CAPSTONE PROJECT 3

EMAIL CAMPAIGN EFFECTIVENESS PREDICTION

By – Wasim Ahmad

Let’s Start Mailing

- Introduction
- Business Context
- Defining Problem Statement
- Data Summary
- Data Wrangling
- Data Visualization
- Analysis of categorical variable
- Analysis of continuous variable
- Heat Map Analysis
- Feature engineering and data preprocessing
- Analysis of dependent variable
- Models Evaluation
- Conclusion and Inference

Introduction

Email marketing is a form of marketing that can make the customers on your email list aware of new products, discounts, and other services. It can also be a softer sell to educate your audience on the value of business brand or keep them engaged between purchases.

In 1978, a marketing manager at Digital Equipment Corp named Gary Thuerk used this new method of direct communication to send out the first commercial email to let people know about a new product.

Business Context

Email Marketing is a powerful tool that small businesses can use to acquire, engage, and retain customers. And having a good clean email list is critical. Therefore, it's always recommended that you clean up any email list you want to use before sending a campaign. Here's an example of a good email verification provider. Below, we've outlined the four types of popular email marketing campaigns and how you can use them to help your business grow.

Email Newsletters.
Acquisition Emails
Retention Emails
Promotional Emails

Problem Statement

- Most of the small to medium business owners are making effective use of Gmail-based Email marketing Strategies for offline targeting of converting their prospective customers into leads so that they stay with them in Business.

- The main objective is to create a machine learning model to characterize the mail and track the mail that is ignored; read; acknowledged by the reader.

Data Summary

The dataset has 68353 rows and 12 columns.

The detailed descriptions are as follows.

- **Email ID** - This column have the email id's of individual customers.
- **Email type** - Email type have two categories 1 and 2. We can consider it as the marketing emails or important updates, notices like emails.
- **Subject Hotness Score** - This column has the subject-line score on the basis of content effectiveness.
- **Email Source** - This column represents the source of the email like sales, marketing or product type email.
- **Email Campaign Type** - The Campaign type of email.

Data Summary continued…

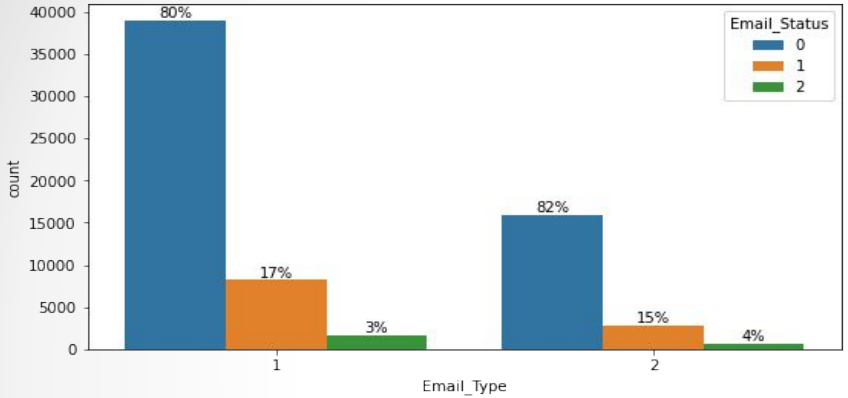
- **Total Past Communications** - This column contains the previous mails from the source.
- **Customer Location** - Categorical data which explains the different demographic location of the customers.
- **Time Email sent Category** - It has 3 categories: 1, 2 and 3 which are considered as morning, evening and night time slot.
- **Word Count** - It contains the no. of words in the mail.
- **Total Links** - Total links in the email body.
- **Total Images** - The total number of image in the email.
- **Email Status** - It is the target variable which contains the characterization of the mail that is ignored; read; acknowledged by the reader

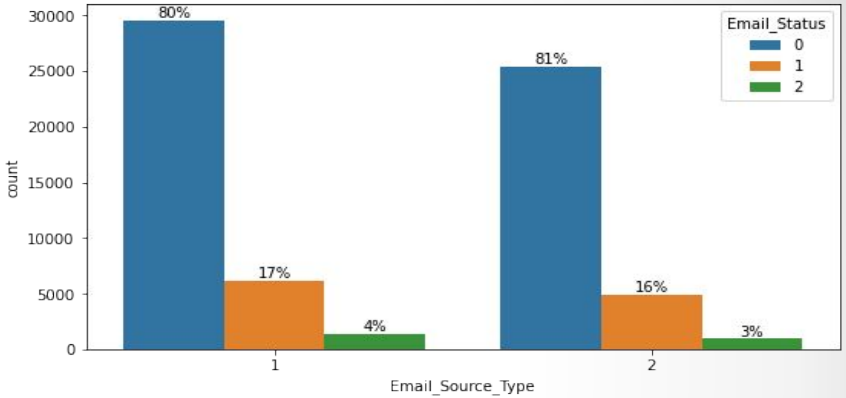
Data Wrangling

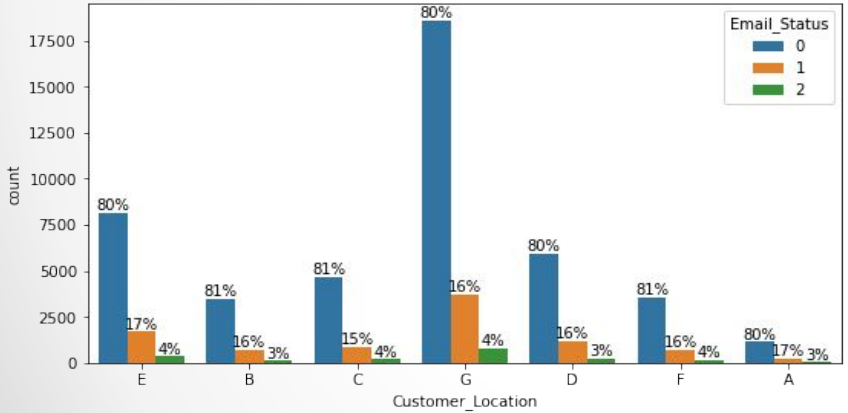
- As we can see we have 17% missing values in Customer Location.
- There is no missing value in any other categorical variable.
- Email Type and Email Source Type have 2 categories.
- Email Campaign Type, Time Email sent Category and Email Status have 3 Categories
- In the Variable 'Customer Location' distinct categories are 8 including one 'nan'
- No. of ignored Emails is highest as 54941
- No. of Emails read by customer is 11039
- No. of Acknowledge Emails is 2373

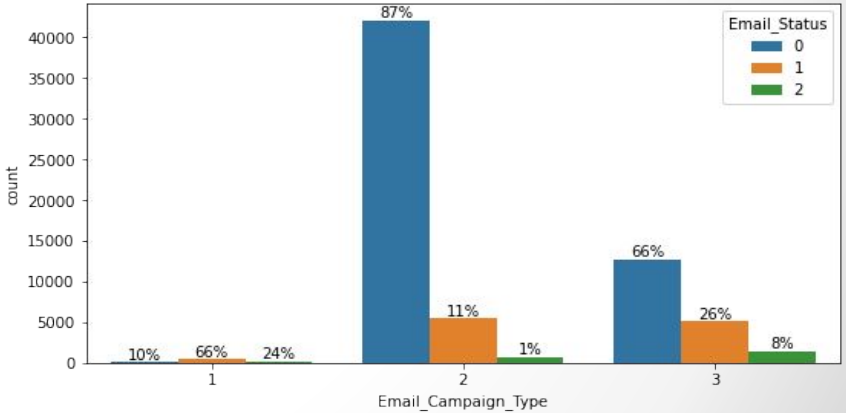
Data Visualization & Storytelling

Analysis of categorical variable







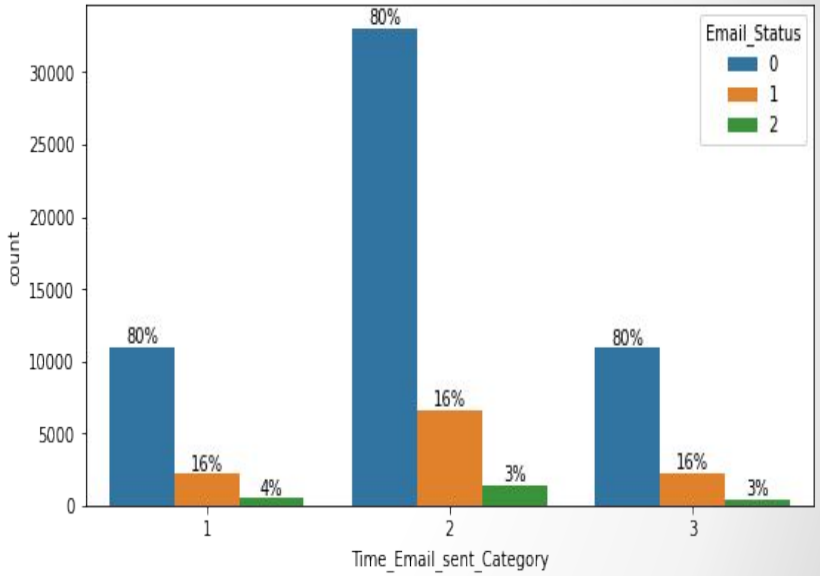


Analysis of categorical variable

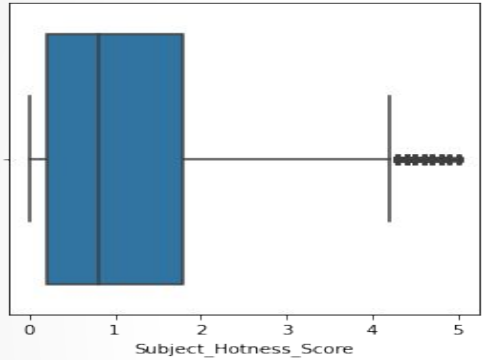
As we can observe the distribution of Email_Status is almost similar in all the categories except in Email_Campaign_Type

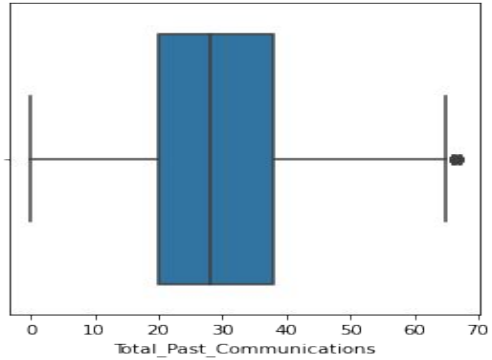
we can also see that it shows a totally different trend . For Email_Campaign_Type=1 we see that only 10% of customers are ignoring the email and for 2 around 87% customer ignore the emails.

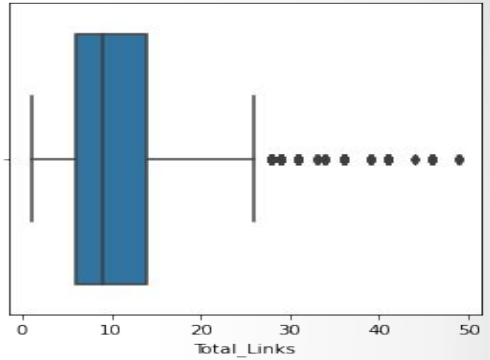
For getting better acknowledged by the reader email should be written Email Type 2, Email source type 1 and Email campaign type 1(Email campaign type 1 are 10% ignored, 66% read and 24% acknowledged by reader)

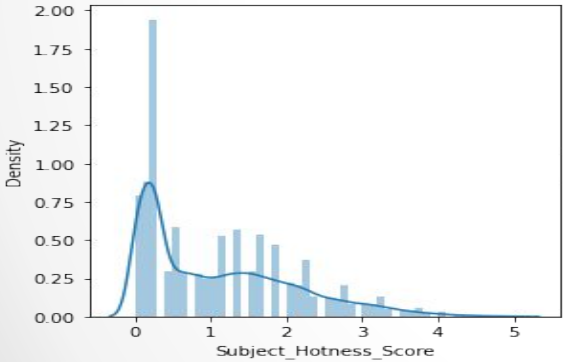


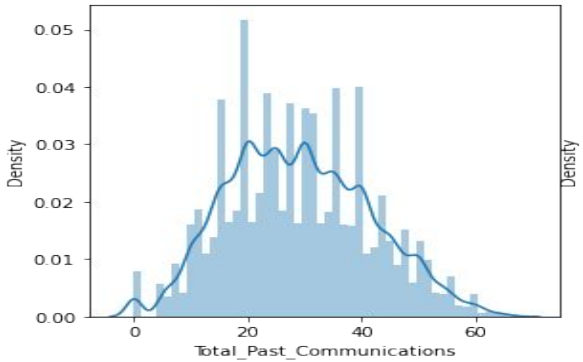
Analysis of continuous variable

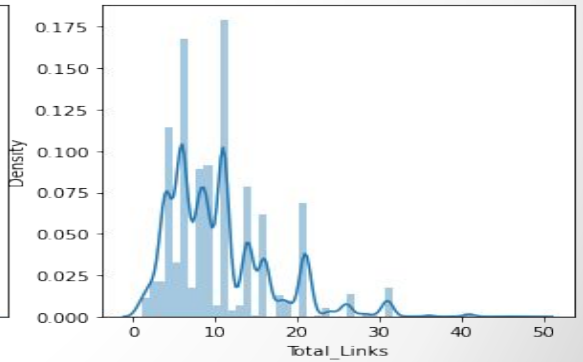












Analysis of continuous variable

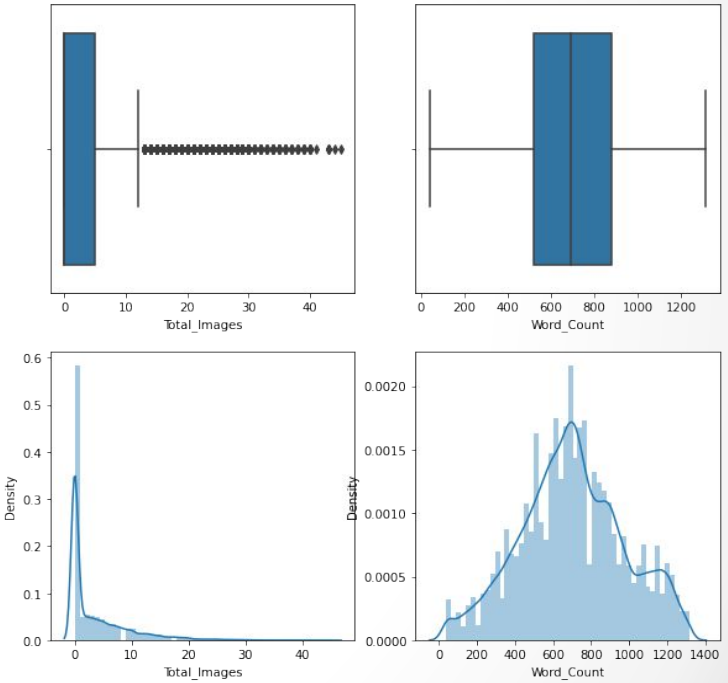
Subject_Hotness_Score All Email_Status i.e 0, 1, 2 have outliers. 0 have highest median and 1, 2 are right skewed. It is observed that the Subject_Hotness_Score for read/acknowledged mails are much lower.

Total_Past_Communications 0, 2 have outliers and 2 have highest median .

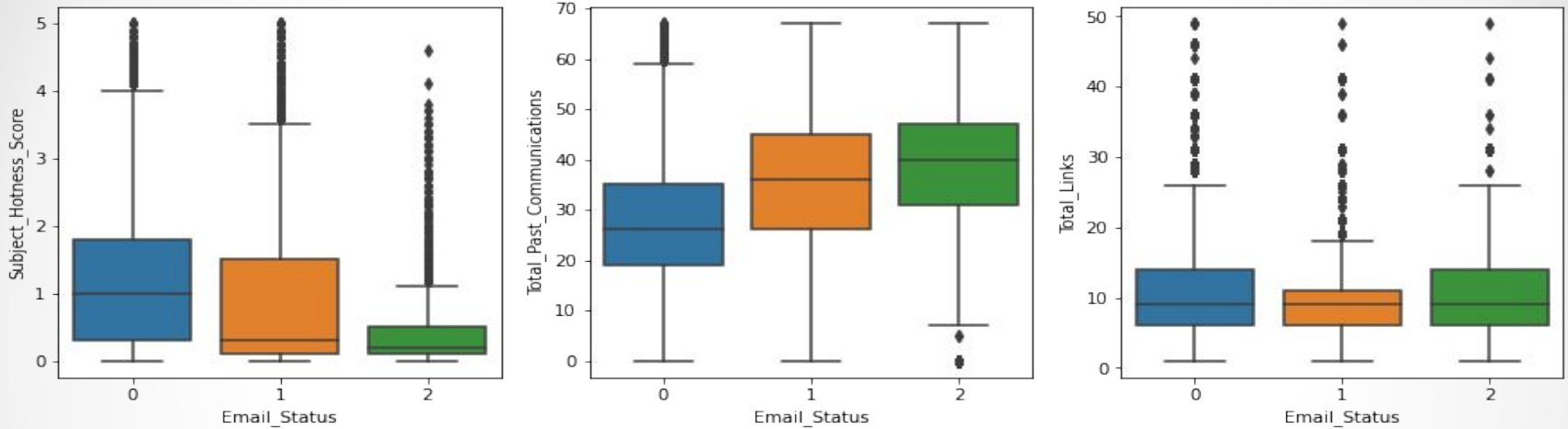
Total_Links 0, 1, 2 all have outliers, All have the same median but 0,2 have higher variance compare to 1.

Total_Images 0, 1, 2 all have outliers and All have the same median. Hence all the mails have the same range of images.

Word_Count Median of 0 is highest. Thus we can understand that ignored mails have higher word count.



Analysis of continuous variable



Subject_Hotness_Score 0, 1, 2 have outliers. 0 have highest median and 1, 2 are right skewed. It is observed that the Subject_Hotness_Score for read/acknowledged mails are much lower.

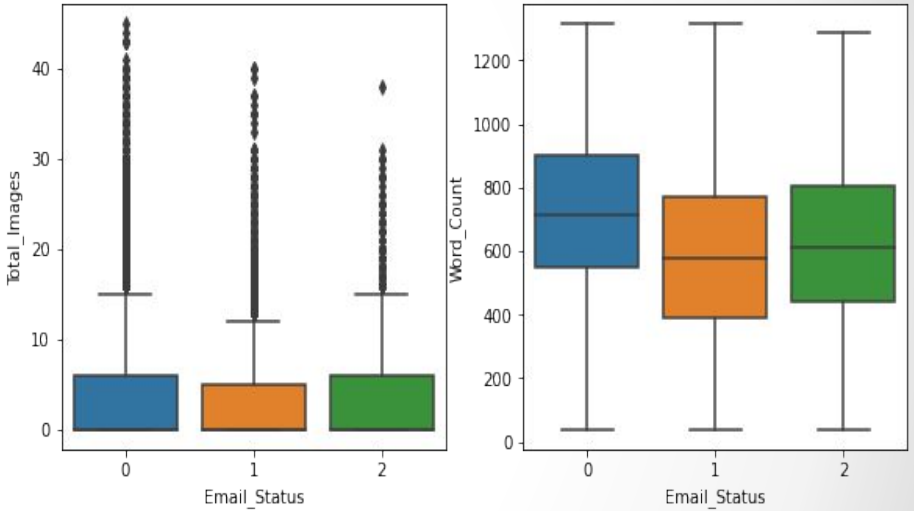
Total_Past_Communications 0, 2 have outliers and 2 have highest median .

Total_Links 0, 1, 2 all have outliers,All have the same median but 0,2 have higher variance compare to 1.

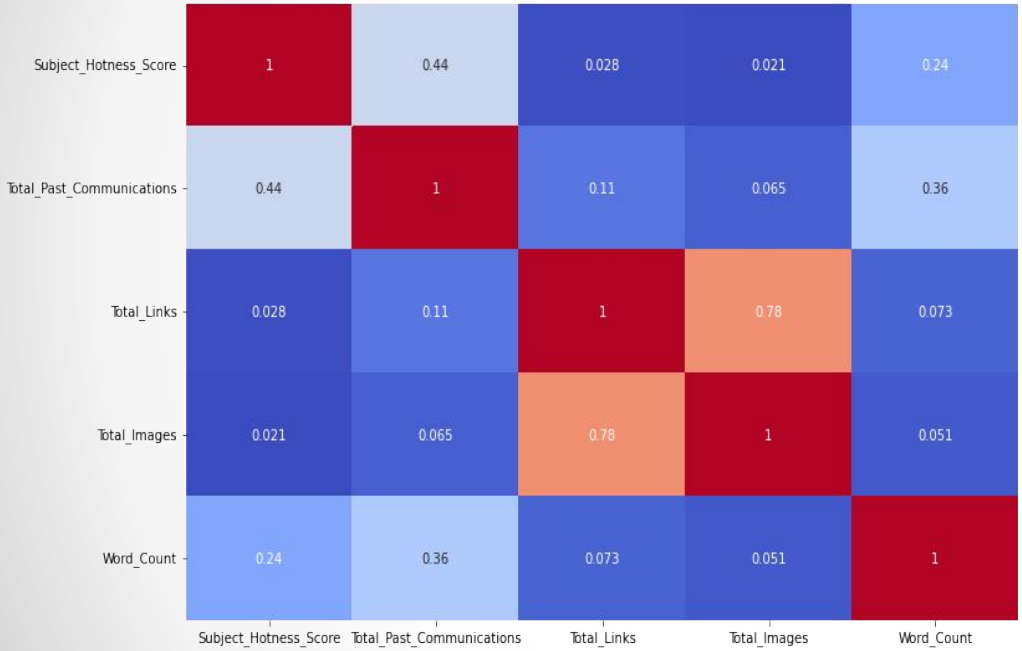
Analysis of continuous variable

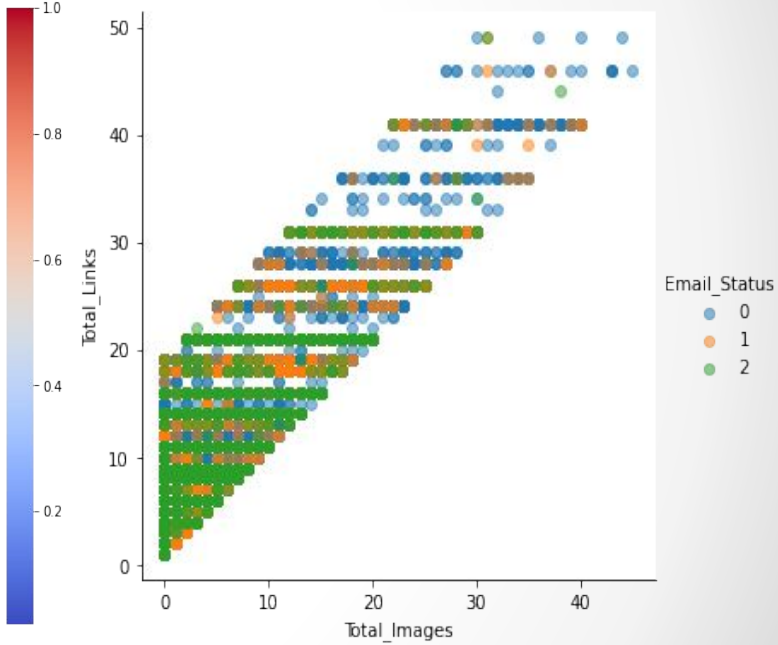
Total_Images 0, 1, 2 all have outliers and All have the same median. Hence all the mails have the same range of images.

Word_Count Median of 0 is highest. Thus we can understand that ignored mails have higher word count.



Heat Map Analysis





Heat Map Analysis

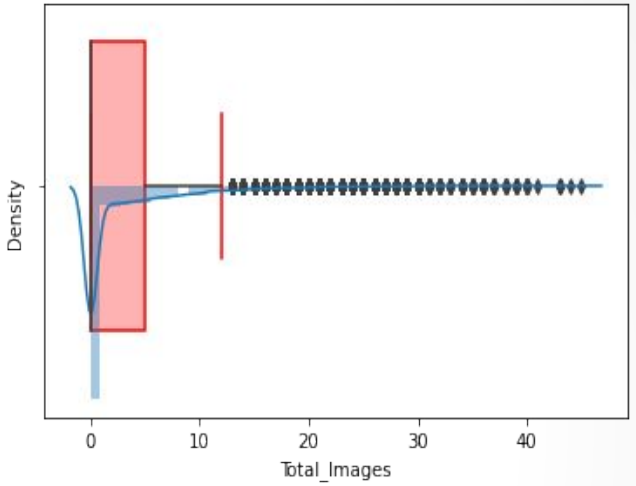
Total_Images and Total_Links, they have 78% positive correlation.

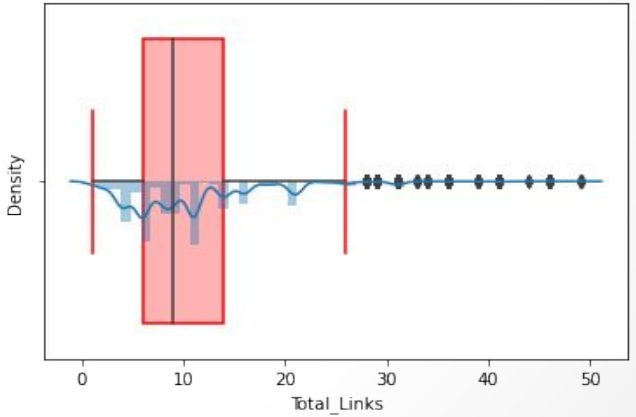
There is a high positive correlation between these two features.

More than 50% of values are 0 and there are there is a presence of outliers in Total_Images

Compared to Total_Images, Total_Links has very few outliers.

Since images and links are correlated and most of the values in Total_Images is 0 we combined both the features.

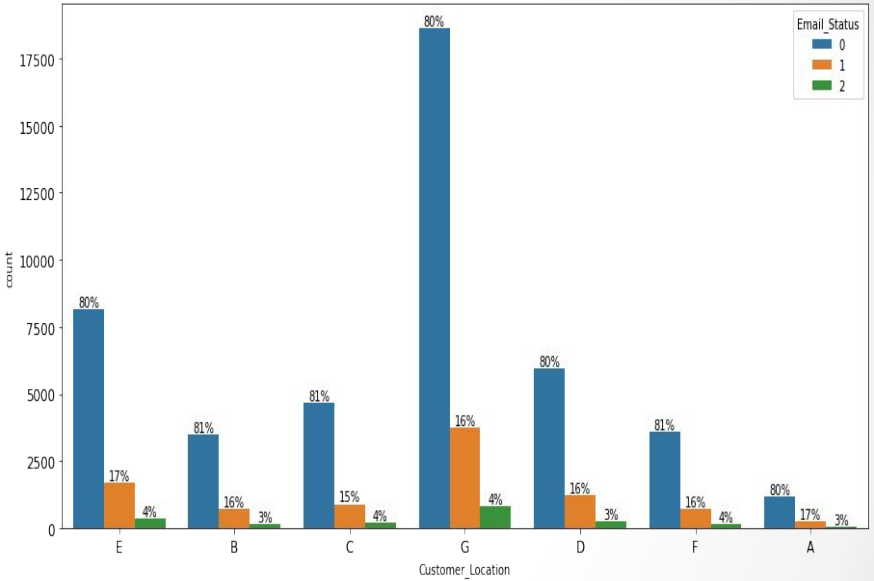




Feature Engineering and Data Preprocessing

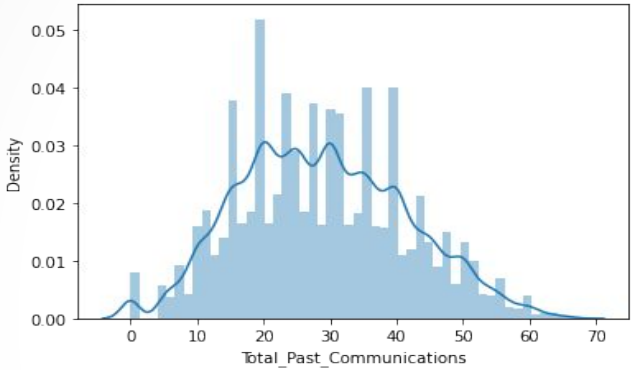
```
# Number of NaNs for each object
email_campaign_data.isnull().sum()
```

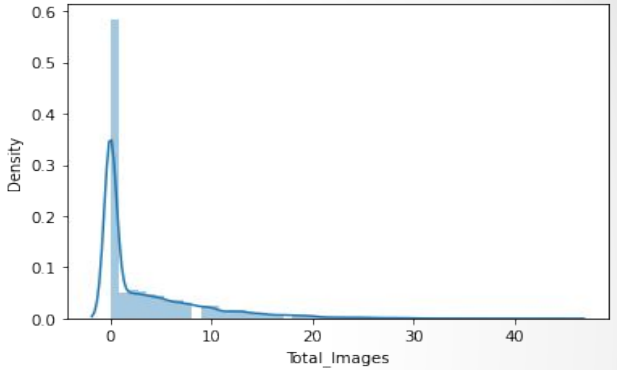
```
Email_ID                        0
Email_Type                      0
Subject_Hotness_Score           0
Email_Source_Type               0
Customer_Location           11595
Email_Campaign_Type             0
Total_Past_Communications    6825
Time_Email_sent_Category        0
Word_Count              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



Irrespective of location the ratio of Email_Status is same throughout . So we dropped this customer location column

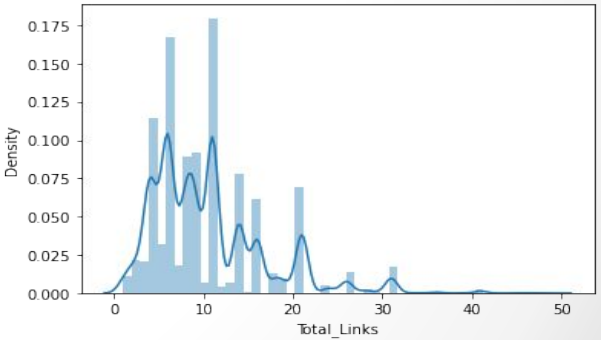
Feature Engineering and Data Preprocessing



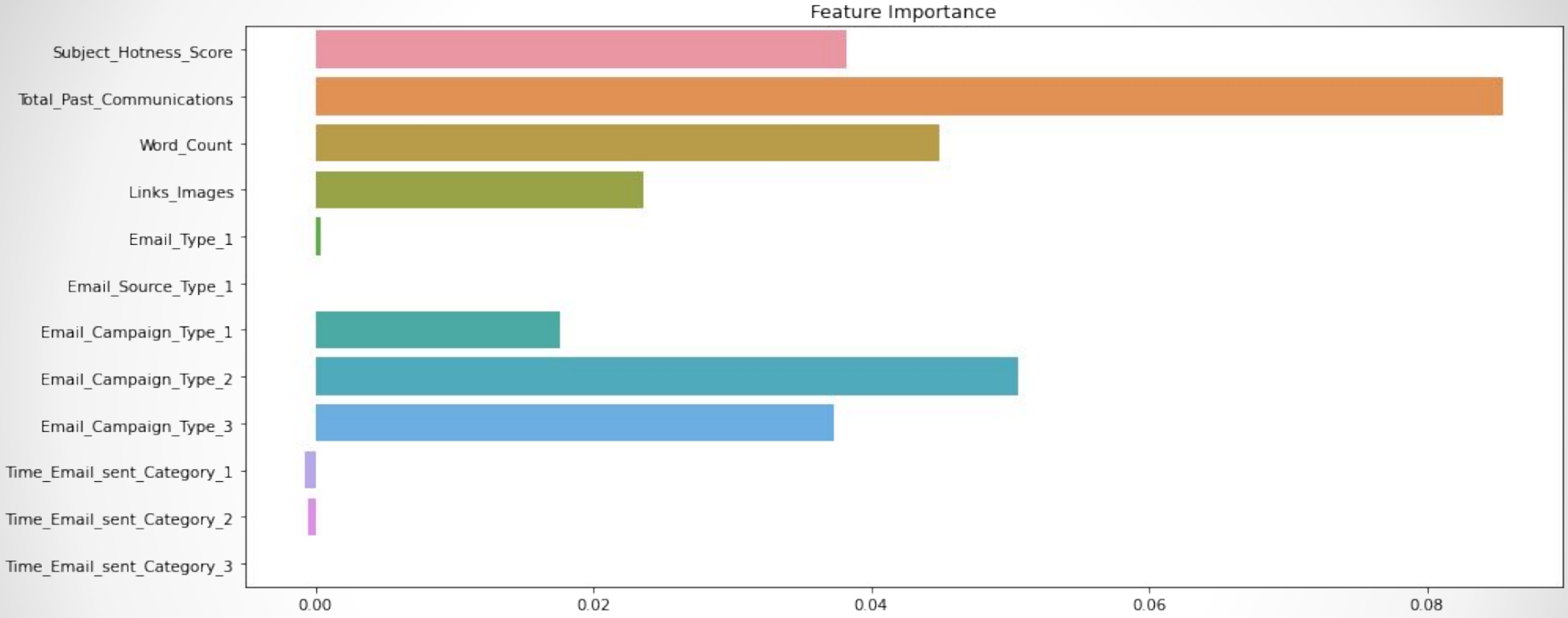


Above data is **symmetrically distributed** so **we can use mean value** for imputing missing values

In 2nd data is **not normally distributed** we will **use Mode** to impute missing value



Total_Links data is **not normally distributed** we will use **Mode** to impute missing value

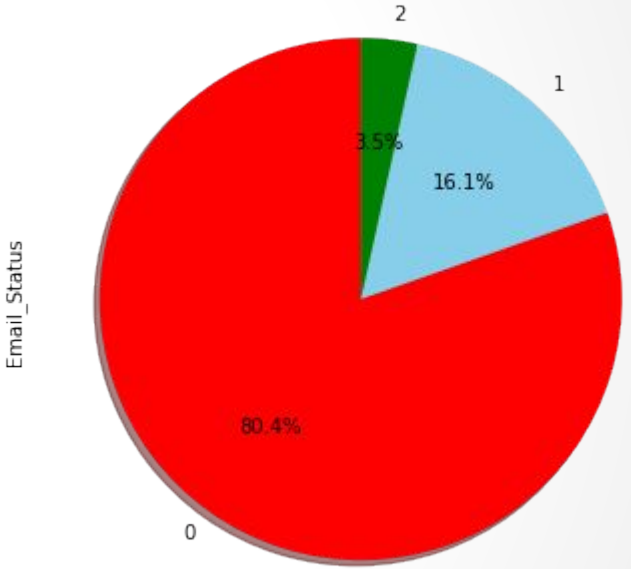


Time_Email_sent_Category_1, Time_Email_sent_Category_2, Time_Email_sent_Category_3 have very less importance we can drop this feature.

Analysis of Dependent variable

In Email_Status

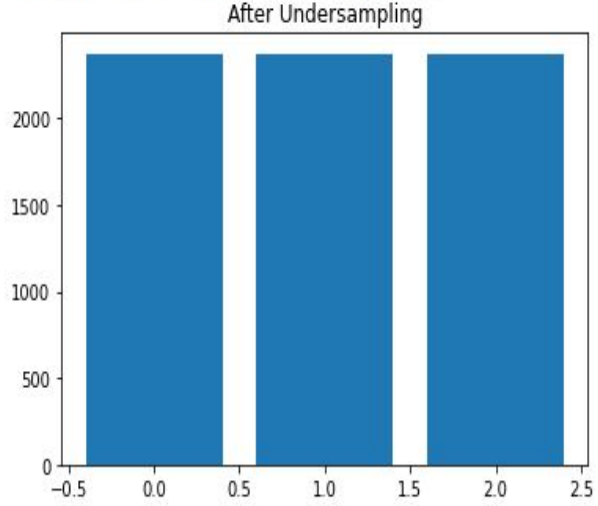
Type 0 - 54941
Type 1 - 11039
Type 2 - 2373



Random Under Sampling can be defined as removing some observations of the majority class. This is done until the majority and minority class is balanced out.

SMOTE (Synthetic Minority Oversampling Technique) works by randomly picking a point from the minority class and computing the k-nearest neighbors for this point. The synthetic points are added between the chosen point and its neighbors.

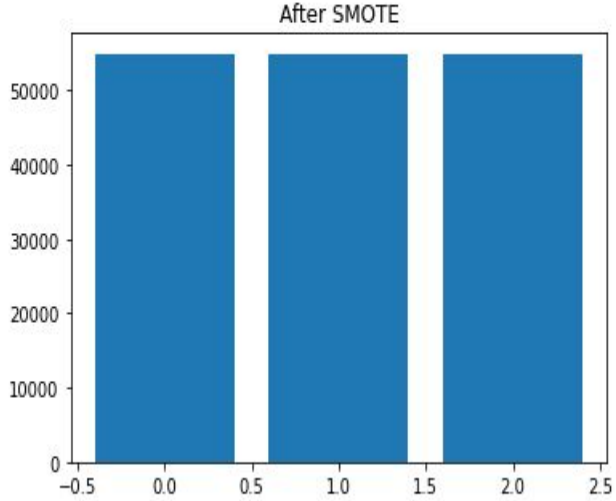
```
Text(0.5, 1.0, 'After Undersampling')
```




```
unique_elements, counts_elements = np.unique(y_rus, return_counts=True)
print("Frequency of unique values of the Email_Status:")
print(np.asarray((unique_elements, counts_elements)))
```

```
Frequency of unique values of the Email_Status:
[[   0    1    2]
 [2373 2373 2373]]
```

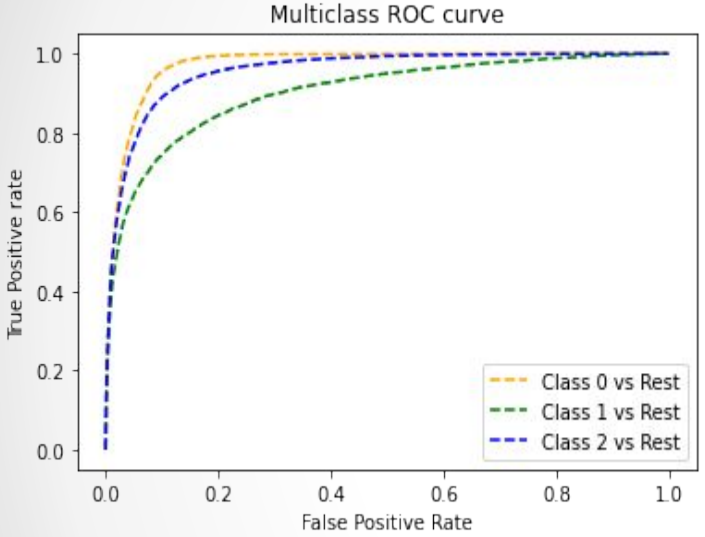
```
Text(0.5, 1.0, 'After SMOTE')
```

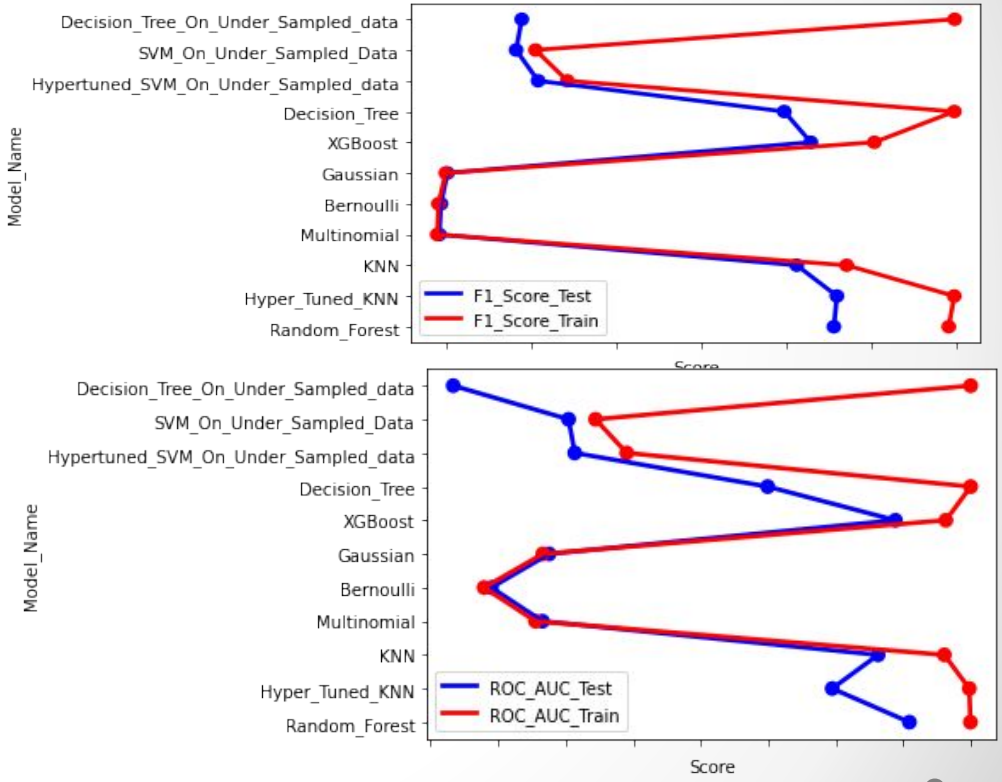



```
unique_elements, counts_elements = np.unique(y_smote, return_counts=True)
print("Frequency of unique values of the Email_Status:")
print(np.asarray((unique_elements, counts_elements)))
```

```
Frequency of unique values of the Email_Status:
[[    0     1     2]
 [54941 54941 54941]]
```

Model Evaluation





Random Forest and **XG – Boost** performed well as compare to all the other models both for train as well as test.

Conclusion and Inference

we observed that Email_Campaign_Type was the most important feature. If your Email_Campaign_Type is 1, there is a 90% likelihood of your Email to be read/acknowledged.

It is observed that both Time_Email_Sent and Customer_Location were insignificant in determining the Email_status. The ratio of the Email_Status was same irrespective of the demographic or the time frame the emails were sent on.

For getting better acknowledged by the reader email should be of Email Type 2, Email source type 1 and Email campaign type 1(Email campaign type 1 are 10% ignored, 66% read and 24% acknowledged by reader)

As the word_count increases beyond the 600 mark we see that there is a high possibility of that email being ignored. The ideal mark is 400-600. No one is interested in reading long Emails.

We also seen that for imbalance handling Oversampling i.e. SMOTE worked way better than undersampling as the latter resulted in a lot of loss of information.

Random Forest and **XG-Boost** performed well as compare to all the other models both for train as well as test.

.

Q & A

THANK YOU