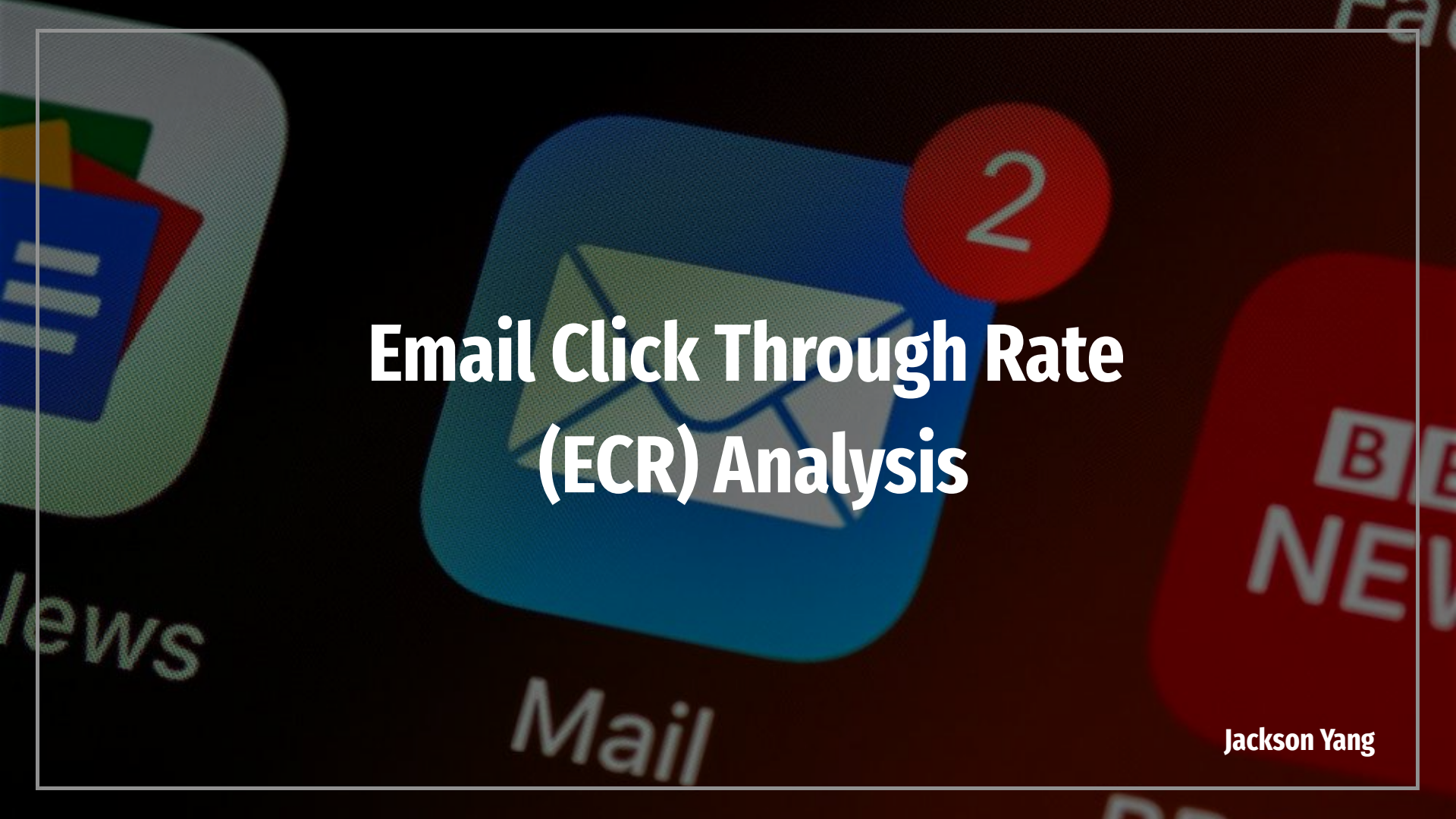


Email Click Through Rate (ECR) Analysis

Jackson Yang

Agenda

01 **Background**

Research Question
and Business Importance

02 **Data**

Data Source
and Data Engineering

03 **Visualisation**

Descriptive Analysis
and Data Exploration

04 **Model**

Logistic Model
and Decision Tree

05 **Takeaways**

Business and Managerial
Implications

06 **Limitations**

Project Constraints
and Further Improvement

Background: Email Marketing

Typical Marketing Channel

Products

Sales

Updates

Background: ECR

The number of clicks through to the related email is called Email Click-through Rate (ECR)

Shows a company the level of engagement their consumers have with their emails and, more crucially, with their products

Background: Research Question

ECR is a powerful metric but it is often difficult for businesses to grasp what truly impacts the ECR and how to increase it efficiently.

Research Question

What are the important factors impacting the percentage of customers who click on the link inside the marketing emails (ECR) and how to improve it?

Data

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

Data Engineering

DV: “clicked” (binary)

IV: Variable	Dummy Variable
”user_past_purchases” (numerical)	-
“user_country” (categorical)	US, ES, FR, UK
“email_text” (categorical)	long, short
“email_version (categorical)	generic, personalised
“hour” (numerical)	night, morning, afternoon, evening
“weekday” (categorical)	Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday

Visualisation: Descriptive Analysis

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

Visualisation: Descriptive Analysis

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

Visualisation: Descriptive Analysis

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

Visualisation: Descriptive Analysis

The MEANS Procedure

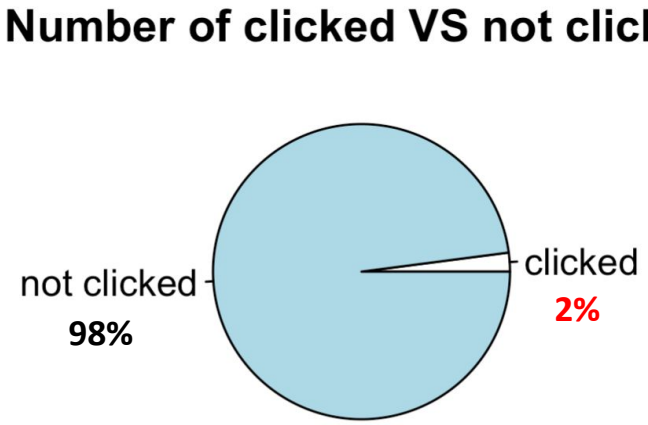
Variable	N	Mean	Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Range	Std Dev
hour	99950	9.06	1.00	6.00	9.00	12.00	24.00	23.00	4.44
user_past_purchases	99950	3.88	0.00	1.00	3.00	6.00	22.00	22.00	3.20
clicked	99950	0.02	0.00	0.00	0.00	0.00	1.00	1.00	0.14

Visualisation: Descriptive Analysis

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

Visualisation: Clicked

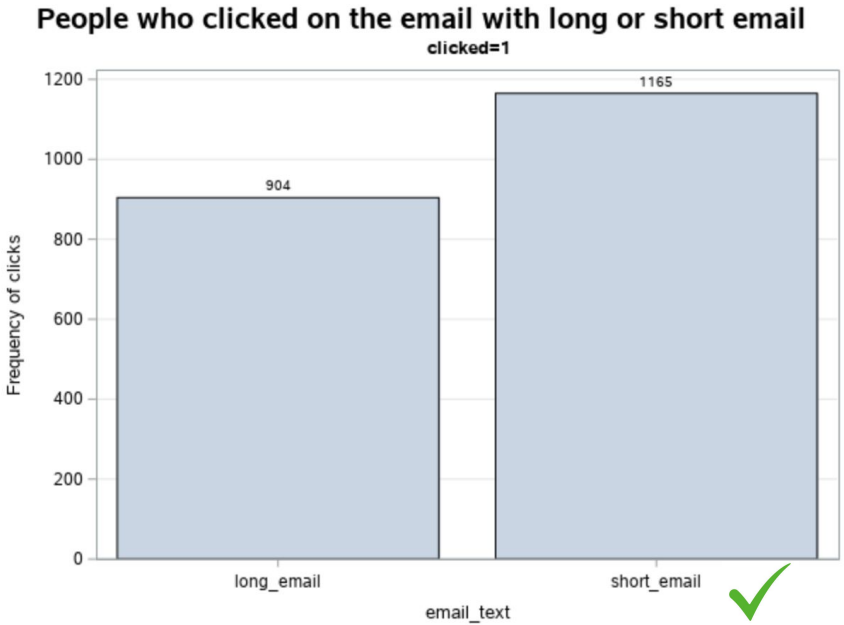


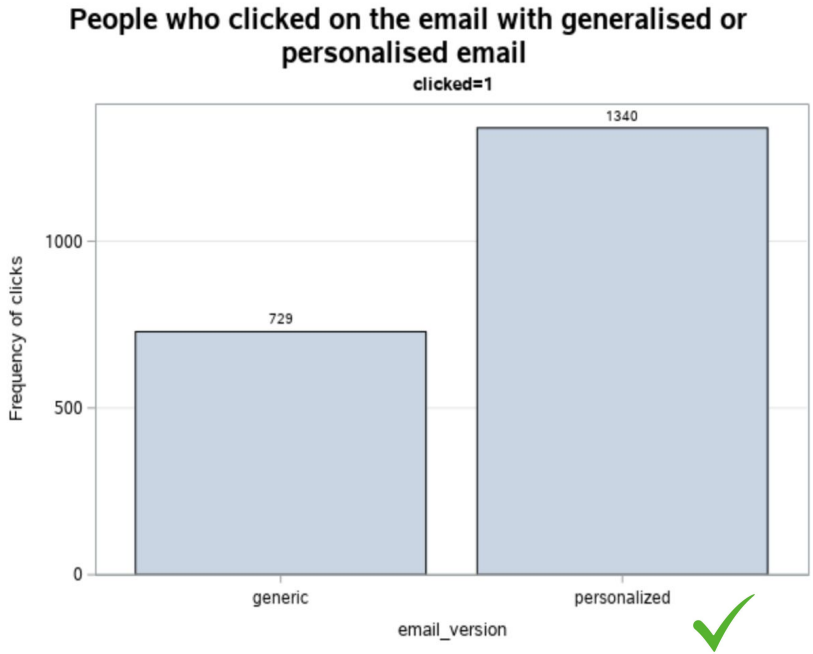
	Number	Percentage
Clicked (1)	2069	2%
Not Clicked (0)	97881	98%
Total	99950	100%

The original dataset is skewed!

Visualisation: Email_text & Email_verion

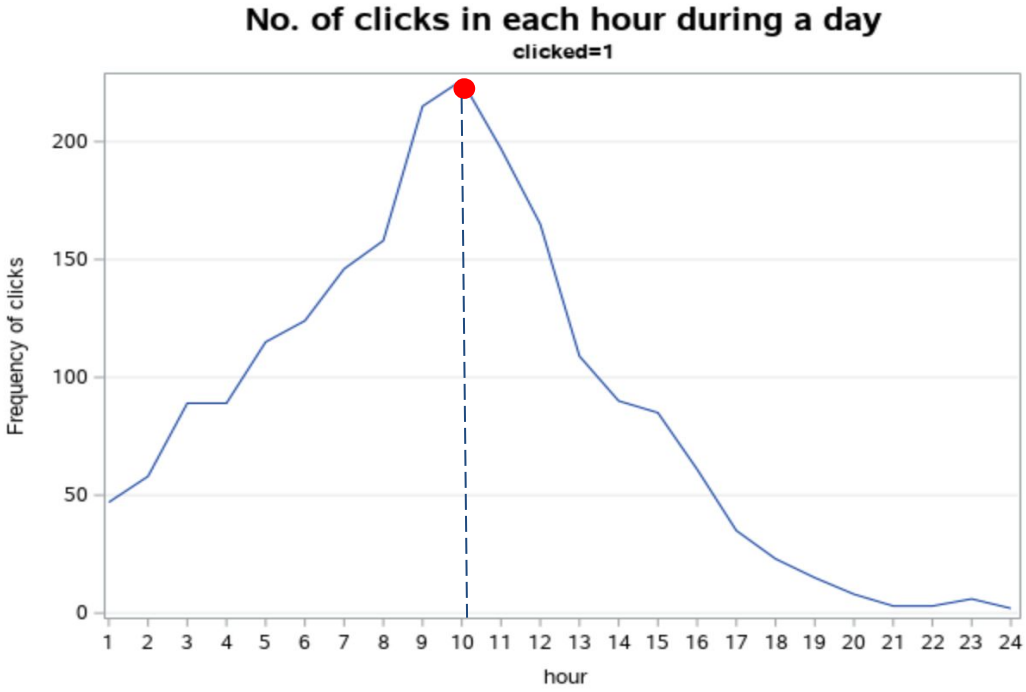
When Clicked=1:





Visualisation: Hour

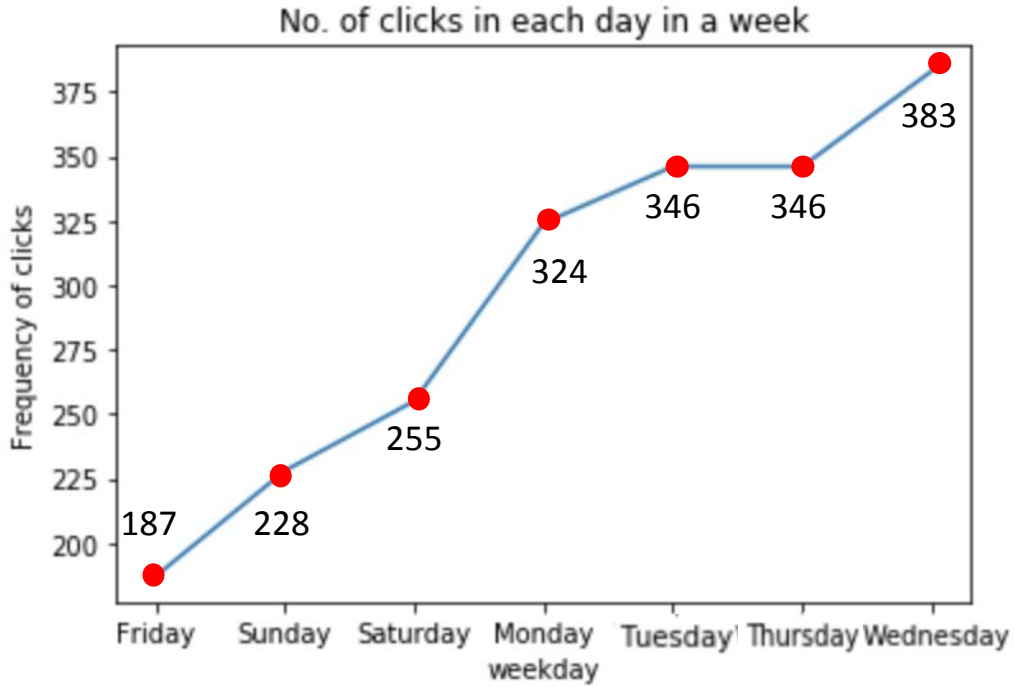
When Clicked=1:

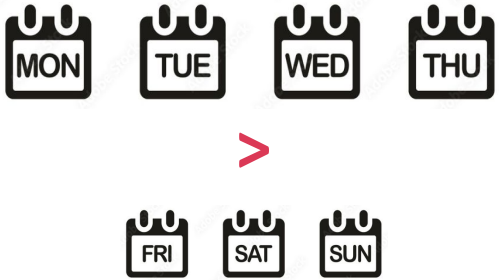


10 am is the peak time

Visualisation: Weekday

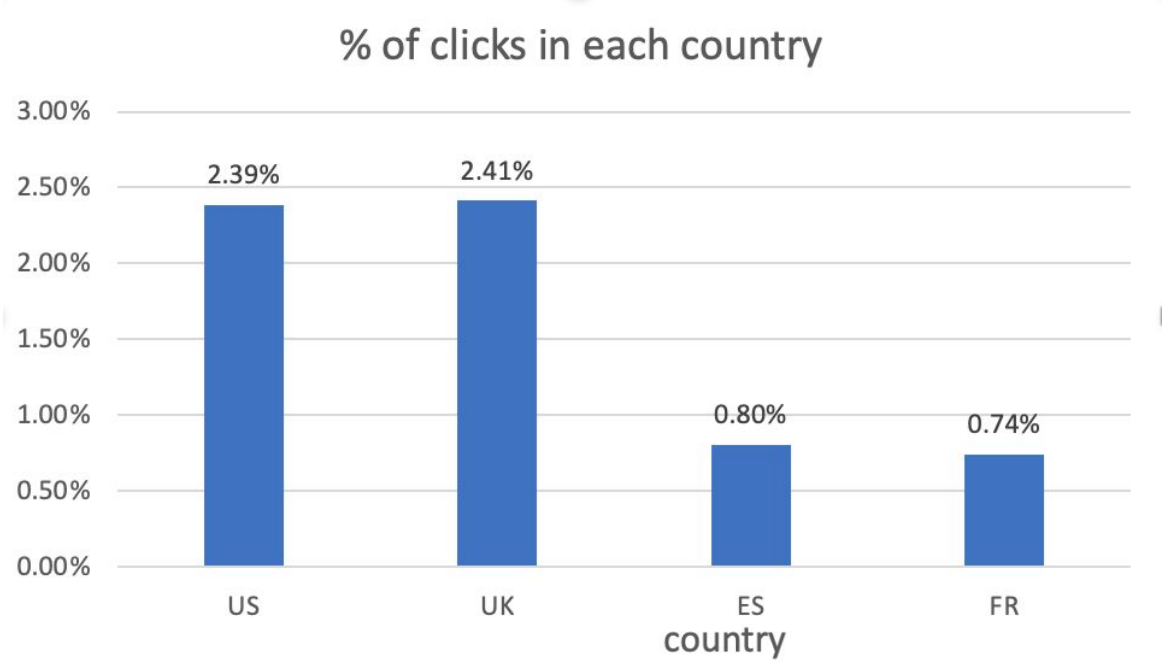
When Clicked=1:





Visualisation: User Countries

When Clicked=1:



English speaking countries

>

Non English speaking countries

Model: Logistic Regression - Model Formulation

DV: "clicked" (binary) → Logistic Regression

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

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

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

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

Model: Logistic Regression - Model Estimation

Training data (79960)	Testing data (19990)

Likelihood Ratio <0.0001

Pr>**ChiSq**: indicates **significance**.

For dummy variables, if the coefficient is larger than 0, this category works better than reference level. Vice versa.

$$log\,(odds) = -\,5.82 + 0.19\,user\_past\_purchases_i - 0.31\,long_i - 0.63\,generic_i - 0.18\,night_i$$

$$+\,0.23\,morning_i - 0.29\,evening_i + 0.49\,Monday_i + 0.57\,Tuesday_i + 0.69\,Wednesday_i$$

$$+\,0.57 Thursday_i + 0.13 Saturday_i + 0.21 Sunday_i + 1.13\,US_i + 1.16 UK_i + \varepsilon_i$$

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

Model: Logistic Regression - Model Estimation

Interpretations of coefficients:

- **User_past_purchases**: bought more in the past, more likely to click.
- **Long**: long emails work worse than short ones.
- **Generic**: generic emails work worse than personalized ones.
- **Hours**: afternoon works better than night but worse than morning.
- **Weekdays**: Monday to Thursday work better than Friday and weekends.
- **User countries**: customers in the UK and US are more likely to click the link than Spanish customers.

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

Model: Logistic Regression - Model Estimation

Interpretations of coefficients:

- **User_past_purchases**: bought more in the past, more likely to click.
- **Long**: long emails work worse than short ones.
- **Generic**: generic emails work worse than personalized ones.
- **Hours**: afternoon works better than night but worse than morning.
- **Weekdays**: Monday to Thursday work better than Friday and weekends.
- **User countries**: customers in the UK and US are more likely to click the link than Spanish customers.

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

Model: Logistic Regression - Model Estimation

Interpretations of coefficients:

- **User_past_purchases**: bought more in the past, more likely to click.
- **Long**: long emails work worse than short ones.
- **Generic**: generic emails work worse than personalized ones.
- **Hours**: afternoon works better than night but worse than morning.
- **Weekdays**: Monday to Thursday work better than Friday and weekends.
- **User countries**: customers in the UK and US are more likely to click the link than Spanish customers.

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

Model: Logistic Regression - Model Estimation

Interpretations of coefficients:

- **User_past_purchases**: bought more in the past, more likely to click.
- **Long**: long emails work worse than short ones.
- **Generic**: generic emails work worse than personalized ones.
- **Hours**: afternoon works better than night but worse than morning.
- **Weekdays**: Monday to Thursday work better than Friday and weekends.
- **User countries**: customers in the UK and US are more likely to click the link than Spanish customers.

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

Model: Logistic Regression - Model Estimation

Interpretations of coefficients:

- **User_past_purchases**: bought more in the past, more likely to click.
- **Long**: long emails work worse than short ones.
- **Generic**: generic emails work worse than personalized ones.
- **Hours**: afternoon works better than night but worse than morning.
- **Weekdays**: Monday to Thursday work better than Friday and weekends.
- **User countries**: customers in the UK and US are more likely to click the link than Spanish customers.

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

Model: Logistic Regression - Prediction

Prediction:

0 customer will click on the link among 19990 customers.

- **Sensitivity=0**
- **Specificity=1**

The model tends to predict that all customers won't click the link.

Classification Matrix:

	Predicted=0	Predicted=1
Actual=0	19584 (TN)	0 (FP)
Actual=1	406 (FN)	0 (TP)
Total	19990	0

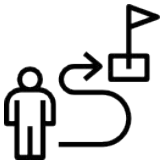
This result deviates from what we have expected!

Why?

Model: Decision Tree - Model Formulation

Relationship between various factors

Good Users and Bad Users



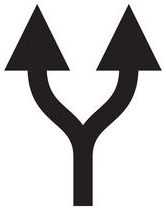
Paths for Optimizing the Bad Users

Explore the Results

Model: Decision Tree - Model Building

Example Tree node

Threshold



Splits a Node Into Two Nodes

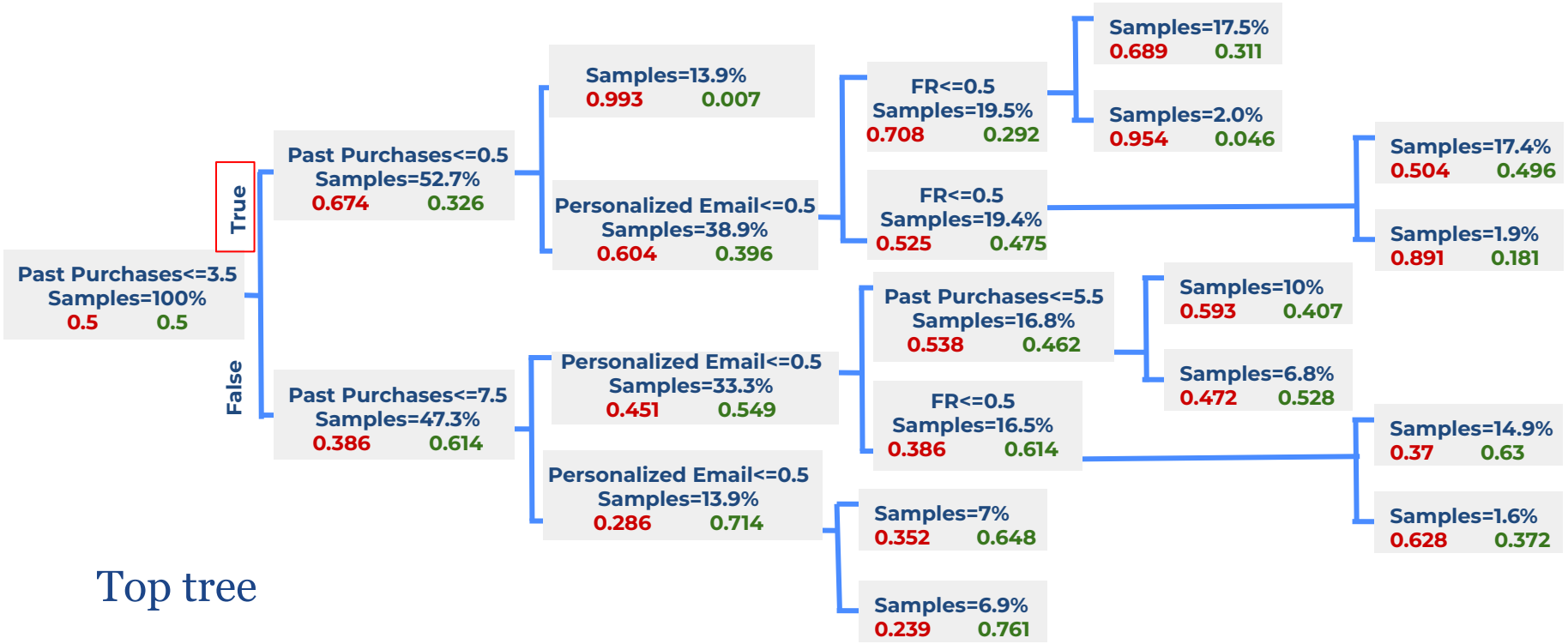
Sample Percentage

Indicates the Percentage of Customers Inside a Node

[0.674, 0.326]

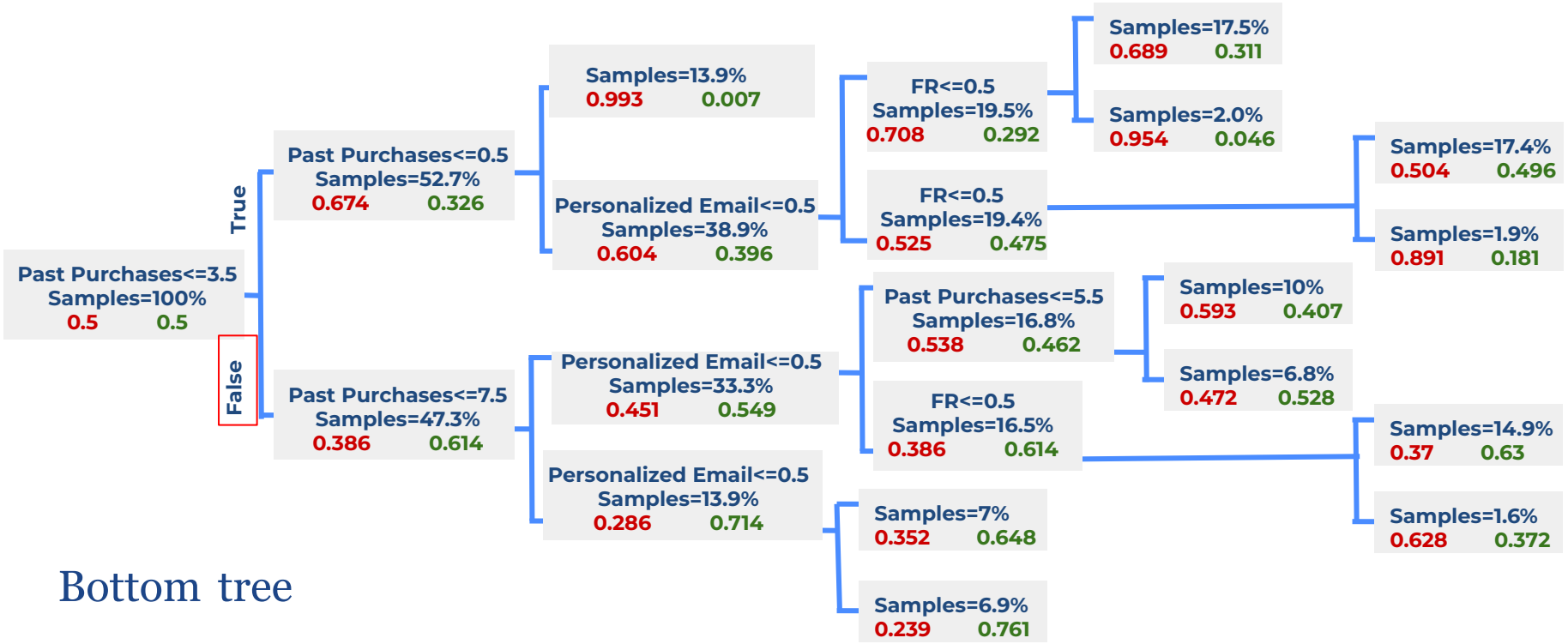
Purity of the Node
If the Left Value is > 0.5, class 0 is assigned; otherwise, the node is class 1

Model: Decision Tree Map



Top tree

Model: Decision Tree Map



Bottom tree

Results: User Country

UK & US > Spain 

Reason

1. English speaking countries are more attracted to the email marketing
2. Bad translation from English to Spanish

What To Do

1. Hire professionals to translate the emails
2. Find another channel to reach out to them

Results: Time to send emails [Weekday]

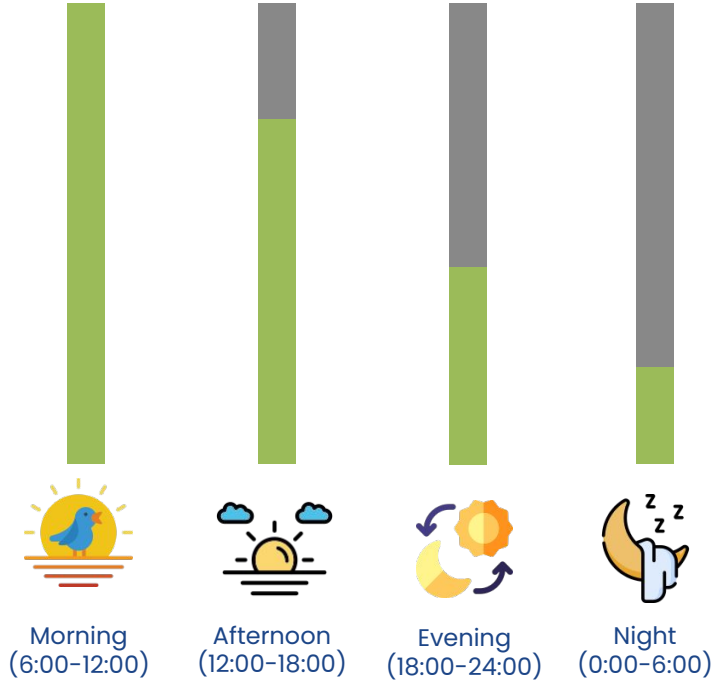
Reason

People check emails more frequently during the week than the weekends

What To Do

Send most of their promotional email during the week

Results: Time to send emails [Time]



Reason

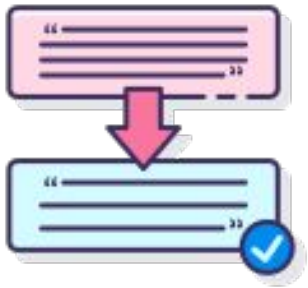
People often check their emails when they just wake up (morning) and then followed by afternoon and evening

What To Do

Send email from Monday to Thursday in the Morning → Generate Most Revenues.

Results: Email length

Short > Long Emails



Reason

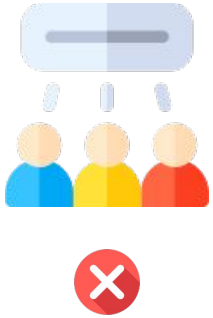
Long emails take longer time to read

What To Do

1. The marketing team should manage to shorten their emails with concise information
2. Try to capture customers with catchy headlines and few sentences

Results: Personalized vs Generic

Personalized > Generic Emails



Reason

Personalized emails make people feel more personal and let people know that they

What To Do

The marketing team should manage to personalize their emails with concise information

Results: User's past purchases

The more the user purchased, the more likely they would click

Reason

1. Customers who buy more are loyal to the company and more likely to click.
2. For those customers who made less than 3 purchases, they were unlikely to click

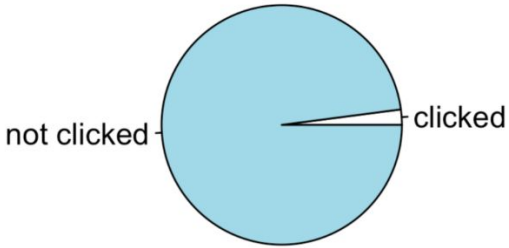
What To Do

Try to make customers buy more

Limitations: Imbalanced Clicks

Imbalanced Class

Number of Clicked VS Not Clicked



Solution

1. **Overdispersion**: Overdispersion is having higher variability in the dataset than is usually expected in a model, in our case, the logistic model.
2. Alternatively, we can use **SMOTE (Synthetic Minority Oversampling Technique)** which is a common approach to resample a dataset, which generates synthetic samples for the minority class.

Thank you!