Ambre Cazaux

Data Analyst Portfolio

Role: Data Analyst / Engineering
Duration: 1 weeks
Tools: GA4, BigQuery, GCP, Python,

Project of building a churn prediction model using BigQuery ML to identify which customers were likely to stop purchasing from us after their first order. The goal was to proactively flag at-risk users and enable marketing to re-engage them before losing them.

GA4 Data Extraction & Cleaning:

```
CREATE OR REPLACE TABLE `bold-bucksaw-446504-a4.ga4_analysis.churn_analysis` AS
  -- CTE 1
WITH ordered_purchases AS (
  SELECT
    user_pseudo_id,
    PARSE_DATE('%Y%m%d', date) AS purchase_date,
    purchase_revenue,
    item_category,
    medium AS campaign_source,
    ROW_NUMBER() OVER (PARTITION BY user_pseudo_id ORDER BY date) AS order_rank
  FROM
    `bold-bucksaw-446504-a4.ga4_analysis.ga4_portfolio_data`
  WHERE
    purchase_revenue IS NOT NULL
    AND item_category IS NOT NULL
),
  -- CTE 2
first_and_second_orders AS (
  SELECT
    user_pseudo_id,
    MAX(CASE WHEN order_rank = 1 THEN purchase_date END) AS first_order_date,
    MAX(CASE WHEN order_rank = 2 THEN purchase_date END) AS second_order_date,
    MAX(CASE WHEN order_rank = 1 THEN purchase_revenue END) AS first_purchase_amount,
    MAX(CASE WHEN order_rank = 1 THEN item_category END) AS product_category,
    MAX(CASE WHEN order_rank = 1 THEN campaign_source END) AS campaign_source
```

```
  FROM
    ordered_purchases
  GROUP BY
    user_pseudo_id
)
  --Final Query
SELECT
  user_pseudo_id,
  first_purchase_amount,
  product_category,
  campaign_source,
  -- Flag churn if the first order is more than 90 days ago and no second purchase
  CASE
    WHEN first_order_date < CURRENT_DATE - INTERVAL 90 DAY AND second_order_date IS
NULL THEN 1
    ELSE 0
  END AS is_churned,
  -- Calculate time to second order only if second order exists
  DATE_DIFF(second_order_date, first_order_date, DAY) AS time_to_second_order
FROM
  first_and_second_orders;
```

Creation of the model:

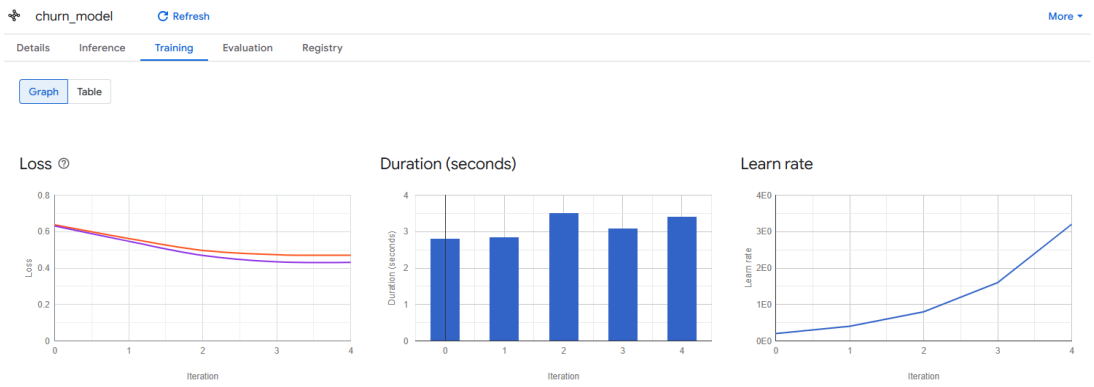
```
CREATE MODEL `bold-bucksaw-446504-a4.ga4_analysis.churn_model`
OPTIONS (
  model_type = 'logistic_reg',
  input_label_cols = ['is_churned']
) AS
SELECT
  first_purchase_amount,
  product_category,
  campaign_source,
  is_churned
FROM
  `bold-bucksaw-446504-a4.ga4_analysis.churn_analysis`;
```

Row	user_pseudo_id	first_purchas...	product_category	campaign_source	is_churned	time_to_sec...
1	50121199.4323445332	3.0	Shop by Brand	(data deleted)	1	null
2	1147529.6357293919	4.0	Accessories	(data deleted)	1	null
3	67251653.8910982513	6.0	Stationery	(data deleted)	0	3
4	9004033.5064832188	6.0	Google	(data deleted)	1	null
5	6135662.0853737840	7.0	Uncategorized Items	(data deleted)	1	null
6	50683324.7270306410	8.0	Accessories	(data deleted)	1	null
7	6436495.4831121763	8.0	Accessories	(data deleted)	1	null
8	7474537327.4725745851	10.0	Shop by Brand	(data deleted)	0	0
9	20687527.9358545857	10.0	Electronics Accessories	(data deleted)	1	null
10	8992560.0877088817	10.0	New	(data deleted)	1	null
11	2346008.0354634581	10.0	Lifestyle	(data deleted)	1	null
12	57736746.4807313510	10.0	Apparel	(data deleted)	1	null

Model Creation:

```
CREATE MODEL `bold-bucksaw-446504-a4.ga4_analysis.churn_model`
OPTIONS (
  model_type = 'logistic_reg',
  input_label_cols = ['is_churned']
) AS
SELECT
  first_purchase_amount,
  product_category,
  campaign_source,
  is_churned
FROM
  `bold-bucksaw-446504-a4.ga4_analysis.churn_analysis`;
```

Model evaluation:



Predict Churn Probability for Each Customer:

Row	user_pseudo_id	predicted_is_chur...	churn_probability....	churn_probability....
1	50121199.4323445332	1	1	0.802726740324...
			0	0.197273259675...
2	1147529.6357293919	1	1	0.835133873913...
			0	0.164866126086...
3	67251653.8910982513	1	1	0.887161038670...
			0	0.112838961329...
4	9004033.5064832188	1	1	0.827162689696...
			0	0.172837310303...
5	6135662.0853737840	1	1	0.820975956928...
			0	0.179024043071...
6	50683324.7270306410	1	1	0.835142602663...
			0	0.164857397336...
7	6436495.4831121763	1	1	0.835142602663...
			0	0.164857397336...
8	7474537327.4725745851	1	1	0.802744308763...

The model had strong interpretability and allowed us to identify key churn predictors, such as users who made small first purchases or waited longer before placing a second order.

For example, users from organic channels churned less, while those from paid campaigns with small first purchases churned more.

We used the model's probabilities to create at-risk user segments, which the CRM team then targeted with re-engagement emails and offers.