Table of Contents

Audience rules in DE and UK:

There is not a clear audience rule and not a clear definition for user activity to be used for building email audience rule.

Latest Finding: email activity (num. opens and clicks) provides a broad cohort of audience to be used for email audience rule.

Problem:

There are about K900 (937867) active user_ids in subscription tables from 2020-01-01 (5 months and 2 weeks) that accounts for 40k users per week, BUT only about 20% (208,643) receive emails (i.e., 9k per week). This means that about 80% (643673) of users (about 30k per week) out of reach.[1]

Question: What are the main reasons behind this mismatch?

Our approach was to identify different audience rules and compare the audience size. The rules are constructed based on user subscription, registration, validation, and website visit behaviour within the time limit of 45 days. Since the rules are under construction, in this document the definitions are in pyspark format. This document uses data from subscription, ssa, registration and website visit tables, it is found based on these data sets; the best scenario we can add 3k (in 22 weeks) to the audience size (using Rule4 in this document). However, if we use email activity (i.e., number of open and click), this will lead us to extensive audience size. This step is under progress.

[1] The code is written in Pyspark and analysis is conducted in EMR, AWS sandbox. This specific query in the code: check_visits_rule.crosstab("SubVal_hasVisit", "isactive").show()

Methodology and findings:

Define different audience rules and compare the audience size, combining data from SSA data (including anonymous users) with registration-website visit-subscription.

Step1: Define and compare rules:

Five rules are introduced and mapped to the active audience. [2]

- Rule1: 'subVal_hasAnySSA_45d': (SSA_any_45d >= 1) & (validated_account > 0) & (sub_state > 0):
    - #isactive == 1 & Rule1 ==1: 185708
    - #isactive == 1 & Rule1 ==0/null: 668875+85722= 754,597
    - #isactive == 0 & Rule1 ==1: 2428
    - #isactive == 0 & Rule1 ==0: 546081

- Rule2: subVal_hasLVSSA_45d: (SSA_LV_45d >= 1) & (validated_account > 0) & (sub_state > 0)
    - #isactive == 1 & Rule2 ==1: 175421
    - #isactive == 1 & Rule2 ==0/null: 679162+85722=764884
    - #isactive == 0 & Rule2 ==1: 2131*
    - #isactive == 0 & Rule2 ==0: 546378

- Rule3: visits45d_sub_date_45d: (visits45d | sub_date_45d)+sub+val: "subVal_hasVisit_45d" >= 1 or "reg_date" 0 to 45 days ago
    - #isactive == 1 & Rule3 ==1: 219032
    - #isactive == 1 & Rule3 ==0/null: 635551+85722 = 721273
    - #isactive == 0 & Rule3 ==1: 5220
    - #isactive == 0 & Rule3 ==0: 543289

- Rule4: ssaAny45d_sub_date_45d: (ssaAny45d | sub_date_45d)+sub+val: "subVal_hasAnySSA_45d" >= 1 or "reg_date" 0 to 45 days ago
    - #isactive == 1 & Rule4 ==1: 235821*
    - #isactive == 1 & Rule4 ==0/null: 618762+85722=704484*
    - #isactive == 0 & Rule4 ==1: 54235
    - #isactive == 0 & Rule4 ==0: 494274*

- Rule5: Rule5: visits45d_sub_date: "subVal_hasVisit_45d" >= 1
    - #isactive == 1 & Rule5 ==1: 219032
    - #isactive == 1 & Rule5 ==0/null:635551+85722=721273
    - #isactive == 0 & Rule5 ==1: 5220
    - #isactive == 0 & Rule5 ==0:543289

[2] Note: the rules are constructed based on user subscription, registration, validation, website visit, within the time limit of 45 days. Since the rules are under construction, in this document the definitions are in pyspark format.

Conclusion:

- Rule4 is comparatively better than the other rules on three bases of: greatest coverage: 23k
- Smallest missing active users: 70k (although the number is still big, it is between 2k to 8k smaller than the other rules)
- Smallest non-active users that are not considered by the rule: 500k (although the number is still big, it is about 6k smaller than the other rules)
    - Rule2 is better than Rule4 on sending about 3k less emails to nun-active users (2k, 5k, comparatively).

Step2: Understand the Rules

Step2.1)- how old are the registrations that are active but are not included in audience rules?

As expected, there is an decremental trend in the age of registrations when considering all five rules (ratio of 2020 to 2019: 3/5):

- 2018: 450k (458199)
- 2019: 460k (459309)
- 2020: 300k (319544)

Step2.2) understand isactive == TRUE & Rule4 == 1 for Rule4

How old are the registrations that are active and covered in Rule4?

There is an incremental trend in the age of registration considering Rule4 (i.e., this rule is covering mostly recent candidates; ratio is about double)

- 2018: 40k (40151)
- 2019: 65k (66207)
- 2020: 130k (129463)

Step2.3) understand isactive == 1 & Rule4 == 0 for Rule4

How old are the registrations that are active and NOT covered in Rule4?

Rule4 is behaving correctly chronologically (ratio of 2020 to 2019: 1/5)

- 2018: 250k (256,724)
- 2019: 280k (282,083)
- 2020: 79k (79,955)*

Conclusion:

Rule4 is covering recently active users better that the combined five rules, as the ratio of active users who are not covered from 2020 to 2019 is 3/5 (i.e., 300k/460k) for the combined five rules but 1/5 (i.e., 79k/280k) for rule 4

Step3: Email Activity

To extend the audience size, email activity data set is analysed to be used in building the rules. The question is: "How is the email activity (i.e., total open and total click) of active users considering the five audience rules."

Finding: There is a lot of activity in number of open and click of email that we can use to build the audience rule.

How many total opens the users have had who were NOT considered in either of the five rules (Rule1,2,3,4,5) were not considered but user were active in last week?

- #total open
    - #IJM total opens in 7days: 48,417,040 (7m per day)
    - #JA total opens in 7 days: 42,258,487
    - #others total opens in 7 days: 68,405
- #total clicks
    - #IJM total click in 7days: 1,866,331
    - #JA total click in 7 days: 17,956,005
    - #other total clicks in 7 days: 3,973

#total open:

- How many total opens users had who were considered?
- in Rule1 and were active in last week?
    - #IJM total open in 7days: 11k (11,349)
    - #JA total open in 7 days: 6k (6,228)
- in Rule2 and were active in last week?
    - #IJM total open in 7days: 11k (10,969)
    - #JA totalo open in 7 days: 5.5k (5,448)
- in Rule3 and were active in last week?
    - #IJM total open in 7days: 51k (51,624)
    - #JA total open in 7 days: 15k (14,721)
- in Rule4 and were active in last week?
    - #IJM total open in 7days: 52k (52,165)
    - #JA total open in 7 days: 16k (16,355)
- in Rule5 and were active in last week?
    - #IJM total open in 7days: 13k (13,558)
    - #JA total open in 7 days: 6k (6,369)

Conclusion:

- Rule4 covers about 40k more total opens than the other rules.
- When we use email event we can cover relatively more number of recent new users.

Notes regarding the logic and data cleaning

1. There are miss matches in different data sets. we use (tracking_my_stepstone_id as user_id) to join data form SSA table that records anonymous behavior to the website.
   # tracking_my_stepstone_id contains the same user id (same integer), but it is for not logged in identified users

2. In this script Considered_date is the maximum of registration dates! This is due to reducing the cost of big-data analysis (memory and time)

3. Why outer join? to show the overlap between web_email_df+ssa and subscription table
   - subscription df - exclude rows with manualunsubscribe (this is done as: (manualunsubscribe is null or not manualunsubscribe)
   - Due to reduce the long time of processing big data in SandBox we filter data on last date, and exclude rows where rule == 0

4. Question:
   #Rule4 is built on (any_ssa > 0) that means all listing types in DE.