INDICE

1.- Introducción

2.- Caso 1: Uso de BQML para predecir usuarios cobrados.

3.- Caso 2: Mejora del modelo de ejemplo de Google sobre la Google Merchandise Store (GMS)

4.- Conclusiones

5.- Documentación

INTRODUCCIÓN

Se van a intentar emplear técnicas de ML para la construcción de modelos de predicción que sean capaces de aumentar el rendimiento de dos casos de negocio. Para esto vamos a usar la aplicación de Machine Learning de BigQuery, se van a explorar los datasets de Google Analytics en BigQuery, se optimizarán y evaluarán distintos modelos hasta conseguir uno aceptable con el cual haremos predicciones de resultados. Después y por último, compararemos estas predicciones con los datos reales y se extraerán conclusiones.

CASO 1: USO DE BIGQUERY ML PARA PREDECIR USUARIOS COBRADOS

El objetivo de este ejercicio es predecir qué usuarios van a ser cobrados y cuáles no en un día concreto. El dataset empleado son los datos reales de Google Analytics de Digital Virgo Spain (DVS).

Como primer acercamiento a los datos, los visitantes del sitio web de DVS pueden suscribirse a sus servicios, lo que se contabilizará como transacción. Después DVS puede conseguir cobrarles la suscripción o no, dependiendo del estado de facturación de su línea telefónica. El % de conversión a transacción es menor a un 1%, y de la transacción al cobro (first billing) un 70% aproximadamente. El objetivo es predecir este first billing para invertir sólo en las transacciones que van a ser cobradas.

Para elaborar este ejercicio se usará la interfaz gráfica de BigQuery ML.

Modelo 1:

Para elaborar el primer modelo, se introdujeron solamente los usuarios que tenían suscripción, además de variables como la fecha, la hora en formato “momento del día”, marca del dispositivo, navegador; y métricas como número de sesiones y transacciones. La consulta quedó así:

Se usó solo a los usuarios suscritos ya que con el ratio de conversión actual el dataset es muy desbalanceado.

-- Modelo1

CREATE OR REPLACE MODEL `wide-oasis-135923.iamarketingdvs.modelodefinitivo1`

OPTIONS(model\_type='logistic\_reg') AS

SELECT

IFNULL(SUM(( SELECT value FROM UNNEST(hits.customMetrics) WHERE index = 2)),0) AS label,

date, case when hits.hour > 7 and hits.hour < 13 then 'manana' when hits.hour > 13 and hits.hour < 20 then 'tarde' else 'noche' end as momentodia, hits.hour as hora, clientId, geoNetwork.city as ciudad, device.deviceCategory as categoriaDispositivo, device.mobileDeviceBranding as marca, device.browser as navegador, trafficSource.source as fuente, trafficSource.campaign as campana, (SELECT value FROM UNNEST(session.customDimensions) WHERE index = 14 GROUP BY 1) AS producto, device.operatingSystem as OS, sum(totals.visits) as visitas, CASE WHEN sum(totals.transactions) < 1 THEN 0 ELSE totals.transactions END as subs

FROM `wide-oasis-135923.140393857.ga\_sessions\_\*` as session,

UNNEST(hits) AS hits

where \_table\_suffix BETWEEN '20190101' AND '20191001'

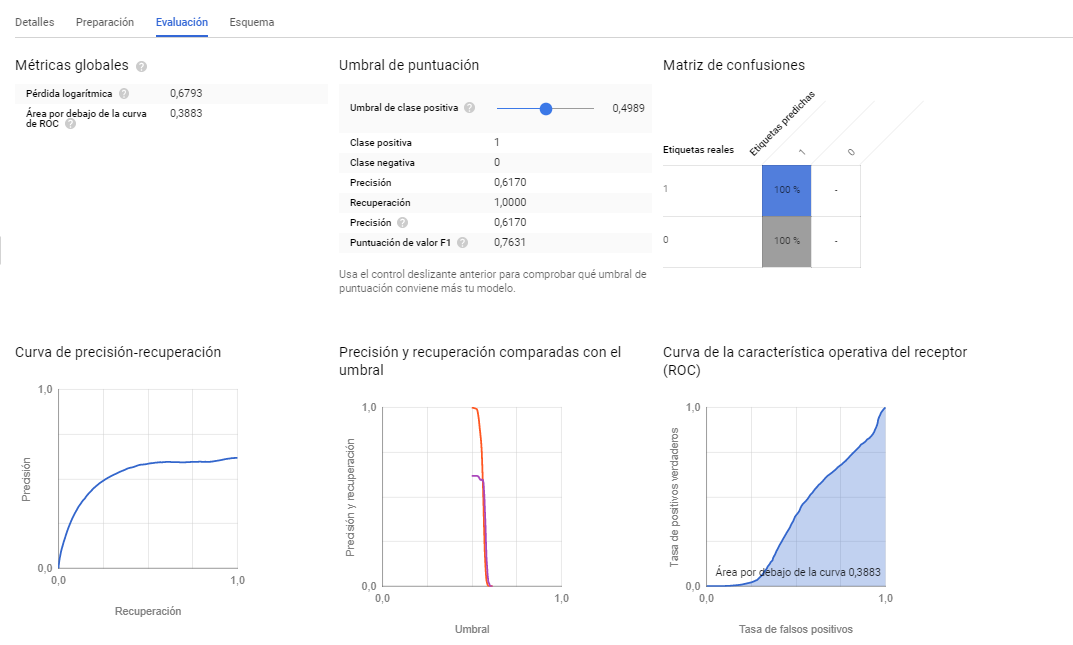
and trafficSource.campaign like '%range%'

and trafficSource.source like '%google%'

and totals.transactions > 0

GROUP BY date, clientId, ciudad, marca, categoriaDispositivo, totals.transactions, navegador, campana, fuente, producto, OS, hora

Los resultados de la evaluación fueron estos:



Como se puede ver, el modelo es malo. Predice el 61% de los casos de cobro sobre los suscritos, sin embargo, lo hace porque siempre dice que el suscrito es un cobro, con lo que acierta siempre el cobro, pero nunca el no cobro (recall 100%).

Lo siguiente fue afinar, se retiraron filas como las suscripciones (siempre 1), la campaña, la fuente de tráfico y la categoría del dispositivo entre otras, eran columnas que siempre tenían el mismo valor, por tanto no aportaban nada al modelo.

Como se puede ver, la evaluación del modelo es mala (y un tanto extraña), basándonos en el área bajo la curva < 0.5, realmente sería peor que lanzar una moneda. Esto es incongruente porque no hay nada más aleatorio que lanzar una moneda, por tanto un área inferior al 50% lo que nos indica es que el modelo está mal formado, hay fallos graves de diseño.

Fijándome en las variables utilizadas, caigo en la cuenta de que algunas son innecesarias (como por ejemplo el nombre de campaña o la fuente de tráfico, casi siempre la misma) y otras siempre tienen el mismo valor, como las transacciones (siempre 1).

Elaboraremos entonces un segundo modelo ajustando los parámetros con los que componemos el dataset de train.

La consulta para la creación del segundo modelo quedó de esta forma:

CREATE OR REPLACE MODEL `wide-oasis-135923.iamarketingdvs.modelodefinitivo2`

OPTIONS(model\_type='logistic\_reg') AS

SELECT

IFNULL(SUM(( SELECT value FROM UNNEST(hits.customMetrics) WHERE index = 2)),0) AS label,

date, case when hits.hour > 7 and hits.hour < 13 then 'manana' when hits.hour > 13 and hits.hour < 20 then 'tarde' else 'noche' end as momentodia, clientId, geoNetwork.city as ciudad, device.mobileDeviceBranding as marca, device.browser as navegador, trafficSource.campaign as campana, (SELECT value FROM UNNEST(session.customDimensions) WHERE index = 14 GROUP BY 1) AS producto, device.operatingSystem as OS, sum(totals.visits) as visitas

FROM `wide-oasis-135923.140393857.ga\_sessions\_\*` as session,

UNNEST(hits) AS hits

where \_table\_suffix BETWEEN '20190101' AND '20191001'

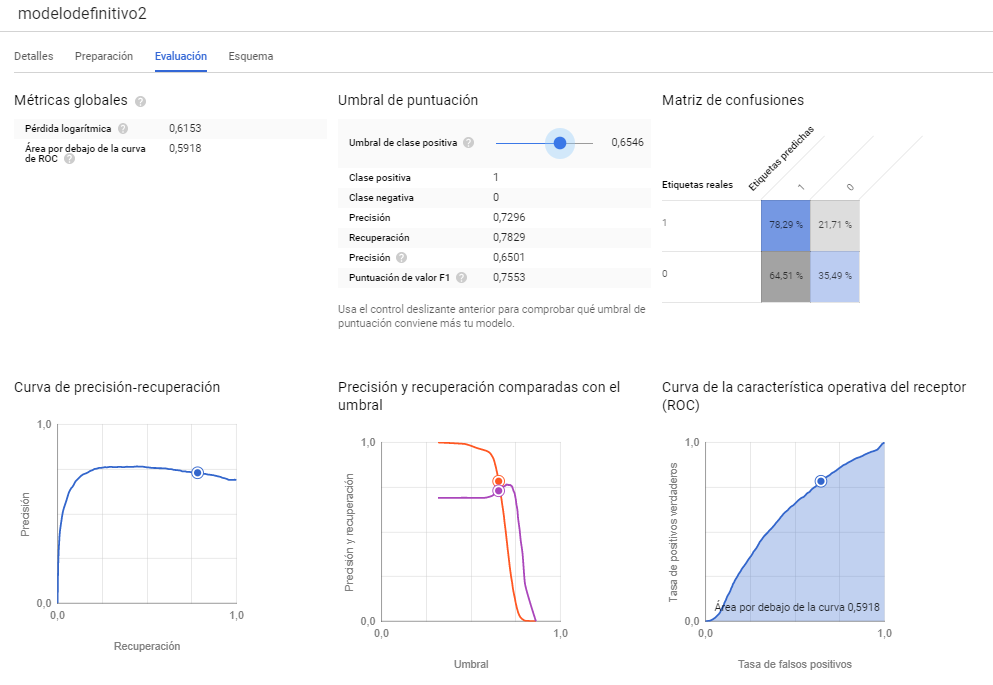
and trafficSource.campaign like '%range%'

and trafficSource.source like '%google%'

and totals.transactions > 0

GROUP BY date, clientId, ciudad, marca, navegador, campana, producto, OS, momentodia

Su evaluación fue esta:



Primero de todo, vemos que un área bajo la curva > 0.5, esto quiere decir que ya tenemos algo mejor que lanzar una moneda al aire, por lo que sabemos positivamente que el ajuste al anterior modelo es positivo, había tanto dimensiones como métricas que no deben estar. Al obtener un modelo con área > 0.5, en la herramienta para evaluar modelos de BQ ML lo primero que vamos a hacer es elegir un umbral de clasificación conveniente para nuestro objetivo (predecir primer cobro en las transacciones). Usando el control deslizante y chequeando cómo va variando la matriz de confusión, decido que 0,6546 es el umbral de clasificación que mejor resultados nos va a reportar, ya que es el que más “distancia” obtiene entre los verdaderos positivos y los falsos positivos 78,29% de verdadero positivo y 64,51% de falso positivo. Es cierto que hay muchos falsos positivos, pero no olvidemos que nuestro modelo tan solo tiene un 0.59 de área bajo la curva ROC, por tanto no parece que podemos aspirar a mucho más.

El recall en ambos modelos es del 100%, sospecho que el dataset tiene mucho que ver. No hay métricas entre la transacción y el billed, solo dimensiones. Cuando vimos el caso práctico en clase, la clave era la construcción de métricas personalizadas basándonos en datos existentes, sin embargo, por el modelo de negocio de este caso, no hay métricas intermedias disponibles entre la suscripción y el cobro.

De todas formas, vamos a hacer el predict para completar el proceso:

SELECT

\*

FROM ML.PREDICT(MODEL `wide-oasis-135923.iamarketingdvs.modelodefinitivo2`, (

SELECT

IFNULL(SUM(( SELECT value FROM UNNEST(hits.customMetrics) WHERE index = 2)),0) AS label,

date, case when hits.hour > 7 and hits.hour < 13 then 'manana' when hits.hour > 13 and hits.hour < 20 then 'tarde' else 'noche' end as momentodia, hits.hour as hora, clientId, geoNetwork.city as ciudad, device.deviceCategory as categoriaDispositivo, device.mobileDeviceBranding as marca, device.browser as navegador, trafficSource.source as fuente, trafficSource.campaign as campana, (SELECT value FROM UNNEST(session.customDimensions) WHERE index = 14 GROUP BY 1) AS producto, device.operatingSystem as OS, sum(totals.visits) as visitas, CASE WHEN sum(totals.transactions) < 1 THEN 0 ELSE totals.transactions END as subs, IFNULL(SUM(( SELECT value FROM UNNEST(hits.customMetrics) WHERE index = 2)),0) AS Fbilled

FROM `wide-oasis-135923.140393857.ga\_sessions\_\*` as session,

UNNEST(hits) AS hits

where \_table\_suffix BETWEEN '20191002' AND '20191030'

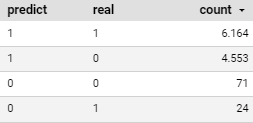
and trafficSource.campaign like '%range%'

and trafficSource.source like '%google%'

and totals.transactions > 0

GROUP BY date, clientId, ciudad, marca, categoriaDispositivo, totals.transactions, navegador, campana, fuente, producto, OS, hora))

Mediante esta matriz de confusión improvisada en datastudio



Podemos comprobar la evaluación, muchas precisión, pero a base de predecir que todos los usuarios son cobrados, acierta un 60% de las veces porque hay un 60 % de cobro aproximadamente.

CASO 2. MEJORA DEL MODELO DE GOOGLE SOBRE EL GOOGLE MERCHANDISE STORE

Para trabajar con un dataset con más riqueza de datos, esto es, con unas métricas más significativas que las del dataset de DVS, vamos a escoger el dataset de la Google Merchandising Store. Vamos a construir un modelo, evaluarlo + mejorarlo y después predecir resultados con el mejor modelo que hayamos sido capaces de construir.

Para poder trabajar en un modelo algo *mejor* *predictor*, hemos escogido el dataset de la Google Merchandise Store, con él pretendemos crear un modelo que funcione “normalmente” (mejor o peor) y predecir los usuarios que van a suscribirse un día concreto. Hay 1 año de datos, por lo que vamos a entrenarlo con los primeros 11 meses y predecir el día siguiente:

La consulta que generaría un primer modelo para hacer un acercamiento a los datos sería esta:

CREATE OR REPLACE MODEL `ml-small.MLsmalldataset.modeloGoogleTiendaGogle1`

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,

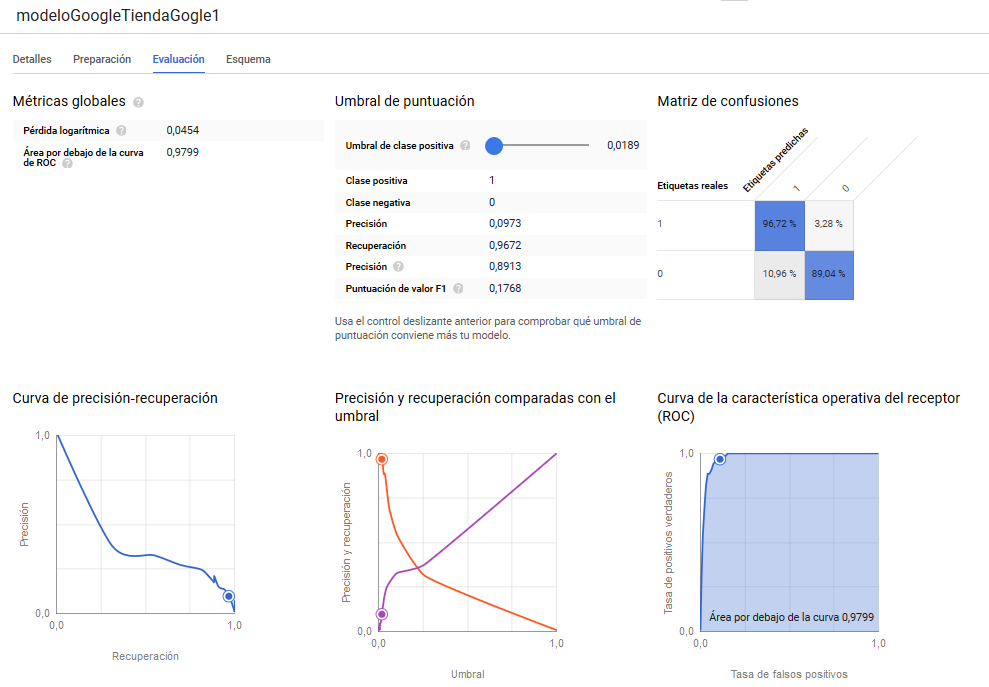
IFNULL(geoNetwork.country, "") AS country,

IFNULL(totals.pageviews, 0) AS pageviews

FROM

`bigquery-public-data.google\_analytics\_sample.ga\_sessions\_\*`

WHERE \_TABLE\_SUFFIX BETWEEN '20160801' AND '20170701'

La evaluación quedó así: 

El modelo es excelente, simplemente con el número de sesiones, el país y el sistema operativo, obtenemos un modelo con un área de 0.9799 bajo la curva ROC. Esto nos permite elegir un umbral de 0.0189, es decir, todas las predicciones con una probabilidad de transacción superior a 1.89%, las vamos a etiquetar como transacción. Debido al gran área bajo la curva, podemos ser muy ambiciosos con la elección de este umbral, ya que tener un 96,72% de verdaderos positivos sólo nos genera un 10.96% de falsos positivos. Este gran modelo es tan difícil de mejorar como poco práctico, ya que tiene muy poco margen de mejora, sin embargo vamos a ver cómo contribuyen algunas modificaciones a su rendimiento.

En un primer paso, vamos a añadir algunas dimensiones y métricas nuevas al modelo, quedando así la consulta:

Los datos en esta ocasión son malos, pero constan de margen de mejora (o eso creo). Este *precission* nos dice que el modelo predice correctamente el 46% de las veces que lo hace, sin embargo, solo ha abarcado un 11% de los positivos reales (recall). Vamos a intentar aumentar estas métricas de rendimiento del modelo añadiendo algunas columnas más al dataset de entrenamiento del modelo:

CREATE OR REPLACE MODEL `ml-small.MLsmalldataset.upgradeTiendaGoogle1`

OPTIONS(model\_type='logistic\_reg') AS

SELECT

IF(totals.transactions IS NULL, 0, 1) AS label,

fullvisitorId as cliente,

IFNULL(device.operatingSystem, "") AS os,

device.isMobile AS is\_mobile,

IFNULL(geoNetwork.country, "") AS country,

IFNULL(geoNetwork.city, "") AS ciudad,

IFNULL(device.deviceCategory, "") AS categoriaDispositivo,

IFNULL(totals.pageviews, 0) AS pageviews,

IFNULL(totals.visits, 0) AS sesiones,

IFNULL(totals.UniqueScreenViews, 0) AS visitasUnicas,

IFNULL(totals.timeOnSite, 0) AS TiempoEnWeb,

IFNULL(totals.sessionQualityDim, 0) AS calidadSesion

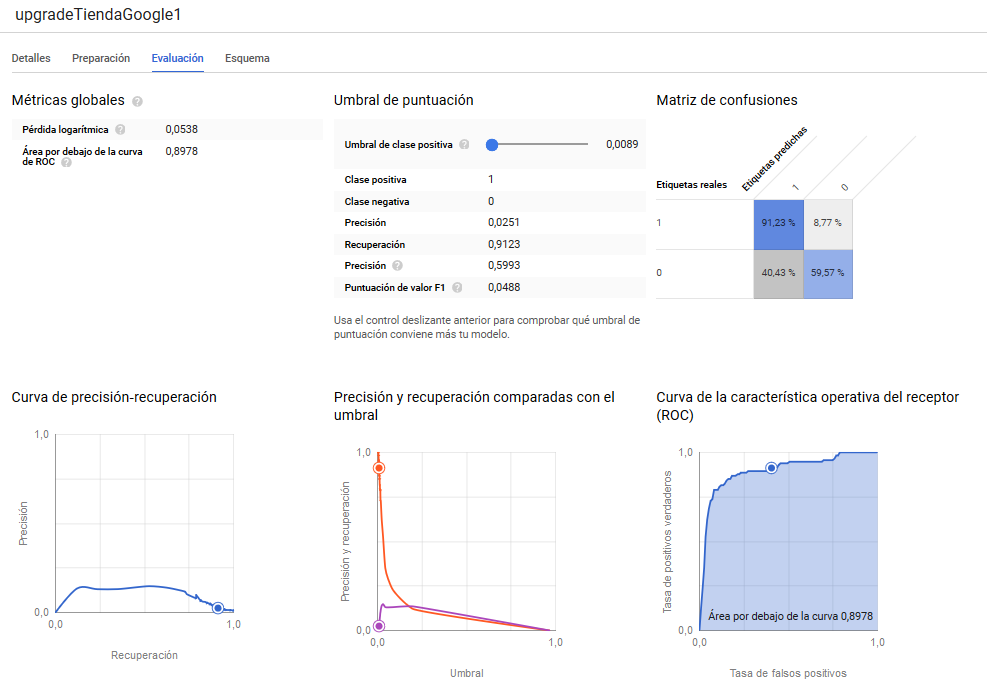
FROM

`bigquery-public-data.google\_analytics\_sample.ga\_sessions\_\*`

WHERE

\_TABLE\_SUFFIX BETWEEN '20160801' AND '20170630'

La evaluación quedó de esta manera:



Parece que hemos empeorado el modelo y hemos disminuido unos 10 puntos el área bajo la curva. Esto ha hecho que eligiendo el umbral que más acerca el número de verdaderos positivos al valor del primer modelo, nos genere un 40,43% de falsos positivos (unos 30 puntos superior en comparación con el primer modelo). Este efecto ha podido ser resultado de haber añadido dimensionalidad intrascendente en vez de métricas relevantes al modelo, vamos a crear un tercero sólo añadiendo las métricas al original:

CREATE OR REPLACE MODEL `ml-small.MLsmalldataset.upgradeTiendaGoogle2`

OPTIONS(model\_type='logistic\_reg') AS

SELECT

IF(totals.transactions IS NULL, 0, 1) AS label,

IFNULL(device.operatingSystem, "") AS os,

IFNULL(totals.pageviews, 0) AS pageviews,

IFNULL(totals.visits, 0) AS sesiones,

IFNULL(totals.UniqueScreenViews, 0) AS visitasUnicas,

IFNULL(totals.timeOnSite, 0) AS TiempoEnWeb,

IFNULL(totals.sessionQualityDim, 0) AS calidadSesion

FROM

`bigquery-public-data.google\_analytics\_sample.ga\_sessions\_\*`

WHERE

\_TABLE\_SUFFIX BETWEEN '20160801' AND '20170630'

Su correspondiente evaluación:

SELECT

\*

FROM ML.EVALUATE(MODEL `ml-small.MLsmalldataset.upgradeTiendaGoogle3`, (

SELECT

IF(totals.transactions IS NULL, 0, 1) AS label,

IFNULL(device.operatingSystem, "") AS os,

IFNULL(totals.pageviews, 0) AS pageviews,

IFNULL(totals.visits, 0) AS sesiones,

IFNULL(totals.UniqueScreenViews, 0) AS visitasUnicas,

IFNULL(totals.timeOnSite, 0) AS TiempoEnWeb,

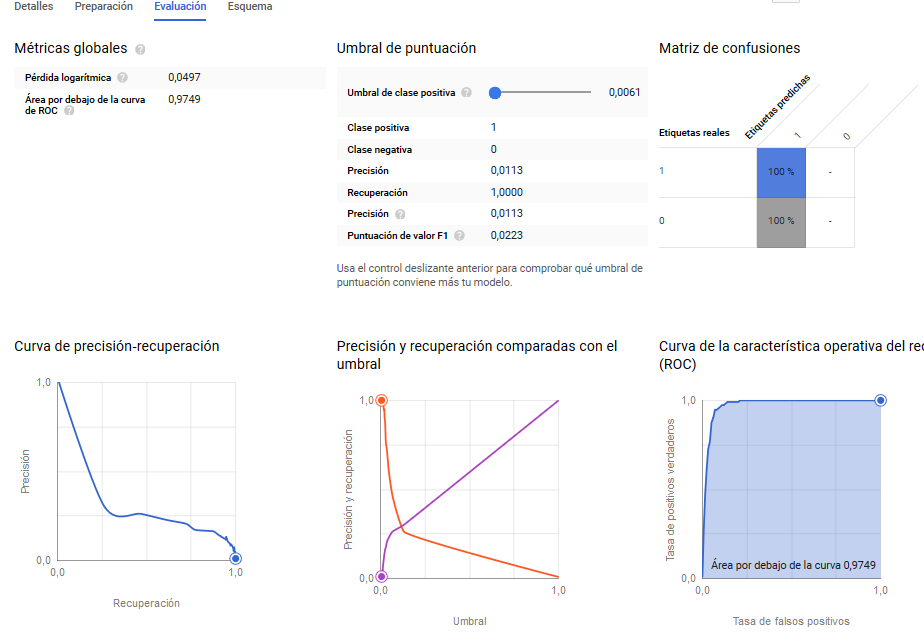
IFNULL(totals.sessionQualityDim, 0) AS calidadSesion

FROM

`bigquery-public-data.google\_analytics\_sample.ga\_sessions\_\*`

WHERE

\_TABLE\_SUFFIX BETWEEN '20170701' AND '20170801'))





Este modelo efectivamente tiene menos dimensiones, sin embargo sigue siendo peor que el primero, el más sencillo.

Como no conseguimos mejorar el modelo añadiendo métricas ni dimensiones, vamos a modificar el primero modelo, es decir, mismo número de columnas, pero con diferentes valores:

Vamos a cambiar el país por ciudad, las visitas a páginas por sesiones y el sistema operativo por el navegador:

CREATE OR REPLACE MODEL `ml-small.MLsmalldataset.upgradeTiendaGoogle4`

OPTIONS(model\_type='logistic\_reg') AS

SELECT

IF(totals.transactions IS NULL, 0, 1) AS label,

IFNULL(device.browser, "") AS navegador,

IFNULL(geoNetwork.city, "") AS city,

IFNULL(totals.visits, 0) AS sesiones

FROM

`bigquery-public-data.google\_analytics\_sample.ga\_sessions\_\*`

WHERE \_TABLE\_SUFFIX BETWEEN '20160801' AND '20170701'

La evaluación queda de esta forma:

SELECT

\*

FROM ML.EVALUATE(MODEL `ml-small.MLsmalldataset.upgradeTiendaGoogle4`, (

SELECT

IF(totals.transactions IS NULL, 0, 1) AS label,

IFNULL(device.browser, "") AS navegador,

IFNULL(geoNetwork.city, "") AS city,

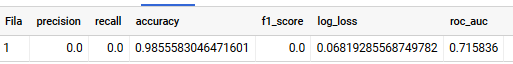
IFNULL(totals.visits, 0) AS sesiones

FROM

`bigquery-public-data.google\_analytics\_sample.ga\_sessions\_\*`

WHERE

\_TABLE\_SUFFIX BETWEEN '20170701' AND '20170801'))



Para encontrarle explicación a estos datos, vamos a explorar las columnas nuevas, es conocido que el dataset de muestra de la Merchandise Store de Google tiene algunos campos no disponibles por temas de privacidad o seguridad:

SELECT

IF(totals.transactions IS NULL, 0, 1) AS label,

IFNULL(device.browser, "") AS navegador,

IFNULL(geoNetwork.city, "") AS city,

IFNULL(totals.visits, 0) AS sesiones

FROM

`bigquery-public-data.google\_analytics\_sample.ga\_sessions\_\*`

WHERE \_TABLE\_SUFFIX BETWEEN '20170801' AND '20170801'



Todo parece correcto, sin embargo, vamos a hacer un último intento cambiando algunas métricas del modelo inicial, hasta el momento el mejor:

CREATE OR REPLACE MODEL `ml-small.MLsmalldataset.upgradeTiendaGogle5`

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 pais,

IFNULL(totals.timeOnSite, 0) AS tiempo,

IFNULL(totals.pageviews, 0) AS pageviews

FROM

`bigquery-public-data.google\_analytics\_sample.ga\_sessions\_\*`

WHERE \_TABLE\_SUFFIX BETWEEN '20160801' AND '20170630'

La evaluación:

SELECT

\*

FROM ML.EVALUATE(MODEL `ml-small.MLsmalldataset.upgradeTiendaGogle5`, (

SELECT

IF(totals.transactions IS NULL, 0, 1) AS label,

IFNULL(device.operatingSystem, "") AS os,

device.isMobile AS is\_mobile,

IFNULL(geoNetwork.country, "") AS pais,

IFNULL(totals.timeOnSite, 0) AS tiempo,

IFNULL(totals.pageviews, 0) AS pageviews

FROM

`bigquery-public-data.google\_analytics\_sample.ga\_sessions\_\*`

WHERE

\_TABLE\_SUFFIX BETWEEN '20170701' AND '20170801'))



No conseguimos añadir métricas y dimensiones que mejoren el modelo inicial. Sin embargo vamos a probar a cambiar los parámetros por defecto de la regresión logística:

CREATE OR REPLACE MODEL `ml-small.MLsmalldataset.upgradeTiendaGoogle6`

OPTIONS(model\_type='logistic\_reg',learn\_rate\_strategy='constant',data\_split\_method= 'random',data\_split\_eval\_fraction = 0.15,learn\_rate= 0.6,l1\_reg=0.15,auto\_class\_weights = true) 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 '20170701'

La evaluacion queda así:

SELECT

\*

FROM ML.EVALUATE(MODEL `ml-small.MLsmalldataset.upgradeTiendaGoogle6`, (

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,

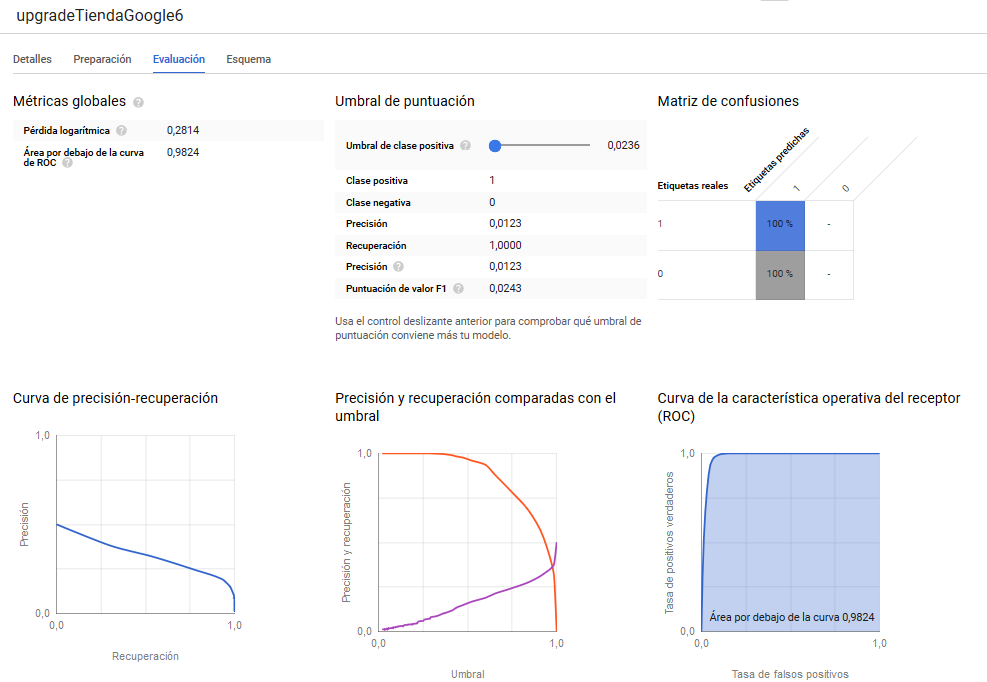
IFNULL(totals.pageviews, 0) AS pageviews

FROM

`bigquery-public-data.google\_analytics\_sample.ga\_sessions\_\*`

WHERE

\_TABLE\_SUFFIX BETWEEN '20170701' AND '20170801'))





Vemos como la precisión ha bajado, sin embargo el recall ha aumentado mucho, esto es, de los que predecimos como suscritos, el 18% lo son realmente, sin embargo, de todos los suscritos reales, hemos acertado un 97%.

Lo que realmente nos gusta es el primer modelo, no hemos conseguido mejorarlo de ninguna manera, así que vamos a predecir transacciones por usuario, sistema operativo y país en el mes de julio.

Para hacerlo usamos la función PREDICT de BQ ML, como ya usamos la de CREATE MODEL o la de EVALUATE.

SELECT

os, country, fullvisitorId,

SUM(predicted\_label) as totalTransactionsPredicted

FROM ML.PREDICT(MODEL `ml-small.MLsmalldataset.modeloGoogleTiendaGogle1`, (

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 '20170731'))

GROUP BY os, country, fullvisitorId

ORDER BY totalTransactionsPredicted DESC

Vamos a hacer ahora un análisis de esta predicción y de la realidad. Para ello vamos a hacer la tabla de las suscripciones reales. Para esto haremos la misma consulta y guardaremos los resultados como otra métrica llamada suscripciones\_reales. Después compararemos.

SELECT

sum(totals.transactions) as realTransactions,

IFNULL(device.operatingSystem, "") AS os,

sum(totals.pageviews) AS pageviews,

IFNULL(geoNetwork.country, "") AS country,

fullvisitorId

FROM

`bigquery-public-data.google\_analytics\_sample.ga\_sessions\_\*`

WHERE

\_TABLE\_SUFFIX BETWEEN '20170701' AND '20170731'

GROUP BY os, country, fullvisitorId

ORDER BY realTransactions DESC

Después de que el Join no funcione y pierda casi 2 horas de tiempo, caigo en la cuenta de que la tabla de transacciones reales tiene valores nulos en vez de ceros ☹ cambio la consulta:

SELECT

case when sum(totals.transactions) > 0 then sum(totals.transactions) else 0 end as transaccionesReales,

IFNULL(device.operatingSystem, "") AS os,

IFNULL(geoNetwork.country, "") AS country,

fullvisitorId

FROM

`bigquery-public-data.google\_analytics\_sample.ga\_sessions\_\*`

WHERE

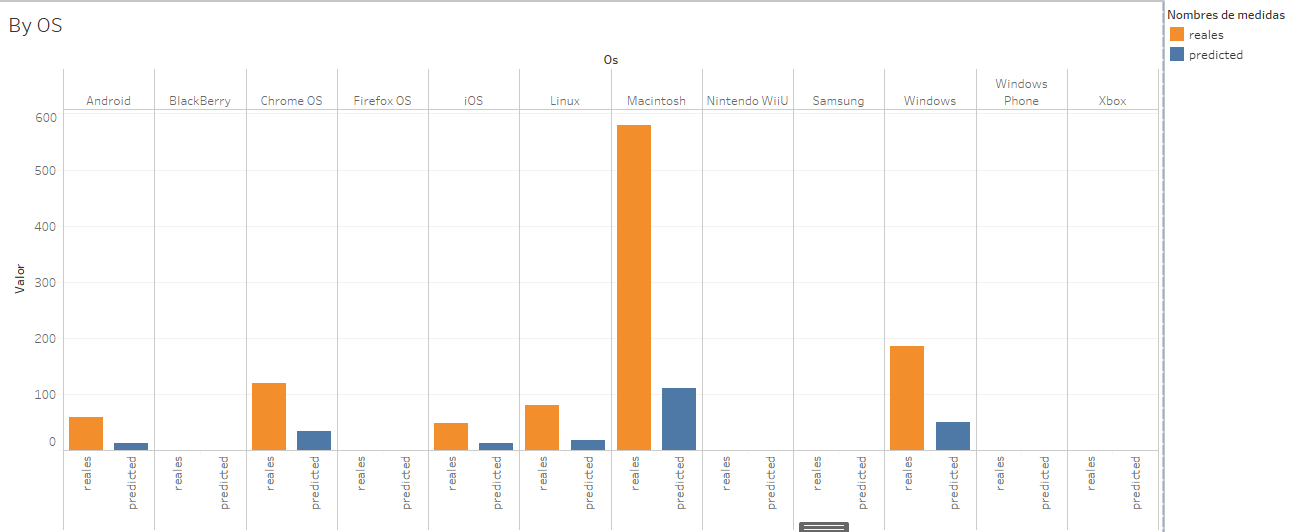
\_TABLE\_SUFFIX BETWEEN '20170701' AND '20170731'

GROUP BY os, country, fullvisitorId

ORDER BY transaccionesReales DESC

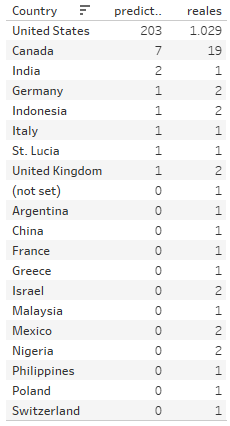
Vamos a guardar los resultados en dos nuevas tablas de BigQuery y las exploraremos en Tableau haciendo con este software la Join usando como clave las tres dimensiones, SO, coutry y fullvisitorId. Lo ideal es llevar la tabla a Tableau lo más limpia posible y preprocesada, pero como todo el proyecto es SQL, he querido hacer algo más en Tableau.

Subsreales vs subs predichas por SO:



Se puede ver, como ya notamos en la fase de evaluación, que hay más transacciones reales que predichas, se nota que el *recall* es relativamente bajo, sin embargo todos los usuarios predichos hicieron transacción en el mes de Julio, una *precission* muy alta. Sin embargo, también puede notarse cierto buen funcionamiento del modelo, ya que no es sólo el número de transacciones lo que predice, sino que las asigna con bastante lógica a cada dimensión, es decir, el mayor número de transacciones predichas está en Macintosh, son menos de las reales, pero también son el máximo y así sucesivamente.

Con respecto a los datos por país, pasa lo mismo, la distribución es algo mayor entre dimensiones, pero los comentarios podrían ser los mismos.



CONCLUSIÓN:

Se han entrenado modelos en dos datasets diferentes, dos modelos de negocio distintos, un modelo de suscripción doublé optin con un 97% de transacciones de usuarios en su primera sesión y un modelo de tienda online clásico como es el de la Google Merchandise Store.

En el primer caso, el modelo impide una predicción real de un objetivo ya que no existen a penas métricas disponibles entre los usuarios que visitan el site y los que son cobrados por las transacciones. El flujo de suscripción en doublé-optin, lo que quiere decir que se necesitan sólo 2 clics para realizar la transacción, lo que es bueno y malo. Bueno por performance de negocio, malo por la casi nula susceptibilidad de optimización aplicando técnicas de machine learning.

En el segundo caso se ha intentado mejorar el modelo de ejemplo que viene en los tutoriales de Google sobre BQ ML, el que se ha versionado para aplicar en el primer caso. En este segundo caso las expectativas eras buenas ya que el modelo usaba solo 3 dimensiones y 2 métricas, pensé que añadiendo más complejidad con nuevas métricas y dimensiones iba a poder conseguir una evaluación mejor que la primera, sin embargo, no fue así. Se aplicaron 3 tácticas, con estos resultados:

1.- Aumentar métricas y dimensiones:



2.- Aumentar sólo métricas:



3.- Mantener dimensionalidad, pero cambiando parámetros:



4.- mantener las dimensiones originales aplicando los parámetros de regresión logística aplicados en el caso práctico de Iberia:



La conclusión final es que no se ha conseguido generar ningún modelo aceptable más que el que venía dado de ejemplo con la Merchandise Store, parece que cuando nos enfrentamos a casos reales, el camino hasta la construcción de modelos no es tan llano y obvio como en los ejemplos preparados de los pétalos o los coches con la cilindrada, la velocidad máxima y la potencia que hemos ido haciendo en clase.

Aún más difícil que la construcción de un buen modelo, es prever y elegir qué tipo de modelo aplicar a cada caso de uso real en el negocio de cada uno.

DOCUMENTACION:

Sintaxis de en BQ ML:

<https://cloud.google.com/bigquery-ml/docs/reference/standard-sql/bigqueryml-syntax-create>

How to query and calculate Google Analytics Data in BigQuery

<https://towardsdatascience.com/how-to-query-and-calculate-google-analytics-data-in-bigquery-cab8fc4f396#f2c8>

Esquema raw de campos de BigQuery Export

<https://support.google.com/analytics/answer/3437719?hl=es>

Primeros pasos con BigQuery ML en la IU web

<https://cloud.google.com/bigquery-ml/docs/bigqueryml-web-ui-start>

W3schools SQL Tutorial

<https://www.w3schools.com/sql/default.asp>

Lucas Díaz García - Diciembre 2019

Modelos de regresión logística en datasets de Google Analytics

TFM Data Science (Kschool)