# ML_with_BigQuery_ML

April 24, 2023

Build logistic regression model

[13]:
```
%%bigquery
# Create the training dataset
SELECT
  IF(arr_delay < 15, 'ontime', 'late') AS ontime,
  dep_delay,
  taxi_out,
  distance,
  IF(is_train_day = 'True', False, True) AS is_eval_day
FROM dsongcp.flights_tzcorr f
JOIN dsongcp.trainday t
ON f.FL_DATE = t.FL_DATE
WHERE
  f.CANCELLED = False AND
  f.DIVERTED = False
LIMIT 5
```

```
Query is running:   0%|          |

Downloading:   0%|          |
```

[13]:

| | ontime | dep_delay | taxi_out | distance | is_eval_day |
|---|---|---|---|---|---|
| 0 | ontime | 7.0 | 12.0 | 351.0 | True |
| 1 | ontime | -3.0 | 11.0 | 351.0 | False |
| 2 | ontime | -7.0 | 10.0 | 351.0 | True |
| 3 | ontime | -16.0 | 13.0 | 351.0 | True |
| 4 | ontime | -4.0 | 8.0 | 351.0 | False |

[14]:
```
%%bigquery
# Create the model
CREATE OR REPLACE MODEL dsongcp.arr_delay_lm
OPTIONS(input_label_cols=['ontime'],
        model_type='logistic_reg',
        data_split_method='custom',
        data_split_col='is_eval_day')
AS
SELECT
  IF(arr_delay < 15, 'ontime', 'late') AS ontime,
```

```
  dep_delay,
  taxi_out,
  distance,
  IF(is_train_day = 'True', False, True) AS is_eval_day
FROM dsongcp.flights_tzcorr f
JOIN dsongcp.trainday t
ON f.FL_DATE = t.FL_DATE
WHERE
  f.CANCELLED = False AND
  f.DIVERTED = False
```

```
Query is running:   0%|          |
```

[14]:
```
Empty DataFrame
Columns: []
Index: []
```

[15]:
```
%%bigquery
SELECT * FROM ML.TRAINING_INFO(MODEL dsongcp.arr_delay_lm)
```

```
Query is running:   0%|          |
```

```
Downloading:   0%|          |
```

[15]:

| | training_run | iteration | loss | eval_loss | learning_rate | duration_ms |
|---|---|---|---|---|---|---|
| 0 | 0 | 10 | 0.169790 | 0.167233 | 25.6 | 14122 |
| 1 | 0 | 9 | 0.171054 | 0.168219 | 25.6 | 13567 |
| 2 | 0 | 8 | 0.173285 | 0.170453 | 51.2 | 14028 |
| 3 | 0 | 7 | 0.190027 | 0.186490 | 25.6 | 13720 |
| 4 | 0 | 6 | 0.219350 | 0.214753 | 12.8 | 13153 |
| 5 | 0 | 5 | 0.261108 | 0.255400 | 6.4 | 15061 |
| 6 | 0 | 4 | 0.321923 | 0.316414 | 3.2 | 12944 |
| 7 | 0 | 3 | 0.415231 | 0.411193 | 1.6 | 13419 |
| 8 | 0 | 2 | 0.521449 | 0.519102 | 0.8 | 13528 |
| 9 | 0 | 1 | 0.606936 | 0.605817 | 0.4 | 14792 |
| 10 | 0 | 0 | 0.661895 | 0.661502 | 0.2 | 9388 |

### 0.1 Evaluate the model

[16]:
```
%%bigquery
SELECT *
FROM ML.EVALUATE(MODEL dsongcp.arr_delay_lm,
                 (

SELECT
  IF(arr_delay < 15, 'ontime', 'late') AS ontime,
  dep_delay,
  taxi_out,
```

```
  distance
FROM dsongcp.flights_tzcorr f
JOIN dsongcp.trainday t
ON f.FL_DATE = t.FL_DATE
WHERE
  f.CANCELLED = False AND
  f.DIVERTED = False AND
    is_train_day = 'False'

                ),
                STRUCT(0.7 AS threshold))
```

```
Query is running:   0%|          |

Downloading:   0%|          |
```

[16]:

| | precision | recall | accuracy | f1_score | log_loss | roc_auc |
|---|---|---|---|---|---|---|
| 0 | 0.964337 | 0.956535 | 0.935174 | 0.96042 | 0.167233 | 0.956269 |

## 0.2 Make prediction from the model

[17]:
```
%%bigquery
SELECT * FROM ML.WEIGHTS(MODEL dsongcp.arr_delay_lm)
```

```
Query is running:   0%|          |

Downloading:   0%|          |
```

[17]:

| | processed_input | weight | category_weights |
|---|---|---|---|
| 0 | dep_delay | -0.132984 | [] |
| 1 | taxi_out | -0.121715 | [] |
| 2 | distance | 0.000223 | [] |
| 3 | __INTERCEPT__ | 4.762572 | [] |

[18]:
```
%%bigquery
SELECT * FROM ML.PREDICT(MODEL dsongcp.arr_delay_lm,
                         (
SELECT 12.0 AS dep_delay, 14.0 AS taxi_out, 1231 AS distance
                         ))
```

```
Query is running:   0%|          |

Downloading:   0%|          |
```

[18]:

| | predicted_ontime | predicted_ontime_probs | \ |
|---|---|---|---|
| 0 | ontime | [{'label': 'ontime', 'prob': 0.850350772376704… | |

| | dep_delay | taxi_out | distance |
|---|---|---|---|
| 0 | 12.0 | 14.0 | 1231 |

[19]:
```
%%bigquery
WITH predictions AS (
SELECT
  *
FROM ML.PREDICT(MODEL dsongcp.arr_delay_lm,
                (
SELECT
  IF(arr_delay < 15, 'ontime', 'late') AS ontime,
  dep_delay,
  taxi_out,
  distance
FROM dsongcp.flights_tzcorr f
JOIN dsongcp.trainday t
ON f.FL_DATE = t.FL_DATE
WHERE
  f.CANCELLED = False AND
  f.DIVERTED = False AND
  t.is_train_day = 'False'
                ),
                STRUCT(0.7 AS threshold))),
stats AS (
SELECT
  COUNTIF(ontime != 'ontime' AND ontime = predicted_ontime) AS correct_cancel
  , COUNTIF(predicted_ontime = 'ontime') AS total_noncancel
  , COUNTIF(ontime = 'ontime' AND ontime = predicted_ontime) AS␣
 ↪correct_noncancel
  , COUNTIF(ontime != 'ontime') AS total_cancel
  , SQRT(SUM((IF(ontime = 'ontime', 1, 0) - p.prob) * (IF(ontime = 'ontime', 1,␣
 ↪0) - p.prob))/COUNT(*)) AS rmse
FROM predictions, UNNEST(predicted_ontime_probs) p
WHERE p.label = 'ontime'
)
SELECT
   correct_cancel / total_cancel AS correct_cancel
   , total_noncancel
   , correct_noncancel / total_noncancel AS correct_noncancel
   , total_cancel
   , rmse
FROM stats
```

```
Query is running:   0%|          |

Downloading:   0%|          |
```

[19]:

| | correct_cancel | total_noncancel | correct_noncancel | total_cancel | rmse |
|---|---|---|---|---|---|
| 0 | 0.836363 | 1301948 | 0.964337 | 283750 | 0.213091 |

### 0.3 Create, evaluate and predict the model by adding additional airport information

[20]:
```
%%bigquery
# To showcase the scalability of BigQuery, add two fields, the origin and␣
 ↪destination airport:
CREATE OR REPLACE MODEL dsongcp.arr_delay_airports_lm
OPTIONS(input_label_cols=['ontime'],
        model_type='logistic_reg',
        data_split_method='custom',
        data_split_col='is_eval_day')
AS
SELECT
  IF(arr_delay < 15, 'ontime', 'late') AS ontime,
  dep_delay,
  taxi_out,
  distance,
  origin,
  dest,
  IF(is_train_day = 'True', False, True) AS is_eval_day
FROM dsongcp.flights_tzcorr f
JOIN dsongcp.trainday t
ON f.FL_DATE = t.FL_DATE
WHERE
  f.CANCELLED = False AND
  f.DIVERTED = False
```

```
Query is running:   0%|          |
```

[20]:
```
Empty DataFrame
Columns: []
Index: []
```

[22]:
```
%%bigquery
SELECT *
FROM ML.EVALUATE(MODEL dsongcp.arr_delay_airports_lm,
                 (
SELECT
  IF(arr_delay < 15, 'ontime', 'late') AS ontime,
  dep_delay,
  taxi_out,
  distance,
  origin,
  dest,
  IF(is_train_day = 'True', False, True) AS is_eval_day
FROM dsongcp.flights_tzcorr f
JOIN dsongcp.trainday t
ON f.FL_DATE = t.FL_DATE
```

```
WHERE
  f.CANCELLED = False AND
  f.DIVERTED = False AND
  t.is_train_day = 'False'
                ),
                STRUCT(0.7 AS threshold))
```

```
Query is running:   0%|          |

Downloading:   0%|          |
```

[22]:

| | precision | recall | accuracy | f1_score | log_loss | roc_auc |
|---|---|---|---|---|---|---|
| 0 | 0.967151 | 0.957706 | 0.938477 | 0.962405 | 0.165557 | 0.960845 |

[21]:

```
%%bigquery
WITH predictions AS (
SELECT
  *
FROM ML.PREDICT(MODEL dsongcp.arr_delay_airports_lm,
                (
SELECT
  IF(arr_delay < 15, 'ontime', 'late') AS ontime,
  dep_delay,
  taxi_out,
  distance,
  origin,
  dest,
  IF(is_train_day = 'True', False, True) AS is_eval_day
FROM dsongcp.flights_tzcorr f
JOIN dsongcp.trainday t
ON f.FL_DATE = t.FL_DATE
WHERE
  f.CANCELLED = False AND
  f.DIVERTED = False AND
  t.is_train_day = 'False'
                ),
                STRUCT(0.7 AS threshold))),
stats AS (
SELECT
  COUNTIF(ontime != 'ontime' AND ontime = predicted_ontime) AS correct_cancel
  , COUNTIF(predicted_ontime = 'ontime') AS total_noncancel
  , COUNTIF(ontime = 'ontime' AND ontime = predicted_ontime) AS␣
 ↪correct_noncancel
  , COUNTIF(ontime != 'ontime') AS total_cancel
  , SQRT(SUM((IF(ontime = 'ontime', 1, 0) - p.prob) * (IF(ontime = 'ontime', 1,␣
 ↪0) - p.prob))/COUNT(*)) AS rmse
FROM predictions, UNNEST(predicted_ontime_probs) p
```

```
WHERE p.label = 'ontime'
)
SELECT
   correct_cancel / total_cancel AS correct_cancel
   , total_noncancel
   , correct_noncancel / total_noncancel AS correct_noncancel
   , total_cancel
   , rmse
FROM stats
```

```
Query is running:   0%|          |

Downloading:   0%|          |
```

[21]:

| | correct_cancel | total_noncancel | correct_noncancel | total_cancel | rmse |
|---|---|---|---|---|---|
| 0 | 0.84953 | 1299749 | 0.967151 | 283750 | 0.209839 |

[ ]: