The Helping Group Assignment

Exercise for Operations Analyst Working Student
We are doing an initiative to reactivate our recurring customers after they cancel their ongoing booking. Your task is to analyse the data using the csv-files provided.

● After a customer cancels a recurring cleaning booking, we try to win the customers back.

● We split the customers in three groups: control group, a group that receives a reactivation email and a group that receives a reactivation call. The customers are assigned randomly to each of these groups.

Please share all the SQL queries and code that you wrote for this. If you have any questions about the case don't hesitate to email us.

Question 1:

How well do the different reactivation initiatives work? Please answer this question using pure SQL.

Answer: Detailed analysis in python file alongwith SQL queries.

Assumption: Even though it is a dummy data, I assume the data to be valid to do statistical analysis.

Underlying Logic: To understand the efficacy of our reactivation initiatives we need to understand that how many customers took reappointments after the concerned initiative. We can know this by **analysing the timestamps** in cancellations **(CTS)**, appointments **(ATS)** and market intervention timestamp **(MTS)** on **Treatment Group (TG).** The different permutations of CTS, ATS and MTS will give different reasons behind consumer decisions.

Also, some customers have taken re-appointments irrespective of getting called or mailed. This is the **Control Group (CG)** and will serve as the **benchmark** to assess the efficacy of our marketing initiative.

The time differential between cancellation timestamp and re-apppontiment timestamp is also crucial. If it is small, it strengthens the hypothesis that our marketing strategy had a causal relationship in the decision to re-activate our plans.

But, if it is more than some other factors might be at play. Eg may be an upcoming birth-day party, anniversary, etc

Note: If this number is substantially large, it will dilute the efficacy of our reactivation initiative and enforce the hypothesis that some other variable is at play.

If number of reactivations in control group = number of reactivations in treatment group, it means that we fail to reject the null hypothesis and the marketing strategy was not effective. Financially speaking ROI (tends to) → 0. Ideally, it should be more.

Total Customers → 5000
Control Group (Neither call nor mail) → 1622
Treatment Group (Calls / Emails) → 1684 (Calls) + 1694 (Emails) → 3378

THERE ARE 6 CASES OVERALL for Market Strategy Treatment group:

- Cancel → Strategy (CALL/Email)→ Appointment → POSITIVE RESULT
- Appointment → Strategy → Cancel → NEGATIVE RESULT

- Cancel → Appointment → Strategy → Positive (No effect)
- Appointment → Cancel → Strategy →NEGATIVE (No effect)

- Strategy → Appointment → Cancel → POSITIVE,
  - **BUT, CHANGED DECISION, may be unsatisfied with the service**

- Strategy → Cancel → Appointment→ Negative,
  - **BUT, AGAIN TOOK THE APPOINTMENT may be an upcoming outlier like birthday party, anniversary party, etc**

AND Two CASES FOR CONTROL GROUP:

- Appointment → Cancel
- Cancel → Appointment

Question 2:

Are the differences between the groups statistically significant? Please answer this question by writing Python code. Please also explain the approach you used for calculating statistical significance.

Answer: **In Python FILE along-with explanations in the markdowns.**

Underlying LOGIC: 'LIKE COMPARES LIKE'

Question 3:

How would you calculate the return on investment for this initiative? What additional information would you need?

Method 1: General view

Efficacy of marketing strategy → directly proportional to ROI

For example, on an average without marketing intervention we were earning 'X' USD. Let us say that the conversion rate increased the bookings 3 times. So it can be inferred that ROI from it increased 3 times than it was before.

METHOD 2: Financial View

ROI = Monetary value gained due to reactivation / Expense on Marketing Intervention

Additional information required :

- Avg. revenue per customer (A.R.)
- Increase in the number of bookings after market intervention (I)
- Expense on Marketing Intervention (E)

$$ROI = (A.R. * I) / E$$

Question 4 - Bonus

How would you improve the design of the experiment or collect additional data to make sure we get as many learnings out of the experiment as possible?

- **Customer Segregation:** Based on age, location, gender, occupation, etc
    - Knowing Occupational diversity of our customers can help us better price targeting. Geographic and Demographic information can help us in improving our services. Eg Improving punctuality and avoiding delay leading to more customer satisfaction.

- **Customers with higher Frequency**: eg. Customer 1329 has the most appointment timestamps (4), these can be well targeted with discounts and different pricing.

- **Timestamps can be more specific by a simple SQL query:**
    - SELECT CURRENT_TIMESTAMP;
    - By getting time specific timestamps we can know better when to call so that more customers pick our call leading to increase in conversion rate.
    - Time differential between cancellation timestamp and re-appointment timestamp can be used as a metric to better understand the efficacy of our marketing strategy.

- **Additional DATA:** Linking data from 'SISTER' companies. Every Industry is linked to another. For example for THE HELPING GROUP, data from Real-Estate Sector can be of great help. It can help in predicting the demand of the appliance retro-fitting, cleaning services, etc. Knowing your position in the entrepreneurial ecosystem can understand the linkages between different sectors.
    - Similar is the case with new development projects in Berlin
    - One industry's demand is another industry's supply and vice-versa

- Further, more analysis on email types can be done.

Question 5 – Bonus

We used fake data for the exercise. How can you tell the data is not real?

- All 5000 customer_ids have done cancellation. A total of 1694 customers are in the email treatment group. But, only 579 have received the re-activation mail.

OR

- Half of the calls not answered is another indicator since a random experiment has a probability of 0.5 and well thought off marketing strategy should have had phone pick-ups sufficiently more.

- If we have email time-stamp with type reactivation before the cancellation time-stamp, the dataset fails the test data validity and can be termed 'dummy' or 'fake'. After all why would we mail for re-activation if he/she has not cancelled the booking.