Getting Started with BigQuery Machine Learning

45 minutes No cost

GSP247

Overview

BigQuery Machine Learning (BigQuery ML) enables users to create and execute machine learning models in BigQuery using SQL queries. The goal is to democratise machine learning by enabling SQL practitioners to build models using their existing tools and to increase development speed by eliminating the need for data movement.

There is a newly available ecommerce dataset that has millions of Google Analytics records for the Google Merchandise Store loaded into BigQuery. In this lab you will use this data to create a model that predicts

whether a visitor will make a transaction.

What you'll learn

In this lab, you learn how to create, evaluate, and use machine learning models in BigQuery.

Prerequisite

To maximize your learning you should have a basic knowledge of SQL or BigQuery.

Setup and requirements

Before you click the Start Lab button

Read these instructions. Labs are timed and you cannot pause them. The timer, which starts when you click **Start Lab**, shows how long Google Cloud resources will be made available to you.

This hands-on lab lets you do the lab activities yourself in a real cloud environment, not in a simulation or demo environment. It does so by giving you new, temporary credentials that you use to sign in and access Google Cloud for the duration of the lab.

To complete this lab, you need:

- Access to a standard internet browser (Chrome browser recommended).

Note: Use an Incognito or private browser window to run this lab. This prevents any conflicts between your personal account and the Student account, which may cause extra charges

incurred to your personal account.

- Time to complete the lab---remember, once you start, you cannot pause a lab.

Note: If you already have your own personal Google Cloud account or project, do not use it for this lab to avoid extra charges to your account.

How to start your lab and sign in to the Google Cloud console

1. Click the **Start Lab** button. If you need to pay for the lab, a pop-up opens for you to select your payment method. On the left is the **Lab Details** panel with the following:

    - The **Open Google Cloud console** button
    - Time remaining
    - The temporary credentials that you must use for this lab
    - Other information, if needed, to step through this lab

2. Click **Open Google Cloud console** (or right-click and select **Open Link in Incognito Window** if you are running the Chrome browser).

    The lab spins up resources, and then opens another tab that shows the **Sign in** page.

    Tip: Arrange the tabs in separate windows, side-by-side.

    Note: If you see the **Choose an account** dialog, click **Use Another Account**.

3. If necessary, copy the **Username** below and paste it into the **Sign in** dialog.

    ```
    student-03-f0e8857b3776@qwiklabs.net
    ```

    content_c

You can also find the **Username** in the **Lab Details** panel.

4. Click **Next**.

5. Copy the **Password** below and paste it into the **Welcome** dialog.

```
Yp7tdb6cSt5b
```

You can also find the **Password** in the **Lab Details** panel.

6. Click **Next**.

Important: You must use the credentials the lab provides you. Do not use your Google Cloud account credentials.

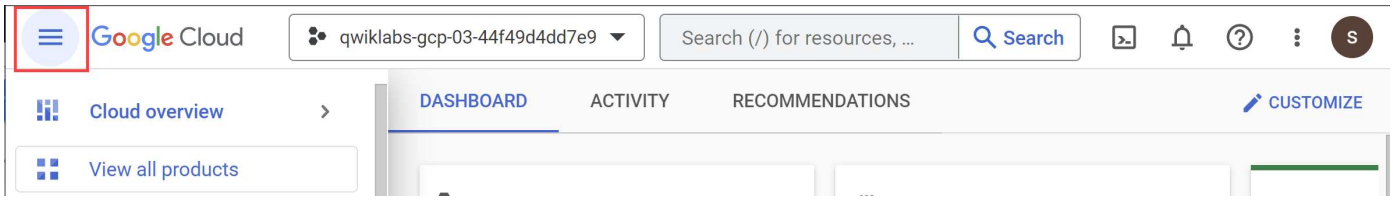
Note: Using your own Google Cloud account for this lab may incur extra charges.

7. Click through the subsequent pages:

   - Accept the terms and conditions.
   - Do not add recovery options or two-factor authentication (because this is a temporary account).
   - Do not sign up for free trials.

After a few moments, the Google Cloud console opens in this tab.

Note: To view a menu with a list of Google Cloud products and services, click the **Navigation menu** at the top-left.



Open the BigQuery console

1. In the Google Cloud Console, select **Navigation menu** > **BigQuery**.

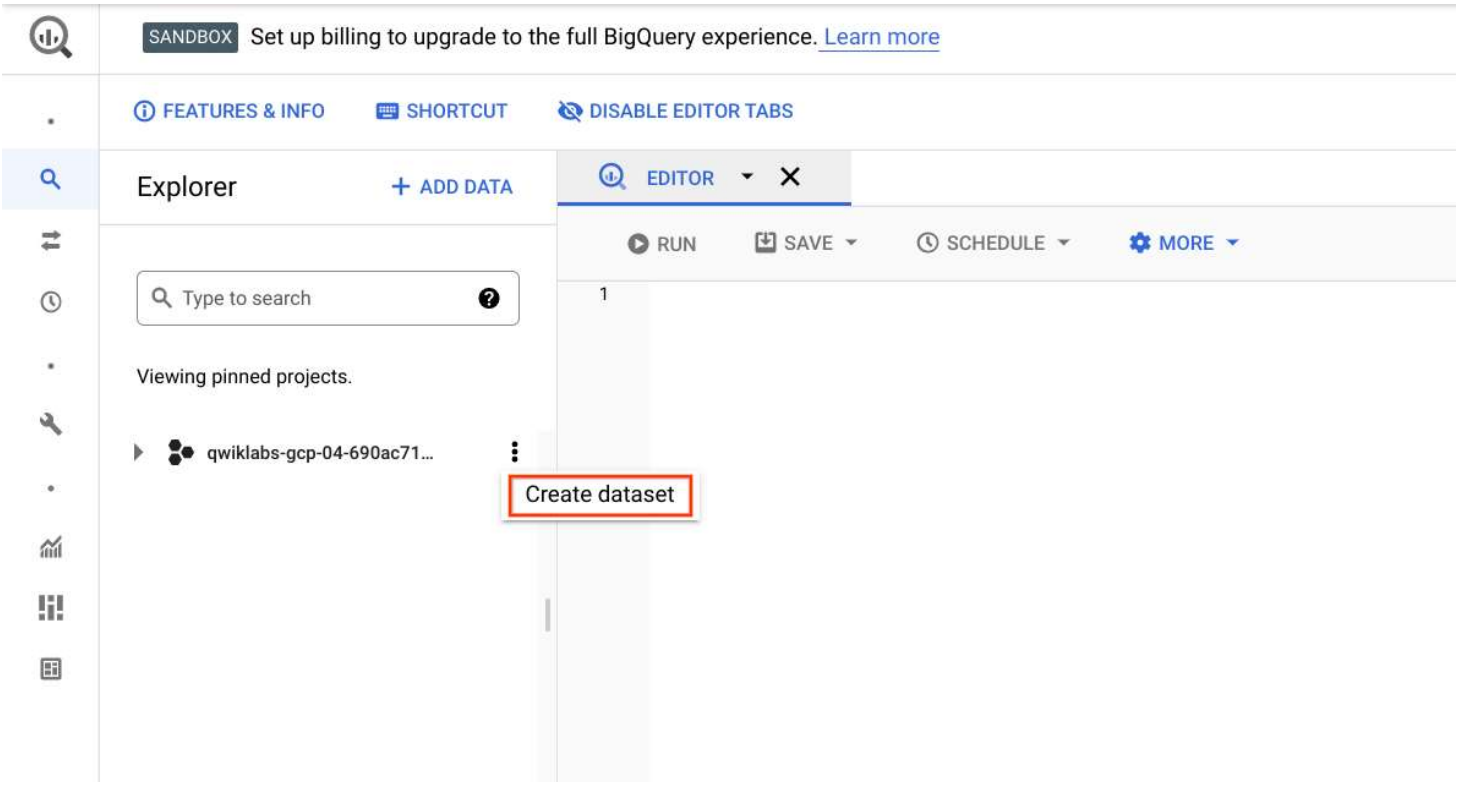
The **Welcome to BigQuery in the Cloud Console** message box opens. This message box provides a link to the quickstart guide and the release notes.

2. Click **Done**.

The BigQuery console opens.

Task 1. Create a dataset

1. To create a dataset, click on the **View actions** icon next to your project ID and select **Create dataset**.



2. Next, name your Dataset ID `bqml_lab` and click **Create dataset**.

Test completed task

Click **Check my progress** to verify your performed task. If you have completed the task successfully you will be granted an assessment score.

Create a BigQuery dataset

Check my progress

Assessment Completed! BigQuery dataset created successfully. Dataset IDs: ["bqml_lab"]

Task 2. Create a model

Now, move on to your task!

1. Go to BigQuery **EDITOR**, type or paste the following query to create a model that predicts whether a visitor will make a transaction:

```
#standardSQL
CREATE OR REPLACE MODEL `bqml_lab.sample_model`
OPTIONS(model_type='logistic_reg') AS
SELECT
  IF(totals.transactions IS NULL, 0, 1) AS label,
  IFNULL(device.operatingSystem, "") AS os,
  device.isMobile AS is_mobile,
  IFNULL(geoNetwork.country, "") AS country,
  IFNULL(totals.pageviews, 0) AS pageviews
FROM
  `bigquery-public-data.google_analytics_sample.ga_sessions_*`
WHERE
  _TABLE_SUFFIX BETWEEN '20160801' AND '20170631'
LIMIT 100000;
```

content_c

2. Click **RUN**.

Here the visitor's device's operating system is used, whether said device is a mobile device, the visitor's country and the number of page views as the criteria for whether a transaction has been made.

In this case, `bqml_lab` is the name of the dataset and `sample_model` is the name of the model. The model type specified is binary logistic regression. In this case, `label` is what you're trying to fit to.

> **Note:** If you're only interested in 1 column, this is an alternative way to setting `input_label_cols`.

The training data is being limited to those collected from 1 August 2016 to 30 June 2017. This is done to save the last month of data for "prediction". It is further limited to 100,000 data points to save some time.

Running the `CREATE MODEL` command creates a Query Job that will run asynchronously so you can, for example, close or refresh the BigQuery UI window.

Test completed task

Click **Check my progress** to verify your performed task. If you have completed the task successfully you will be granted an assessment score.

Create a model to predict visitor transaction

Check my progress

Assessment Completed!

(Optional) Model information & training statistics

If interested, you can get information about the model by expanding `bqml_lab` dataset and then clicking the `sample_model` model in the UI. Under the **Details** tab you should find some basic model info and training options used to produce the model. Under **Training**, you should see a table either a table or graphs, depending on your *View as* settings:

sample_model

QUERY MODEL DELETE MODEL EXPORT MODEL

DETAILS TRAINING EVALUATION SCHEMA

View as

- ○ Graphs
- ◉ Table

Iteration	Training Data Loss	Evaluation Data Loss	Learn Rate	Duration (seconds)
10	0.0467	0.0342	25.6	4.63
9	0.0470	0.0343	12.8	4.70
8	0.0475	0.0350	25.6	5.31
7	0.0482	0.0354	25.6	5.03
6	0.0511	0.0393	12.8	5.05
5	0.0583	0.0471	6.4	6.10
4	0.0724	0.0624	3.2	6.96
3	0.1017	0.0934	1.6	5.93
2	0.1732	0.1673	0.8	6.01
1	0.3231	0.3197	0.4	6.33
0	0.5227	0.5214	0.2	5.04

Rows per page: 50 1 – 11 of 11

Task 3. Evaluate the model

- Replace the previous query with the following and then click **Run**:

```
#standardSQL
SELECT
  *
FROM
  ml.EVALUATE(MODEL `bqml_lab.sample_model`, (
SELECT
  IF(totals.transactions IS NULL, 0, 1) AS label,
  IFNULL(device.operatingSystem, "") AS os,
  device.isMobile AS is_mobile,
  IFNULL(geoNetwork.country, "") AS country,
  IFNULL(totals.pageviews, 0) AS pageviews
FROM
  `bigquery-public-data.google_analytics_sample.ga_sessions_*`
WHERE
  _TABLE_SUFFIX BETWEEN '20170701' AND '20170801'));
```

content_c

If used with a linear regression model, the above query returns the following columns:

- `mean_absolute_error`, `mean_squared_error`, `mean_squared_log_error`,
- `median_absolute_error`, `r2_score`, `explained_variance`.

If used with a logistic regression model, the above query returns the following columns:

- `precision`, `recall`
- `accuracy`, `f1_score`
- `log_loss`, `roc_auc`

Please consult the machine learning glossary or run a Google search to understand how each of these metrics are calculated and what they mean.

You'll realize the `SELECT` and `FROM` portions of the query are identical to that used during training. The `WHERE` portion reflects the change in time frame and the `FROM` portion shows that you're calling `ml.EVALUATE`.

You should see a table similar to this:

Row	precision	recall	accuracy	f1_score	log_loss	roc_auc
1	0.47368421052631576	0.10893854748603352	0.9853834982788297	0.17713853141559424	0.04552280390355375	0.9773986013986014

Test completed task

Click **Check my progress** to verify your performed task. If you have completed the task successfully you will be granted an assessment score.

Evaluate the Model

Check my progress

Assessment Completed!

Task 4. Use the model

Predict purchases per country

With this query you will try to predict the number of transactions made by visitors of each country, sort the results, and select the top 10 countries by purchases:

- Replace the previous query with the following and then click **Run**:

```
#standardSQL
SELECT
  country,
  SUM(predicted_label) as total_predicted_purchases
FROM
  ml.PREDICT(MODEL `bqml_lab.sample_model`, (
SELECT
  IFNULL(device.operatingSystem, "") AS os,
  device.isMobile AS is_mobile,
  IFNULL(totals.pageviews, 0) AS pageviews,
  IFNULL(geoNetwork.country, "") AS country
FROM
  `bigquery-public-data.google_analytics_sample.ga_sessions_*`
WHERE
  _TABLE_SUFFIX BETWEEN '20170701' AND '20170801'))
GROUP BY country
```

content_c

```
ORDER BY total_predicted_purchases DESC
LIMIT 10;
```

This query is very similar to the evaluation query demonstrated in the previous section. Instead of `ml.EVALUATE`, you're using `ml.PREDICT` and the BigQuery ML portion of the query is wrapped with standard SQL commands. For this lab you're interested in the country and the sum of purchases for each country, so that's why `SELECT`, `GROUP BY` and `ORDER BY`. `LIMIT` is used to ensure you only get the top 10 results.

You should see a table similar to this:

Row	country	total_predicted_purchases
1	United States	140
2	Taiwan	5
3	India	2
4	Turkey	1
5	Venezuela	1
6	United Kingdom	1
7	Japan	1
8	Indonesia	1
9	Canada	1
10	St. Lucia	1

Test completed task

Click **Check my progress** to verify your performed task. If you have completed the task successfully you will be granted an assessment score.

Predict purchases per country

Check my progress

Assessment Completed!

Predict purchases per user

Here is another example. This time you will try to predict the number of transactions each visitor makes, sort the results, and select the top 10 visitors by transactions:

- Replace the previous query with the following and then click **Run:**

```
#standardSQL
SELECT
  fullVisitorId,
  SUM(predicted_label) as total_predicted_purchases
FROM
  ml.PREDICT(MODEL `bqml_lab.sample_model`, (
SELECT
  IFNULL(device.operatingSystem, "") AS os,
  device.isMobile AS is_mobile,
  IFNULL(totals.pageviews, 0) AS pageviews,
  IFNULL(geoNetwork.country, "") AS country,
  fullVisitorId
FROM
  `bigquery-public-data.google_analytics_sample.ga_sessions_*`
WHERE
  _TABLE_SUFFIX BETWEEN '20170701' AND '20170801'))
GROUP BY fullVisitorId
ORDER BY total_predicted_purchases DESC
LIMIT 10;
```

content_c

You should see a table similar to this:

Row	fullVisitorId	total_predicted_purchases
1	9417857471295131045	3
2	8069922490326866 50	2
3	0576935009275810 77	2
4	29694186761262587 98	2
5	0376394056092189113	2
6	83889310329550527 46	2
7	74203005015230124 60	2
8	12809936612043474 50	2
9	11228833092889594 2	2
10	86395516253142188 23	1

Test completed task

Click **Check my progress** to verify your performed task. If you have completed the task successfully you will be granted an assessment score.

Predict purchases per user

Check my progress

Assessment Completed!

Task 5. Test your understanding

Below are multiple choice questions to reinforce your understanding of this lab's concepts. Answer them to the best of your abilities.

BigQuery is a fully-managed enterprise data warehouse that enables super-fast SQL queries.

- check True
- False

Which option best describes what BigQuery ML does?

- close ~~Exports data from the warehouse, reformats the data, then executes the model using standard SQL queries.~~
- Creates machine learning models using Python or Java in BigQuery, then executes the model using standard SQL queries.
- check Creates and executes machine learning models in BigQuery using standard SQL queries.
- Creates machine learning models so you can export and use the model to re-evaluate the accuracy of other models.

Submit

Congratulations!

This concludes the self-paced lab, Getting Started with BigQuery Machine Learning. You created a binary logistic regression model, evaluated the model, and used the model to make predictions.