# BKMS Assignment 4

작성자

## A. Data Pre-processing

- You should pre-process data for training the BigQuery ML model
- Pre-processing options:
  – Transform columns or filter rows using SQL
  – Create view for create the train set
  – One-hot encoding for a categorical column
  – etc.

*Amongst public data available on public datasets, I chose **“covid19_open_data”** as the data to run machine learning model on.*

```
select distinct cumulative_persons_fully_vaccinated, population,  cumulative_deceased,
case
when mod(population,10)<8 THEN 'training'
when mod(population,10)=8 THEN 'evaluation'
when mod(population,10)=9 THEN 'prediction'
end as dataframe
from `bigquery-public-data.covid19_open_data.covid19_open_data`
where date = '2022-01-05' and cumulative_deceased is not null and cumulative_persons_f
ully_vaccinated is not null and population is not null
```

*To pre-process datasets, 1) excluding rows that contain null data and 2) split rows into training, evaluation, and prediction data were carried out. The result is as below.*

Query results SAVE RESULTS EXPLORE DATA

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS

| Row | cumulative_persons_fully_vaccinated | population | cumulative_deceased | dataframe |
|---|---|---|---|---|
| 1 | 253318 | 600229 | 1188 | prediction |
| 2 | 10284 | 23432 | 75 | training |
| 3 | 39048 | 87185 | 231 | training |
| 4 | 137013 | 322375 | 1255 | training |
| 5 | 8543 | 20919 | 50 | prediction |
| 6 | 87345 | 126837 | 126 | training |
| 7 | 277644 | 520316 | 1740 | training |
| 8 | 5749833 | 12345411 | 17880 | training |
| 9 | 18726 | 24052 | 6 | training |
| 10 | 1228 | 2227 | 0 | training |
| 11 | 3221 | 5211 | 1 | training |
| 12 | 1857216 | 2508900 | 1408 | training |
| 13 | 53836 | 83250 | 57 | training |
| 14 | 4063 | 7576 | 18 | training |
| 15 | 20335 | 40215 | 95 | training |
| 16 | 14115 | 21684 | 70 | training |
| 17 | 8895 | 14454 | 33 | training |
| 18 | 17478 | 48969 | 57 | prediction |
| 19 | 11211 | 26764 | 115 | training |

Results per page: 50 1 – 50 of 4213

# B. BigQuery ML Model

- Choose a BigQuery ML model for the data and scenario

→ A l*inear regression model that “cumulative person who are vaccinated” and “population” predicts “cumulative deceased”*

- Train/Evaluate the model
  - Training

```
CREATE OR REPLACE MODEL
`covid.model_real`
```

```
OPTIONS
( model_type='LINEAR_REG',
input_label_cols=['cumulative_deceased'],
DATA_SPLIT_METHOD = 'NO_SPLIT',
LS_INIT_LEARN_RATE=.15,
L1_REG=1,
max_iterations=10) AS
SELECT
* EXCEPT(dataframe)
FROM
`covid.input_view_real`
WHERE
dataframe = 'training'
```

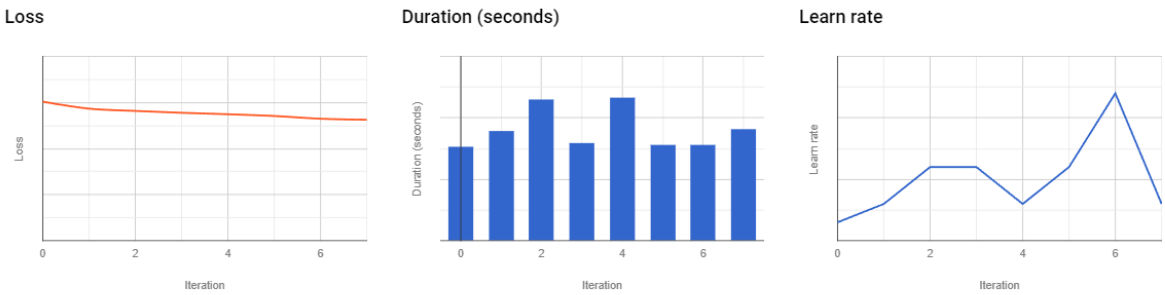
## Model Details EDIT

| | |
|---|---|
| Model ID | credible-rex-349204:covid.model_real |
| Description | |
| Labels | |
| Date created | Friday, May 6, 2022 at 11:00:46 PM GMT+09:00 |
| Model expiration | Never |
| Date modified | Friday, May 6, 2022 at 11:00:46 PM GMT+09:00 |
| Data location | US |
| Model type | LINEAR_REGRESSION |
| Loss type | Mean squared loss |

## Training Options

Training options are the optional parameters that were added in the script to create this model.

| | |
|---|---|
| Max allowed iterations | 10 |
| Actual iterations | 8 |
| L1 regularization | 1.00 |
| L2 regularization | 0.00 |
| Early stop | true |
| Min relative progress | 0.01 |
| Learn rate strategy | Line search |
| Line search initial learn rate | 0.15 |
| Calculate P Values | false |
| Data split method | No Split |




| | |
|---|---|
| Mean absolute error | 2,323.3977 |
| Mean squared error | 262,624,056.0195 |
| Mean squared log error | 10.8255 |
| Median absolute error | 1,197.0482 |
| R squared | 0.2661 |

- Evaluate

```
SELECT
*
FROM
ML.EVALUATE (MODEL `covid.model_real`,
(
SELECT
*
FROM
`covid.input_view_real`
WHERE
dataframe = 'evaluation'
)
)
```

Query results

SAVE RESULTS EXPLORE DATA

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS

| Row | mean_absolute_error | mean_squared_error | mean_squared_log_error | median_absolute_error | r2_score | explained_variance |
|---|---|---|---|---|---|---|
| 1 | 1714.3767593038706 | 24434590.023899265 | 10.457465510156053 | 1189.161541945504 | 0.21762288868787905 | 0.22140393605194642 |

- Perform prediction using the trained model

```
SELECT
*
FROM
```

```
ML.PREDICT (MODEL `covid.model_real`,
(
SELECT
*
FROM
`covid.input_view_real`
WHERE
dataframe = 'prediction'
)
)
```

| Row | predicted_cumulative_deceased | cumulative_persons_fully_vaccinated | population | cumulative_deceased | dataframe |
|---|---|---|---|---|---|
| 1 | 1357.0634714087694 | 436529 | 493559 | 484 | prediction |
| 2 | 1263.9265965909685 | 20958 | 38529 | 70 | prediction |
| 3 | 1262.5724602256064 | 20334 | 34179 | 109 | prediction |
| 4 | 1262.8010022003698 | 32027 | 39759 | 20 | prediction |
| 5 | 1315.9443381722153 | 126947 | 239929 | 656 | prediction |
| 6 | 1254.612276425539 | 671 | 1919 | 11 | prediction |
| 7 | 1254.2124030596756 | 115 | 479 | 1 | prediction |
| 8 | 1255.5395076627065 | 2082 | 5309 | 18 | prediction |
| 9 | 1258.6884949973114 | 6628 | 16719 | 44 | prediction |
| 10 | 1265.0140136394853 | 19658 | 41269 | 102 | prediction |
| 11 | 1256.2364213427461 | 5180 | 8709 | 24 | prediction |
| 12 | 1257.3602407633475 | 4478 | 11809 | 9 | prediction |
| 13 | 1268.5924762855907 | 26531 | 54949 | 116 | prediction |
| 14 | 1254.882061225363 | 2262 | 3399 | 0 | prediction |
| 15 | 1256.5236580461451 | 5545 | 9729 | 20 | prediction |
| 16 | 1258.3099094564734 | 10103 | 17029 | 80 | prediction |
| 17 | 1262.7553175305961 | 22194 | 35509 | 80 | prediction |
| 18 | 1267.0958842269338 | 15888 | 45979 | 162 | prediction |
| 19 | 1258.6784754845476 | 6796 | 16759 | 50 | prediction |
| 20 | 1274.2117360251762 | 39400 | 77299 | 262 | prediction |
| 21 | 1267.2522225925936 | 21359 | 48739 | 195 | prediction |

# C. Analysis

- The analysis part must include an analysis of the following:
  - Reasons for applying the pre-processing method

    In part A, two pre-processing methods were applied.

    1) Deleting null→ If there is a null value in the regression model, it is impossible to train the linear regression model. Initially, I tried to make an ML model which includes more predictors- population density, nurses per capita, doctors per capita et cetera. However, there were many null values in those columns and to gain ‘big’ data, only two predictors were available-population and vaccinated population.

2) Split datasets into training, evaluation, and test data → validation-set approach provides an estimate of test error. By doing so, linear regression is fit on the training set, and performance is evaluated on the estimation set. To randomize, splitting was performed by the remainder of the population by 10.

- Reasons for selecting the ML model

  Linear regression is the most powerful and simple ML model to understand the dependence of the result on predictors. Intuitively, the total population and the cumulative persons fully vaccinated seem to have an effect on the number of deceased by COVID-19. To figure out such effects, linear regression of the population and vaccinated population predicts the casualty from COVID-19.

- Model evaluation & prediction result analysis

  R-squared, which shows the prediction accuracy of linear regression was too low in evaluation sets. It can be interpreted in both ways: (1) The assumption that the relationship between predictors and results does not exist. (2) linear dependence between predictors may hamper the model from predicting an accurate result.

  As low R-squared implies, the gap between predicted deaths and real deaths was so high, which is can be found in the last picture in part B.

- How to improve the result
  - Add more predictors: The more predictors, the more it can interpret the variance.
  - Considering the interaction between predictors: If there’s an interaction between predictors, adding an interaction term may increase the performance
  - Try the polynomial model: The actual relationship may be based on a more complex form. i.e. high order polynomials.