Dissertation Type : Research

-

Time Series Analysis of Mortality Rate of the COVID-19

By

Abhishek Singh

Supervisor

Prof. Nello Cristianini

Department of Engineering Mathematics

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Data Science (MSc) in the Faculty of Engineering.

Tuesday, 29 June 2021

Word count: 7738

Acknowledgement

I would like to thank my excellent supervisor for his great domain knowledge and consistent support.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ABHISHEK SINGH DATE: 29-06-2021

Contents

List of Figures

Abstract

Before closing their borders and locking down their citizens, many world leaders downplayed the new virus sweeping the globe. The first case of COVID-19 was reported in the central Chinese city of Wuhan in late 2019. The world hasn't been the same since. The disease quickly became a pandemic, with the death total currently over 3.3 million worldwide. It feels life is rapidly changing as a pandemic goes on; there are new rules put in place every other week. New variants are coming in day after day, and the most important thing we are still in it. Nobody was prepared for the last two years. The widespread adoption of electronic health records (EHRs) has enabled machine learning in patient healthcare. This has allowed researchers to predict the cases [8] and deaths [12] and the probability when the pandemic will end [3]. One of the issue with most of the early studies is that they have used a very small dataset. One of the first work to predict future cases was using a stacker of six different algorithms. [25] Their results were better than most of the ARIMA models used in [8], [1], [18], which tells us that we can use the Machine Learning algorithms efficiently to predict the time series problems. However, we believed we could obtain better results by applying more modern machine learning techniques. The dataset is taken from Our world in data,[1] which is maintained and updated daily by Johns Hopkins University. They have built 207 country profiles which allow us to explore the statistics on the coronavirus pandemic for every country in the world. It is not a simple matter to identify the most successful countries in making progress against it in a fast-evolving pandemic.[2] For a comprehensive assessment, we track the impact of the pandemic across our report, and we built country profiles for five countries to study in-depth the statistics on the coronavirus pandemic. The countries considered are the USA, UK, India, Canada and Italy. Each profile includes interactive visualizations, explanations of the presented metrics, and details on the data sources. We have predicted the Mortality Rate based on the population, ICU patients, deaths, cases and Reproduction Rate. We have tried several models, and the best four are mentioned here. The models are compared based on the R^2 score, Mean square error and the mean absolute error. We have developed software that processes EHRs so that they can be analyzed with Machine learning techniques. The software forms its predictions based mainly on the deaths and population. The software can take care of the missing values and preprocess the data when required.

The rest of the report is divided into Chapters. The first chapter is the Contextual Background that provides the Motivations of the project and the proposed system. The next is the Technical Background that provides technical content of the works related to this project and the information about the dataset in detail. The next is the Project Implementation which discusses the data pipeline and the Model construction. The next is the Project Results which discusses the R^2 score, Mean Absolute error (MAE) and Mean square error (MSE). The next is the critical evaluation, where there is a comparison between the models we have used. The final chapter is the conclusion which discusses the further work that is possible.

[1] https://ourworldindata.org
[2] https://ourworldindata.org/coronavirus

Chapter 1

Contextual Background

1.1 Introduction

Firstly, this chapter provides the Motivations of the project, demonstrating why it is an essential field of study. Secondly, a description of the proposed system is stated, and the developmental challenges are highlighted. Finally, the key objectives of the project are identified. These objectives form the critical requirements for the system and are used to evaluate its success in Chapter N.

1.2 Motivations

Viruses were one of the first living things on earth, but they are not alive. They need to hijack other living cells to reproduce, and that's their only goal, to survive and replicate themselves. The official name of this virus is SARS-CoV-2. 'COVID-19' is the name of the disease it causes, which stands for 'Coronavirus disease 2019'. Corona, as in 'crown'. The virus is named for its crown-like spikes. It spreads through droplets when we sneeze, cough, or speak and enter us directly through our eyes, nose, or mouth. The virus can also live on many surfaces for hours, so people can pick it up on their hands and infect themselves if they touch their face, something the average person does 20 times an hour. And you can be infected and spread it without any symptoms, or they can be mistaken for the flu. That's what makes this coronavirus so devious. [1]

1.2.1 This pandemic

Going about their life, a person with this coronavirus likely infects a couple other people, and each of those people infects a couple more, and so on, and so on, which is why the number of cases increases on an exponential curve, doubling every single days.

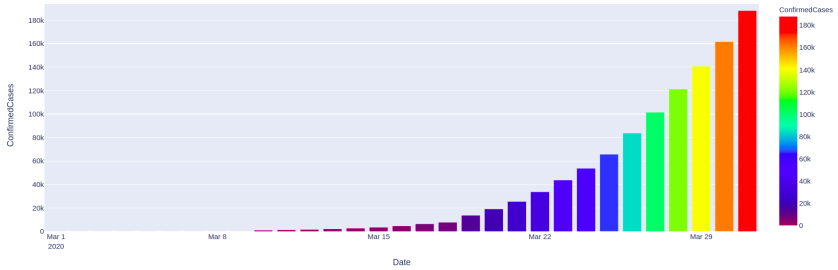

Figure 1.1: Confirmed COVID-19 cases per day in US

1.2.2 Misdiagnosis

A medical condition is considered to be misdiagnosed when the proposed line of treatment differs largely from the correct line of treatment. The misdiagnosis of conditions can have fatal consequences for patients and severe

[1] https://srgeyehospital.com/coronavirus-and-ocular-manifestation/

financial and legal consequences for medical practices. Misdiagnosis can arise when diagnostic tests falsely rule out correct diagnoses or when diagnostic tests produce incorrect diagnoses. [7] There are different kinds of tests-

1. Polymerase chain reaction (PCR), tests are used to directly screen the presence of viral RNA, which will be detectable in the body before antibodies form or symptoms of the disease are present. This means the tests can tell whether or not someone has the virus very early on in their illness. False negatives can occur up to 30% of the time with different PCR tests, meaning they're more useful for confirming the presence of an infection than giving a patient the all-clear. They can also provide false positive results, as they're so sensitive they can potentially signal a positive result upon detecting dead, deactivated virus still present in the body of someone who has recovered from Covid-19. These types of Covid-19 test need to be sent away to a laboratory for analysis, meaning it can take days for people to find out their results.
2. Lateral flow tests (LFTs), can diagnose Covid-19 on the spot, but aren't as accurate as PCR tests. The major benefit of LFTs over PCRs is that they do not need to be sent away for confirmation, and instead provide results within 15 to 30 minutes. However, what they gain in speed they sacrifice in accuracy. In [14], showed a wide variance in accuracy between different brands of LFT. The review also found that the tests were far better at identifying Covid-19 in people who had symptoms than those who did not. LFT sensitivity in symptomatic people ranged from 34% to 88%, with an average accuracy of 72%. In people without symptoms the LFTs correctly identified an average of 58% of those who were infected.
3. Antibody (or serology) tests can't diagnose active infection, but they can help to tell if a person has immunity to Covid-19. Wright says: "An antibody test tells us what proportion of the population has been infected. It won't tell you who is infected, because the antibodies are generated after a week or two, after which time the virus should have been cleared from the system. But it tells you who's been infected and who should be immune to the virus." In [26], the authors have mentioned that the people who recover from even mild cases of COVID-19 produce antibodies for at least five to seven months, and could do so for much longer. [2]

The Lateral flow tests are widely used as they give quick results. The risk of False Negative more in this, but when we have so many people to tests, this is the best option. Also, the PCR tests are pretty expensive, which many people cannot afford.

1.2.3 Early Identification

The conditions which can be identified early are usually ones that have been studied heavily. Thus, rarer conditions of which less is known are less likely to be identified early. On the other hand, a case can be made that the more common conditions affect more patients, hence making information about them more valuable. Ideally, rare diseases should be correctly diagnosed. A system that automatically calculates patients at risk of developing certain conditions is helpful in this case.

Methods used to identify early signs of conditions often include lab tests. Despite their inherent cost, these lab tests also require the patient's presence, which means that the time signs can be detected depending on when the patient visits the clinician. Remote testing can allow better chances for early identification.

1.2.4 Existing work

The authors of [15] have discussed several contemporary, innovative, and important models and discussed their possible applications. They have discussed the data preprocessing in details which are proved to be quite helpful for us. Time series algorithms like ARIMA and Deep AR forecasting is concerned with the limitations.

The authors of [4], have made a disease outbreak detection system using autoregressive moving average in time series analysis. The tool was developed for the Philippine setting, specifically for the use of outbreak monitoring agencies such as the National Epidemiological Center (NEC), and thus uses Philippine health data which basically comes from two major sources: surveys and censuses, as well as from administrative records of health and health related agencies.[3]

In [16], the authors have modelled the evolution of the COVID-19 outbreak and performed the prediction using the ARIMA and Prophet time series forecasting models. They used day level information of the COVID-19 outbreak of the top 10 most affected countries; US, Spain, Italy, France, Germany, Russia, Iran, United

[2] https://www.medicaldevice-network.com/features/types-of-covid-19-test-antibody-pcr-antigen/
[3] https://ieeexplore.ieee.org/document/7388087

Kingdom, Turkey, and India. They used data from January 22, 2020, to May 20, 2020. Their studies include the Confirmed cases, death cases, recovered cases and active cases for the adopted countries. The historical data depicts that the COVID-19 badly affected the countries which do not impose lockdowns or do not follow social distancing. Some variations in virus spread rate, recovery rate, and death rate can be seen in different countries based on population density, the available health system in a country, testing capability, and action taken to contain the outbreak. The prediction is made for the confirmed cases and death cases, and in both cases, it is observed that FBProphet prediction have high error factors than the ARIMA model. This model could have been accurate, but soon, a new variant of the COVID was detected, and then the rate of spread of the virus changed utterly. Also, as there is a vaccine now, fewer people are dying, so this study is not valid anymore. Even if we consider that the conditions would have stayed the same, they haven't added enough variables which are possible reasons for the spread of the virus, like population density, health system, patient history etc.

In [20], the authors used the Kaggle dataset from January 2020 to June 2020. They made the prediction using four methods, namely NAÏVE, Holt's linear trend method, Holt's Winter seasonal method, and ARIMA. They predicted the confirmed cases. The Naïve model is the best as the error in this model is less compared to the other models. The confirmed cases highly depend on the testing rate of the countries. Most countries are using the Lateral Flow test, which 50% of the time gives the wrong result. Also, countries like the United Kingdom does not do enough tests on Sundays, due to which the cases were very few on Sundays. The other model is India, which stayed on Lockdown on weekends for a very long time, resulting in fewer COVID cases on Mondays. The dataset is taken when most countries were in Lockdown, and some were going through the peak in their patients. Though this model will not be effective enough for predicting the new cases now, the good performance of the NAIVE model, especially against the ARIMA, opens the idea of using the NAIVE model in an ensemble model.

The [3] is one of the few studies which have used a deep learning approach. They have used ARIMA, CNNs and LSTM and compared them at the end. In this paper, the datasets used are obtained from the John Hopkins University's publicly available datasets to develop a state-of-the-art forecasting model of COVID-19 outbreak. They have incorporated data-driven estimations and time series analysis to predict the trends in coming days such as the confirmed cases, deaths and the recovered cases. In all cases, the LSTM performed better than the other two models. The problem with this paper is that they have not mentioned the amount of dataset used, which makes it difficult to judge the performance of ARIMA over LSTM, as we know that ARIMA requires a lot of data.

In [27], the authors have used Facebook Prophet on the World Health Organization data for COVID-19 to forecast the spread of COVID-19 for April 7 until May 3 2020. The forecast model was further used to forecast the trend of the virus for the 8^{th} until May 14 2020. They have predicted the confirmed cases and deaths. They got an accuracy of 79.6%. In [16], Facebook Prophet is used, which gave almost the same accuracy.

In [9], the authors have tried to highlight the importance of country lockdown and self-isolation in control the disease transmissibility among the Italian population through data-driven model analysis. They have adopted a seasonal ARIMA forecasting package with R statistical model. The data was taken from the Italian Health Ministry website, including the registered cases and the recovered cases. The dataset was just 45 days, which is from mid-February to the end of March. Predictions were made with 93.75% of accuracy for registered case models and 84.4% of accuracy for recovered case models. This was the time when Italy was the epicentre of the pandemic.

In [22], they have proposed a Bayesian optimization guided shallow LSTM for predicting the country-specific risk of COVID-19. They have used the trend data to predict different parameters for the risk classification task. Combining the overall optimized LSTMs, it can be seen that a shallow network performs better compared to deep learning networks. They have also analyzed the prediction performance combining the weather data. In many articles, it is claimed that the weather has a role in the Virus outbreak. Eventually, it was found to be False, and the authors also got the same results. They did multivariate time series, which was a good start, as, during this time, not enough data was available. But to check on any possible reason was a good option.

In [25], autoregressive integrated moving average (ARIMA), cubist regression (CUBIST), random forest (RF), ridge regression (RIDGE), support vector regression (SVR), and stacking-ensemble learning are evaluated in the task of time series forecasting with one, three, and six-days ahead the COVID-19 cumulative confirmed cases in ten Brazilian states with a high daily incidence. The CUBIST regression, RF, RIDGE, and SVR

models are adopted as base-learners and Gaussian processes (GP) as meta-learners in the stacking-ensemble learning approach. In most cases, the SVR and stacking-ensemble learning reach a better performance regarding adopted criteria than compared models. In general, the developed models can generate accurate forecasting, achieving errors in a range of 0.87%–3.51%, 1.02%–5.63%, and 0.95%–6.90% in one, three, and six-days-ahead, respectively. [4] This is one of the reasons why the Machine learning algorithms are not recommended for time series task. The better performance of a simple SVR model gives an option of using Machine Learning models for this task.

In [10], the authors have tried to predict the peak, duration and the attack rate (the percentage of the total population that will be infected throughout the outbreak) of COVID-19 outbreaks in the six Western countries of the Group of Seven, namely, Canada, France, Germany, Italy, UK and USA. They incorporated the governments' interventions (stay-at-home advises/orders, lockdowns, quarantines and social distancing) against COVID-19 into consideration. To identify the turning point and predict the further spread of COVID-19 outbreaks, they use the segmented Poisson model. They combined the power-law with the exponential law for the daily new cases based on a segmented Poisson model. They took the data from January to July'2020 from worldometer.[5]

In [13], the authors took temporal data from 22^{th} of January'2020 to 15^{th} of March'2020. They predicted the confirmed cases and recovered of China, Italy and France using the SIRD model. Their studies tell that the recovered are related to the cultural cases than the country. The confirmed cases were easy to predict using SIRM, but the recovered gave bad results, and there was no relation between the countries. The SIRM is used in [6] as well for the prediction of Italy cases. They worked on the data when there was Lockdown which predicted the cases would eventually decrease. They compared the results with the ARIMA model. The observations they took was just for a month, and as we know that ARIMA requires atleast 50 observations to give better results. [17].

In [21], they developed a statistical model to analyze the role of temperature and humidity in the modulation of the doubling time of COVID-19 cases. The statistical model developed was implemented in two steps: firstly, an exponential model relating the accumulated number of confirmed cases and time was considered. Secondly, the rate of spread was used as the dependent variable in a linear model that took as independent variables temperature, humidity, precipitation and wind speed. Results suggest that temperature correlates positively with the doubling time and negatively with humidity. This means that, with spring and summer, the rate of progression of COVID-19 is expected to be slower. Still, these two variables contribute at maximum to 18% of the variation, being the remaining 82% related to other factors such as containment measures, general health policies, population density, transportation, cultural aspects, etc. Besides, the direct impact is also small: for example, if the temperature raises 20°C, it is expected that the doubling time increases on average 1.8 days in the best-case scenario.[6] This is a study which was published on March 8'2020. We know after so much research that now there is no relationship with temperature.

> This virus demonstrates no seasonal pattern as such so far. What it clearly demonstrates is that if you take the pressure off the virus —the virus bounces back. That's the reality, that's the fact.
>
> – Dr. Michael Ryan, WHO press briefing 8/10/2020

The authors of [19] collected the daily reported cumulative number of infected cases and deaths from January 30 to December 20, 2020, from the COVID-19-India API website[7] State-level data for the total number of confirmed cases were collected from March 14 to December 20 2020. The 15-days-ahead forecast of COVID-19 for India was generated using four different methods, the exponential growth model, logistic growth model, Gompertz model and ARIMA model. They took 95% of the confidence level of Interval. A 95% confidence interval is a range of values that you can be 95% certain contains the population's true mean. This is not the same as a range that has 95% of the values. [8] The results of the models were analyzed using MAPE and RMSE. The four models show that ARIMA fitted values nearly coincide with the actual reported values (infections and deaths) from January 30 to December 20 2020, defining a better fit of the forecast using the ARIMA model. Thus the ARIMA model was employed for forecasting the cumulative number of infected cases at the regional level. They mentioned that the number of cases is directly related to an increase in the number of testing

[4] https://blogs.oracle.com/ai-and-datascience/post/7-ways-time-series-forecasting-differs-from-machine-learning/
[5] https://www.worldometers.info/
[6] https://www.medrxiv.org/content/10.1101/2020.03.05.20031872v1.full
[7] https://api.covid19india.org/documentation/csv/
[8] https://www.graphpad.com/guides/prism/latest/statistics/stat_more_about_confidence_interval.htm

facilities and the interstate movement of people.

In [28], the authors have predicted the cases in Italy for the next ten days. This study uses the complete periodic data of developing the COVID-19 epidemic in Hubei, China, to establish the ARIMA models. Though the population of Hubei and Italy may be the same, the conditions in both countries were very different. Italy was the first country that went short of Hospital beds, whereas, on the other hand, China was well prepared for the pandemic. The test in Italy may be calculated as correct, but the death data is crucial as most people were dying in their homes.

1.3 Proposed system

Building on the work in [25] and the ARIMA model, this project aims to apply modern machine learning techniques to the COVID-19 data available on Our World in Data[9] to identify and make predictions of the Mortality Rate of five countries namely, the US, UK, Canada, India and Italy. By empirically comparing multiple models in terms of their forecasting accuracy, we intend to suggest an appropriate model, which can be readily used by society, organizations, or governments to assess near futures of this outbreak.

1.3.1 Challenges

We acknowledge that this is a difficult forecasting problem, since this pandemic continues and there are many factors that we cannot presently control.

1.3.1.1 Actual Cases

Despite very massive testing, it is impossible to test the whole population. Hence we do not know the actual infections. We only know that the confirmed cases are a fraction of existing conditions. To estimate existing infections and other vital metrics, several research groups have developed epidemiological models of COVID-19. These models use the data like confirmed cases and deaths, testing rates, and more – plus a range of assumptions and epidemiological knowledge. [10]

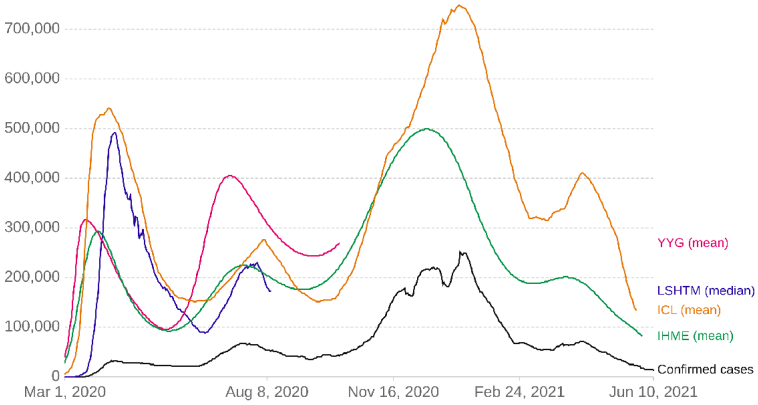

Figure 1.2: Mean of estimated true infections in the US

The above Figure shows the mean estimates of the true number of daily infections in the United States from four of the most prominent models.

1. Imperial College London (ICL)
2. The Institute for Health Metrics and Evaluation (IHME)
3. Youyang Gu (YYG)
4. The London School of Hygiene Tropical Medicine (LSHTM)

Two things are clear from this chart: All four models agree that true infections far outnumber confirmed

[9] https://ourworldindata.org/coronavirus
[10] https://srgeyehospital.com/coronavirus-and-ocular-manifestation/

cases. But the models disagree by how much, and how infections have changed over time. [11]

When the number of confirmed cases in the US reached a peak in late July 2020, the IHME and LSHTM models estimated that the true number of infections was about twice as high as confirmed cases, the ICL model estimated it was nearly three times as high, and Youyang Gu's model estimated it was more than six times as high. Back in March the estimated discrepancy between confirmed cases and true infections was even many times higher.

1.3.1.2 Actual Tests

The number of tests does not refer to the same in each country – one difference is that some countries report the number of people tested, while others report the number of tests performed (which can be higher if the same person is tested more than once). And other countries report their testing data in a way that leaves it unclear what the test count refers to exactly.[12]

1.3.1.3 Rare Conditions

Rare conditions are more challenging for GPs to identify due to a combination of the lack of knowledge around them and the fact that they appear less frequently. This may propagate into the prediction system in several ways. Firstly, the lack of diagnosed patients will mean that there will be few records to learn from. Secondly, misdiagnosis is higher for rarer diseases, and if this is encoded in the prediction will again be more challenging.

1.3.2 Why Mortality Rate?

Mortality rate or death rate is a measure of the number of deaths (in general, or due to a specific cause) in a particular population, scaled to the size of that population, per unit of time. [5] Most of the early work is done to predict the confirmed cases, but it is inefficient to do so, as it is impossible to know the actual cases. The only solution to this problem is to tests the whole population in a day. There are some models, as shown in section 1.2, but there is no model to measure the truth. Most of the countries increased their hospital facilities so that the vulnerable people can be given treatment immediately. Hence, the deaths reported can be more accurate, as the hospitals provide this information. One model is of the United Kingdom. They managed their facilities that they never got the peak when they outrun the hospital facilities. The other model is of India, where they outrun the hospitals in their second wave. India's actual covid toll is expected to be around 26% more than the official.[13] The best that we can get is the deaths. This is the reason we will look at the Mortality Rate.

1.3.3 Objectives

To produce a system that predicts the Mortality Rate of different countries, we produced a list of objectives. We have predicted the Mortality Rate as this is the best

1. The system should be open source software that processes the deaths, confirmed cases, population, ICU patients of the specific country.
2. The system should operate primarily based on deaths and not require any identifying feature (for privacy considerations)
3. It should show that the system scales evenly and predictably with data magnitude.
4. The results should improve on those provided in [25]

[11] https://ourworldindata.org/covid-models
[12] https://ourworldindata.org/coronavirus-testing#the-positive-rate-a-crucial-metric-for-understanding-the-pandemic
[13] https://www.nytimes.com/interactive/2021/05/25/world/asia/india-covid-death-estimates.html

Chapter 2

Technical Background

2.1 Introduction

This chapter provides the background required to understand the technical content of the works related to this project. First is the statistical background, which introduces raised condition likelihoods and provides definitions to understand the analysis and results. Following is the machine learning background. Next to it are the Time series models are discussed. Finally, deep learning models are discussed.

2.2 Statistical Background

2.2.1 Introduction

In order to calculate the performance of the model, some statistical definitions are required.

2.2.2 Regression

In statistical modelling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables.[1] A system that performs regression is known as a regressor. The model needs to be evaluated. It helps you understand your model's performance and makes it easy to present your model to other people. In this context, the regressor's purpose is to predict the Mortality Rate. The metrics used for model evaluation are : R^2, Root Mean Square Error (MSE) and Mean Absolute Error (MAE)

2.2.3 R Square (R^2)

R Square measures how much variability in dependent variable can be explained by the model. It is the square of the Correlation Coefficient(R) and that is why it is called R Square.[2]

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \widehat{y_i})^2}{\sum_i (y_i - \widehat{y})^2} \tag{2.1}$$

R Square is calculated by the sum of squared of prediction error divided by the square's total sum, which replaces the calculated prediction with mean. R Square value is between 0 to 1, and a more significant value indicates a better fit between prediction and actual value. [11]

2.2.4 Mean Square Error (MSE)

While R Square is a relative measure of how well the model fits dependent variables, Mean Square Error is an absolute measure of the goodness for the fit.

[1] https://www.erim.eur.nl/necessary-condition-analysis

[2] https://towardsdatascience.com/what-are-the-best-metrics-to-evaluate-your-regression-model-418ca481755b

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2 \tag{2.2}$$

Mean Square Error is calculated by the sum of the square of prediction error, which is actual output minus predicted output and then divide by the number of data points. It gives you an absolute number on how much your predicted results deviate from the actual number. One cannot interpret many insights from one result, but it gives an actual number to compare against other model results and help you select the best regression model. [8]

2.2.5 Mean Absolute Error (MAE)

Mean Absolute Error(MAE) is similar to Mean Square Error(MSE). However, instead of the sum of square of error in MSE, MAE is taking the sum of the absolute value of error.[3]

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y_i}| \tag{2.3}$$

Compare to MSE or RMSE, MAE is a more direct representation of sum of error terms. MSE gives larger penalization to big prediction error by square it while MAE treats all errors the same. [8]

2.3 Machine Learning models

2.3.1 Introduction

Methods which feature in the development of this system include RandomForest Regressor, LASSO, Linear Regression and XGBoost. Some prerequisite knowledge to understand the machine learning techniques below is described in the following definitions.

2.3.2 Machine Learning

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed.[4]

2.3.3 Supervised Learning

Supervised learning is a machine learning technique with input variables (X) and an output variable (Y). We use an algorithm to learn the mapping function from the input to the output.[5]

$$Y = F(X) \tag{2.4}$$

The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process.[6] We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance.

[3] https://towardsdatascience.com/what-are-the-best-metrics-to-evaluate-your-regression-model-418ca481755b
[4] https://www.geeksforgeeks.org/machine-learning/
[5] https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/
[6] https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

2.3.4 Unsupervised Learning

Unsupervised learning is a type of algorithm that learns patterns from untagged data. The hope is that, through mimicry, the machine is forced to build a compact internal representation of its world and then generate imaginative content.

2.3.5 Ensemble Learning

Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in order to decrease variance (bagging), bias (boosting), or improve predictions (stacking).[7]

2.3.6 RandomForest Regression

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max_samples parameter if bootstrap = True (default), otherwise the whole dataset is used to build each tree.[8]

2.3.7 Linear Regression

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

$$y = b + \sum_{i=1}^{n} x_i \cdot w_i \tag{2.5}$$

where,
y = output
x_i = the inputs
n = number of data points

2.3.8 Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. This type of regression is useful when we want to automate certain parts of model selection, like variable selection/parameter elimination. The goal of algorithm is to minimize :

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2.6}$$

where,
λ = amount of shrinkage

2.3.9 XGBoost

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. Gradient boosting is a machine learning technique for regression, classification and other tasks, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. [23]When a decision tree is the weak learner, the resulting algorithm is called gradient boosted trees, which usually outperforms random forest. [23] [24] It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.[9]

[7] https://blog.statsbot.co/ensemble-learning-d1dcd548e936
[8] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
[9] https://en.wikipedia.org/wiki/Gradient_boosting

2.4 Time series models

2.4.1 Introduction

Time series forecasting can be framed as a supervised learning problem. This re-framing of your time series data allows you access to the suite of standard linear and nonlinear machine learning algorithms on your problem.[10] The time series algorithms used are ARIMA, SARIMA and DeepAR Forecasting algorithm.

2.4.2 Autoregressive Integrated Moving Average (ARIMA)

ARIMA is a statistical analysis of model that uses time series data to either better understand the dataset or to predict the future trends. It makes use of lagged moving averages to smooth time series data.

2.4.3 Seasonal Autoregressive Integrated Moving Average (SARIMA)

It is an extension to ARIMA that supports the direct modelling of the seasonal component of the series is called SARIMA. A seasonal ARIMA model is formed by including additional seasonal terms in the ARIMA. The seasonal part of the model consists of terms that are very similar to the non-seasonal components of the model, but they involve backshifts of the seasonal period. [1]

2.4.4 DeepAR Forecasting Algorithm

The Amazon SageMaker DeepAR forecasting algorithm is a supervised learning algorithm for forecasting scalar (one-dimensional) time series using recurrent neural networks (RNN). Classical forecasting methods, such as autoregressive integrated moving average (ARIMA) or exponential smoothing (ETS), fit a single model to each individual time series. They then use that model to extrapolate the time series into the future.[11]

2.5 Deep learning models

2.5.1 Introduction

Two deep learning models are used for this project, namely LSTM and ConvLSTM1. They are defined here.

2.5.2 Long short-term memory (LSTM)

Long short-term memory (LSTM) networks4 are recurrent neural networks (RNN) widely used in deep learning. LSTMs were designed to process sequences of data and improved upon traditional RNN by using memory cells that can store information in memory for long sequences, and a set of gates to control the flow of this memory information. These innovations allow LSTM to learn longer term dependencies in sequential data. [2]

2.5.3 Convolutional LSTM (ConvLSTM)

Convolutional LSTM architecture allows convolutions at the gates of the LSTM to capture spatio-temporal patterns. Although it learns spatio-temporal patterns similar to CNN LSTM, the manner in which it does so is different (by using convolutions at the gates).[12]

[10] https://machinelearningmastery.com/time-series-forecasting-supervised-learning/
[11] https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html
[12] https://arxiv.org/pdf/2006.13852.pdf

Bibliography

[1] J. Armstrong. How to make better forecasts and decisions: Avoid face-to-face meetings. *Foresight: The International Journal of Applied Forecasting*, 5:3–15, 09 2006.

[2] Arko Barman. Time series analysis and forecasting of COVID-19 cases using LSTM and ARIMA models. *CoRR*, abs/2006.13852, 2020.

[3] Suraj Bodapati, Harika Bandarupally, and M Trupthi. Covid-19 time series forecasting of daily cases, deaths caused and recovered cases using long short term memory networks. In *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, pages 525–530, 2020.

[4] Richard John M. Buendia and Geoffrey A. Solano. A disease outbreak detection system using autoregressive moving average in time series analysis. In *2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–5, 2015.

[5] Adriana Burlea-Schiopoiu and Koudoua Ferhati. The managerial implications of the key performance indicators in healthcare sector: A cluster analysis. *Healthcare*, 9(1), 2021.

[6] Giuseppe C. Calafiore, Carlo Novara, and Corrado Possieri. A time-varying sird model for the covid-19 contagion in italy. *Annual Reviews in Control*, 50:361–372, 2020.

[7] M. Carrington, U. Choe, S. Ubillos, D. Stanek, M. Campbell, L. Wansbrough, P. Lee, G. Churchwell, K. Rosas, S. R. Zaki, C. Drew, C. D. Paddock, M. DeLeon-Carnes, M. Guerra, A. R. Hoffmaster, R. V. Tiller, and B. K. De. Fatal case of brucellosis misdiagnosed in early stages of brucella suis infection in a 46-year-old patient with marfan syndrome. *Journal of Clinical Microbiology*, 50(6):2173–2175, 2012.

[8] Tianfeng Chai and R. Draxler. Root mean square error (rmse) or mean absolute error (mae)? *Geosci. Model Dev.*, 7, 01 2014.

[9] Nalini Chintalapudi, Gopi Battineni, and Francesco Amenta. Covid-19 virus outbreak forecasting of registered and recovered cases after sixty day lockdown in italy: A data driven model approach. *Journal of Microbiology, Immunology and Infection*, 53(3):396–403, 2020.

[10] Yagmur Cinar, Hamid Mirisaee, Parantapa Goswami, Eric Gaussier, Ali Ait-Bachir, and Vadim Strijov. Time series forecasting using rnns: an extended attention mechanism to model periods and handle missing values. 03 2017.

[11] A. Colin Cameron and Frank A.G. Windmeijer. An r-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, 77(2):329–342, 1997.

[12] Duccio Fanelli and Francesco Piazza. Analysis and forecast of covid-19 spreading in china, italy and france. *Chaos, Solitons Fractals*, 134:109761, 05 2020.

[13] Duccio Fanelli and Francesco Piazza. Analysis and forecast of covid-19 spreading in china, italy and france. *Chaos, Solitons Fractals*, 134:109761, 05 2020.

[14] Shaun Griffin. Covid-19: Lateral flow tests are better at identifying people with symptoms, finds cochrane review. *BMJ*, 372, 2021.

[15] Mirjana Ivanović and Vladimir Kurbalija. Time series analysis and possible applications. In *2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 473–479, 2016.

[16] Naresh Kumar and Seba Susan. Covid-19 pandemic prediction using time series forecasting models. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–7, 2020.

[17] Ariel Linden. What should be the minimum number of observations for a time series model?, 03 2015.

[18] Martin Längkvist, Lars Karlsson, and Amy Loutfi. A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters*, 42:11–24, 2014.

[19] Sherry Mangla, Ashok Kumar Pathak, Mohd Arshad, and Ubydul Haque. Short-term forecasting of the COVID-19 outbreak in India. *International Health*, 06 2021. ihab031.

[20] Sujeet Maurya and Shikha Singh. Time series analysis of the covid-19 datasets. In *2020 IEEE International Conference for Innovation in Technology (INOCON)*, pages 1–6, 2020.

[21] B Oliveiros, L Caramelo, N C Ferreira, and F Caramelo. Role of temperature and humidity in the modulation of the doubling time of covid-19 cases. *medRxiv*, 2020.

[22] Ratnabali Pal, Arif Ahmed Sekh, Samarjit Kar, and Dilip K. Prasad. Neural network based country wise risk prediction of covid-19. *Applied Sciences*, 10(18):6448, Sep 2020.

[23] S. Madeh Piryonesi and Tamer E. El-Diraby. Data analytics in asset management: Cost-effective prediction of the pavement condition index. *Journal of Infrastructure Systems*, 26(1):04019036, 2020.

[24] S. Madeh Piryonesi and Tamer E. El-Diraby. Using machine learning to examine impact of type of performance indicator on flexible pavement deterioration modeling. *Journal of Infrastructure Systems*, 27(2):04021005, 2021.

[25] Matheus Henrique Dal Molin Ribeiro, Ramon Gomes da Silva, Viviana Cocco Mariani, and Leandro dos Santos Coelho. Short-term forecasting covid-19 cumulative confirmed cases: Perspectives for brazil. *Chaos, Solitons Fractals*, 135:109853, Jun 2020.

[26] Tyler J. Ripperger, Jennifer L. Uhrlaub, Makiko Watanabe, Rachel Wong, Yvonne Castaneda, Hannah A. Pizzato, Mallory R. Thompson, Christine Bradshaw, Craig C. Weinkauf, Christian Bime, Heidi L. Erickson, Kenneth Knox, Billie Bixby, Sairam Parthasarathy, Sachin Chaudhary, Bhupinder Natt, Elaine Cristan, Tammer El Aini, Franz Rischard, Janet Campion, Madhav Chopra, Michael Insel, Afshin Sam, James L. Knepler, Andrew P. Capaldi, Catherine M. Spier, Michael D. Dake, Taylor Edwards, Matthew E. Kaplan, Serena Jain Scott, Cameron Hypes, Jarrod Mosier, David T. Harris, Bonnie J. LaFleur, Ryan Sprissler, Janko Nikolich-Žugich, and Deepta Bhattacharya. Orthogonal sars-cov-2 serological assays enable surveillance of low-prevalence communities and reveal durable humoral immunity. *Immunity*, 53(5):925–933.e4, 2020.

[27] Akpojoto Siemuri, Rasheed Omobolaji Alabi, and Mohammed Elmusrati. Covid-19: Easing the coronavirus lockdowns with caution. *medRxiv*, 2020.

[28] Qiuying Yang, Jie Wang, Hongli Ma, and Xihao Wang. Research on covid-19 based on arima model—taking hubei, china as an example to see the epidemic in italy. *Journal of Infection and Public Health*, 13(10):1415–1418, 2020.

Appendix A

Project Plan (PP)

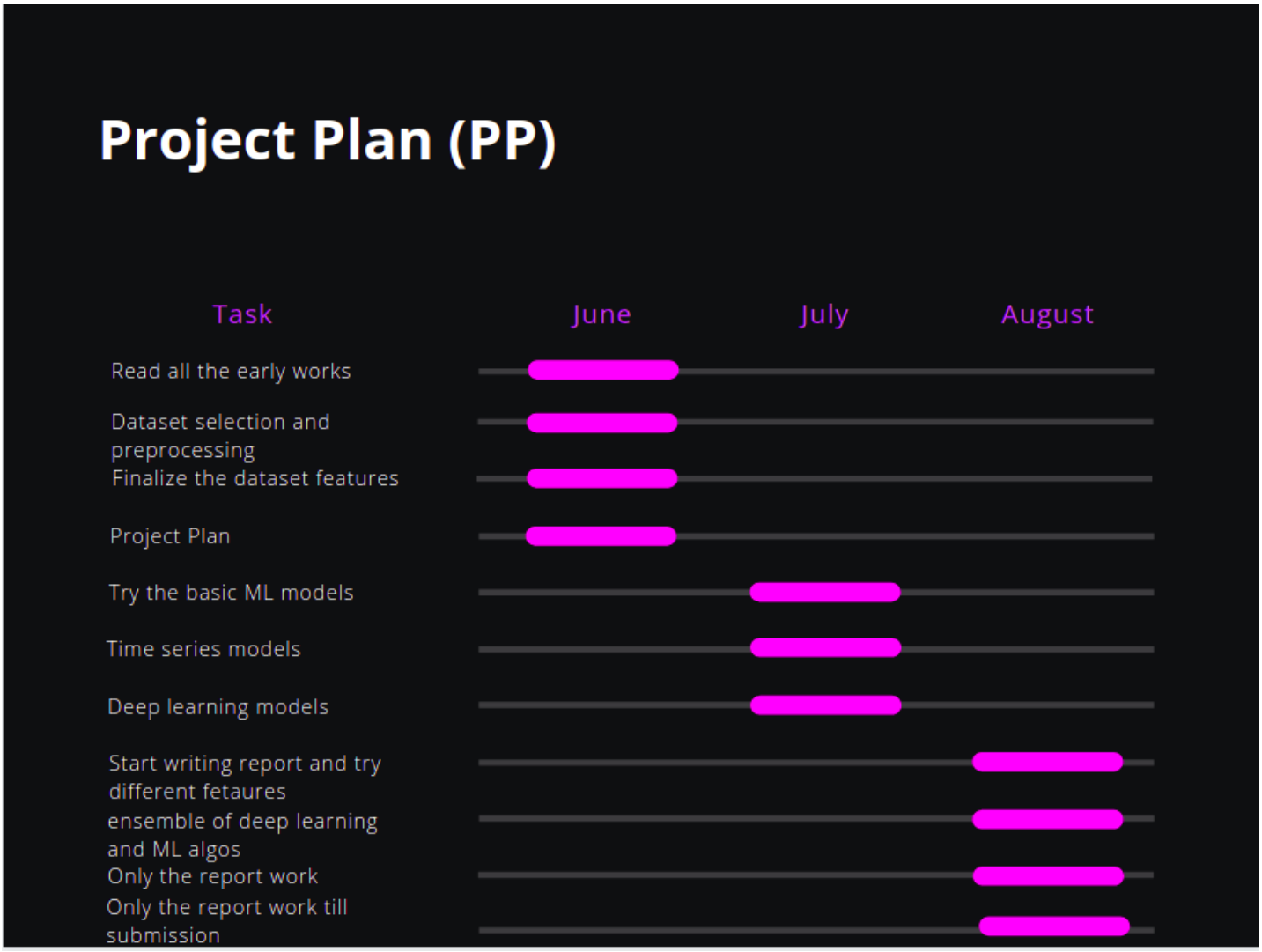

Figure A.1: Weekly Task

Appendix B

Risk Assessment

Risk Event	Probability	Risk Response
Many researchers in the early studies made prediction considering the cases and the deaths. However, these all things changed as most of the countries faced new variants.	70%	This problem has already been discussed with Professor Nello, and there is not much we can do about this. So, considering this, we have taken five countries, out of which two are in international lockdown. Therefore, it is less likely to get a new variant in these two countries. India is taken as a more significant challenge, as this is the most challenging country to make a prediction.
It has been claimed by many countries that China is hiding their cases. China is not on our list, but the same things can be done by others as well.	30%	The countries do this in fear that others might stop international trade, which can impact their economy. In this case, if we get any reliable evidence, we will eliminate that country from the time series. The elimination of any country can be the worst scenario as this will lead to the deletion of atleast a chapter and a lot of content in the middle. The other thing that we can do is make a prediction based on the data provided by them, but this will increase the chances of more errors in the prediction of that particular country. The decision depends on how close the submission date is.
Delay in making the code, in case a lot of debugging is required	80%	This will be a significant problem, as the plan I have in mind is to apply the deep learning algorithms, which are challenging to implement. The best possible solution for this problem can be to avoid the Deep learning models. This is less likely that I will fail to implement the Machine learning and Time-series algorithms as they are easy to implement, and most of the codes are made using the sklearn library.
The change in the pattern of the datasets on Our world in data	20%	The plan is to make a code that can be used by directly downloading the dataset, and then it should work by changing just one line of the code. The preprocessing steps are in the code, but they deleted a column of the Recovered patients in the early days because it was not reliable anymore as the patients were advised to stay in-home quarantine. Hence, this information was impossible to calculate. In this case, the early work of prediction of recovered cases became not applicable. In this case, we will need to divide the data we are working on and check for the predictions. Hence, we will not consider the cases after June 4th.
Time, cost and scope deviation to be expected	40%	One of the algorithms I am using is taken from Amazon Web services which are not costly to implement. But to get comfortable with this might take a bit of time, resulting in the delay of other works. So, I will stop this work in this case, and I will concentrate on the additional work first.