Take-Home Test

[Data Analyst]

Question 1 - a:
Assume that you work for an eCommerce business and want to extract the data to identify how much would be an appropriate cost of the customer acquisition (CAC). First of all, the CAC can depend on how many items the customers have bought in their first purchase and how much the value of the items was. Write a query to print customer_id, name, the date of their first purchase, the number of the items that the customer purchased on that day, and the total price of the items.

Input Format

Customers

Column	Type	Description
id (pk)	Integer	Primary key of customers
name	String	Customer name
created_at	Timestamp	Timestamp the customer signed up

Items

Column	Type	Description
id (pk)	Integer	Primary key of items
name	String	Name of items
price	Integer	Price of items
created_at	Timestamp	Timestamp the item was uploaded

Orders

Column	Type	Description
id (pk)	Integer	Primary key of order
customer_id (fk)	Integer	Foreign key from **Customers**
item_id (fk)	Integer	Foreign key from **Items**
created_at	Timestamp	datestamp the order was made

Examples

Customers

id	name	created_at
1	Sam	2018-07-12
2	Jimmy	2019-01-22

items

id	name	price	created_at
1	Watch	$200	2007-04-03
2	Belt	$75	2004-04-06
3	Wallet	$150	2010-07-21

Orders

id	customer_id	item_id	created_at
1	1	1	2018-08-01
2	1	2	2018-08-11
3	2	1	2019-01-22
4	2	3	2019-01-22
5	2	2	2019-03-03

Expected Result

customer_id	name	first_purchase_date	n_items	total_price
1	Sam	2018-08-01	1	$200
2	Jimmy	2019-01-22	2	$350

Answer:

Query

```
SELECT X.CUSTOMER_ID,
            MAX(X.NAME)             AS NAME,
            MAX(X.CREATED_AT) AS FIRST_PURCHASE_DATE,
            COUNT(1)                AS N_ITEMS,
            SUM(X.PRICE)            AS TOTAL_PRICE
FROM (
   SELECT A.CUSTOMER_ID,
               B.NAME,
               C.PRICE,
               A.CREATED_AT,
               CASE WHEN DENSE_RANK() OVER(PARTITION BY A.CUSTOMER_ID
ORDER BY A.CREATED_AT) = 1 THEN 1 ELSE 0 END AS IS_FIRST_ORDER
   FROM     ORDERS A,
               CUSTOMERS B,
               ITEMS C
   WHERE  A.CUSTOMER_ID = B.ID
               AND A.ITEM_ID    = C.ID
) X
WHERE      X.IS_FIRST_ORDER = 1
GROUP BY X.CUSTOMER_ID
;
```

* Please feel free to answer the question with any SQL grammar that you are most familiar with.

Question 1 - b:
You got the data from Question 1 - a and now are thinking that the CAC should be different from month to month. Assuming that you would like to validate your questioning by hypothesis testing, write two hypotheses to identify that the total price of the customers' first purchase will be different from month to month and also a statistical methodology to test it.

H0: $\mu_1 = \mu_2 = \mu_3 \ldots = \mu_k$
(If data from January to December are given, $\mu_1 = \mu_2 = \mu_3 \ldots = \mu_{12}$)
HA: Means are not all equal.

How to test:

First, a normality test for each group is performed.

Second, if the normality test is satisfied, ANOVA(parametric test) is performed. Otherwise, Kruskal-Wallis Test(non-parametric test) is performed.

Question 1 - c:
Explain the possible outputs from your approach and how to interpret the results.

Answer:

In the situation of ANOVA, the F-value and the p-value for the corresponding value are calculated, and in the situation of the Kruskal-Wallis Test, the H-value and the p-value for the corresponding value are calculated.

If the calculated p-value is less than the predetermined significance level, the null hypothesis is rejected. This means that the total price of the customer's first purchase is different from month to month. Therefore, it is necessary to do a post-hoc test to see how it differs from month to month.

If the p-value is greater than the predetermined significance level, the null hypothesis is accepted. This means that the total price of the customer's first purchase is the same.

Question 1 - d:

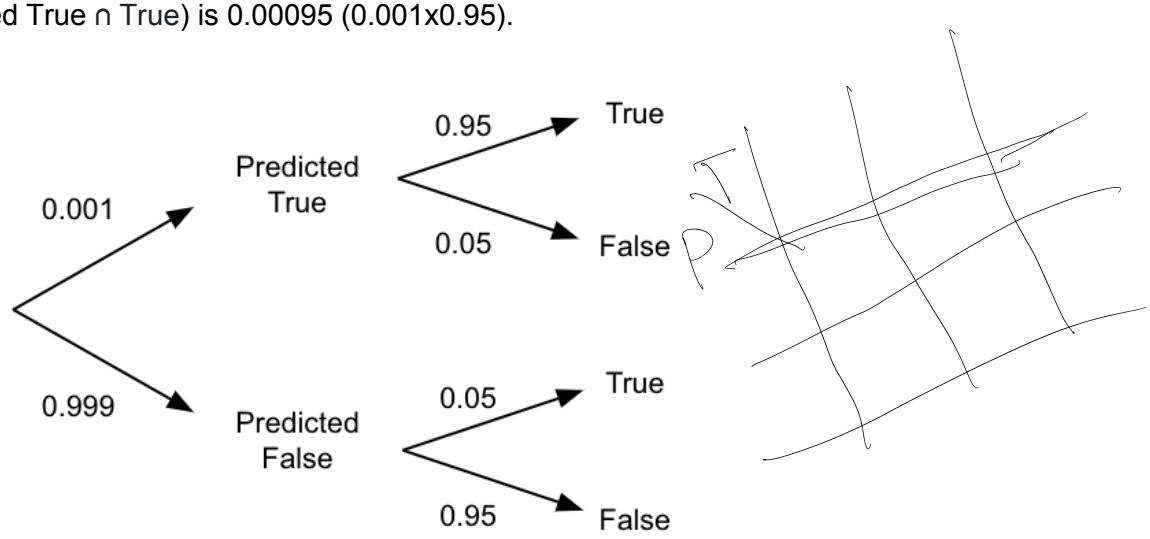
Assume that you got "$115" as an average of the customer's first purchase value for this month. Please write what you think of this number. Your answer could deal with whether the CAC should be higher or lower than $115, or what other assumptions or numbers do you need to decide the appropriate CAC?

Answer:

X _____ : 추리할 필요가 없는 오소

- 왜냐면 an average of the customer's first purchase value가 같다는 가정이 필요함.
- LTV = 115 × Expected Number of Payment periods × average gross margin
- LTV : CAC = 3 : 1 라고 선정함
- 115 × 12 × 0.

Question 2:
The below flowchart shows the probability of the cancer tumor classifier. The three-digit decimals at first indicate the probability of each predicted output that the classification model will return. The two-digit decimals after that represent if the predicted output is actually true or false. For instance, the probability of the case the classifier returns true and it is actually true (true positive, or P(Predicted True ∩ True) is 0.00095 (0.001x0.95).



What is the probability that someone who actually has the cancer tumor but also got the output True from the classifier at first. In other words, calculate P(Predicted True | True).

Answer:

The intersection of events Predicted True and Predicted False is the empty set(Predicted True ∩ Predicted False = ∅), and the union of the two events is the sample space(Predicted True ∪ Predicted False = Ω). Therefore, it is possible to utilize Bayesian rule to which the law of overall probability is added. In the end, the probability for a given event can be expressed in the following way.

$$P(Predicted\ True \mid True) = \frac{P(True \mid Predicted\ True) * P(Predicted\ True)}{P(True,\ \ Predicted\ True) + P(True,\ \ Predicted\ False)}$$

Based on the above equation, the answer is 0.01866(0.00095 / (0.00095 + 0.04995)).

$$P(PT \mid T) = \frac{P(T \mid PT) * P(PT)}{P(T \mid PT) + P(T \mid \overline{PT})}$$

$$P(T \mid PT) = \frac{P(T \cap PT)}{P(PT)} \qquad P(T \mid \overline{PT}) = \frac{P(T \mid \overline{PT})}{P(\overline{PT})}$$