Predicting Visitor Purchases with a Classification Model with BigQuery ML

1 hour 20 minutesFree

Rate Lab

**Overview**

[BigQuery ML](https://cloud.google.com/bigquery/docs/bigqueryml-analyst-start) (BigQuery machine learning) is a new feature in BigQuery where data analysts can create, train, evaluate, and predict with machine learning models with minimal coding.

There is a newly available [ecommerce dataset](https://www.blog.google/products/marketingplatform/analytics/introducing-google-analytics-sample) 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 run some typical queries that businesses would want to know about their customers' purchasing habits.

Objectives

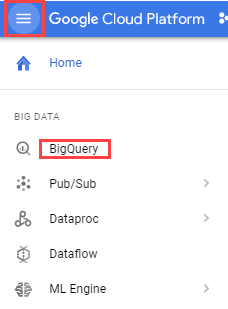
In this lab, you learn to perform the following tasks:

* Use BigQuery to find public datasets
* Query and explore the ecommerce dataset
* Create a training and evaluation dataset to be used for batch prediction
* Create a classification (logistic regression) model in BigQuery ML
* Evaluate the performance of your machine learning model
* Predict and rank the probability that a visitor will make a purchase

**Set up your environments**.

.  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 lists UI updates.

Click **Done**.

Access the course dataset

Once BigQuery is open, open on the below direct link in a new browser tab to bring the public **data-to-insights** project into your BigQuery projects panel:

<https://console.cloud.google.com/bigquery?p=data-to-insights&d=ecommerce&t=web_analytics&page=table>

The field definitions for the **data-to-insights** ecommerce dataset are [here](https://support.google.com/analytics/answer/3437719?hl=en). Keep the link open in a new tab for reference.

**Task 1. Explore ecommerce data**

**Scenario:** Your data analyst team exported the Google Analytics logs for an ecommerce website into BigQuery and created a new table of all the raw ecommerce visitor session data for you to explore. Using this data, you'll try to answer a few questions.

**Question:** Out of the total visitors who visited our website, what % made a purchase?

1. Click the **Query editor**.
2. Add the following to the New Query field:

*#standardSQL*

WITH visitors AS(

SELECT

COUNT(DISTINCT fullVisitorId) AS total\_visitors

FROM `data-to-insights.ecommerce.web\_analytics`

),

purchasers AS(

SELECT

COUNT(DISTINCT fullVisitorId) AS total\_purchasers

FROM `data-to-insights.ecommerce.web\_analytics`

WHERE totals.transactions IS NOT NULL

)

SELECT

total\_visitors,

total\_purchasers,

total\_purchasers / total\_visitors AS conversion\_rate

FROM visitors, purchaserscontent\_copy

1. Click **Run**.

The result: 2.69%

**Question:** What are the top 5 selling products?

Add the following query in the **Query editor**, and then click **Run**:

SELECT

p.v2ProductName,

p.v2ProductCategory,

SUM(p.productQuantity) AS units\_sold,

ROUND(SUM(p.localProductRevenue/1000000),2) AS revenue

FROM `data-to-insights.ecommerce.web\_analytics`,

UNNEST(hits) AS h,

UNNEST(h.product) AS p

GROUP BY 1, 2

ORDER BY revenue DESC

LIMIT 5;content\_copy

The result:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Row** | **v2ProductName** | **v2ProductCategory** | **units\_sold** | **revenue** |
| 1 | Nest® Learning Thermostat 3rd Gen-USA - Stainless Steel | Nest-USA | 17651 | 870976.95 |
| 2 | Nest® Cam Outdoor Security Camera - USA | Nest-USA | 16930 | 684034.55 |
| 3 | Nest® Cam Indoor Security Camera - USA | Nest-USA | 14155 | 548104.47 |
| 4 | Nest® Protect Smoke + CO White Wired Alarm-USA | Nest-USA | 6394 | 178937.6 |
| 5 | Nest® Protect Smoke + CO White Battery Alarm-USA | Nest-USA | 6340 | 178572.4 |

**Question:** How many visitors bought on subsequent visits to the website?

Run the following query to find out:

*# visitors who bought on a return visit (could have bought on first as well*

WITH all\_visitor\_stats AS (

SELECT

fullvisitorid, *# 741,721 unique visitors*

IF(COUNTIF(totals.transactions > 0 AND totals.newVisits IS NULL) > 0, 1, 0) AS will\_buy\_on\_return\_visit

FROM `data-to-insights.ecommerce.web\_analytics`

GROUP BY fullvisitorid

)

SELECT

COUNT(DISTINCT fullvisitorid) AS total\_visitors,

will\_buy\_on\_return\_visit

FROM all\_visitor\_stats

GROUP BY will\_buy\_on\_return\_visitcontent\_copy

The results:

|  |  |  |
| --- | --- | --- |
| **Row** | **total\_visitors** | **will\_buy\_on\_return\_visit** |
| 1 | 729848 | 0 |
| 2 | 11873 | 1 |

Analyzing the results, you can see that (11873 / 729848) = 1.6% of total visitors will return and purchase from the website. This includes the subset of visitors who bought on their very first session and then came back and bought again.

What are some of the reasons a typical ecommerce customer will browse but not buy until a later visit? Choose all that could apply.



The customer wants to comparison shop on other sites before making a purchase decision.



The customer is waiting for products to go on sale or other promotion



The customer is doing additional research

Submit

This behavior is very common for luxury goods where significant up-front research and comparison is required by the customer before deciding (think car purchases) but also true to a lesser extent for the merchandise on this site (t-shirts, accessories, etc).

In the world of online marketing, identifying and marketing to these future customers based on the characteristics of their first visit will increase conversion rates and reduce the outflow to competitor sites.

**Task 2. Select features and create your training dataset**

Now you will create a Machine Learning model in BigQuery to predict whether or not a new user is likely to purchase in the future. Identifying these high-value users can help your marketing team target them with special promotions and ad campaigns to ensure a conversion while they comparison shop between visits to your ecommerce site.

Google Analytics captures a wide variety of dimensions and measures about a user's visit on this ecommerce website. Browse the complete list of fields [here](https://support.google.com/analytics/answer/3437719?hl=en) and then [preview the demo dataset](https://bigquery.cloud.google.com/table/data-to-insights:ecommerce.web_analytics?tab=preview) to find useful features that will help a machine learning model understand the relationship between data about a visitor's first time on your website and whether they will return and make a purchase.

Your team decides to test whether these two fields are good inputs for your classification model:

* totals.bounces (whether the visitor left the website immediately)
* totals.timeOnSite (how long the visitor was on our website)

What are the risks of only using the above two fields?



Whether a user bounces is highly correlated with their time on site (e.g. 0 seconds)



Only using time spent on the site ignores other potential useful columns (features)



Both of the above

Submit

Machine learning is only as good as the training data that is fed into it. If there isn't enough information for the model to determine and learn the relationship between your input features and your label (in this case, whether the visitor bought in the future) then you will not have an accurate model. While training a model on just these two fields is a start, you will see if they're good enough to produce an accurate model.

In the **Query editor**, add the following query:

SELECT

\* EXCEPT(fullVisitorId)

FROM

*# features*

(SELECT

fullVisitorId,

IFNULL(totals.bounces, 0) AS bounces,

IFNULL(totals.timeOnSite, 0) AS time\_on\_site

FROM

`data-to-insights.ecommerce.web\_analytics`

WHERE

totals.newVisits = 1)

JOIN

(SELECT

fullvisitorid,

IF(COUNTIF(totals.transactions > 0 AND totals.newVisits IS NULL) > 0, 1, 0) AS will\_buy\_on\_return\_visit

FROM

`data-to-insights.ecommerce.web\_analytics`

GROUP BY fullvisitorid)

USING (fullVisitorId)

ORDER BY time\_on\_site DESC

LIMIT 10;content\_copy

Then click **Run**.

Results:

|  |  |  |  |
| --- | --- | --- | --- |
| **Row** | **bounces** | **time\_on\_site** | **will\_buy\_on\_return\_visit** |
| 1 | 0 | 15047 | 0 |
| 2 | 0 | 12136 | 0 |
| 3 | 0 | 11201 | 0 |
| 4 | 0 | 10046 | 0 |
| 5 | 0 | 9974 | 0 |
| 6 | 0 | 9564 | 0 |
| 7 | 0 | 9520 | 0 |
| 8 | 0 | 9275 | 1 |
| 9 | 0 | 9138 | 0 |
| 10 | 0 | 8872 | 0 |

Which fields are the model features? What is the label (correct answer)?



The features are bounces and will\_buy\_on\_return\_visit. The label is time\_on\_site



The features is will\_buy\_on\_return\_visit. The labels are bounces and time\_on\_site



The features are bounces and time\_on\_site. The label is will\_buy\_on\_return\_visit

Submit

Which fields are known after a visitor's first session? (Check all that apply)



visitId



time\_on\_site



will\_buy\_on\_return\_visit



bounces

Submit

Which field isn't known until later in the future after their first session?



bounces



will\_buy\_on\_return\_visit



visitId



time\_on\_site

Submit

**Discussion:** **will\_buy\_on\_return\_visit** is not known after the first visit. Again, you're predicting for a subset of users who returned to your website and purchased. Since you don't know the future at prediction time, you cannot say with certainty whether a new visitor come back and purchase. The value of building a ML model is to get the probability of future purchase based on the data gleaned about their first session.

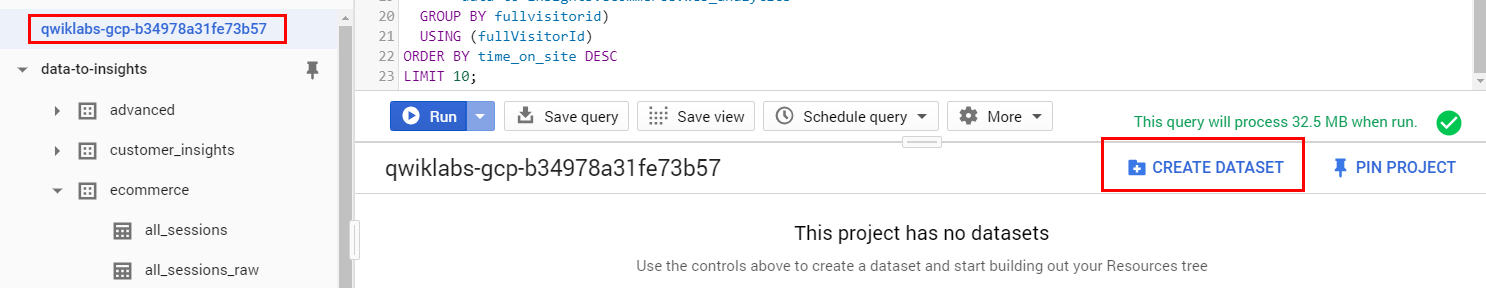
**Question:** Looking at the initial data results, do you think **time\_on\_site** and **bounces** will be a good indicator of whether the user will return and purchase or not?

**Answer:** It's often too early to tell before training and evaluating the model, but at first glance out of the top 10 time\_on\_site, only 1 customer returned to buy, which isn't very promising. Let's see how well the model does.

**Task 3. Create a BigQuery dataset to store models**

Next, create a new BigQuery dataset which will also store your ML models.

1. In the left pane, click on your project name, and then click **Create Dataset**.



1. In the **Create Dataset** dialog:

* For **Dataset ID**, type **ecommerce**.
* Leave the other values at their defaults.

1. Click **Create dataset**.

**Task 4. Select a BigQuery ML model type and specify options**

Now that you have your initial features selected, you are now ready to create your first ML model in BigQuery.

There are the two model types to choose from:

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Model Type** | **Label Data type** | **Example** |
| Forecasting | linear\_reg | Numeric value (typically an integer or floating point) | Forecast sales figures for next year given historical sales data. |
| Classification | logistic\_reg | 0 or 1 for binary classification | Classify an email as spam or not spam given the context. |

**Note:** There are many additional model types used in Machine Learning (like Neural Networks and decision trees) and available using libraries like [TensorFlow](https://www.tensorflow.org/tutorials/). At the time of writing, BigQuery ML supports the two listed above.

Which model type should you choose for will buy or wont buy?



Classification model (like logistic\_reg etc.)



Recommendation model (like matrix\_factorization etc.)



Forecasting model (like linear\_reg etc.)

Submit

Enter the following query to create a model and specify model options:

CREATE OR REPLACE MODEL `ecommerce.classification\_model`

OPTIONS

(

model\_type='logistic\_reg',

labels = ['will\_buy\_on\_return\_visit']

)

AS

*#standardSQL*

SELECT

\* EXCEPT(fullVisitorId)

FROM

*# features*

(SELECT

fullVisitorId,

IFNULL(totals.bounces, 0) AS bounces,

IFNULL(totals.timeOnSite, 0) AS time\_on\_site

FROM

`data-to-insights.ecommerce.web\_analytics`

WHERE

totals.newVisits = 1

AND date BETWEEN '20160801' AND '20170430') *# train on first 9 months*

JOIN

(SELECT

fullvisitorid,

IF(COUNTIF(totals.transactions > 0 AND totals.newVisits IS NULL) > 0, 1, 0) AS will\_buy\_on\_return\_visit

FROM

`data-to-insights.ecommerce.web\_analytics`

GROUP BY fullvisitorid)

USING (fullVisitorId)

;content\_copy

Next, click **Run** to train your model.

Wait for the model to train (5 - 10 minutes).

**Note:** You cannot feed all of your available data to the model during training since you need to save some unseen data points for model evaluation and testing. To accomplish this, add a WHERE clause condition is being used to filter and train on only the first 9 months of session data in your 12 month dataset.

After your model is trained, you will see the message "This statement created a new model named qwiklabs-gcp-xxxxxxxxx:ecommerce.classification\_model".

Click **Go to model**.

Look inside the ecommerce dataset and confirm **classification\_model** now appears.

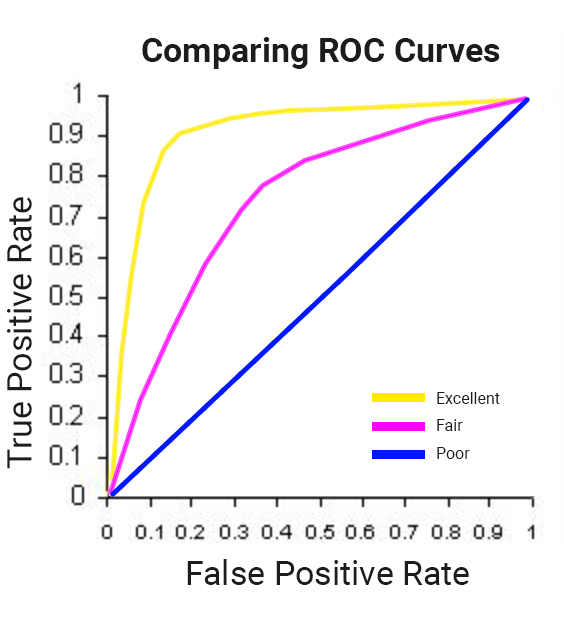
Next, you will evaluate the performance of the model against new unseen evaluation data.

**Task 5. Evaluate classification model performance**

Select your performance criteria

For classification problems in ML, you want to minimize the False Positive Rate (predict that the user will return and purchase and they don't) and maximize the True Positive Rate (predict that the user will return and purchase and they do).

This relationship is visualized with a ROC (Receiver Operating Characteristic) curve like the one shown here, where you try to maximize the area under the curve or AUC:



In BigQuery ML, **roc\_auc** is simply a queryable field when evaluating your trained ML model.

Now that training is complete, you can evaluate how well the model performs with this query using ML.EVALUATE:

SELECT

roc\_auc,

CASE

WHEN roc\_auc > .9 THEN 'good'

WHEN roc\_auc > .8 THEN 'fair'

WHEN roc\_auc > .7 THEN 'not great'

ELSE 'poor' END AS model\_quality

FROM

ML.EVALUATE(MODEL ecommerce.classification\_model, (

SELECT

\* EXCEPT(fullVisitorId)

FROM

*# features*

(SELECT

fullVisitorId,

IFNULL(totals.bounces, 0) AS bounces,

IFNULL(totals.timeOnSite, 0) AS time\_on\_site

FROM

`data-to-insights.ecommerce.web\_analytics`

WHERE

totals.newVisits = 1

AND date BETWEEN '20170501' AND '20170630') *# eval on 2 months*

JOIN

(SELECT

fullvisitorid,

IF(COUNTIF(totals.transactions > 0 AND totals.newVisits IS NULL) > 0, 1, 0) AS will\_buy\_on\_return\_visit

FROM

`data-to-insights.ecommerce.web\_analytics`

GROUP BY fullvisitorid)

USING (fullVisitorId)

));content\_copy

You should see the following result:

|  |  |  |
| --- | --- | --- |
| **Row** | **roc\_auc** | **model\_quality** |
| 1 | 0.724588 | not great |

After evaluating your model you get a **roc\_auc** of 0.72, which shows the model has not great, predictive power. Since the goal is to get the area under the curve as close to 1.0 as possible, there is room for improvement.

**Task 6. Improve model performance with feature engineering**

As was hinted at earlier, there are many more features in the dataset that may help the model better understand the relationship between a visitor's first session and the likelihood that they will purchase on a subsequent visit.

Add some new features and create a second machine learning model called classification\_model\_2:

* How far the visitor got in the checkout process on their first visit
* Where the visitor came from (traffic source: organic search, referring site etc.)
* Device category (mobile, tablet, desktop)
* Geographic information (country)

Create this second model by running the below query:

CREATE OR REPLACE MODEL `ecommerce.classification\_model\_2`

OPTIONS

(model\_type='logistic\_reg', labels = ['will\_buy\_on\_return\_visit']) AS

WITH all\_visitor\_stats AS (

SELECT

fullvisitorid,

IF(COUNTIF(totals.transactions > 0 AND totals.newVisits IS NULL) > 0, 1, 0) AS will\_buy\_on\_return\_visit

FROM `data-to-insights.ecommerce.web\_analytics`

GROUP BY fullvisitorid

)

*# add in new features*

SELECT \* EXCEPT(unique\_session\_id) FROM (

SELECT

CONCAT(fullvisitorid, CAST(visitId AS STRING)) AS unique\_session\_id,

*# labels*

will\_buy\_on\_return\_visit,

MAX(CAST(h.eCommerceAction.action\_type AS INT64)) AS latest\_ecommerce\_progress,

*# behavior on the site*

IFNULL(totals.bounces, 0) AS bounces,

IFNULL(totals.timeOnSite, 0) AS time\_on\_site,

totals.pageviews,

*# where the visitor came from*

trafficSource.source,

trafficSource.medium,

channelGrouping,

*# mobile or desktop*

device.deviceCategory,

*# geographic*

IFNULL(geoNetwork.country, "") AS country

FROM `data-to-insights.ecommerce.web\_analytics`,

UNNEST(hits) AS h

JOIN all\_visitor\_stats USING(fullvisitorid)

WHERE 1=1

*# only predict for new visits*

AND totals.newVisits = 1

AND date BETWEEN '20160801' AND '20170430' *# train 9 months*

GROUP BY

unique\_session\_id,

will\_buy\_on\_return\_visit,

bounces,

time\_on\_site,

totals.pageviews,

trafficSource.source,

trafficSource.medium,

channelGrouping,

device.deviceCategory,

country

);content\_copy

**Note:** You are still training on the same first 9 months of data, even with this new model. It's important to have the same training dataset so you can be certain a better model output is attributable to better input features and not new or different training data.

A key new feature that was added to the training dataset query is the maximum checkout progress each visitor reached in their session, which is recorded in the field hits.eCommerceAction.action\_type. If you search for that field in the [field definitions](https://support.google.com/analytics/answer/3437719?hl=en) you will see the field mapping of 6 = Completed Purchase.

As an aside, the web analytics dataset has nested and repeated fields like [ARRAYS](https://cloud.google.com/bigquery/docs/reference/standard-sql/arrays) which need to broken apart into separate rows in your dataset. This is accomplished by using the UNNEST() function, which you can see in the above query.

Wait for the new model to finish training (5-10 minutes).

Evaluate this new model to see if there is better predictive power:

*#standardSQL*

SELECT

roc\_auc,

CASE

WHEN roc\_auc > .9 THEN 'good'

WHEN roc\_auc > .8 THEN 'fair'

WHEN roc\_auc > .7 THEN 'not great'

ELSE 'poor' END AS model\_quality

FROM

ML.EVALUATE(MODEL ecommerce.classification\_model\_2, (

WITH all\_visitor\_stats AS (

SELECT

fullvisitorid,

IF(COUNTIF(totals.transactions > 0 AND totals.newVisits IS NULL) > 0, 1, 0) AS will\_buy\_on\_return\_visit

FROM `data-to-insights.ecommerce.web\_analytics`

GROUP BY fullvisitorid

)

*# add in new features*

SELECT \* EXCEPT(unique\_session\_id) FROM (

SELECT

CONCAT(fullvisitorid, CAST(visitId AS STRING)) AS unique\_session\_id,

*# labels*

will\_buy\_on\_return\_visit,

MAX(CAST(h.eCommerceAction.action\_type AS INT64)) AS latest\_ecommerce\_progress,

*# behavior on the site*

IFNULL(totals.bounces, 0) AS bounces,

IFNULL(totals.timeOnSite, 0) AS time\_on\_site,

totals.pageviews,

*# where the visitor came from*

trafficSource.source,

trafficSource.medium,

channelGrouping,

*# mobile or desktop*

device.deviceCategory,

*# geographic*

IFNULL(geoNetwork.country, "") AS country

FROM `data-to-insights.ecommerce.web\_analytics`,

UNNEST(hits) AS h

JOIN all\_visitor\_stats USING(fullvisitorid)

WHERE 1=1

*# only predict for new visits*

AND totals.newVisits = 1

AND date BETWEEN '20170501' AND '20170630' *# eval 2 months*

GROUP BY

unique\_session\_id,

will\_buy\_on\_return\_visit,

bounces,

time\_on\_site,

totals.pageviews,

trafficSource.source,

trafficSource.medium,

channelGrouping,

device.deviceCategory,

country

)

));content\_copy

(Output)

|  |  |  |
| --- | --- | --- |
| **Row** | **roc\_auc** | **model\_quality** |
| 1 | 0.910382 | good |

With this new model you now get a **roc\_auc** of 0.91 which is significantly better than the first model.

Now that you have a trained model, time to make some predictions.

**Task 7. Predict which new visitors will come back and purchase**

Next you will write a query to predict which new visitors will come back and make a purchase.

The prediction query below uses the improved classification model to predict the probability that a first-time visitor to the Google Merchandise Store will make a purchase in a later visit:

SELECT

\*

FROM

ml.PREDICT(MODEL `ecommerce.classification\_model\_2`,

(

WITH all\_visitor\_stats AS (

SELECT

fullvisitorid,

IF(COUNTIF(totals.transactions > 0 AND totals.newVisits IS NULL) > 0, 1, 0) AS will\_buy\_on\_return\_visit

FROM `data-to-insights.ecommerce.web\_analytics`

GROUP BY fullvisitorid

)

SELECT

CONCAT(fullvisitorid, '-',CAST(visitId AS STRING)) AS unique\_session\_id,

*# labels*

will\_buy\_on\_return\_visit,

MAX(CAST(h.eCommerceAction.action\_type AS INT64)) AS latest\_ecommerce\_progress,

*# behavior on the site*

IFNULL(totals.bounces, 0) AS bounces,

IFNULL(totals.timeOnSite, 0) AS time\_on\_site,

totals.pageviews,

*# where the visitor came from*

trafficSource.source,

trafficSource.medium,

channelGrouping,

*# mobile or desktop*

device.deviceCategory,

*# geographic*

IFNULL(geoNetwork.country, "") AS country

FROM `data-to-insights.ecommerce.web\_analytics`,

UNNEST(hits) AS h

JOIN all\_visitor\_stats USING(fullvisitorid)

WHERE

*# only predict for new visits*

totals.newVisits = 1

AND date BETWEEN '20170701' AND '20170801' *# test 1 month*

GROUP BY

unique\_session\_id,

will\_buy\_on\_return\_visit,

bounces,

time\_on\_site,

totals.pageviews,

trafficSource.source,

trafficSource.medium,

channelGrouping,

device.deviceCategory,

country

)

)

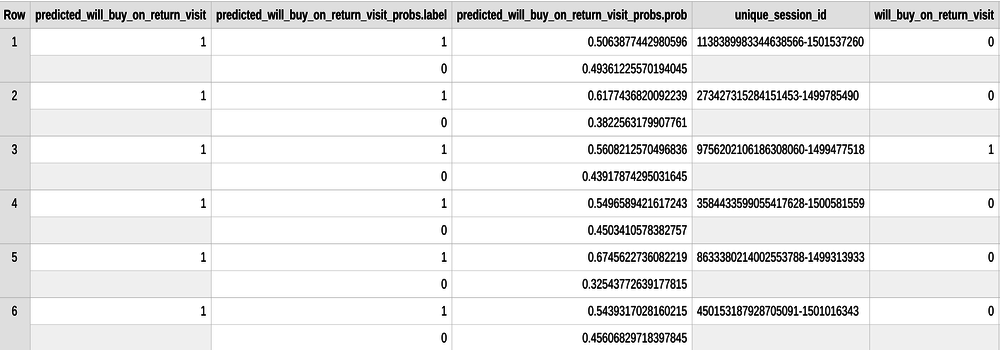
ORDER BY

predicted\_will\_buy\_on\_return\_visit DESC;content\_copy

The predictions are made on the last 1 month (out of 12 months) of the dataset.

Your model will now output the predictions it has for those July 2017 ecommerce sessions. You can see three newly added fields:

* predicted\_will\_buy\_on\_return\_visit: whether the model thinks the visitor will buy later (1 = yes)
* predicted\_will\_buy\_on\_return\_visit\_probs.label: the binary classifier for yes / no
* predicted\_will\_buy\_on\_return\_visit.prob: the confidence the model has in it's prediction (1 = 100%)

**Results**

* Of the top 6% of first-time visitors (sorted in decreasing order of predicted probability), more than 6% make a purchase in a later visit.
* These users represent nearly 50% of all first-time visitors who make a purchase in a later visit.
* Overall, only 0.7% of first-time visitors make a purchase in a later visit.
* Targeting the top 6% of first-time increases marketing ROI by 9x vs targeting them all!

Additional information

**roc\_auc** is just one of the performance metrics available during model evaluation. Also available are [accuracy, precision, and recall](https://en.wikipedia.org/wiki/Precision_and_recall). Knowing which performance metric to rely on is highly dependent on what your overall objective or goal is.

Congratulations!

You created a machine learning model using just SQL.