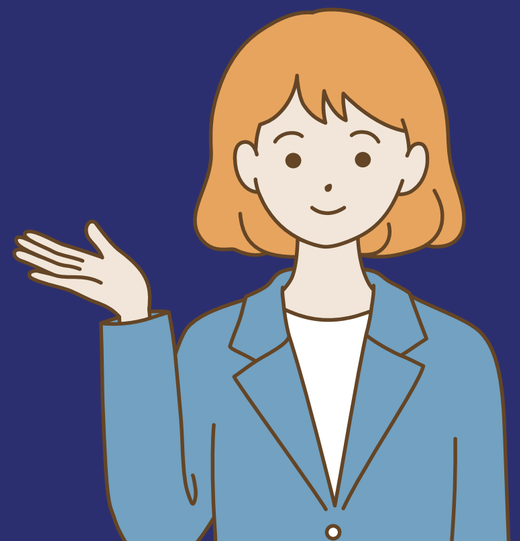
Data science Email campaign analysis

Presented By Rifqi arrayan

Metadata

Email ID: unique for each email
Email_Type: Category or type of email (1 for promotion, 2 for notification)
Subject_Hotness_Score: Score of how interesting the email is.
Email_Source_Type: Type of email source (1 internal or 2 external).
Email_Campaign_Type: The type of email campaign.
Customer_Location: The location of the customer.
Total_Past_Communications: Number of previous communications with the customer.
Time_sent_category:email sending time categories (1 in the morning, 2 in the afternoon, 3 in the evening)
Word_Count: The number of words in the email.
Total_Links: The number of links in the email.
Total_Images: The number of images in the email.
Email_status: email status (0 for unopened email, 1 for opened email, 2 for replied email).

Table of content



Background

In the ever-evolving digital era, companies of all sizes-small, medium, and large-are striving to increase their customer base through various marketing strategies. One effective and measurable method is email marketing. Small companies may see email as a cost-effective way to reach potential customers, while medium and large companies can utilize email to run more complex and segmented campaigns.

However, the success of an email campaign is not always guaranteed. Factors such as email content, campaign type, email source, and subscriber characteristics can affect subscriber response. Therefore, it is important for companies to understand the elements that contribute to the success of their email campaigns. With this understanding, they can design more effective emails, increase engagement, and ultimately increase customer conversion and loyalty.

Goals

1.Identifying Campaign Success Factors:
Find and understand the key factors that contribute to the success or failure of email campaigns. This includes analysis of numerical and categorical features that influence campaign results.

2.Provide Recommendations for Future Campaigns:
Based on the analysis results, provide actionable recommendations for designing more effective email campaigns in the future. This could include suggestions on email content, subscriber segmentation, and delivery strategies.

3.Increase Customer Loyalty and Retention:
By understanding what makes an email campaign successful, companies can design strategies that not only attract new customers but also retain existing ones, increasing customer loyalty.

4.Optimize Marketing Resources:
Helps companies allocate their marketing resources more efficiently, ensuring that the effort and budget spent on email campaigns generate maximum ROI.

TOOLS

Google colab is used as platform for analyzing data

Google Sheets is used to make data easier preprocessing is like removing duplicates

pyhton is used data for data analysis processes

Analysis process

1	Data Preparation
2	Exploratory Data Analysis
3	processing Data
4	Data inshigt
5	Correlation
6	Modelling data

Data preparation

This email campaign data is obtained from kaggle which functions to determine the effectiveness of promotion via email.

Rows

68353

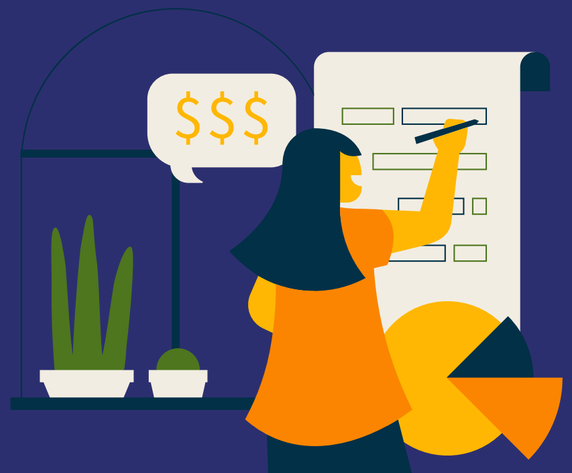
Columns

12

```
from google.colab import files
df = files.upload()
```

Choose Files No file chosen Upload widget i

```
Saving Train_psolI3n.csv to Train_psolI3n.csv
```




```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 68353 entries, 0 to 68352
Data columns (total 12 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   Email_ID                     68353 non-null  object 
 1   Email_Type                   68353 non-null  int64  
 2   Subject_Hotness_Score        68353 non-null  float64
 3   Email_Source_Type            68353 non-null  int64  
 4   Customer_Location            56758 non-null  object 
 5   Email_Campaign_Type          68353 non-null  int64  
 6   Total_Past_Communications    61528 non-null  float64
 7   Time_Email_sent_Category     68353 non-null  int64  
 8   Word_Count                   68353 non-null  int64  
 9   Total_Links                  66152 non-null  float64
 10  Total_Images                 66676 non-null  float64
 11  Email_Status                 68353 non-null  int64  
dtypes: float64(4), int64(6), object(2)
memory usage: 6.3+ MB
```

Processing data

BEFORE

```
[ ] df.duplicated().sum()

0

df.isna().sum().sort_values(ascending=False)

Customer_Location             11595
Total_Past_Communications      6825
Total_Links                    2201
Total_Images                   1677
Email_ID                          0
Email_Type                        0
Subject_Hotness_Score             0
Email_Source_Type                 0
Email_Campaign_Type               0
Time_Email_sent_Category          0
Word_Count                        0
Email_Status                      0
dtype: int64
```

AFTER

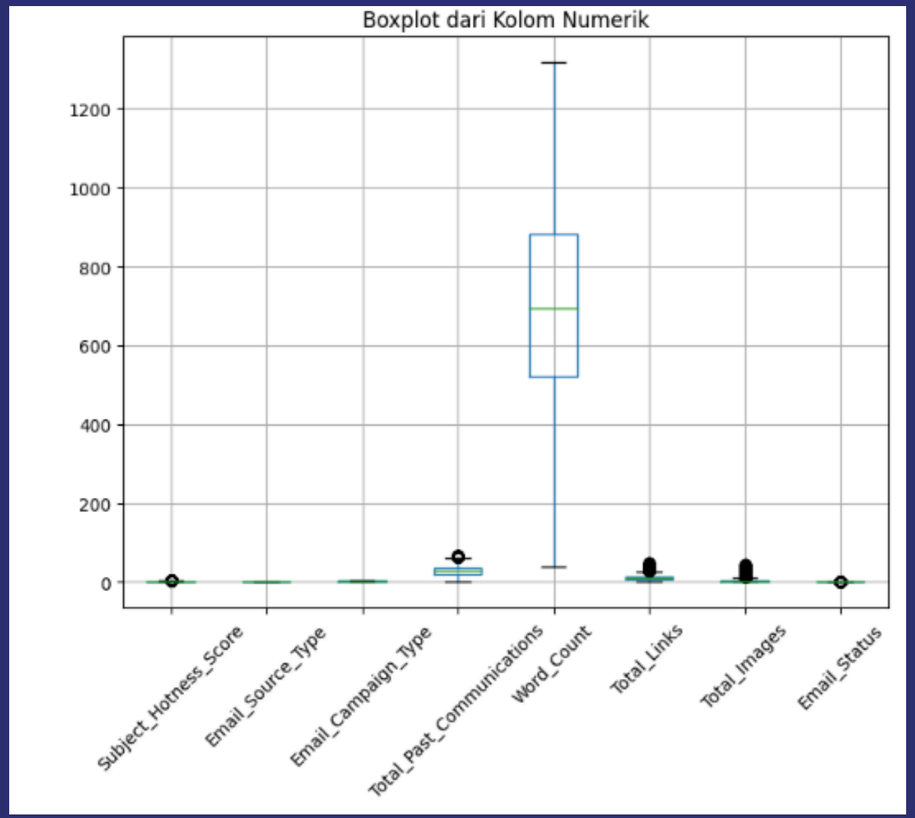
```
Email_ID                      0
Email_Type                    0
Subject_Hotness_Score         0
Email_Source_Type             0
Customer_Location             0
Email_Campaign_Type           0
Total_Past_Communications     0
Time_Email_sent_Category      0
Word_Count                    0
Total_Links                   0
Total_Images                  0
Email_Status                  0
dtype: int64
```

because there are too many missing data, therefore I fill the missing values in categorical data with mode while in numerical data with median.

outlier handling

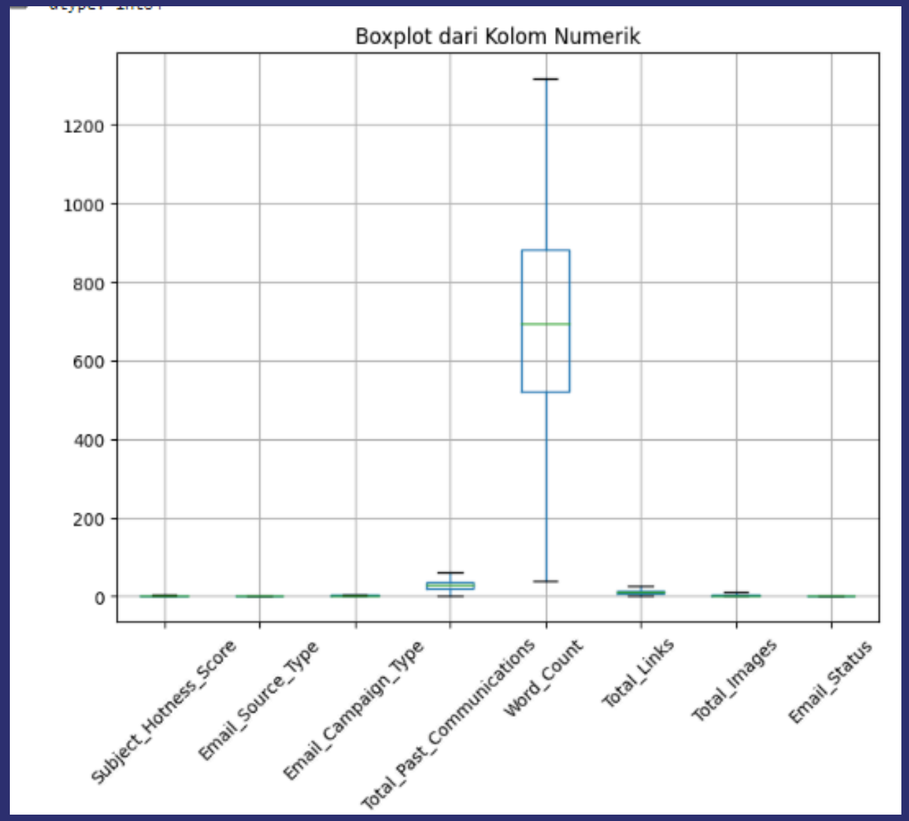
BEFORE

```
Data outlier:
Subject_Hotness_Score            247
Email_Source_Type                  0
Email_Campaign_Type                0
Total_Past_Communications        136
Word_Count                         0
Total_Links                     1608
Total_Images                    5585
Email_Status                   13412
dtype: int64
```



AFTER

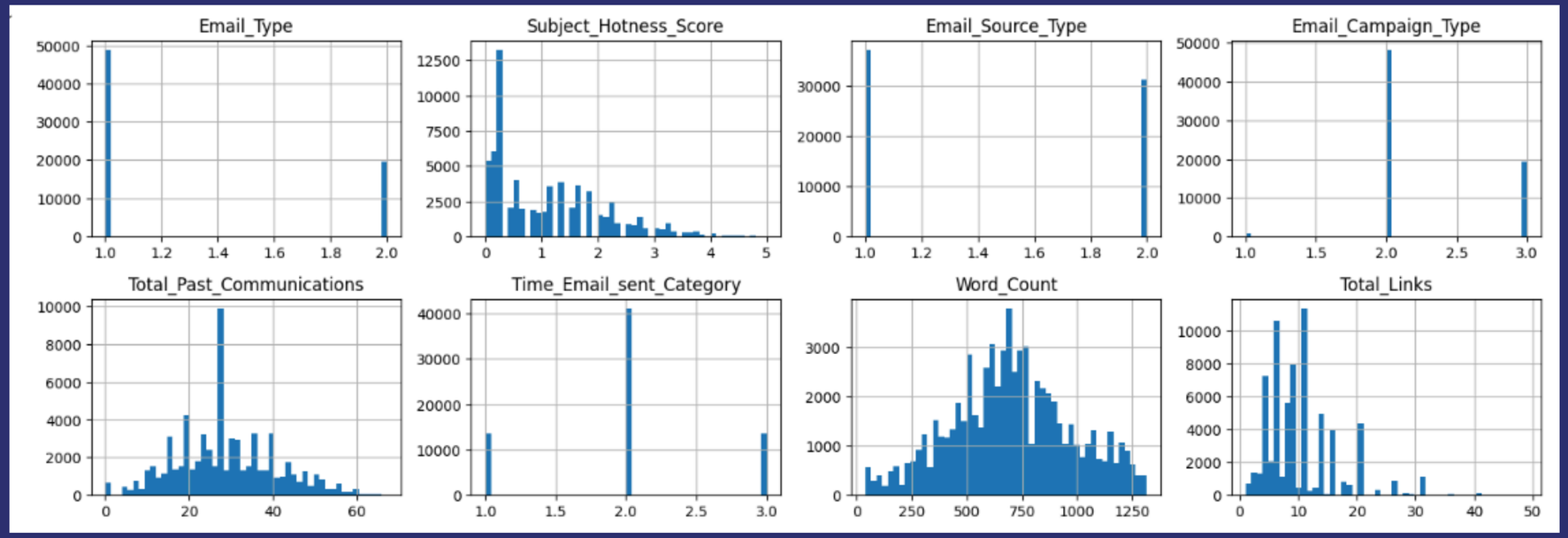
```
Data outlier setelah mengganti dengan Q1 dan Q3:
Subject_Hotness_Score          0
Email_Source_Type              0
Email_Campaign_Type            0
Total_Past_Communications      0
Word_Count                     0
Total_Links                    0
Total_Images                   0
Email_Status                   0
dtype: int64
```

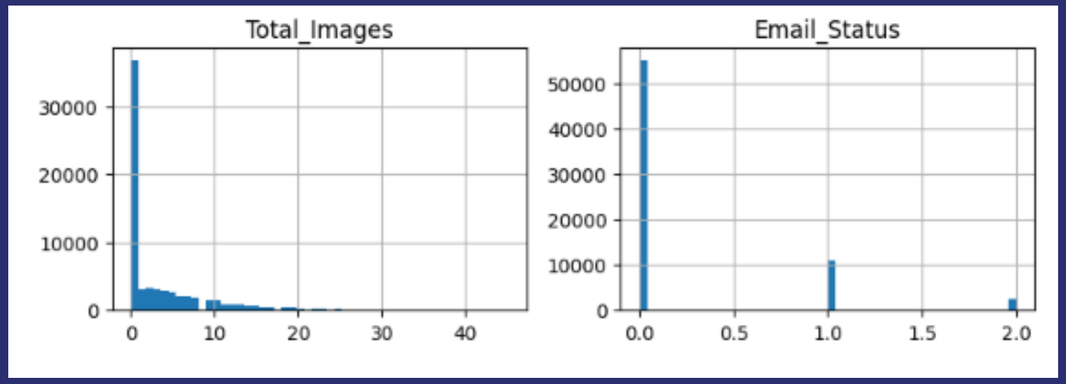


because the number of outliers was too much, I solved it by replacing the outlier values with Q1 and Q3.

Data inshigt

distribution checking

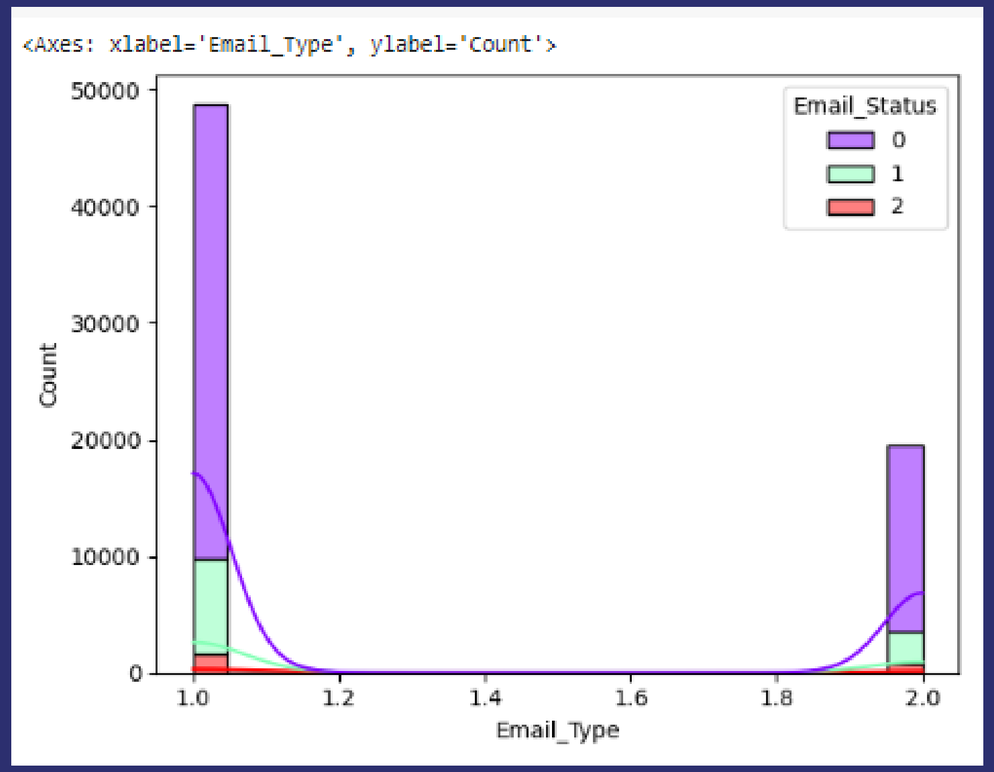




distribution checking conclusion

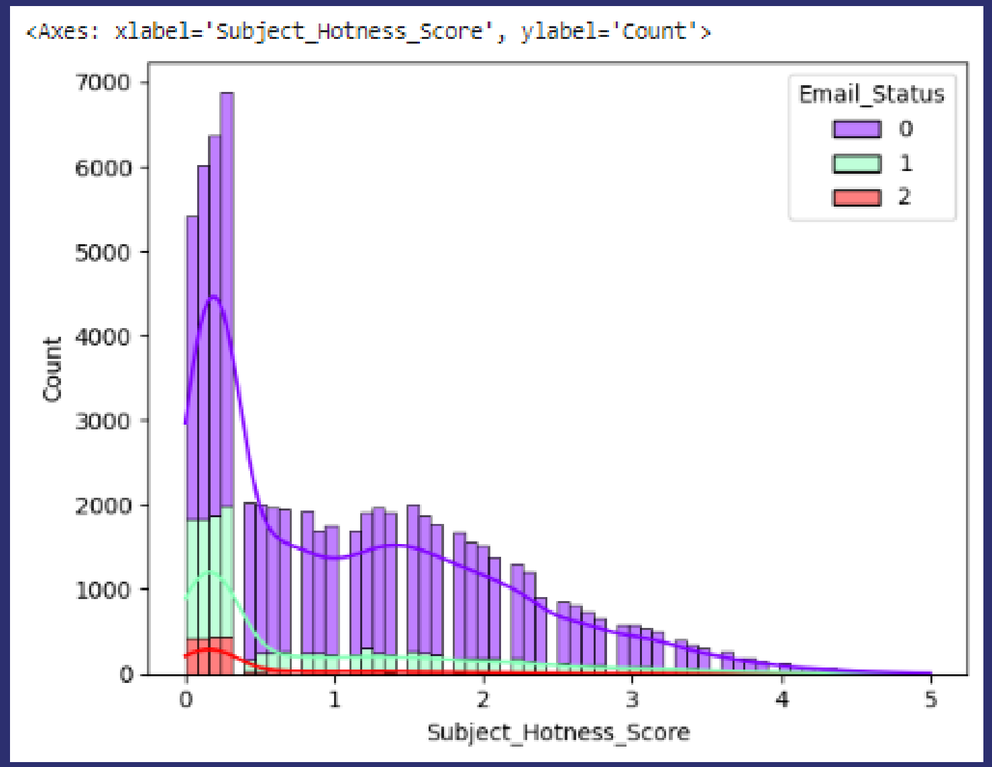
The distribution of features in the email campaign dataset seems reasonable and provides relevant insights. An even distribution of campaign types indicates balanced usage, while skewness in subject attractiveness scores may indicate highly attractive or less attractive subjects. A distribution of email source types dominated by certain categories and geographic concentration of subscribers is also reasonable. Different subscriber interaction patterns, variations in content length, number of links, and images in emails indicate varied content strategies and conform to common practices in email marketing.

email type



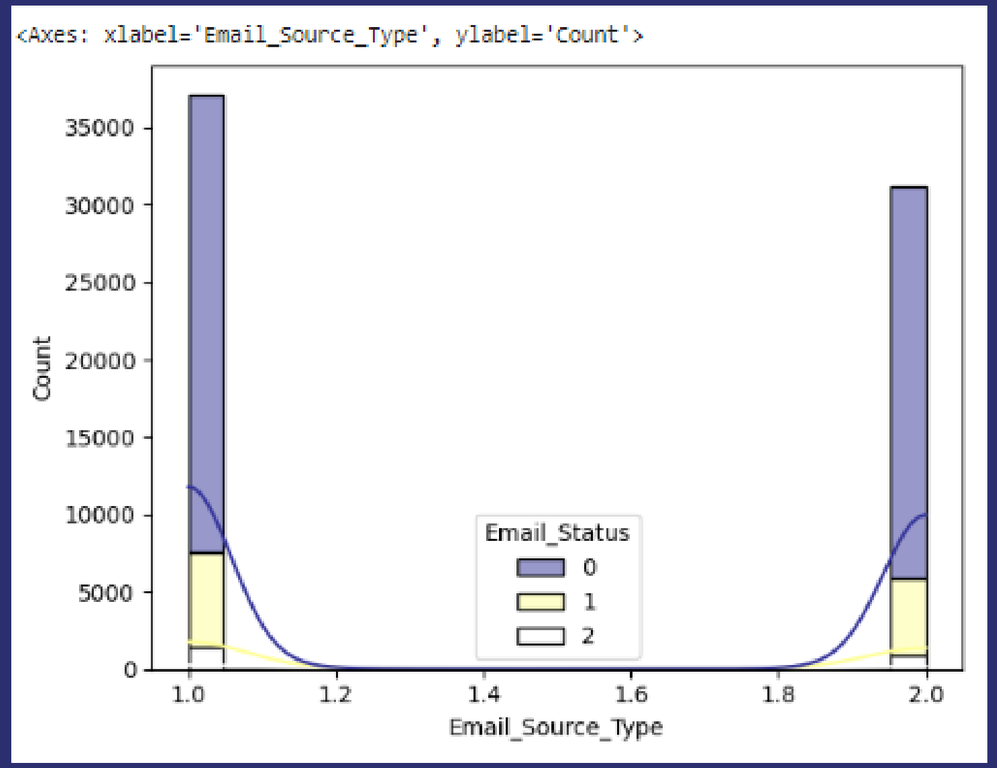
-Plot showing Email_Type distribution differentiated by Email_Status
- this plot shows a fairly even distribution with little difference between response statuses.

Subject honest



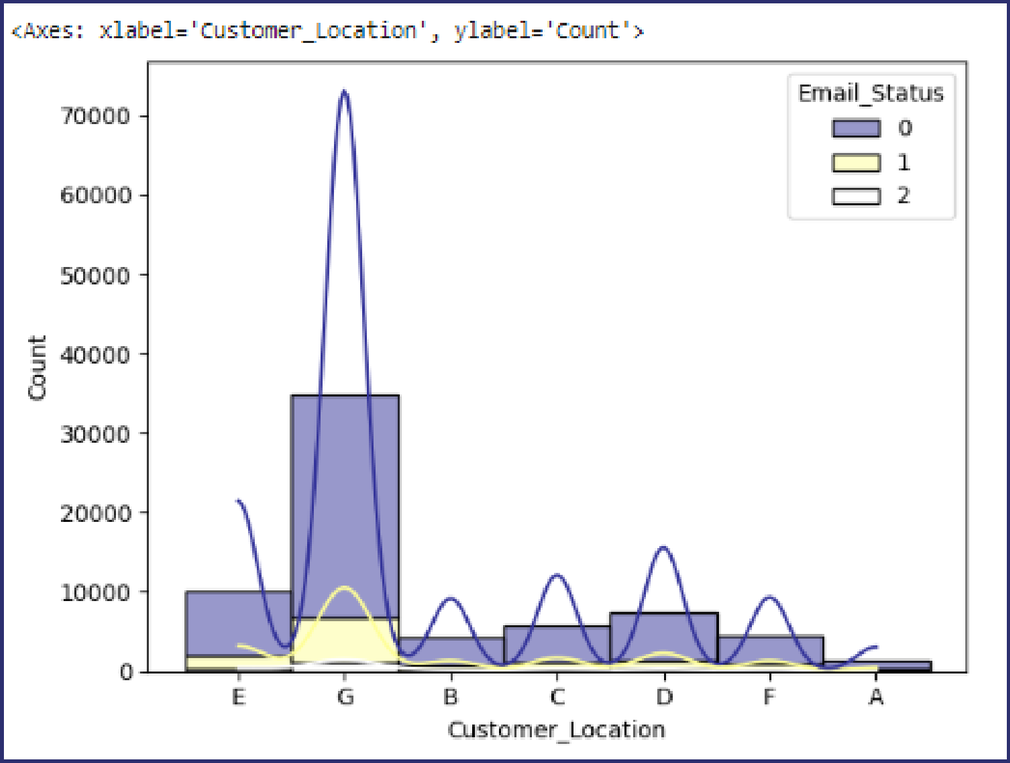
-The plots show the distribution of Subject_Hotness_Score differentiated by Email_Status.
-It appears that higher subject hotness scores are more likely that the email will be opened (email_status=1)
strategy:
Focus on improving email subject quality. Use interesting and relevant subjects to increase hotness scores.

email source type



-This plot will show the distribution of Email_Source_Type differentiated by Email_Status.
Strategy:
-Evaluate Email Sources: Focus on email sources that have higher response rates.

Customer location

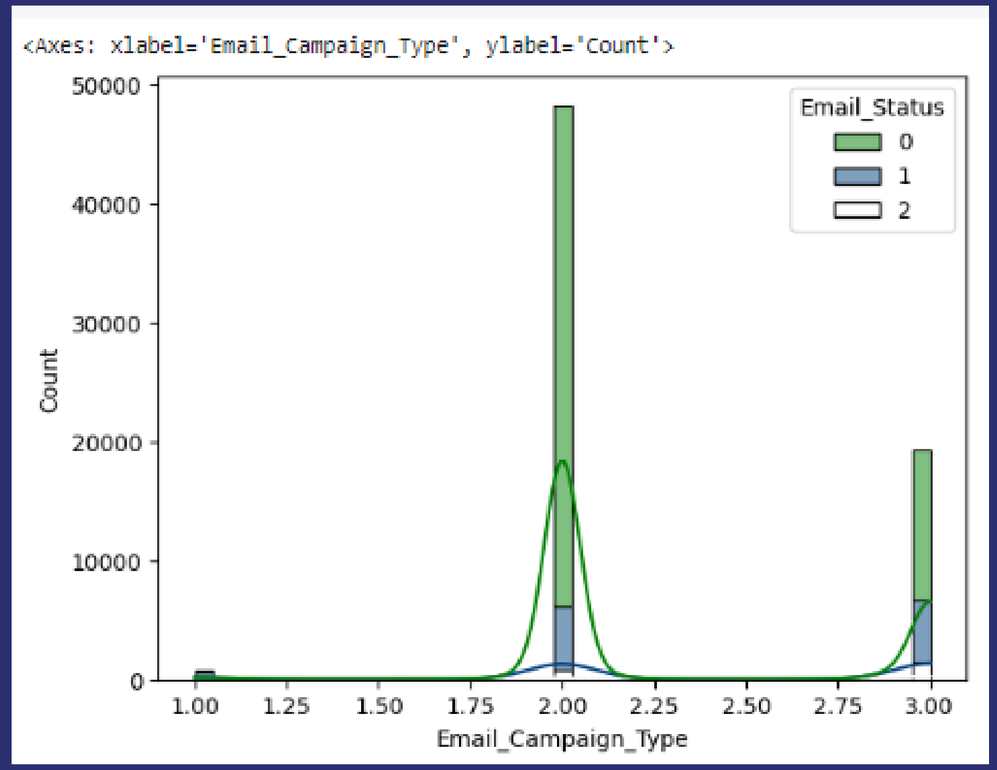


-Location G showed a high response rate compared to other locations.

Strategies and Suggestions:

-Try the strategies that worked at location G to increase response at other locations.

email campaign type



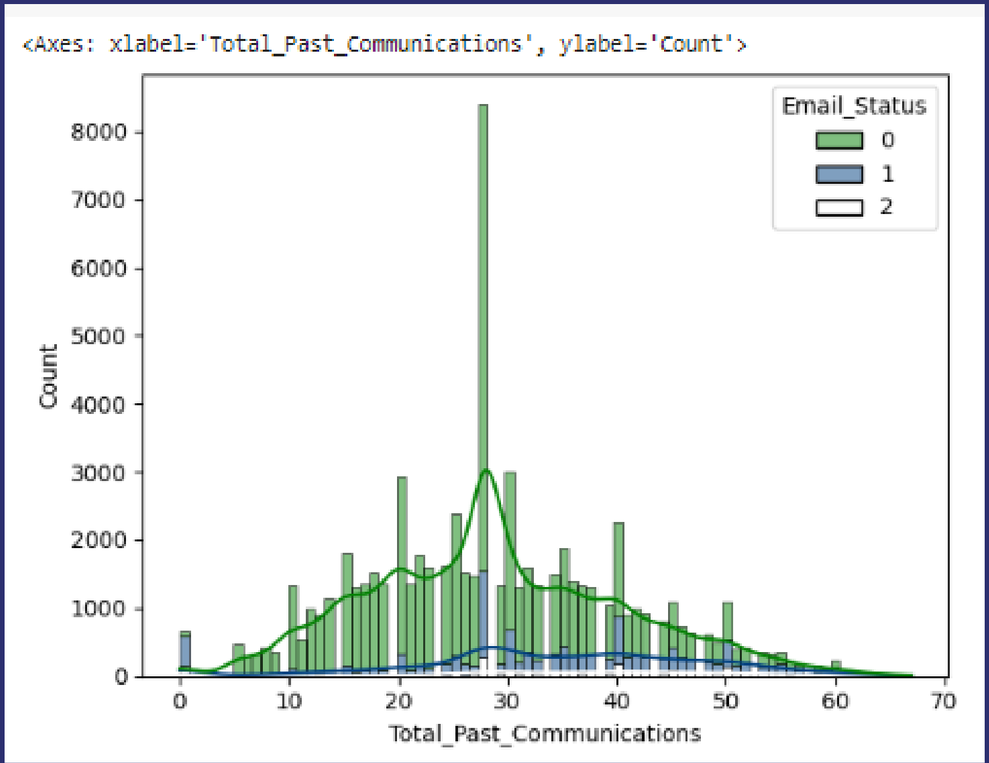
-Email_Campaign_Type 2 has the highest response rate.

Strategy and Advice:

-Campaign Optimization: Focus more campaigns on type 2 that show the best results.

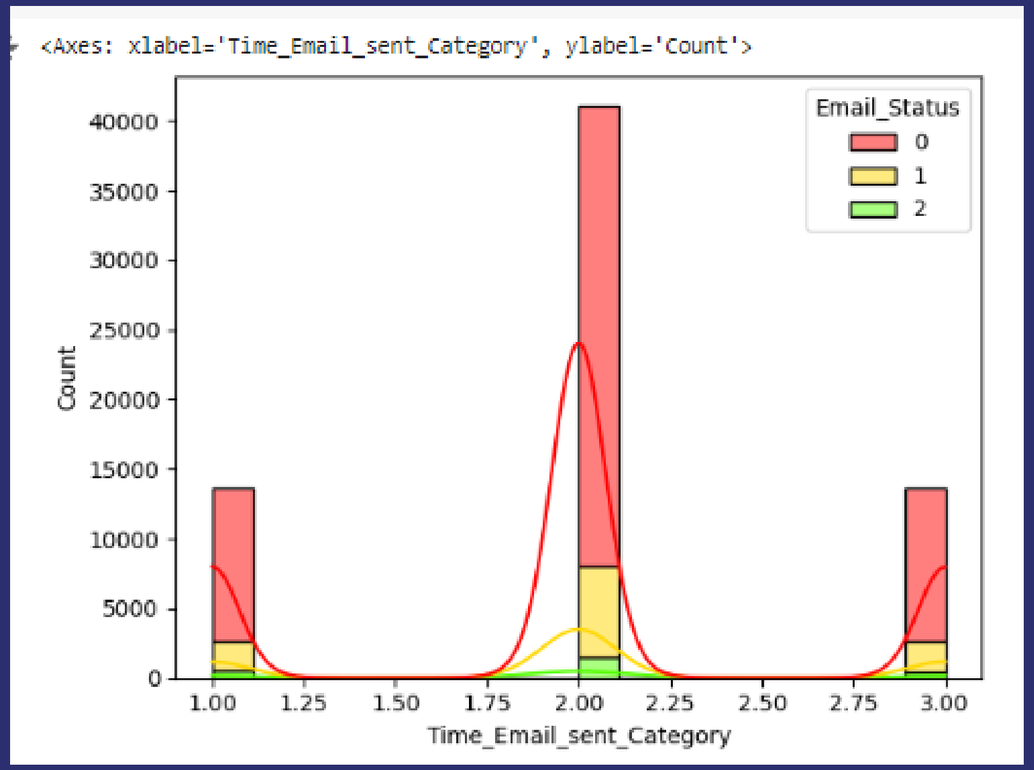
-Campaign Type Testing: Continuously test and evaluate other campaign types to find potential improvements.

Total past communication



-Higher number of previous communications (30)
Strategy and Advice:
-Increase Frequency: Increase the number of communications with customers who have shown prior interest.
-Personalization: Create more personalized and relevant communications to maintain interest and increase response.

Time email sent category



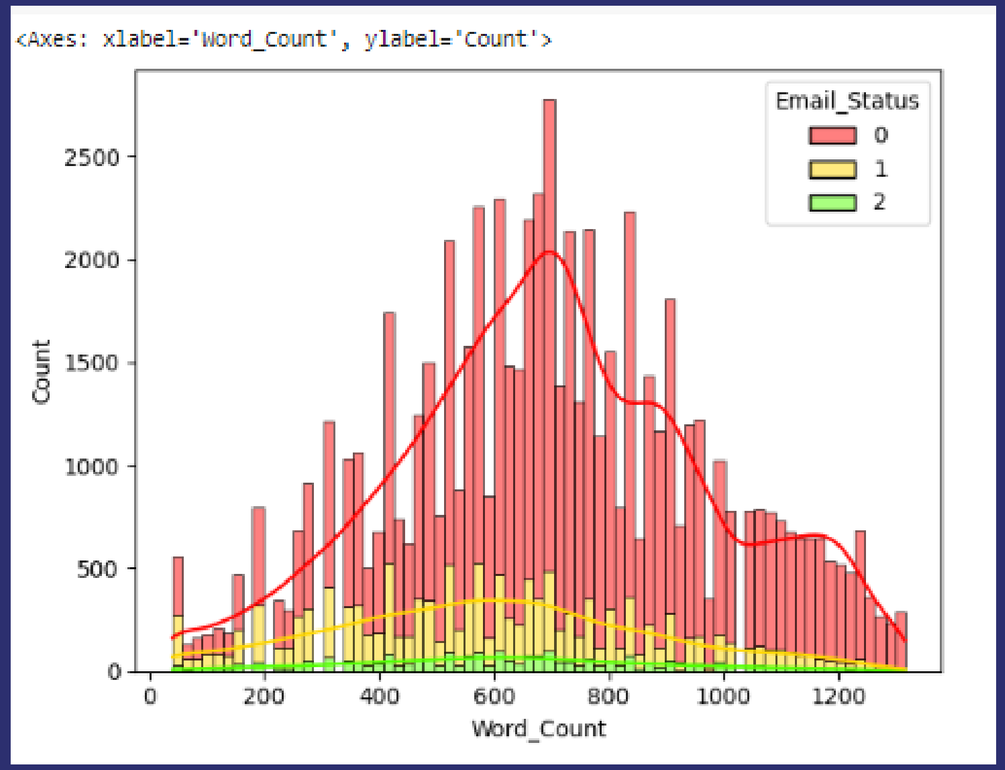
-Emails sent in time category 3 (evening) have higher open and reply rates.

Strategies and Suggestions:

-Optimal Scheduling: Send emails more frequently at time category 3 to maximise response.

-Sending Time Experiments: Conduct tests by sending emails at various times to find the best time that suits the audience.

Word count



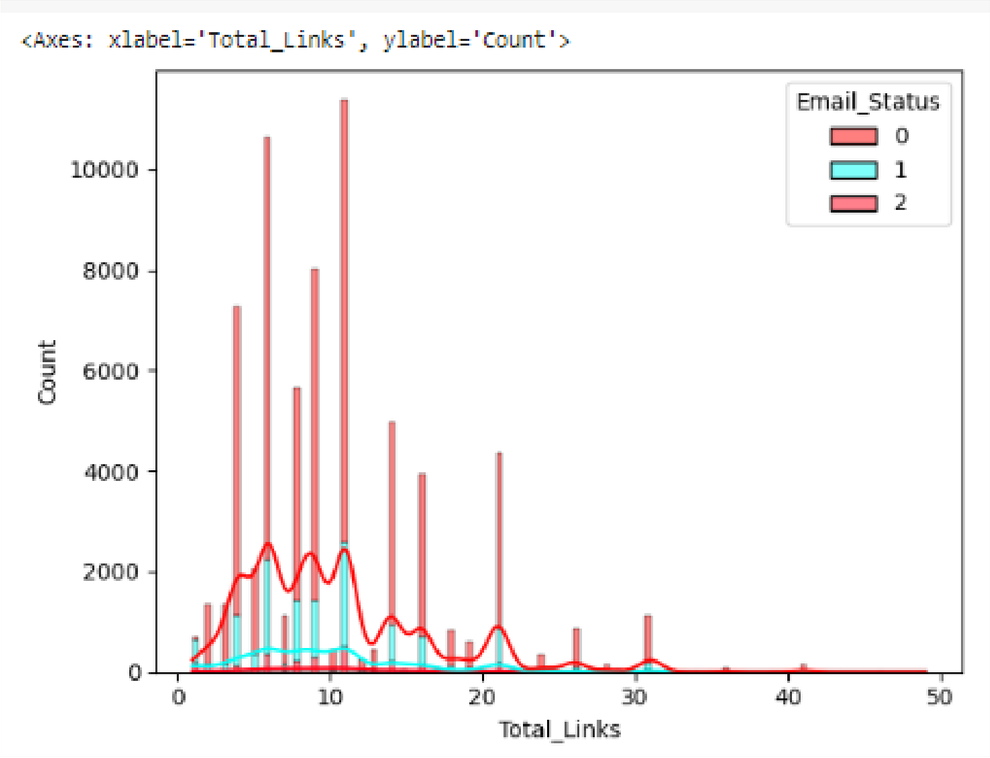
-Emails with a word count between 300-500 words have the highest response rate.

Strategies and Suggestions:

-Content Length Optimisation: Keep the email content length in the range of 300-500 words for maximum effectiveness.

-Clarity and Relevance: Ensure the content is concise, clear, and relevant to capture the reader's interest.

Total links



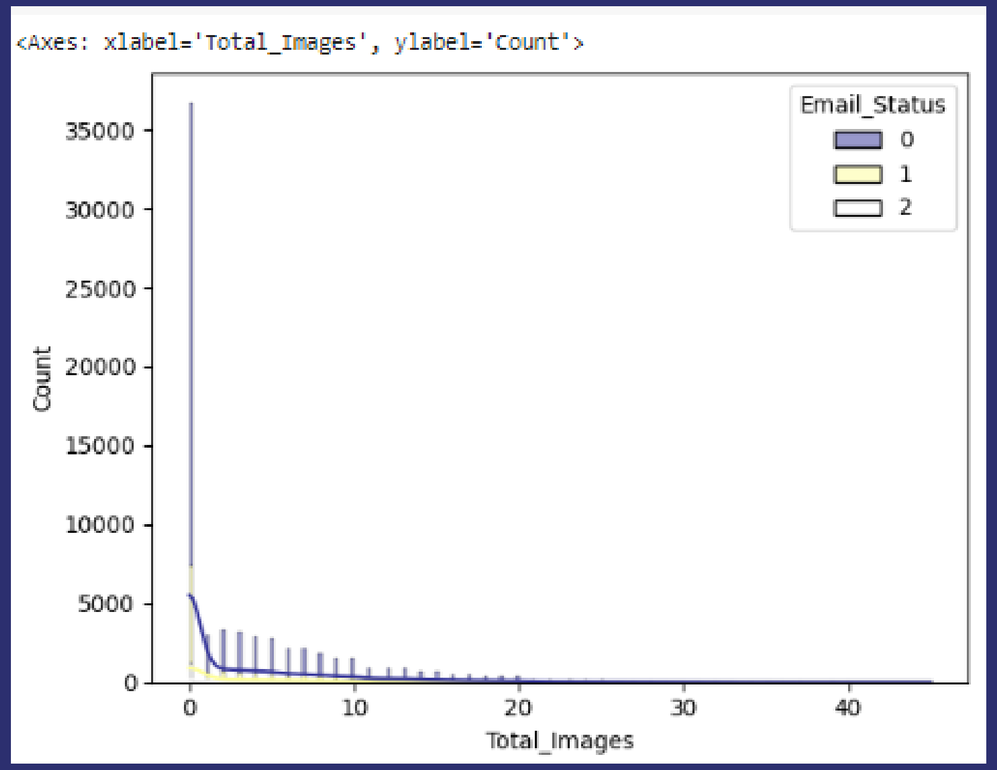
-Emails with 2-4 links have the highest response rate.

Strategy and Advice:

-Optimise the Number of Links: Keep the number of links in an email between 2-4 to increase response.

-Link Relevance: Make sure the links are relevant to the content and easy for readers to find.

Total images



-Emails with 1-2 images have the highest response rate.

Strategies and Suggestions:

-Optimise the Number of Images: Use 1-2 images in each email for best results.

-Visual Design: Use images that are attractive and support the message to increase interest.

Correlation

	Email_Status
Email_Status	1.00000
Email_Campaign_Type	0.18551

	Email_Status
Email_Status	1.000000
Customer_Location_Encoded	0.001459

	Email_Status
Email_Status	1.000000
Subject_Hotness_Score	-0.146531

	Email_Status
Email_Status	1.00000
Email_Campaign_Type	0.18551

	Email_Status
Email_Status	1.000000
Email_Source_Type	-0.024527

	Email_Status
Email_Status	1.000000
Total_Past_Communications	0.233065

	Email_Status
Email_Status	1.000000
Word_Count	-0.171116

	Email_Status
Email_Status	1.000000
Total_Links	-0.027846

	Email_Status
Email_Status	1.000000
Total_Images	-0.017392

correlation analysis conclusion

Based on the correlation analysis, it can be concluded that Total_Past_Communications and Email_Campaign_Type show a weak positive correlation to Email_Status, which means that more past communications and an effective email campaign type slightly increase the response rate. In contrast, Subject_Hotness_Score and Word_Count show a weak negative correlation, indicating that email subjects that are too high in score and emails that are too long tend to decrease response rates. Other factors such as Email_Source_Type, Customer_Location_Encoded, Total_Links, and Total_Images have very weak correlations, indicating that they hardly affect email response rates significantly. Therefore, marketing strategies should focus on increasing the frequency of effective communication, optimising campaign types, and customising the subject and length of emails to increase customer response rates.

Modelling data

Logistic Regression

	precision	recall	f1-score	support
0	0.82	0.98	0.89	16505
1	0.46	0.11	0.17	3312
2	0.00	0.00	0.00	689
accuracy			0.81	20506
macro avg	0.43	0.36	0.36	20506
weighted avg	0.74	0.81	0.75	20506

Accuracy: 0.8075197503169804

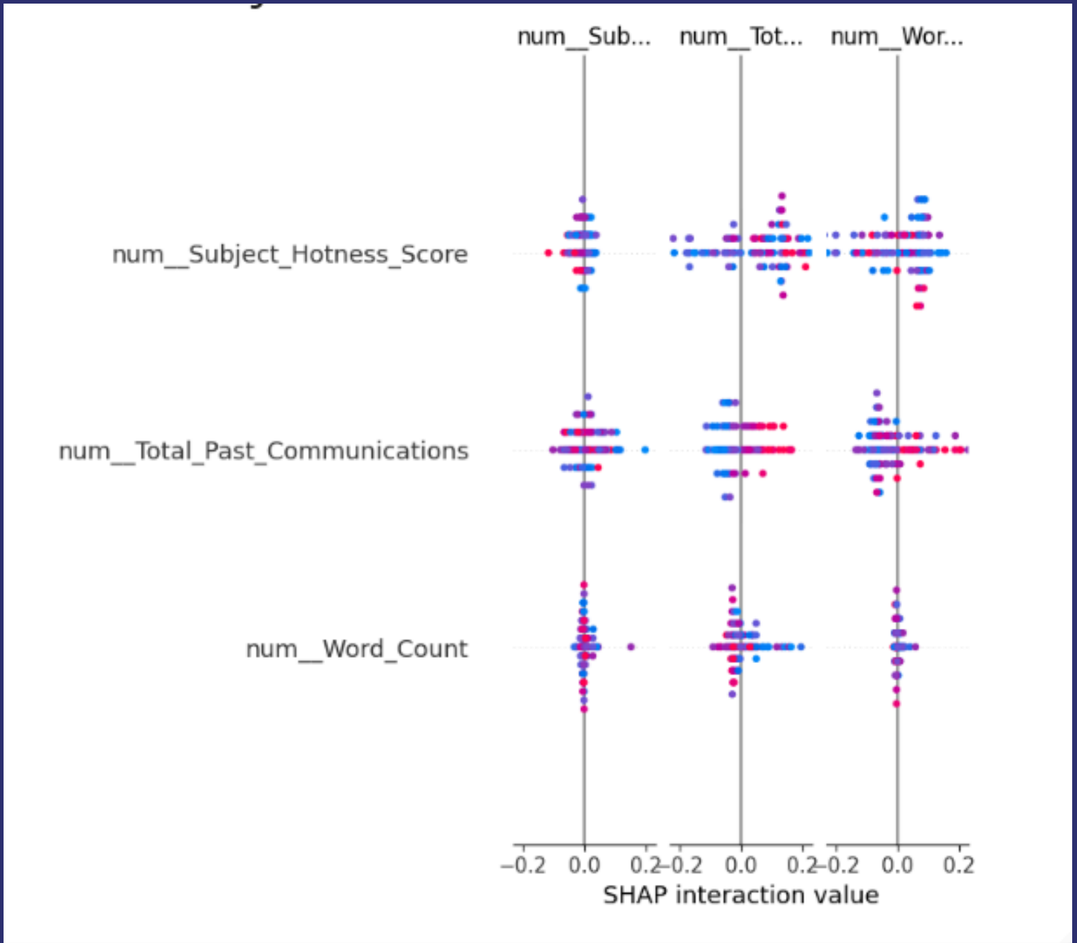
Random forest

	precision	recall	f1-score	support
0	0.83	0.96	0.89	16505
1	0.43	0.17	0.25	3312
2	0.17	0.03	0.06	689
accuracy			0.80	20506
macro avg	0.48	0.39	0.40	20506
weighted avg	0.75	0.80	0.76	20506

Accuracy: 0.8025455964108066

Of the two models, the best model is logistic regression even though the difference in accuracy is not too significant.

Shap value



Subject_Hotness_Score and Total_Past_Communications are two features that have more significant influence than Word_Count.

Subject_Hotness_Score has a balanced influence, with high values tending to improve model predictions.

Total_Past_Communications shows that more previous communications tends to improve model predictions.

Word_Count has a less significant influence on model predictions.

VIEW IN COLAB

VIEW IN GITHUB


Thank
You!