Getting Started with BigQuery Machine Learning

45 minutesFree

**GSP247**



**Overview**

[BigQuery Machine Learning](https://cloud.google.com/bigquery/docs/bigqueryml-intro) (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](https://support.google.com/analytics/answer/7586738?hl=en&ref_topic=3416089#zippy=%2Cin-this-article) that has millions of Google Analytics records for the [Google Merchandise Store](https://shop.googlemerchandisestore.com/) 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

How to create, evaluate and use machine learning models in BigQuery

What you'll need

* A Browser, such as [Chrome](https://www.google.com/chrome/browser/desktop/) or [Firefox](https://www.mozilla.org/firefox/)
* Basic knowledge of SQL or BigQuery

**Setup and requirements**

Setup

**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 a panel populated with the temporary credentials that you must use for this lab.



1. Copy the username, and then click **Open Google Console**. The lab spins up resources, and then opens another tab that shows the **Sign in** page.



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

If you see the **Choose an account** page, click **Use Another Account**. 

1. In the **Sign in** page, paste the username that you copied from the left panel. Then copy and paste the password.

***Important:*** You must use the credentials from the left panel. Do not use your Google Cloud Training credentials. If you have your own Google Cloud account, do not use it for this lab (avoids incurring charges).

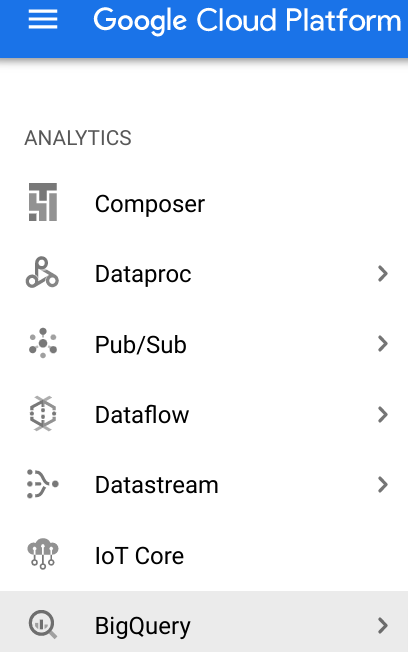
1. Click through the subsequent pages:
   * Accept the terms and conditions.
   * Do not add recovery options or two-factor authentication (because this is a temporary account).
   * Do not sign up for free trials.

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

**Note:** You can view the menu with a list of Google Cloud Products and Services by clicking the **Navigation menu** at the top-left. 

Open BigQuery Console

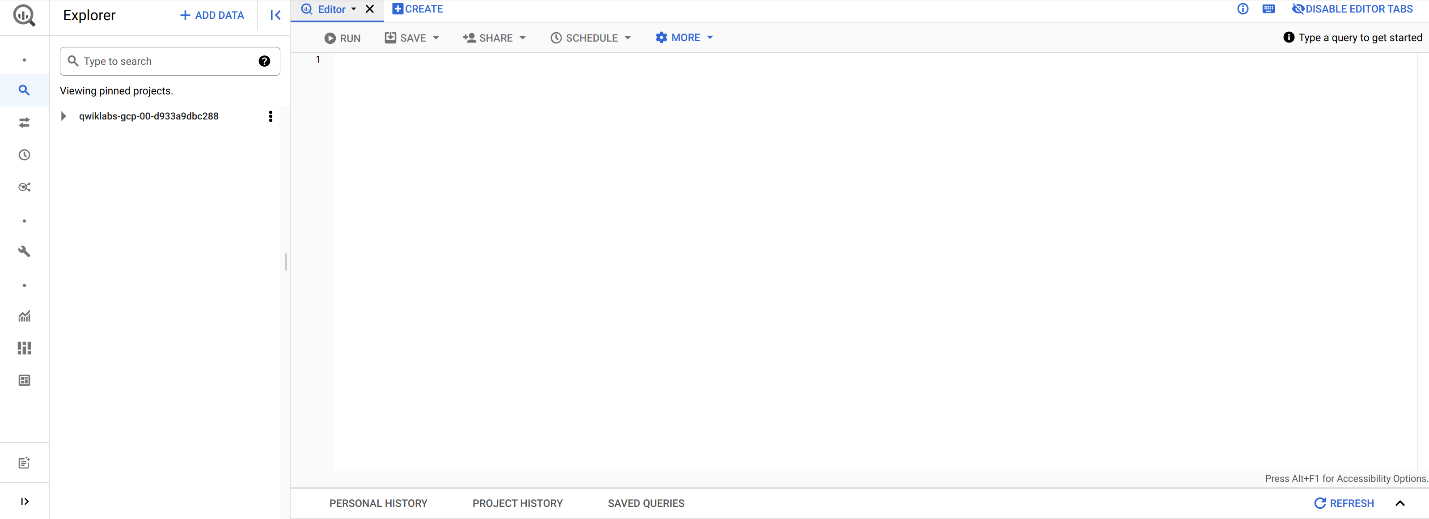
In the Google Cloud Console, select **Navigation menu** > **BigQuery**:



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.

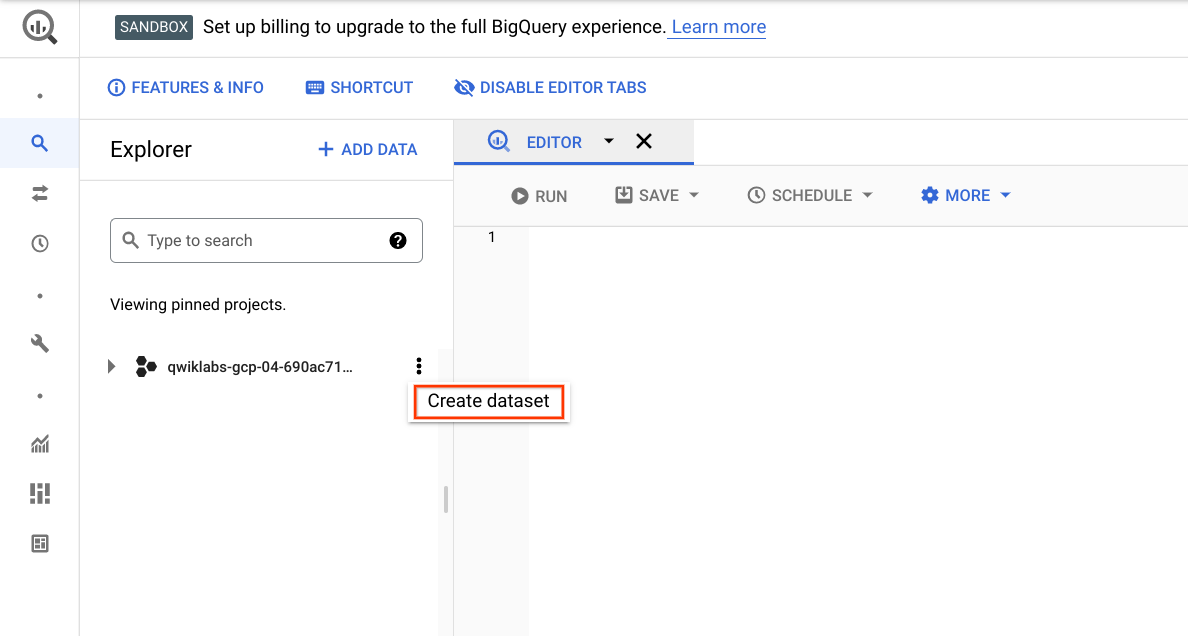
Click **Done**.

The BigQuery console opens.

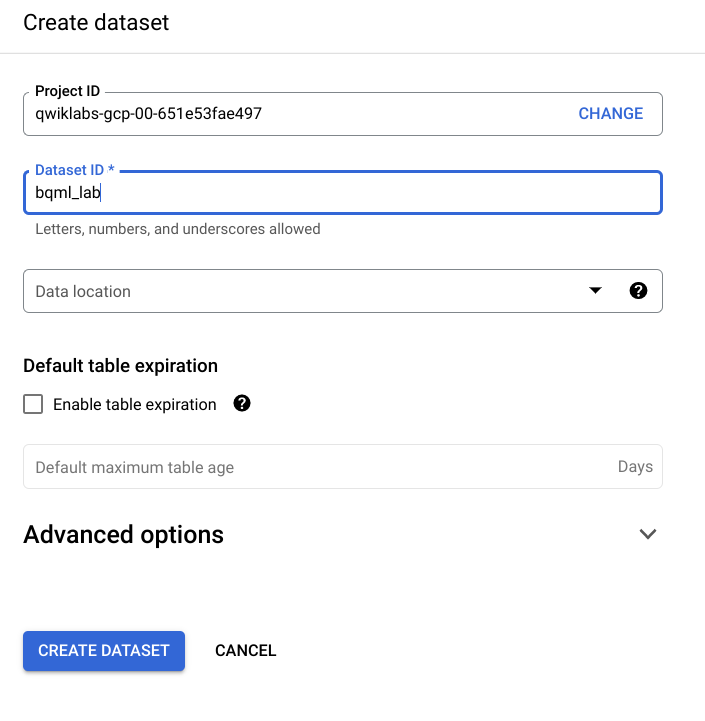


**Create a dataset**

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



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 with an assessment score.

Create a BigQuery dataset

Check my progress

**Create a model**

Now, move on to your task!

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;

Copied!

content\_copy

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

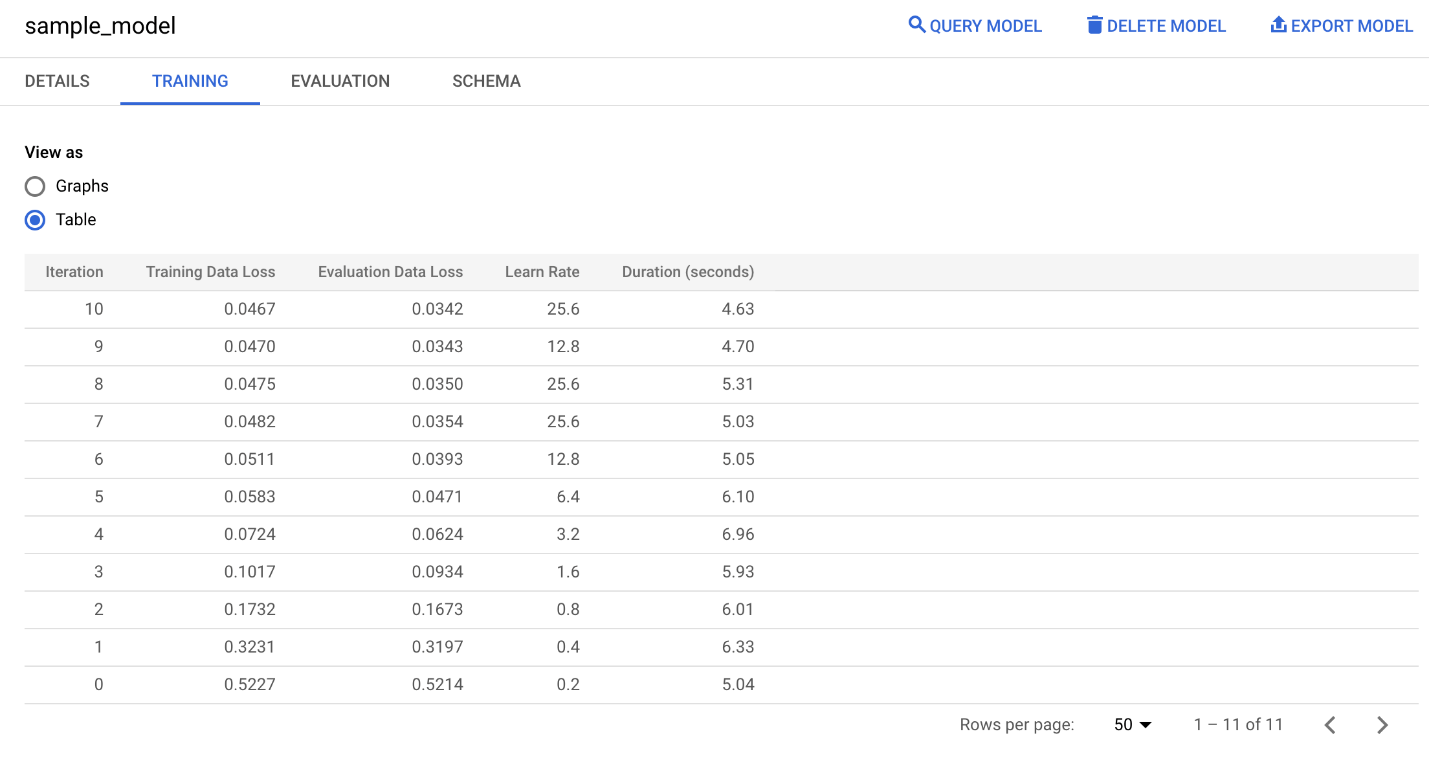
Click **Check my progress** to verify your performed task. If you have completed the task successfully you will be granted with an assessment score.

Create a model to predict visitor transaction

Check my progress

**(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:





**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'));

Copied!

content\_copy

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.

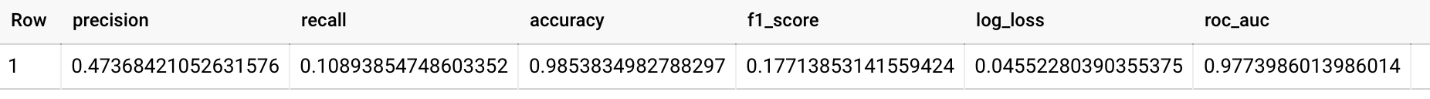
If used with a logistic regression model, the above query returns the following columns:

* precision, recall
* accuracy, f1\_score
* log\_loss, roc\_auc

Please consult the [machine learning glossary](https://developers.google.com/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 is 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:



Test Completed Task

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

Evaluate the Model

Check my progress

**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

ORDER BY total\_predicted\_purchases DESC

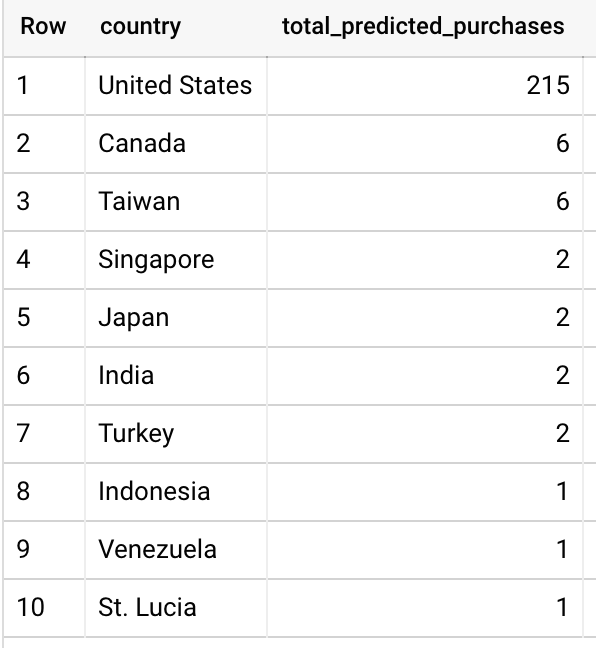
LIMIT 10;

Copied!

content\_copy

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:



Test Completed Task

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

Predict purchases per country

Check my progress

**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

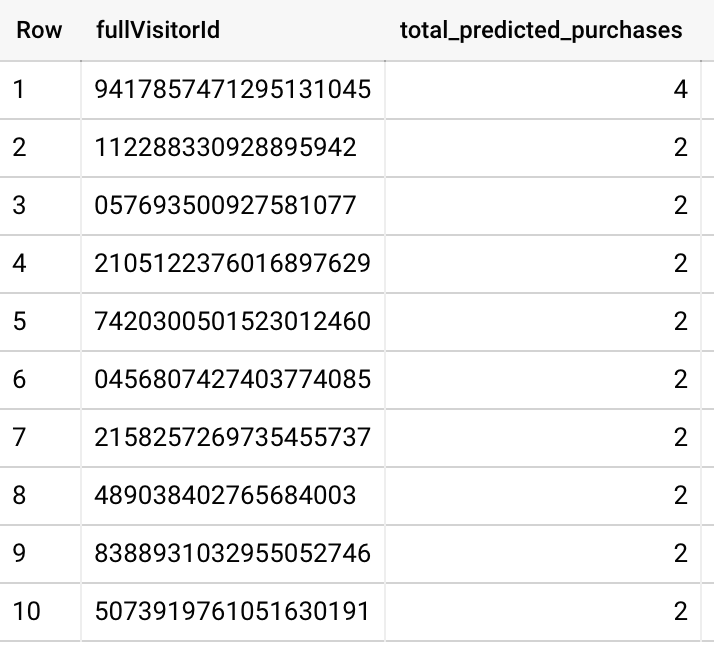
ORDER BY total\_predicted\_purchases DESC

LIMIT 10;

Copied!

content\_copy

You should see a table similar to this:



Test Completed Task