Email Click-Through Rate (ECR) Analysis

I. INTRODUCTION

1.1 Research Question and Business Importance

Email marketing is a typical marketing channel that lets businesses share new products, sales, and updates with customers on their contact list. The number of clicks through to the related email is one of the important metrics measuring how successful email marketing campaigns are. ECR is one of the quickest ways to announce new products' information and one of the most efficient strategies for drawing potential purchases from customers and engaging them.

However, despite the fact that ECR is a powerful metric for assessing how engaged a company's customers are, it is often difficult for businesses to grasp what truly impacts the ECR and how to increase it efficiently, given that numerous factors influence whether a consumer will click on an email. Therefore, our research question is: what are the important factors impacting the percentage of customers who click on the link inside the marketing emails, so-called the email click-through-rate (ECR), and how to improve it?

1.2 Key expectations

Numerous variables can have a significant influence on the ECR. We expect our project outcome will inform us which area the company should focus on to improve ECR. We would also like to have a model that can correctly predict the ECR of customers based on their features.

1.3 How would the project be useful to managers/stakeholders

Managers and other stakeholders can benefit from the results of an ECR study by using the information to adjust their marketing strategies. It will always be important to measure the performance of email marketing to identify areas for improvement.

II. DATA

2.1 The source of data

The data is from one of our teammates' previous projects, and its source cannot be shared due to confidentiality.

2.2 The data and the variables and how they are constructed

Our dataset consists of 99950 customer-level records of and 8 variables: email_id, email_text, email_version, hour, weekday, user_country, user_past_purchases and clicked. This dataset records the click of the link in a marketing email within a week, including customers' information. At midnight each day, the same number of emails was randomly sent to the company's customers. Please refer to [Appendix 1] for details.

2.3 Data engineering

"Hour" is an issue in logistic regression since the model may attempt to identify a linear link between time and the ECR, despite the fact that the relationship is frequently nonlinear. Therefore, we classified 0 am to 6 am as the night, 6 am to 12 pm as the morning, 12 pm to 18 pm as the afternoon, and 18 pm to 24 am as the evening. Also, we converted all the categorical variables such as email_text, email_version, weekday, and user_counrty to dummy variables. We chose short email, personalized, Friday, ES, and afternoon as the reference levels [Appendix 2].

III. EXPLORATORY DATA ANALYSIS

3.1 Descriptive statistics

There are 99950 pieces of customer-level data. The mean of 'hour' is 9 [9 am morning] with a standard deviation of 4.44. We can see that there are emails sent out at midnight (24). For user_past_purchases, the average number of transactions made by a consumer is around four, with some customers making no purchases at all and others making 22. Finally, the average click rate is very small (0.0207 around 2.1%). Only 2% of customers clicked the email [Appendix 3].

3.2 Visualization of the data [Details in Appendix]

3.2.1 Number of emails that are clicked vs not clicked

In the dataset, there are 2069 customers who clicked the link, accounting for around 2% of total customers. This indicates that the original dataset is skewed [Appendix 4].

3.2.2 Email_text

There are slightly more long emails (50248) than short ones (49702). We can probably infer that those who received the shorter email were more likely to open it than those who received the longer email [Appendix 5].

3.2.2 Email_version

Personalized email (hello John...) is much more effective. This makes sense because recipients are more likely to read and receive emails with customized headlines [Appendix 6].

3.2.3 Hour/Weekday

Users are more likely to click on the link around 10 a.m. The click-through rate declines after 10 a.m. The company can try to send out the majority of its emails around 10 a.m. in order to raise the click-through rate [Appendix 7]. The company should try to start the marketing campaign on Wednesday, Thursday, or Tuesday and should avoid Friday, Saturday, and Sunday because almost nobody opens their emails those days [Appendix 7].

3.2.4 User Countries

The US has the highest number of clicks, while France and ES have the lowest. The marketing team should either find a way to increase the ECR in these countries or stop trying to sell to them because they don't seem interested in our products [Appendix 8].

IV. MODEL ESTIMATION

4.1 The Logistic Model

4.1.1 Model Formulation

We partitioned the 99950 pieces of customer-level data into two groups, with 80% in the training set and 20% in the testing set. We chose to build a logistic regression model because the dependent variable "clicked" is binary. "Clicked" equals 1 when a certain customer clicked on the link, and 0 otherwise. Our goal is to predict the probability that a customer i clicks on the link, denoted as $P(clicked_i = 1)$, using independent variable user_past_purchases and all the dummy variables we created . Model Formula is in [Appendix 9].

4.1.2 Model Estimation

The convergence criterion for the model is satisfied [Appendix 10]. The likelihood ratio of this model is smaller than 0.0001, which implies that the overall model is significant. The model suggests that, apart from "evening", "Sunday", "Saturday" and "FR", the other variables' P values are smaller than 0.05, which means they are significant in influencing the customers' choice on click. Detailed interpretation of the model in [Appendix 10].

4.1.3 Prediction and Classification Matrix (similarity and sensitivity)

After deriving the final model, we used the test set to assess the model's predictive performance. The prediction shows nobody among the 19990 customers clicked on the link. The sensitivity is 0, and specificity is 1. This model tends to predict that all the customers will not click on the link[Appendix 11]. This output deviates from our expectation. We realized this is because the dataset itself is unbalanced, in which customers who clicked the link only accounted for 2 % of the total 99950 customers. We will discuss this issue in the project limitations part.

4.2 Decision tree

4.2.1 Model Building

We hope to use a decision tree to investigate the relationships between factors, develop metrics for differentiating between quality users and bad ones, and uncover paths for optimizing those bad customers (clicked = 0). Detailed interpretation of the tree is in [Appendix 12].

V. RESULTS AND MANAGERIAL TAKEAWAYS

5.1 User_country

Click rates in the UK and US perform significantly better than in Spain and France. It implies that people from English-speaking countries are more easily attracted by the email marketing strategy or these companies' products. As a result, the company can target customer groups from English-speaking countries in the future. Another explanation could be that the marketing emails that were translated from English to Spanish were inaccurate. The company's marketing team should try to improve the translation accuracy of marketing materials. They should hire a better team to compose emails.

5.2 Time to send emails

The emails sent from Monday to Thursday worked better consistently than those sent on Friday and weekends. This might be because people check emails more frequently from Monday to Thursday while at work. The marketing team can send promotional emails between Monday and Thursday and avoid Friday and weekends. When looking at the time categories during a day, emails sent in the morning (6:12 am) performed the best. Then, the afternoon (12–18 pm) and evening (18–24 pm) perform better than those sent during the night (0–6 am). As a result, the

marketing team should be able to send emails in the morning. In summary, weekday mornings are most likely to generate high click-through rates.

5.3 Email length

Long emails did not work as well as short emails. Both logistic and tree models show that a long email decreases the likelihood of a customer clicking compared to a short email. As a result, the marketing team should manage to shorten their marketing emails. Concise sentences should be presented to attract customers. If an email cannot capture customers' attention within a few sentences, the customers are likely to discard the email.

5.4 Personalized

Personalized email works much better than generic email. For instance, an email that begins with "Hello xx (name of the customer)" will always draw more attention than one that begins with "Dear customer." The company can extract the names that customers fill in when signing up for membership and use algorithms to add each person's name to the email. Such personalized details can surprise customers and help companies gain a competitive edge.

5.5 User's past purchases

In general, the more customers have purchased in the past, the more likely they are to click on the link. The company should segment customers according to their number of past orders and target loyal customers when sending marketing emails.

VI. LIMITATIONS

The dataset is imbalanced because clicked customers only account for 2 % of the total 99950 customers. When predicting using a model built on an unbalanced dataset, most of the predictions will be made corresponding to the majority class (clicked =0), and the minority class (clicked the link) features will be treated as noise and ignored, which can result in huge bias in the model. In our project, the model is overfitting towards records with an output of not clicking by regarding the "clicked" category as outliers. This is the reason why the testing dataset indeed included 406 customers who clicked on the link, but our prediction is that no customer would click on the link. To deal with the skewed dataset, the Synthetic Minority Oversampling Technique (SMOTE) is a common approach to resampling a dataset, which generates synthetic samples for the minority class.

TABLE OF CONTENT

VII. APPENDIX

Link to our SAS code [PDF version]

[Appendix 1] Snapshot of the dataset

email_id	email_text	email_version	hour	weekday	user_country	user_past_purchases	clicked
8	short_email	generic	9	Thursday	US	3	0
33	long_email	personalized	6	Monday	US	0	0
46	short_email	generic	14	Tuesday	US	3	0
49	long_email	personalized	11	Thursday	US	10	0
65	short_email	generic	8	Wednesday	UK	3	0
66	long_email	generic	12	Wednesday	US	0	0
72	short_email	generic	4	Saturday	US	0	0
73	long_email	generic	18	Thursday	FR	5	0
82	long_email	personalized	17	Thursday	ES	0	0
114	short_email	personalized	5	Wednesday	US	2	0

- **email_id** : the id of the email that was sent. It is unique by email. Nominal.
- **email_text**: two different versions of the email have been sent: one has "long text" (i.e. has 4 paragraphs) and one has "short text" (only has two paragraphs). Nominal.
- **email_version**: some emails were "personalized" (i.e. they had the name of the user receiving the email in the incipit, such as "Hi John,"), while some emails were "generic" (the incipit was just "Hi,"). Nominal.

- **hour**: the local time on which the email was sent. Numeric.
- **weekday**: the weekday on which the email was sent. Nominal.
- **user_country**: the country where the user receiving the email is based. It comes from the user ip address when they created the account. Nominal.
- **user_past_purchases**: how many items in the past were bought by the user receiving the email. Numeric.
- **clicked**: Whether the user has clicked on the link inside the email. This is our label and, most importantly, the goal of the project is to increase this. Numeric.

[Appendix 2] Dummy Variable Reference Level

In specific, a new dummy variable was created named “length”, which equals 1 when email type is long, and 0 when email type is short. The new dummy variable “type” equals 1 if the email version is generic, and equals 0 if the email version is personalized. The variable “weekday” was transformed into six dummy variables: “Monday”, “Tuesday”, “Wednesday”, “Thursday”,“Friday”, “Saturday” and “Sunday”. The variable “user_country” was transformed into three dummy variables: “FR”, “ES”, “US” and “UK”. We chose short email, personalized, Friday, ES and afternoon as the reference levels. All the dummy variables were compared with their respective reference level when estimating the models.

Code:

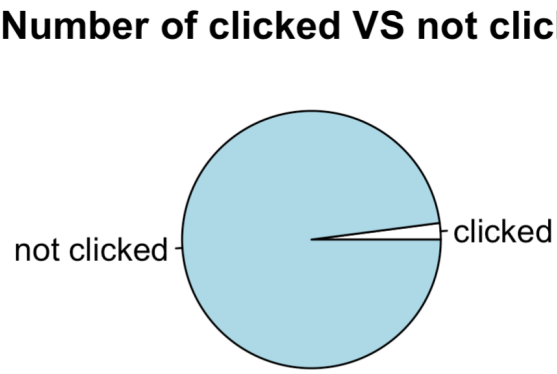
```
data click_r; set click_r;
if hour <= 6 then night = 1; else night = 0;
if hour <= 12 and hour >7 then morning = 1; else morning = 0;
if hour <= 18 and hour >13 then afternoon = 1; else afternoon = 0;
if hour <= 24 and hour >19 then evening = 1; else evening = 0;
if email_text = "long_email" then long = 1; else long = 0;
if email_version = "generic" then generic = 1; else generic = 0;
     if weekday = "Monday" then Monday = 1;
      else Monday= 0;
    if weekday = "Tuesday" then Tuesday = 1;
      else Tuesday= 0;
     if weekday = "Wednesday" then Wednesday = 1;
      else Wednesday= 0;
    if weekday = "Thursday" then Thursday = 1;
      else Thursday= 0;
        if weekday = "Friday" then Friday = 1;
      else Friday= 0;
        if weekday = "Saturday" then Saturday = 1;
      else Saturday= 0;
     if weekday = "Sunday" then Sunday = 1;
      else Sunday= 0;
     if user_country = "FR" then FR = 1;
      else FR= 0;
     if user_country = "US" then US = 1;
      else US= 0;
         if user_country = "UK" then UK = 1;
      else UK= 0;
run;
```

[Appendix 3] Summary Statistics

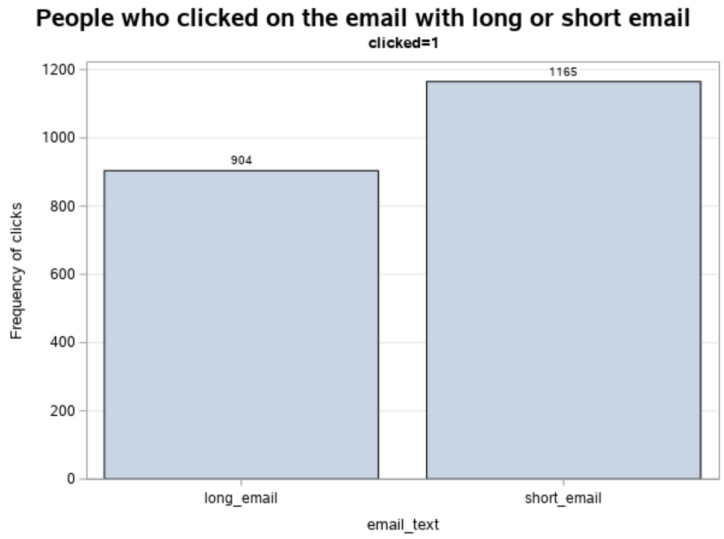
The MEANS Procedure

Variable	N	Mean	Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Range	Std Dev
hour	99950	9.06	1.00	6.00	9.00	12.00	24.00	23.00	4.44
user_past_purchases	99950	3.88	0.00	1.00	3.00	6.00	22.00	22.00	3.20
clicked	99950	0.02	0.00	0.00	0.00	0.00	1.00	1.00	0.14

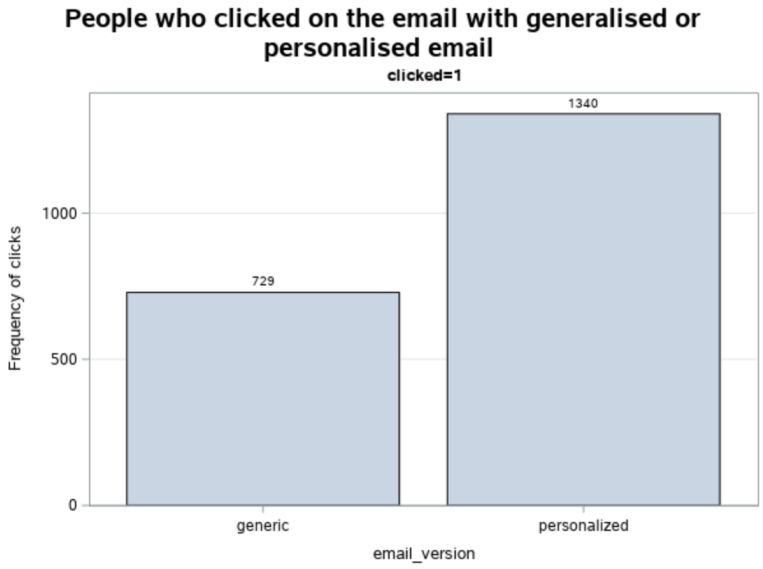
[Appendix 4] Number of emails that were clicked vs not clicks



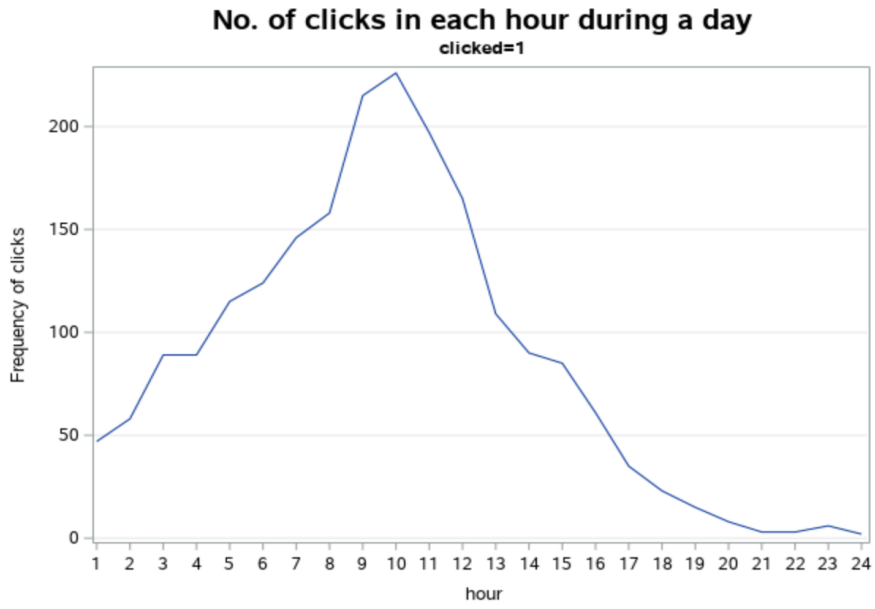
[Appendix 5] Email Text

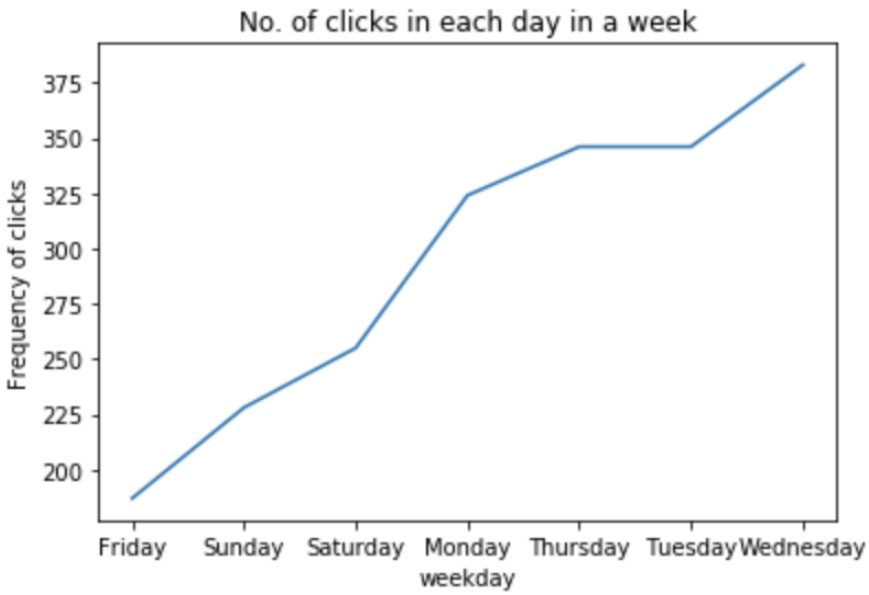


[Appendix 6] Personalized Email

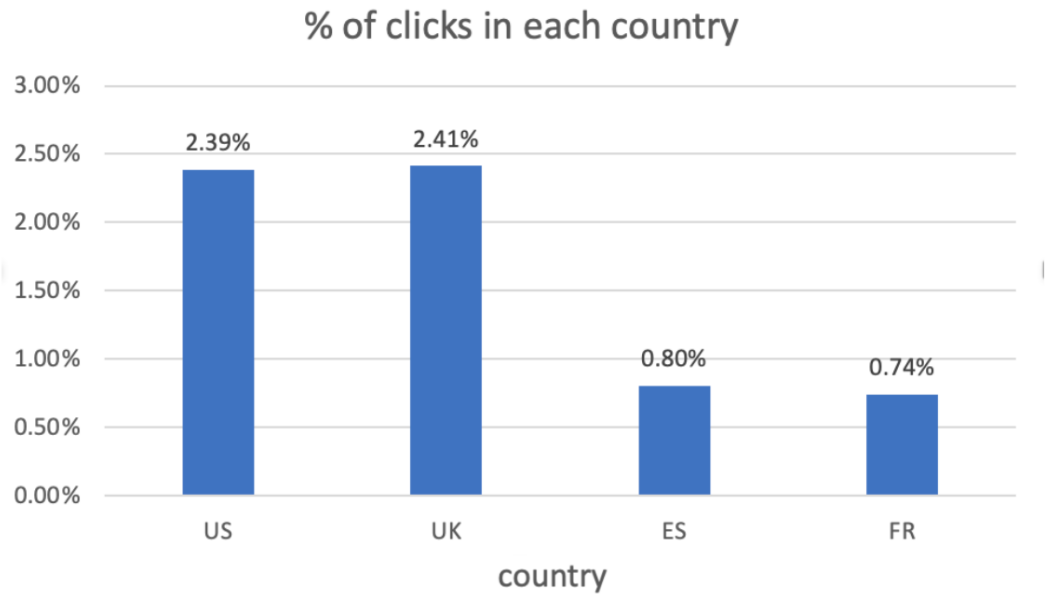


[Appendix 7] Weekday/Hour





[Appendix 8] Number of clicks in each country



[Appendix 9] Logistic Regression Formulation

$$odds = \frac{P(clicked_i=1)}{P((clicked_i=0)} = \frac{P\ (clicked_i=1)}{1-P\ (clickedi=1)}$$

$$logit = log(odds) = log(\frac{P(clicked_i=1)}{1-P(clickedi=1)})$$

$$= \beta_0 + \beta_1 user\_past\_purchases_i + \beta_2 long_i + \beta_3 generic_i + \beta_4 night_i + \beta_5 morning_i$$

$$+ \beta_6 evening_i + \beta_7 Monday_i + \beta_8 Tuesday_i + \beta_9 Wednesday_i + \beta_{10} Thursday_i + \beta_{11} Saturday_i$$

$$+\beta_{12} Sunday_i + \beta_{13} FR_i + \beta_{14} US_i + \beta_{15} UK_i + \varepsilon_i$$

[Appendix 10] Logistic Regression Output

SAS code:

```
proc logistic data=train descending out=train_c outmodel=model;
model clicked = user_past_purchases night morning evening length type  Monday Tuesday
Wednesday Thursday Sunday Saturday FR US UK;
run;
```

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1323.5407	15	<.0001
Score	1426.9818	15	<.0001
Wald	1308.6902	15	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.8168	0.1607	1310.1690	<.0001
user_past_purchases	1	0.1892	0.00638	878.6676	<.0001
long	1	-0.3074	0.0507	36.8040	<.0001
generic	1	-0.6315	0.0524	145.2720	<.0001
night	1	-0.1755	0.0693	6.4171	0.0113
morning	1	0.2332	0.0602	14.9867	0.0001
evening	1	-0.2868	0.2444	1.3768	0.2406
Monday	1	0.4889	0.1026	22.7178	<.0001
Tuesday	1	0.5670	0.1014	31.2754	<.0001
Wednesday	1	0.6862	0.1001	47.0227	<.0001
Thursday	1	0.5742	0.1013	32.1100	<.0001
Sunday	1	0.1275	0.1106	1.3297	0.2489
Saturday	1	0.2123	0.1082	3.8509	0.0497
FR	1	-0.00270	0.1783	0.0002	0.9879
US	1	1.1304	0.1296	76.0219	<.0001
UK	1	1.1575	0.1364	72.0163	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
user_past_purchases	1.208	1.193	1.224
long	0.735	0.666	0.812
generic	0.532	0.480	0.589
night	0.839	0.733	0.961
morning	1.263	1.122	1.421
evening	0.751	0.465	1.212
Monday	1.631	1.334	1.994
Tuesday	1.763	1.445	2.151
Wednesday	1.986	1.632	2.417
Thursday	1.776	1.456	2.166
Sunday	1.136	0.915	1.411
Saturday	1.237	1.000	1.529
FR	0.997	0.703	1.414
US	3.097	2.402	3.993
UK	3.182	2.436	4.157

$$log\ (odds) = -\ 5.82\ +\ 0.19\ user\_past\_purchases_i\ -\ 0.31\ long_i\ -\ 0.63\ generic_i\ -\ 0.18\ night_i$$
$$+\ 0.23\ morning_i\ -\ 0.29\ evening_i\ +\ 0.49\ Monday_i\ +\ 0.57\ Tuesday_i\ +\ 0.69\ Wednesday_i$$
$$+\ 0.57 Thursday_i\ +\ 0.13 Saturday_i\ +\ 0.21 Sunday_i\ +1.13\ US_i\ \ +\ 1.16 UK_i\ +\ \varepsilon_i$$

The estimated coefficients β_i is the expected change in the log odds if increasing the predictor X_i by 1 unit. In other words, if the predictor X_i increases by 1, the odds will be multiplied by e^{β_i}. If the coefficient β_i is positive, there is a positive relationship between X_i and the probability of click, because a positive β_i makes e^{β_i} larger than 1.

The coefficient of "user_past_purchases" is 0.1892, which means if the customer placed one more order in the past, the odds will be multiplied by $e^{0.1892}$ =1.21, and the customer is more likely to click on the link.

The coefficient of "long" is -0.3074 and the odds ratio is $e^{-0.3074} = 0.735$, indicating that customers receiving long emails have 0.735 times the odds of customers who receive short emails. Since the odds equals probability of clicking divided by probability of not clicking, a long email decreases the likelihood of a customer clicking on the link by 26.5% compared to a short email.

[Appendix 11] Logistic Regression Prediction

SAS code:

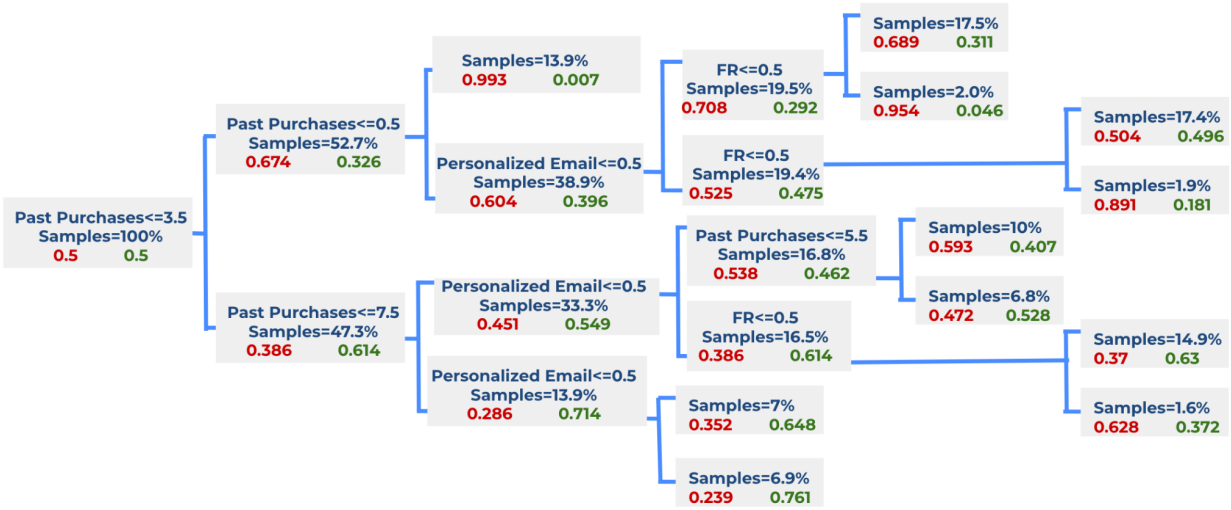
```
*threshold value=0.5;
data predict_new; set test;
if P_1 ge 0.5 then Pre_click=1 ; else Pre_click=0;
run;

proc freq data=predict_new;
table Pre_click;
run;

*2x2 metrix (classification metrix);
proc freq data=predict_new;
tables clicked*Pre_click;
run;
```

The FREQ Procedure

$$Sensitivity = \frac{TP}{TP+FN} = 0$$

$$Specificity = \frac{TN}{TN+FP} = 1$$

[Appendix 12] Tree and its interpretation



The first is the split threshold which splits a node into two nodes to the right. The segment that is true will go to the top and false will go to the bottom. The 'sample =' indicates the percentage of customers inside a node. If the value is high, we can see that a lot of customers are captured in that node, which is optimal for our analysis. The [..., ...] indicate the purity of the

node. If the first value is greater than 0.5, class 0 is assigned to the node. Otherwise, the node is class 1.

From the leftmost node, we can see that this node contains the entire population, which makes sense because we have not yet divided the data. Then, for the top node, those who made fewer than 3.5 purchases will be routed there. 52.7% of users in this node had made fewer than 3.5 purchases. 67.4% of them did not click on the email, whereas 32.6% did. Then, we moved to the next top node, where 13.9% of consumers made fewer than 0.5 purchases and 99.3% of them did not click on the email. In this part, we can infer that individuals who did not make a purchase (= 0.5) almost never clicked.

Next, we shift to the bottom node, where 38.9% of customers made between 0.5 and 3.5 purchases, 60% did not click, and 40% did click. Then, 19.5% of them were not receiving personalized email (personalized email = 0.5 is true). Moreover, the majority of recipients (70.8%) did not click on the email if it was generic (Hi instead of Hello John!). This finding may suggest that personalized communications should be prioritized over generic emails in order to entice customers to click on the email.

Then, proceed to the node on the right where the top tree ends. If a customer meets all of the requirements of the preceding nodes (past purchases [0.5, 3.5] and personalized email = 0.5 is false) and is from France (FR = 0.5 is false), there is a 95.4% chance that they did not click on the emails. In other words, if a French client who has not made many transactions receives a generic email, it is extremely unlikely that the customer will click on the email. For French customers who made purchases [0.5, 3.5] and received a personalized email (personalized email = 0.5 false), there is an 89.1% chance that they did not click on the email. We can likely assume that the email sent to France in general did not perform well, and this is where organizations may consider how to improve the content they send to France. Possible cause could be that the staff responsible for composing the email in French did not translate its content appropriately.

Let's start at the bottom of the first split (past purchases = 3.5 is false). Overall, this pattern is similar to the result of the analysis performed on the top nodes (previous purchases = 3.5) The more a customer's past purchases, the greater the likelihood that they will click.

Additionally, personalized email is more effective than generic email at generating clicks. Lastly, if the clients are French, it is unlikely that they will click.