-- date\_labels is 1 day after date\_candidates (which is the current batch run's start date)

DECLARE date\_candidates, date\_labels DATE;

DECLARE positive\_rate FLOAT64;

SET date\_candidates = (SELECT DATE(TIMESTAMP\_MILLIS($start\_time$)));

SET date\_labels = DATE\_ADD(date\_candidates, INTERVAL 1 DAY);

CREATE TABLE IF NOT EXISTS `twttr-recos-ml-prod.realgraph.labeled\_candidates$table\_suffix$` AS

SELECT

0 AS source\_id,

1 AS destination\_id,

1 AS label,

1 AS num\_days,

1 AS num\_tweets,

1 AS num\_follows,

1 AS num\_favorites,

1 AS num\_tweet\_clicks,

1 AS num\_profile\_views,

1 AS days\_since\_last\_interaction,

1 AS label\_types,

DATE("2023-01-08") AS ds;

-- delete any prior data to avoid double writing

DELETE

FROM `twttr-recos-ml-prod.realgraph.labeled\_candidates$table\_suffix$`

WHERE ds = date\_candidates;

-- join labels with candidates with 1 day attribution delay and insert new segment

INSERT INTO `twttr-recos-ml-prod.realgraph.labeled\_candidates$table\_suffix$`

WITH label\_positive AS (

SELECT source\_id, destination\_id

FROM `twttr-bq-cassowary-prod.user.interaction\_graph\_labels\_daily`

WHERE DATE(dateHour)=date\_labels

), label\_negative AS (

SELECT source\_id, destination\_id

FROM `twttr-bq-cassowary-prod.user.interaction\_graph\_agg\_negative\_edge\_snapshot`

) SELECT

F.source\_id,

F.destination\_id,

CASE WHEN P.source\_id IS NULL THEN 0 ELSE 1 END AS label,

num\_days,

num\_tweets,

num\_follows,

num\_favorites,

num\_tweet\_clicks,

num\_profile\_views,

days\_since\_last\_interaction,

label\_types,

date\_candidates AS ds

FROM `twttr-recos-ml-prod.realgraph.candidates\_sampled` F

LEFT JOIN label\_positive P USING(source\_id, destination\_id)

LEFT JOIN label\_negative N USING(source\_id, destination\_id)

WHERE N.source\_id IS NULL AND N.destination\_id IS NULL

AND F.ds=date\_candidates

;

-- get positive rate

SET positive\_rate =

(SELECT SUM(label)/COUNT(label) AS pct\_positive

FROM `twttr-recos-ml-prod.realgraph.labeled\_candidates$table\_suffix$`

);

-- create training dataset with negative downsampling (should get ~50-50 split)

-- this spans over the cumulative date range of the labeled candidates table.

CREATE OR REPLACE TABLE `twttr-recos-ml-prod.realgraph.train$table\_suffix$` AS

SELECT \* FROM `twttr-recos-ml-prod.realgraph.labeled\_candidates$table\_suffix$`

WHERE CASE WHEN label = 0 AND RAND() < positive\_rate THEN true WHEN label = 1 AND RAND() < (1-positive\_rate) THEN true ELSE false END

;