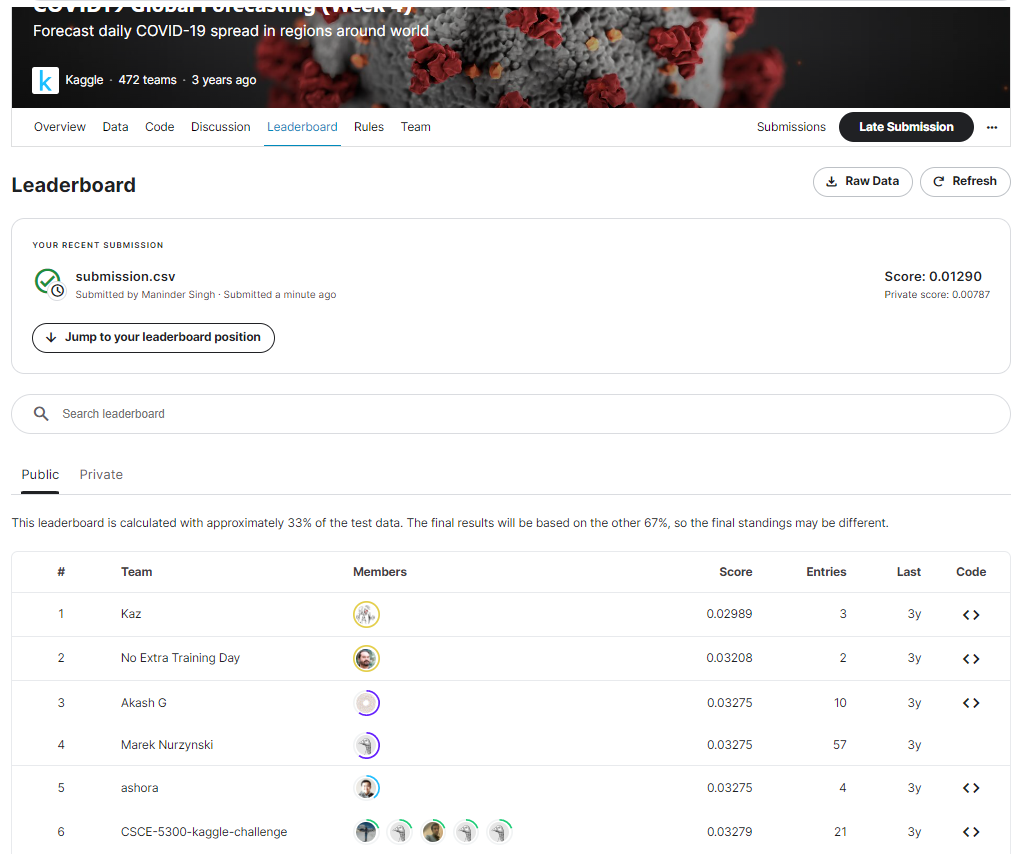


Figure 23

1. **Results & Future Work**

I have analysed the Covid-19(week 4) dataset and tried using two different models Linear Regression and XGBoost with different features wherein I tried creating new features also in the current notebook using shift index by desired number to find which model predicts the better results. And to conclude with training the model with data and predicting output XGBoost has better and accurate prediction values for the current dataset used. The best score achieved from the model is 0.00787

XGBoost regression models are more powerful and accurate than Logistic regression models because it can capture more complex non-linear relationships between features and the target variables, which allows them to make more accurate predictions.

In future I’ll be trying to predict with the increase in growth rate on everyday basis in different country regions so that the prediction using ensemble learning can get even better considering growth rate feature.

**References**

1. Cutler, D., 2020, April. How will COVID-19 affect the health care economy?. In *JAMA Health Forum* (Vol. 1, No. 4, pp. e200419-e200419). American Medical Association.
2. Akbulaev, N., Mammadov, I. and Aliyev, V., 2020. Economic impact of COVID-19. *Sylwan*, *164*(5).
3. Dong, E., Du, H. and Gardner, L., 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases*, *20*(5), pp.533-534.
4. Syeda, H.B., Syed, M., Sexton, K.W., Syed, S., Begum, S., Syed, F., Prior, F. and Yu Jr, F., 2021. Role of machine learning techniques to tackle the COVID-19 crisis: systematic review. *JMIR medical informatics*, *9*(1), p.e23811.
5. Sujath, R.A.A., Chatterjee, J.M. and Hassanien, A.E., 2020. A machine learning forecasting model for COVID-19 pandemic in India. *Stochastic Environmental Research and Risk Assessment*, *34*(7), pp.959-972.
6. "An interactive web-based dashboard to track COVID-19 in real time ...." <https://europepmc.org/article/MED/32087114>.
7. Dong, C., Qiao, Y., Shang, C., Liao, X., Yuan, X., Cheng, Q., Li, Y., Zhang, J., Wang, Y., Chen, Y. and Ge, Q., 2022. Non-contact screening system based for COVID-19 on XGBoost and logistic regression. *Computers in biology and medicine*, *141*, p.105003.
8. Nunez, E., Steyerberg, E.W. and Nunez, J., 2011. Regression modeling strategies. *Revista Española de Cardiología (English Edition)*, *64*(6), pp.501-507.
9. Lin, W., Wu, Z., Lin, L., Wen, A. and Li, J., 2017. An ensemble random forest algorithm for insurance big data analysis. *Ieee access*, *5*, pp.16568-16575.
10. Liu, Y., Wang, Y. and Zhang, J., 2012, September. New machine learning algorithm: Random forest. In *International Conference on Information Computing and Applications* (pp. 246-252). Springer, Berlin, Heidelberg.
11. Swetapadma, A. and Yadav, A., 2016. A novel decision tree regression-based fault distance estimation scheme for transmission lines. *IEEE transactions on power delivery*, *32*(1), pp.234-245.
12. Richards, N.M. and King, J.H., 2014. Big data ethics. *Wake Forest L. Rev.*, *49*, p.393.
13. Albrecht, J.P., 2016. How the GDPR will change the world. *Eur. Data Prot. L. Rev.*, *2*, p.287.