Christopher McGahon

s2102466

BDL CW2

Table of Contents

[Introduction 2](#_Toc61868378)

[Task 1 2](#_Toc61868379)

[Creating the models 2](#_Toc61868380)

[Linear Regression 2](#_Toc61868381)

[Create Model 2](#_Toc61868382)

[Logistic Regression 3](#_Toc61868383)

[Pre-Processing 3](#_Toc61868384)

[Create Model 5](#_Toc61868385)

[The domains in which the dataset might be used or useful 5](#_Toc61868386)

[Any 'digitized' business/enterprise processes, real or suggested, into which the ML 6](#_Toc61868387)

[models might be infused 6](#_Toc61868388)

[The value that predictions using the ML models will bring 6](#_Toc61868389)

[Task 2 7](#_Toc61868390)

[Linear Regression 7](#_Toc61868391)

[Model Metrics 7](#_Toc61868392)

[Mean-absolute-error (MEA) 7](#_Toc61868393)

[Root-mean-square error (MSE) 8](#_Toc61868394)

[Root-mean-squared logarithmic error (RMSE) 8](#_Toc61868395)

[Explained variance (explained variation) 8](#_Toc61868396)

[Explain ability 8](#_Toc61868397)

[Inclusivity 8](#_Toc61868398)

[Logistic Regression 8](#_Toc61868399)

[Model Metrics 9](#_Toc61868400)

[Precision 9](#_Toc61868401)

[Recall 9](#_Toc61868402)

[Accuracy 9](#_Toc61868403)

[F1 score 9](#_Toc61868404)

[Log loss 9](#_Toc61868405)

[AUC ROC 9](#_Toc61868406)

[Explain ability 9](#_Toc61868407)

[Inclusivity 9](#_Toc61868408)

[Task 3 10](#_Toc61868409)

[Linear Regression 10](#_Toc61868410)

[Logistic Regression 10](#_Toc61868411)

[Conclusion 10](#_Toc61868412)

[References 11](#_Toc61868413)

# Introduction

For this project we will use regression functions to model the growth of confirmed Covid-19 related deaths as well as the percentile growth rate of cases on any given day. The regression functions are commonly used in classification problems, and in this project, we will examine how they fare as prediction tools. Each cumulative case counts over time and logistic regression curves have a sigmoid shape. We shall try to fit a theoretically predicted curve over the actual cumulative case counts over time to reach certain conclusions about the case count growth. For example, the time of peak daily new cases and the total cases that may be reached during this outbreak. In a typical machine learning problem, one generally looks for datasets, utilizes Jupyter Notebook or Spyder to import the data, train, and test the model requiring wide-ranging programming and knowledge of machine learning frameworks. However, Google’s BigQuery ML makes it very simple. BigQuery ML enables data analysts to use machine learning through existing SQL tools and techniques (Google, 2020). It minimizes the task of importing huge datasets. With the help of simple SQL queries and Data Studio, one can visualize the data in a matter of seconds and can model the data using the CREATE MODEL statement, by specifying the MODEL\_TYPE. Once the model is trained, it can then be predicted using ML.PREDICT and evaluated using ML.EVALUATE.

# Task 1

Using your existing allocated dataset, create two models, one a Linear Regression model and the other a Logistic Regression model. Your decision on which specific ML models to create should be informed by consideration of any of the following

# Creating the models

## Linear Regression

### Create Model

CREATE or REPLACE MODEL `strategic-howl-293309.BDL\_CW2.Linear\_Model`

OPTIONS

  (model\_type='linear\_reg',

    input\_label\_cols=['deaths'],

    data\_split\_method='seq',

     data\_split\_eval\_fraction = 0.1,

    data\_split\_col = 'numDay') AS

SELECT

  date,

  numDay,

  country\_name,

  deaths,

  confirmed\_cases,

FROM

  `strategic-howl-293309.BDL\_CW2.Linear\_Model\_Data`

WHERE

  confirmed\_cases IS NOT NULL and

  deaths IS NOT NULL and

  deaths > 0              and

  confirmed\_cases > 0

Since the dataset was already well laid out there was no need to do any preprocessing to get a good linear regression model, values found to be relevant to the prediction were grabbed. Deaths were chosen to be prediction rather than cases as non as cases deaths cannot go up or down. To get a more accurate evaluation, data was split to allow a certain set of the data to be used for evaluation. Date was used as the value to compare against deaths, since this way it may be possible to predict future casualty rates.

## Logistic Regression

### Pre-Processing

CREATE OR REPLACE VIEW

  `strategic-howl-293309.BDL\_CW2.Logistic\_Data\_UUID` As (

SELECT GENERATE\_UUID () uuid

 , \*

FROM strategic-howl-293309.BDL\_CW2.Data

);

To create my model a new record called ID was created to make sure every record in the dataset was unique with its own merit for prediction, this was done by using the inbuilt GENERATE\_UUID () function to do this.

CREATE OR REPLACE TABLE

  `strategic-howl-293309.BDL\_CW2.Data\_ID` As (

SELECT

  date,

  country\_name,

  deaths,

  confirmed\_cases,

  RANK () OVER (ORDER BY uuid) unique\_id,

    ROUND (- 100 \* (1 - LEAD (confirmed\_cases) OVER (ORDER BY country\_name) / confirmed\_cases), 2) AS Growth\_percent

FROM

  strategic-howl-293309.BDL\_CW2.Data\_UUID

  WHERE

   confirmed\_cases IS NOT NULL and

  deaths IS NOT NULL and

  deaths > 0              and

  confirmed\_cases > 0

  );

Following the creation of a new record, the unique ID’s were transformed into more readable ID. Dataset records were trimmed down into more relevant data for predictions. Here the growth percentage of each day to day in the data set was recorded and made into a new field and placed into a new dataset called Data ID.

CREATE OR REPLACE view

  `strategic-howl-293309.BDL\_CW2.Logistic\_Model\_DF` AS

SELECT

\*,

    CASE

    WHEN MOD(unique\_id, 10) < 8 THEN 'training'

    WHEN MOD(unique\_id, 10) = 8 THEN 'evaluation'

    WHEN MOD(unique\_id, 10) = 9 THEN 'prediction'

  END AS dataframe,

  CASE

    WHEN (Growth\_percent) > 0 THEN False

    WHEN (Growth\_percent) <= 0 THEN True

  END AS posGrowth

FROM

  `strategic-howl-293309.BDL\_CW2.Data\_ID`

Then in the final step of processing, a view was created, using each of the ID’s. Using two different cases each record in the data set was given one of each new field, called ‘Training’, ‘Evaluation’, or ‘Prediction’. These were used to allocate parts of the dataset to each of these respective machine learning activities. The other case, was used as a Boolean to determine if a positive percentile growth of cases had occurred on a given date, useful for future predictions.

### Create Model

CREATE OR REPLACE MODEL

  `strategic-howl-293309.BDL\_CW2.Logistic\_Model`

OPTIONS

  (model\_type='LOGISTIC\_REG',

    auto\_class\_weights=TRUE,

    input\_label\_cols=['posGrowth'],

  ) AS

SELECT

  \* EXCEPT (dataframe)

FROM

  `strategic-howl-293309.BDL\_CW2.Logistic\_Model\_DF`

WHERE

  dataframe = 'training'

and posGrowth is not null

Finally, our model was ready to be created, like how the linear model made use of model\_type to identify itself as linear regression here, we made the model logistic. Since a lot of pre-processing was carried out up to this step there was not much more to add. We made use of the posGrowth Boolean we had created earlier as our main outlier for prediction and all our values assigned the training value were used to train the model.

## The domains in which the dataset might be used or useful

COVID‐19 has generated an extraordinary amount of information, data that can be used by researchers, physicians, and governments. Machine learning is something that can and is utilized to better anticipate and respond to such one in a lifetime events such as seen with the COVID-19 pandemic (Ienca et al, 2020). So far there have been 26 million cases, upwards of 800,000 deaths. It has affected nearly 200 countries at the time of writing (Worldometers, 2020). Governments have been forced to adopt more data‐powered decisions for responding to the challenges brought on. Advancements of different technologies have led to the vast collection of structured and unstructured data to be mined by enterprises and governments for timely decision‐making abilities (Wu et al, 2020). Organizations are bolstering their capabilities in a matter of weeks to inform business responses to COVID‐19 challenges and prepare for the future. Data analytics facilities could offer nearly up to $15.5 trillion in annual economic value to organizations (Henkeet al, 2020). From the 1800’s to 2020/2021, the time for new technology to diffuse has shrunk from 100 years to within a decade for technologies, calling in in a new environment where access to data has become increasingly common (Comin et al, 2008). The ability to predict trends using data collection by organizations & nations across the world will led to a decline in the spread of the virus and hopefully return the world to a more normal situation.

## Any 'digitized' business/enterprise processes, real or suggested, into which the ML

## models might be infused

Day to day the rate of data collected is growing exponentially across a multitude of different sectors. As these quantities of data grow so do new and interesting ways interact with it and gain useful insights from the data are being discovered. In the past, tedious and long work was performed by humans to parse through all the new information; however, it was found to ineffective as hidden meanings or patterns could not be recognized (Villars et al, 2011). This is vital for sectors like healthcare. People can usually create about or two decent models a week whereas a computer can create thousands of models a week (Davenport et al, 2016).

For the finance sector, ML is utilized to anticipate exchanges that are probably going to be fake; utilizing calculations, the machine can consequently make an impression on card holders to check if the charge was false. Moreover, it is frequently utilized in the retail business. By utilizing prescient investigation controlled by ML, makers can figure out what clients may need straightaway and can make an online encounter that considers the shoppers' need (Villars et al, 2011). For instance, Amazon.com uses ML to show certain items to clients dependent on what they have seen previously. Model: suggesting protein powder after a client has purchased a bunch of weights. By using the logistic curve on different types of commercial goods it may be used to calculate trends in buying, or profits.

## The value that predictions using the ML models will bring

Covid-19 has placed an enormous burden on healthcare professionals around the world. By estimating the number of cases has become an important task for the public’s health so that countermeasures can be put in place for planning by increasing medical responses accordingly. Different nations have their own interventions such as lockdown, social distancing, closing of schools etc. to slowdown the spread of Covid-19. The effect of these interventions can be used to estimate how health policies can be modified to be more effective.

The models are built using data by Oxford, collected from publicly released data from countries and by utilizing the data and creating models to predict the new outcome. As it is noted regression, deep learning etc. have shown that some of the machine learning techniques which have been successfully used in the past for vital healthcare work such as image analysis, speech recognition, & health informatics (Goodfellow et al, 2016).

Regression uses supervised machine learning to calculate the connection between the dependent and independent variables. The relationship between these variables is processed by the machine and are used for prediction. Regression analyses models, trying to find the best parameters to fit the data to those models; e.g, multivariate linear regression (Fig. 1) the dependent variable *y* is dependent on *j* predictor variables xixi (*i* = 1 to *j*). The job is to find out the values of aiai and a0a0 to find the best fit of the data. where ai (i=1 to j) and a0 are constant.

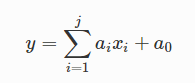


Fig. 1

Some methods have been used to predicts out comes so far. The previous days confirmed cases are used to predict the next day confirmed cases. Linear regression has used to predict the confirmed cases in the future (Gupta R et al, 2020)). Furthermore, logistic regression has beeen used to predict the number of confirmed cases at a given day (Pavlyshenko BM ,2020). However, some countries have better facilities of collecting data. This data can be used for countries with less facilities of collecting data which tend to give a lot of prediction noise when trying to collect values.

# Task 2

Using BigQuery ML Evaluation Functions and Model Feature Inspection Functions discuss and critically appraise the models created in terms of:

## Linear Regression

   SELECT

  \*

  FROM

    ML.EVALUATE (MODEL `strategic-howl-293309.BDL\_CW2.Linear\_Model`,

(

  SELECT

   \*

    FROM `strategic-howl-293309.BDL\_CW2.Data`

    ))



## Model Metrics

### Mean-absolute-error (MEA)

The mean average between the predicted and the actual. This metric can range from 0 and is indefinite. Playing around with the value it was found that by filtering by certain countries, it was found that the error would change, this indicates that the noise that is gathered from having so many different countries recording their own values. A more consistent dataset would lead to the lowering of the MEA, or any other errors. However, there is still a lot of work needed on the model before it can be drastically lowered. The lower the score is, the higher the quality of the model is.

### Root-mean-square error (MSE)

Used to measure the differences between the values predicted by a model or values observed. The low MSE indicates that there would be a match between the actual and the predicted datasets used.

### Root-mean-squared logarithmic error (RMSE)

Is a good metric of how accurately the model predicts the response, and is vital for accurate prediction, which appears to be the case for the model created, as the value is low.

### Explained variance (explained variation)

This is used to measure the discrepancy between a model and actual data. In other words, it is the part of the model's total variance that is explained by factors that are present and is not due to error variance. Since the variance is high, we can assume that there is a strong level of association.

## Explain ability

The model is rather easily read, as thanks to the simple layout of the dataset, there is not anything that could not be explained to a person who does not know anything about machine learning. The model simply grabs the number of deaths a country has, then based on the country, the number of cases etc. linear regression is used to predict the deaths with regards to deaths, hopefully giving an indication and prediction into future trends.

## Inclusivity

The model affects every person regardless of who they are. The dataset doesn’t identify who the individuals who have been confirmed or died so there is no need to model the data to work around this bias.

## Logistic Regression

  SELECT

  \*

FROM

  ML.EVALUATE (MODEL `strategic-howl-293309.BDL\_CW2.Logistic\_Model`,

    (

    SELECT

       \*

    FROM

      `strategic-howl-293309.BDL\_CW2.Logistic\_Model\_DF`

 WHERE

       dataframe = 'evaluation'

    )

  )



## Model Metrics

### Precision

The model created has a score of 0.734 or, when it predicts that there is a growth of corona cases, it is correct around 73% of the time. Considering there is a lot of noise around the data this is a good value, further processing could however, make it smaller.

### Recall

The fraction of rows with this label that the model correctly predicted. Also called "True positive rate". Recall = 0.77 and indicates how accurately the model can identify the relevant data.

### Accuracy

The fraction of classification predictions produced by the model that were correct, the model seems to correct 0.74% of the time.

### F1 score

F1 is a useful metric if you are looking for a balance between precision and recall and there is an uneven class distribution. This is easier to work with since now, instead of balancing precision and recall, a good F1-score and that would indicate a good Precision and Recall.

### Log loss

This ranges from zero to infinity, where a lower value indicates a higher-quality model. Since the log loss is 0.51 it would be assuming that the model is that of a binary problem, which it is.

### AUC ROC

This ranges from zero to one, where a higher value indicates a higher-quality model. Since the scoring is large, it can assume that the model is good at doing its job.

## Explain ability

The regression model is a bit harder to explain than the linear regression, as there was not set categories in the dataset that appealed to revealing interesting information. As so several layers of pre-processing were needed to shape the data into something more malleable. But after this it becomes easy enough. Like with the linear regression the model was made using simply, human readable fields, using a logic curve to determine if positive, or negative growth of corona cases have occurred, giving insights to trends and hopefully prediction to when cases would drop.

## Inclusivity

Same as with the linear model.

# Task 3

Create and execute two BigQuery ML queries for each of your models that demonstrate how each model can be used for prediction.

## Linear Regression

SELECT \* FROM

ML.PREDICT (MODEL `strategic-howl-293309.BDL\_CW2.Linear\_Model`,

(

SELECT

\*

FROM `strategic-howl-293309.BDL\_CW2.Data`

This prediction model can be used to base a prediction using all the information on the dataset. It might be useful on a global scale but due incorrect data collection in certain countries and missing information in some fields for other it suffers from a lot of noise issues.

SELECT \* FROM

ML.PREDICT(MODEL `strategic-howl-293309.BDL\_CW2.Linear\_Model`,

(

SELECT

\*

FROM `strategic-howl-293309.BDL\_CW2.Data`

WHERE country\_name = 'United States'

AND date >= '2020-12-01'

AND date <= '2020-12-31'

))

By using this query, we can filter the result by country and date, giving a better overall result. However it still does change from country to country.

## Logistic Regression

 SELECT

  \*

FROM

  ML.PREDICT (MODEL `strategic-howl-293309.BDL\_CW2.Logistic\_Model`,

    (

    SELECT

       \*

    FROM

      `strategic-howl-293309.BDL\_CW2.Logistic\_Model\_DF`

 WHERE

       dataframe = 'prediction'

    )

  )

Here we grab all the data and base our prediction, which does a good job based on the metrics discussed earlier on. It does make it hard to make any informed choices for any individuals, however.

   SELECT

  \*

FROM

  ML.PREDICT  (MODEL `strategic-howl-293309.BDL\_CW2.Logistic\_Model`,

    (

    SELECT

       \*

    FROM

      `strategic-howl-293309.BDL\_CW2.Logistic\_Model\_DF`

 WHERE

       dataframe = 'prediction'

       and country\_name = 'Ireland'

    )

  )

But by again filtering by country a better insight can be gained from the data, allowing more informed choices based on the curve of the data to determine when the climb of cases will plateau.

# Conclusion

COVID‐19 is a once in a lifetime event, and as so it is difficult to predict the future. Data analytics offers an insight into the pandemic and how organizations & nations can effectively strategize and respond using big data technics such as Machine learning. Machine learning offers important and significant opportunities for researchers across the globe interested in studying future employment trends, global crises and their impact on innovation and knowledge sharing. It also offers information regarding business resilience driven by the digital and analytics capabilities, and the future functioning and sustainability of global supply chains. Machine learning has the potential to help both patents and providers in terms of better care and lower costs. Several companies and organization have already taken the first step in this industry and have helped facilitate the transition to patient and evidence-orientated care. The data is there, we just must figure out how to interpret it-companies like the ones mentioned above are just a small number of the entities taking us one step closer to that vision.

# References

*Pattern recognition and machine learning*. Springer

‘The effects of a “black swan” event (COVID‐19) on herding behavior in cryptocurrency markets: evidence from cryptocurrency USD, EUR, JPY and KRW markets’, EUR, JPY and KRW Markets, 27 April 2020.

https://cloud.google.com/bigquery-ml/docs/bigqueryml-intro

Seir and regression model based covid-19 outbreak predictions in India. medRxiv

Regression approach for modeling covid-19 spread and its impact on stock market

Big data: What it is and why you should care. *White paper, IDC*, *14*, 1-14.

*What is Machine Learning: A Tour of Authoritative Definitions and a Handy One-Liner You Can Use*

*Wikipedia: Machine Learning*.

Data analytics, innovation, and firm productivity. *Management Science*, *66*(5), 2017-2039

A new approach to measuring technology with an application to the shape of the diffusion curves. *The Journal of Technology Transfer*, *33*(2), 187-207.

Just how smart are smart machines? *MIT Sloan Management Review*, *57*(3), 21.