Capstone Project
Email Campaign Effectiveness Prediction

CONTRIBUTOR:

Yogesh K

Content

- Problem statement
- Data Summary
- Data Cleaning
- Imputing missing values
- Analysis of categorical features
- Analysis of Continuous features
- Outlier treatment
- Feature Engineering
- Understanding target variable
- Handling Imbalanced data.
- Different Models
- Challenges
- Conclusion

Mail 1

mariam Abacha
To Me

Dec 11 at 4:10 PM

MY BROTHER
THANK YOU FOR YOUR MAIL ,THE REASON WHY YOUR PASSPORT IS NEEDED IS FOR ME TO KNOW WHOM I AM DEALING WITH AND FOR THE AMERICAN FEMALE AGENT MRS SANDRA NELSON TO RECOGNIZE YOU WHEN BOTH OF YOU WILL MEET ONE ON ONE IN NEW YORK . MY BROTHER EVERY ARRANGEMENT HAS BEEN MADE IN OTHER TO MOVE THE CONSIGNMENT TO NEW YORK ,THE LONDON AGENT WILL PROCEED TO AMERICA TOMORROW WITH THE CONSIGNMENT FUNDS UPON HIS ARRIVAL HE WILL HAND OVER THE CONSIGNMENT TO THE AMERICAN FEMALE AGENT MRS SANDRA NELSON ,IN THESE ARRANGEMENT I WILL LIKE YOU TO SEND YOUR PASSPORT ,DIRECT TELEPHONE NUMBER ,AND COMPLETE ADDRESS WHICH WILL ENABLE MY LAWYER TO OBTAIN THE POWER OF ANTONY AND MOU ON YOUR BEHALF BECAUSE THE DOCUMENT WILL EMPOWER YOU TO RECEIVE THE CONSIGNMENT WITHOUT ANY PROBLEM . AS REGARDS TO THE DELIVERY AGENT HE HAS HUMILITY THAT WILL PROTECT HIM AND THE CONSIGNMENT ,THERE IS NO PROBLEM ABOUT THAT . ALL I NEED FROM YOU IS FOR YOU TO ASSURE ME THAT YOU WILL NOT BETRAY ME OR SIT ON THE FUNDS WHEN IT GETS TO YOU AND I WANT YOU TO LET ME KNOW THE AREA YOU WILL INVEST THESE FUNDS FOR ME AFTER YOU MUST HAVE TAKEN THE 30% OF THE TOTAL SUM WHICH I OFFER YOU ,ALL YOU NEED TO DO BY THE TIME YOU RECEIVE THE CONSIGNMENT FROM THE AGENT MRS SANDRA NELSON IS TO DEPOSIT THE FUNDS TO YOUR LOCAL ACCOUNT IN NEW YORK AND RETURN BACK TO YOUR CITY AFTER THEN I CAN SEND YOU INFORMATION FOR YOU TO SEND MONEY TO ME TO FLY DOWN AND MEET YOU ONE ON ONE ,I HEREBY ATTACH THE PICTURE OR THE MONEY FOR YOU TO SEE HOW THE FUNDS WAS PACKAGE AND DEPOSITED TO THE SECURITY COMPANY ,AS SOON AS YOU GET BACK TO THIS MAIL I WILL LET YOU KNOW HOW MUCH IT WILL COST YOU FOR THE CLEARANCE IN NEW YORK BECAUSE I WILL TAKE CARE THE SHIPMENT IN LONDON TO NEW YORK ,WHILE YOU WILL TAKE CARE OF THE CLEARANCE IN NEW YORK . I WAIT FOR THE INFORMATION SO THAT MY LAWYER CAN OBTAIN THE DOCUMENTS TOMORROW.
MARIAM .

Mail 2

Greetings Karen

It was great to meet you last week at Zuora Day and/or Dreamforce. I look forward to connecting how we may take next steps regarding your interest and initiative to improve subscription commerce, billing and payment capabilities.

Perhaps you may be interested in a FREE Trial?

I am including a few items for your review:

1/ Zuora Day highlights
http://www.youtube.com/watch?v=FhhgF4dkXEU

2/ The Definitive Guide to Subscription Commerce. In this guide you will learn the new rules required to run a successful subscription business utilizing Zuora + SalesForce.

Click here to get "The Definitive Guide to Subscription Commerce"
http://info.zuora.com/DefinitiveGuideDownload.html

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 comprised of 12 features including the target variable **Email_Status**.
- The **5 numerical variables** were :

  Word_Count

  Total_Past_Communications

  Subject_Hotness_Score

  Total_Links

  Total_Images
- The **5 categorical variables** were:

  Email_Type

  Email_Source_Type

  Customer_Location

  Email_Campaign_Type

  Time_Email_Sent_Catergory
- The total no. of records in our dataset is 68353

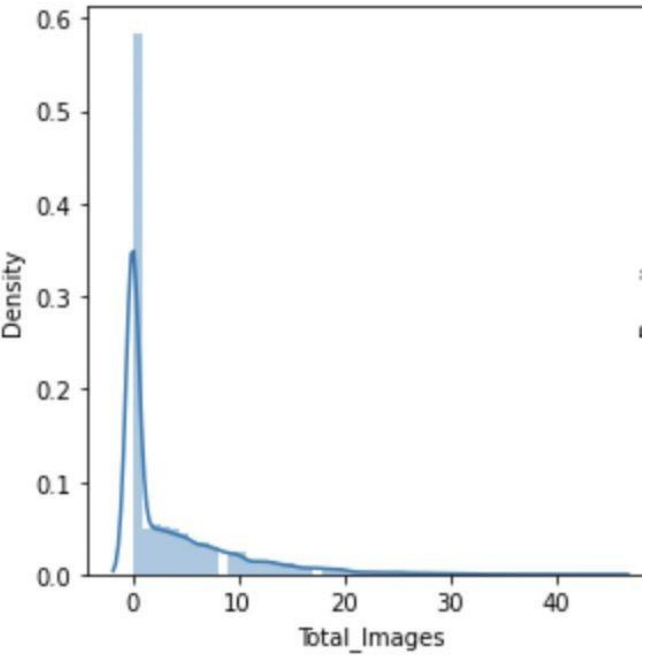
Data Cleaning

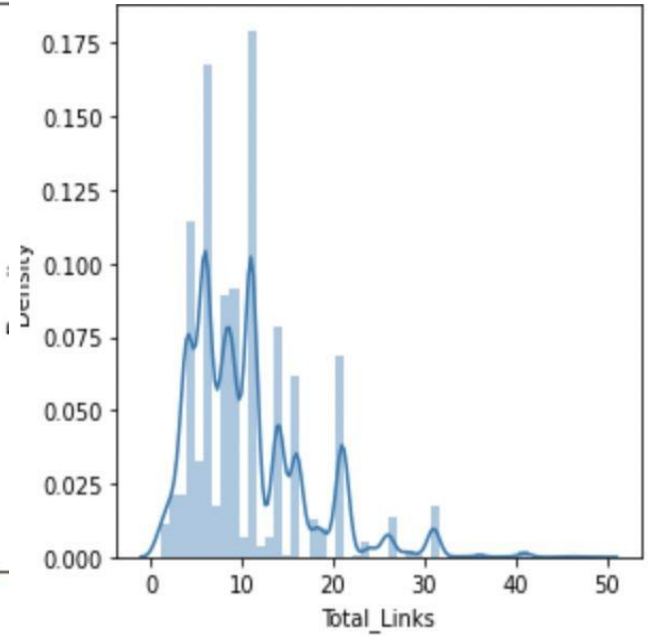
1. Null Value Imputation:

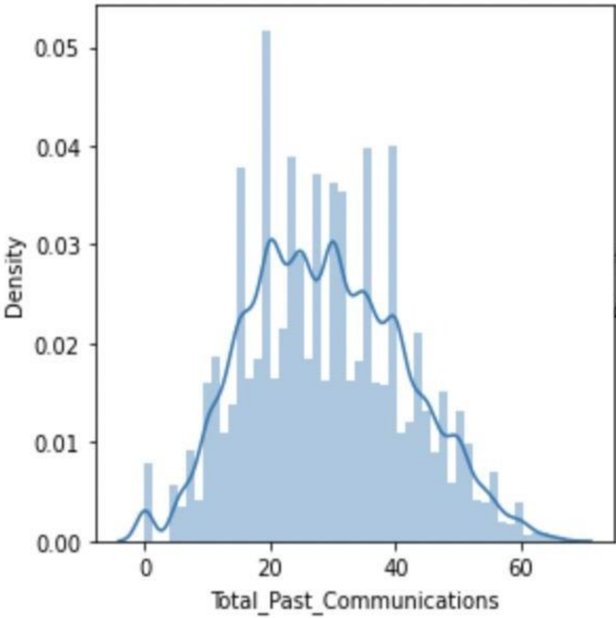
```
Email_ID                        0
Email_Type                      0
Subject_Hotness_Score           0
Email_Source_Type               0
Customer_Location           11595
Email_Campaign_Type             0
Total_Past_Communications    6825
Time_Email_sent_Category        0
Word_Count              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

Imputing missing values

- Impute the missing values for Total_Past_Communication by the mean
- Impute the missing values for Total_Links & Total_Images by the mode



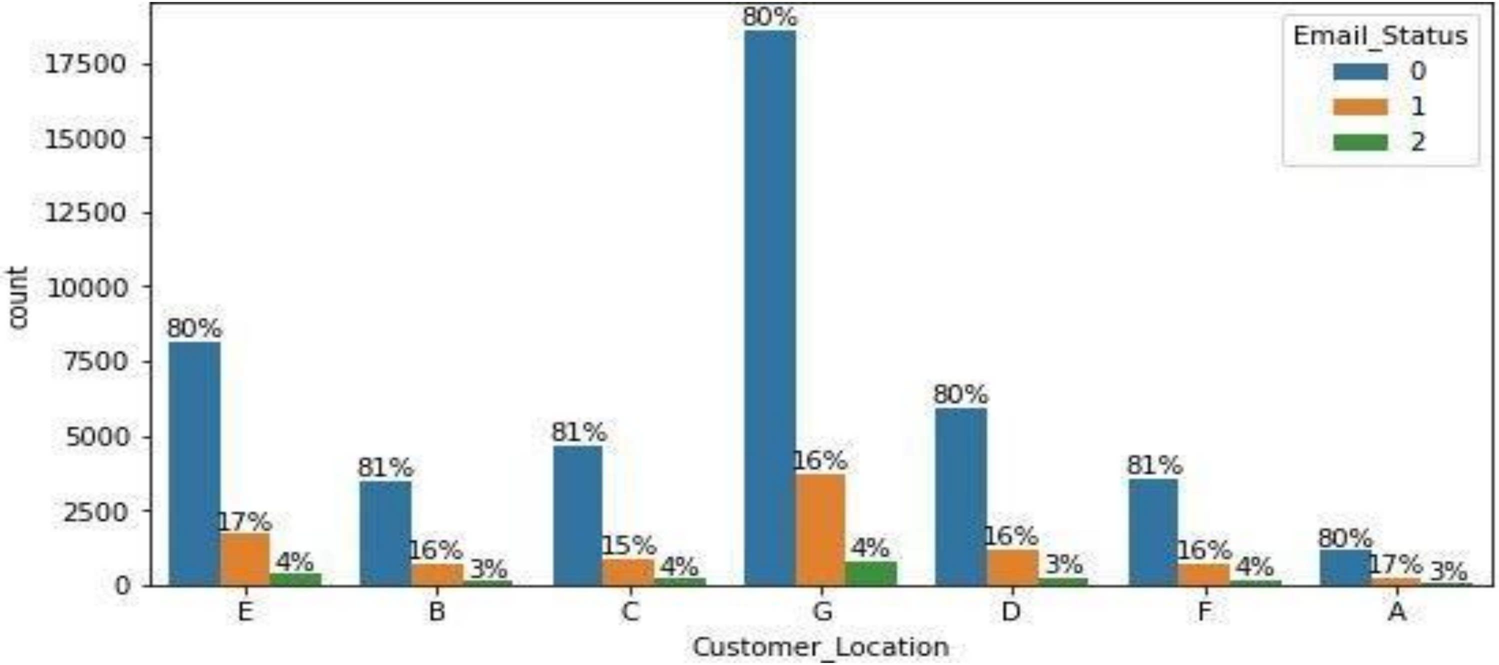




Analysis of Categorical features

- **Customer_Location w.r.t Email_Status**

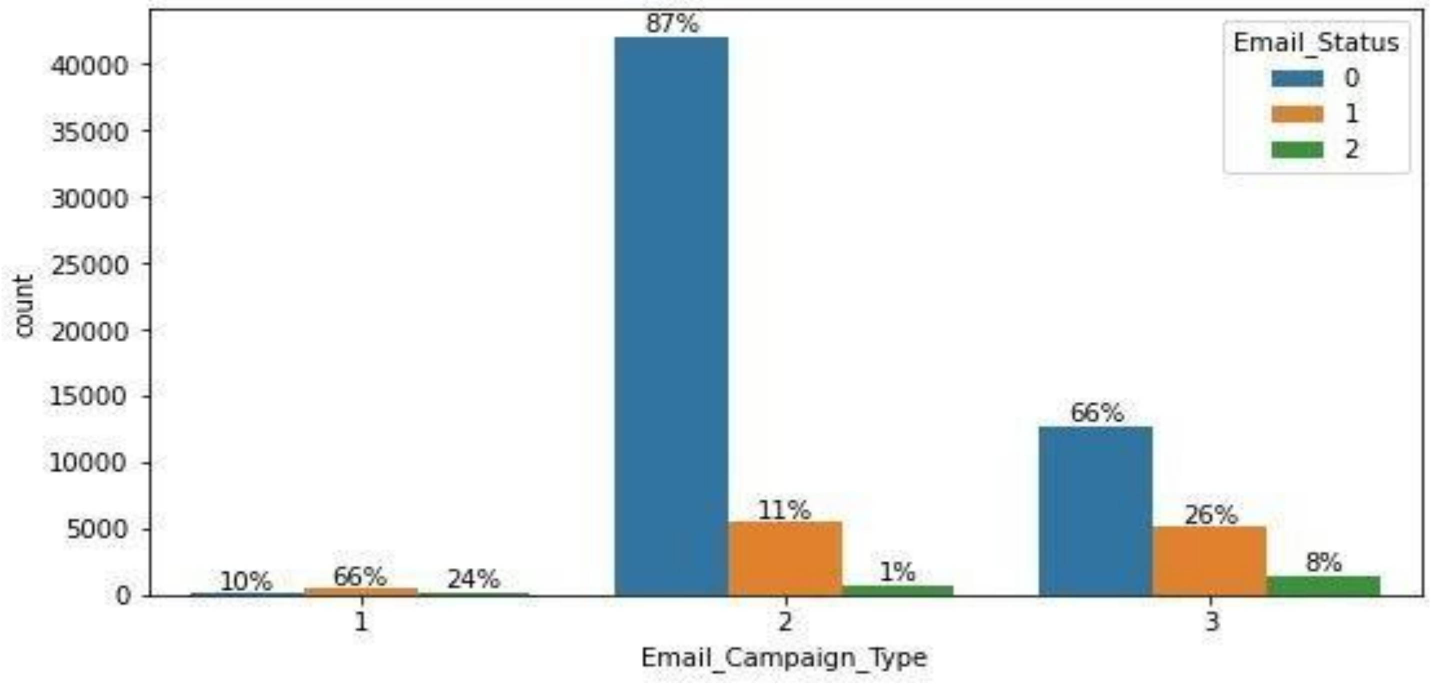
  Inference: same ratio of Email_Status for different demographics



Analysis of Categorical features

- **Email_Campaign_Type w.r.t. Email_Status**

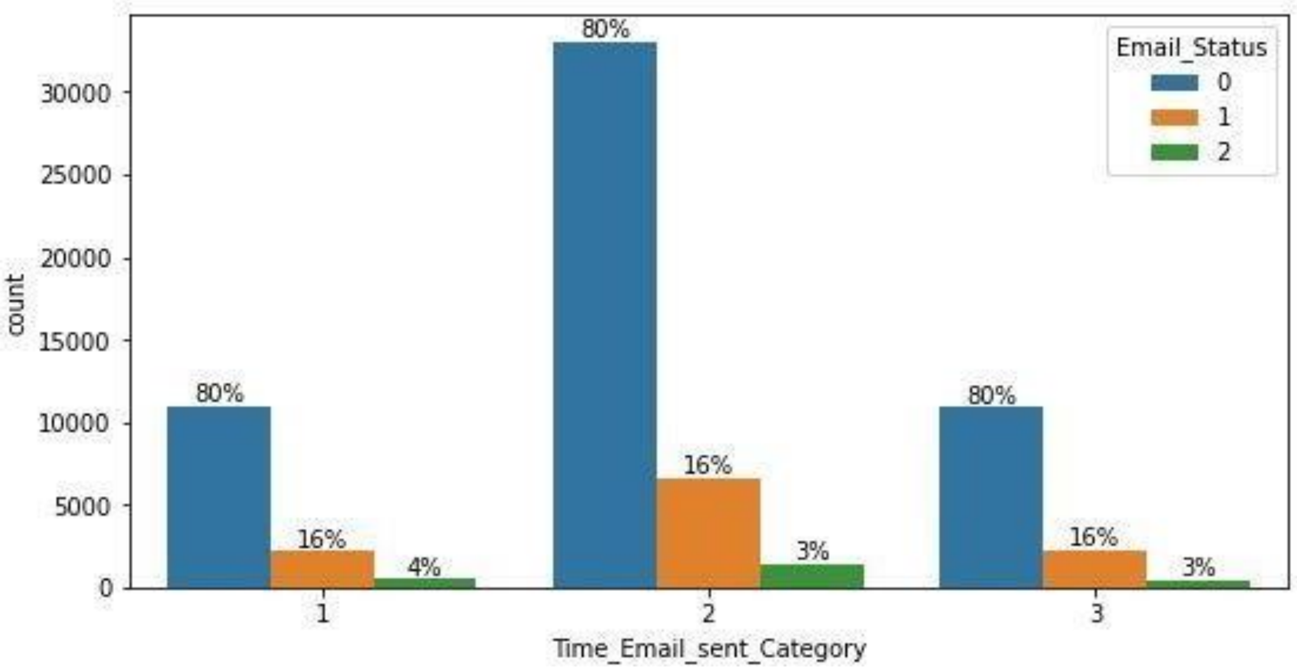
  90% of the time Email gets read or acknowledged if Campaign_Type is 1



Analysis of Categorical features

- **Time_Email_Sent_Catagory**

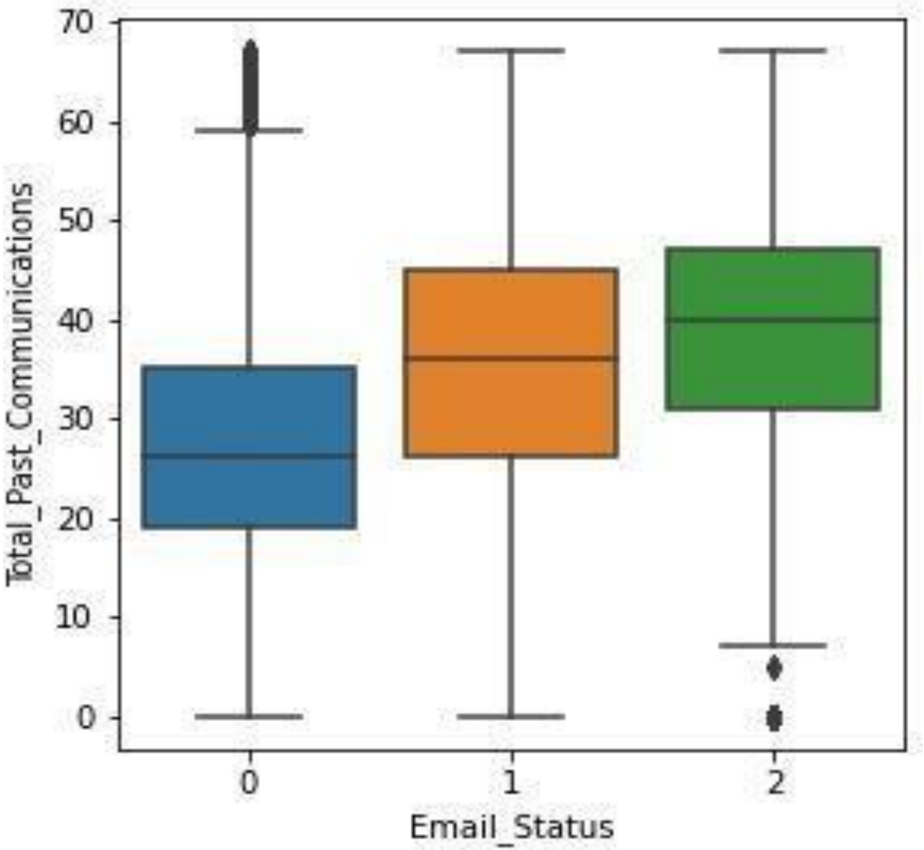
  Time Email Sent has no influence over Email_Status



Analysis of Continuous features

- **Total_Past_Communications**

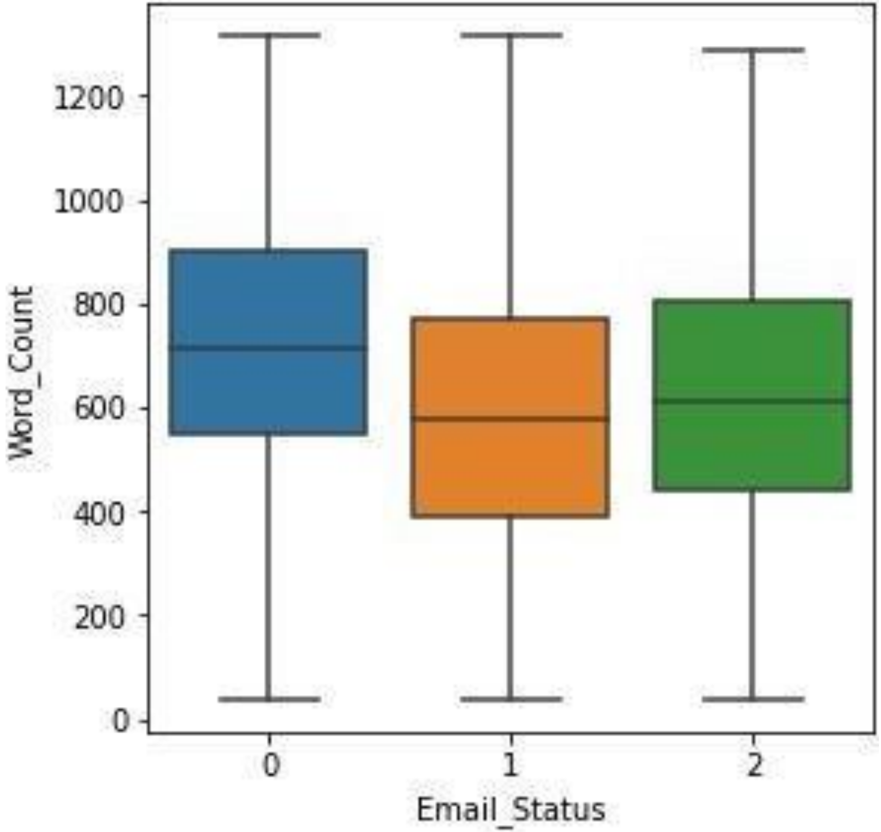
  As no. of past communication is increasing, Email is less ignored.



Analysis of Continuous features

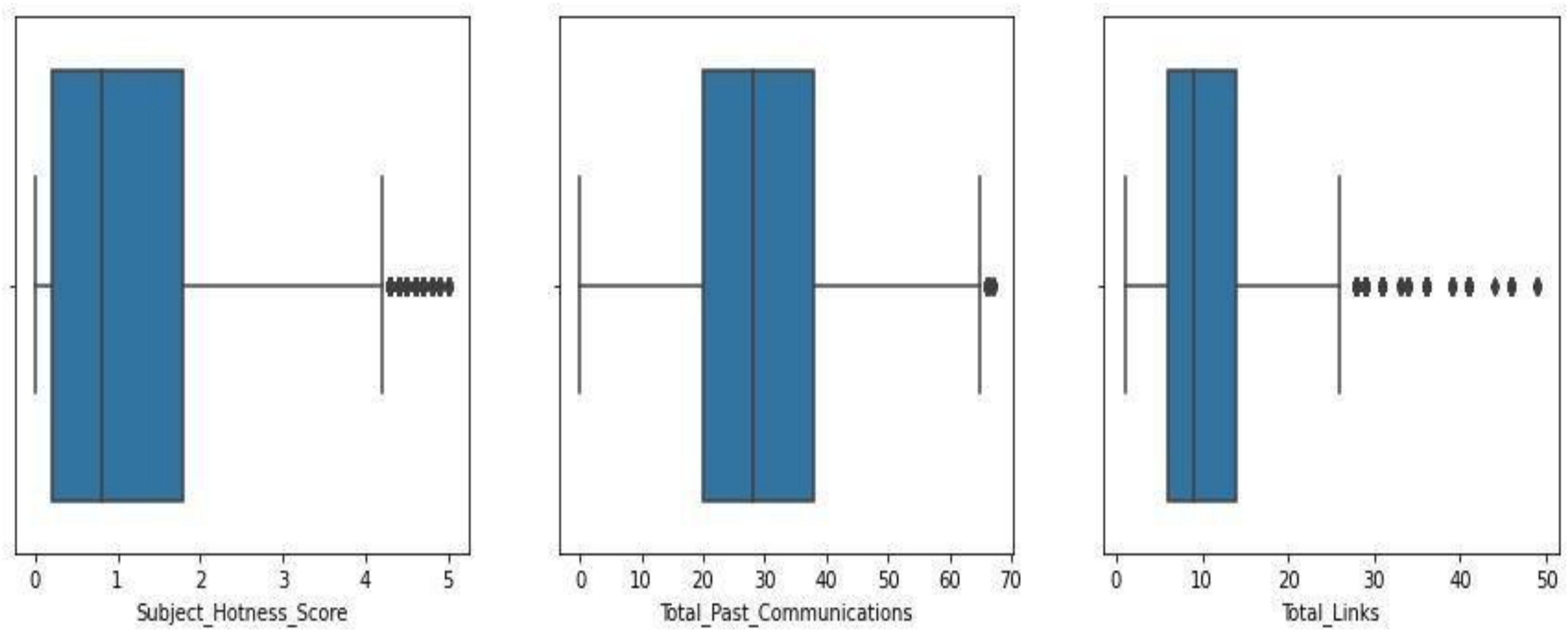
- **Word_Count**

No one is interested in reading Emails that are too long!!



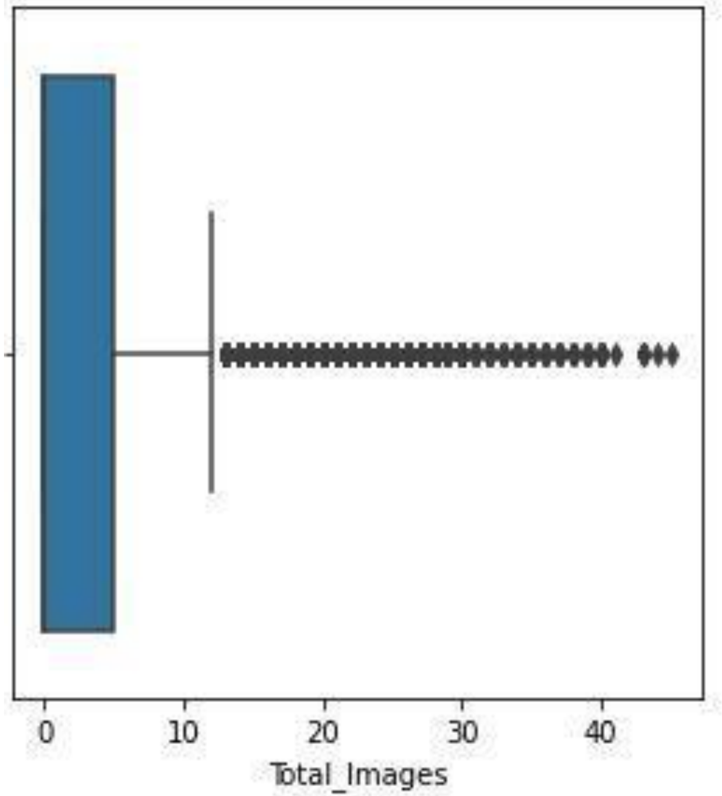
Analysis of Continuous features

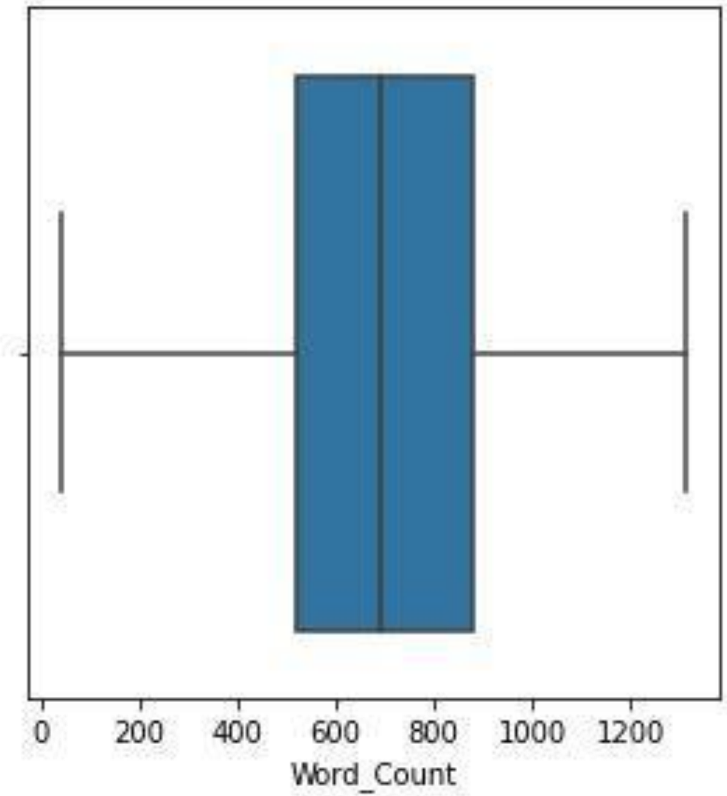
- **Outliers in different continuous features**



Analysis of Continuous features

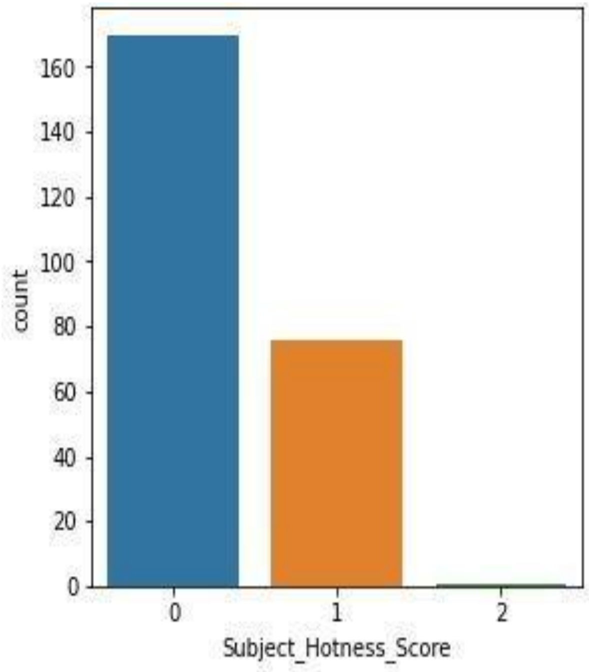
- **Outliers in different continuous features**

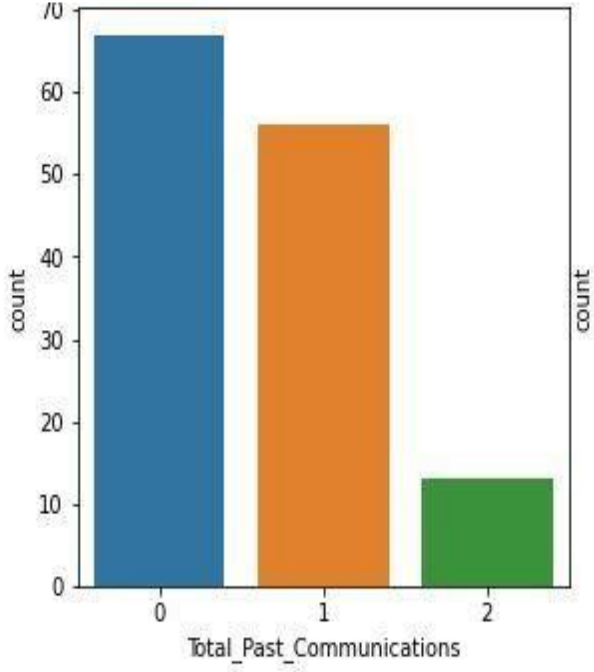


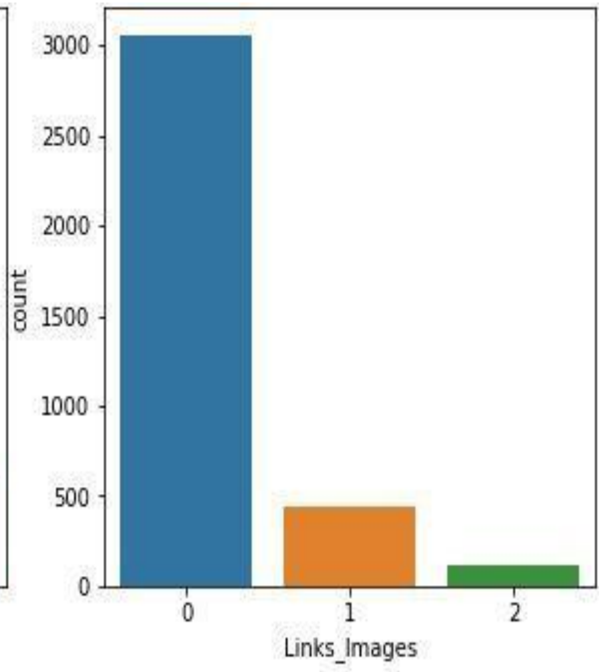


Outlier Treatment

- More than 5% of data in minority classes is outlier

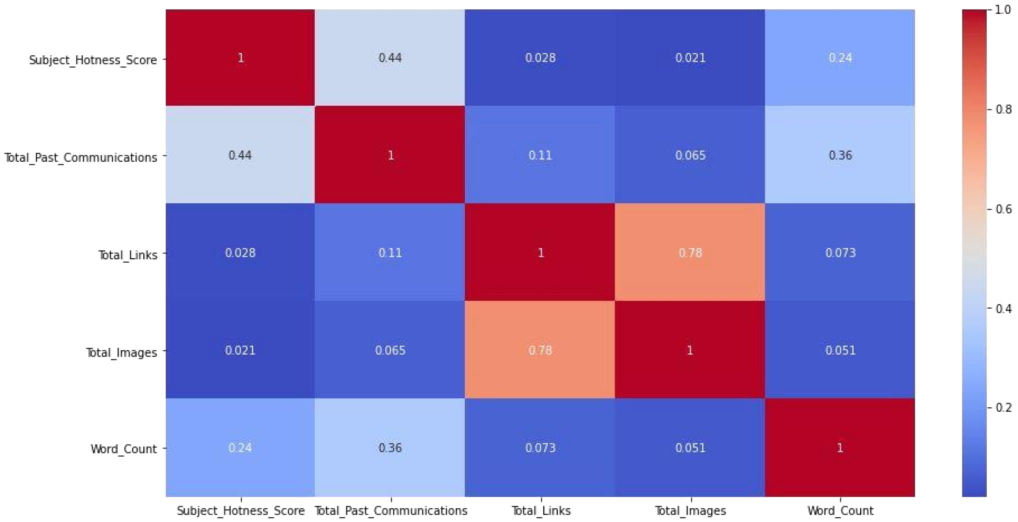


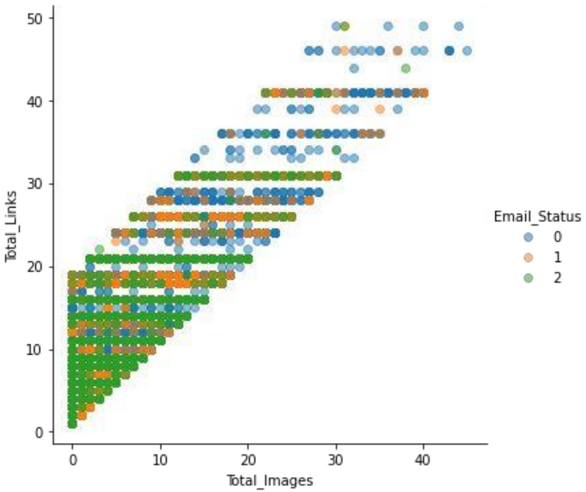




Feature Engineering

1. **Combining Total_Images and Total_Links:**





High **positive correlation** observed and hence **Links_Images = Total_Images + Total_Links**

Feature Engineering

2. Multicollinearity Check:

- Multicollinearity checked using **VIF Factor**

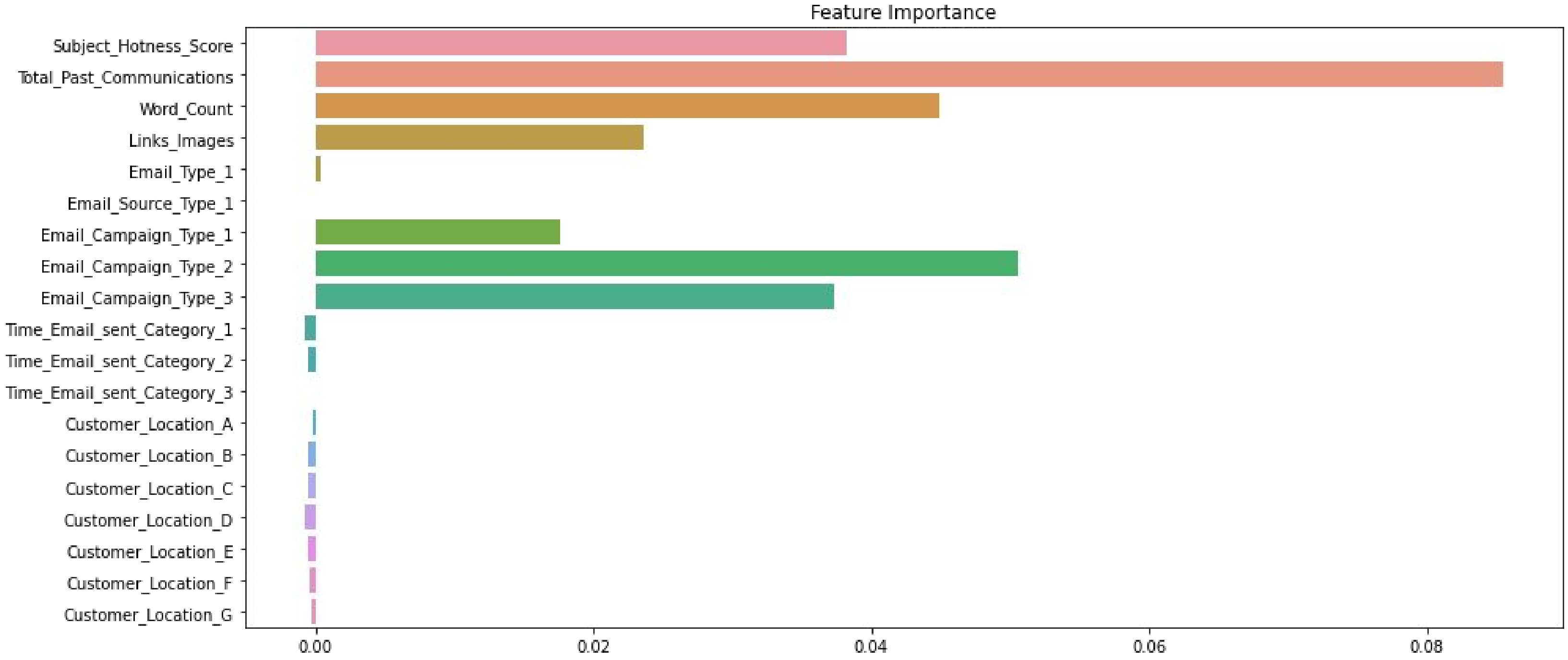
  Why ?
- Variables with high multicollinearity can adversely affect the model and removing highly correlated independent variables can help in reducing curse of dimensionality as well

- We can observe that all numerical variables are within the threshold(i.e. 5).

	variables	VIF
0	Subject_Hotness_Score	1.734531
1	Total_Past_Communications	3.430879
2	Word_Count	3.687067
3	Links_Images	2.629047

Feature Engineering

3. Understanding Feature Importance:



Feature Engineering

3. Understanding Feature Importance:

- The concept used to understand feature importance is **Information Gain**.

  Why?

- It explains which feature has maximum impact in classification based on the **notion of Entropy**.
- It works well for **numeric** as well as **categorical** data

- From the graph we understand that **Total_Past_Communications** and **Email_Campaign_Type** have **high importance**.
- **Time_Email_Sent_Category and Customer_Location are not important** and hence we decide to drop the feature.

Feature Engineering

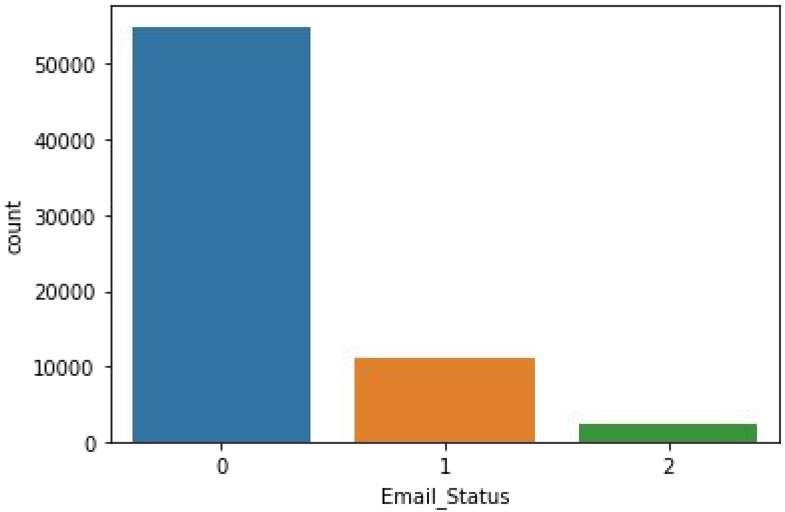
- **Numerical variables** were scaled using **MinMaxScaler**.

  Why?
  The numerical features of the dataset do not have a certain range and they differ from each other.

- **Categorical variables** were encoded using **One-Hot Encoding**.

  Why?
  This method changes categorical data to a numerical format and enables you to group your categorical data without losing any information.

Understanding Target Variable



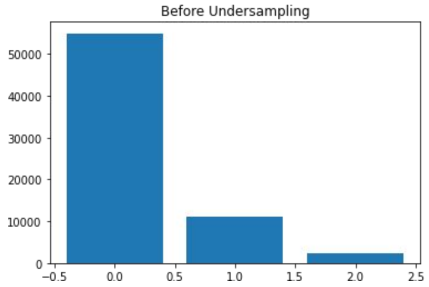
The target variable consists of 3 classes:

- 0 - ignored - 54941
- 1 - read - 11039
- 2 - acknowledged - 2373

Target Variable was **highly imbalanced**.

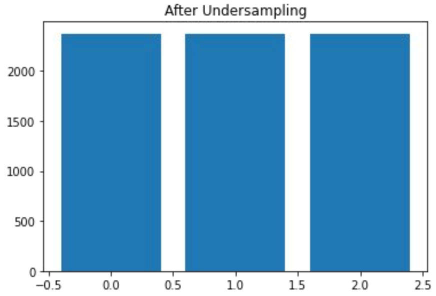
Handling Imbalanced data

1. **Undersampling Technique:**



- Technique used was **Random UnderSampler**
- Created balanced data with **2373** records for each class.

Why it didn’t work?

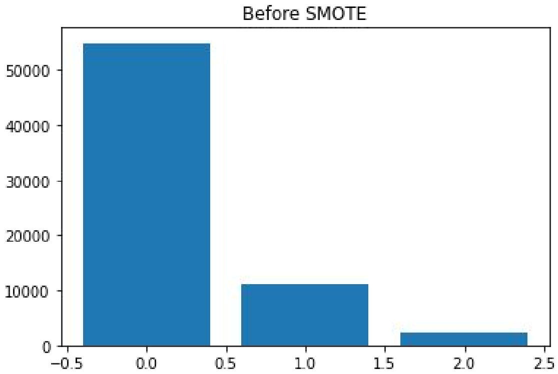


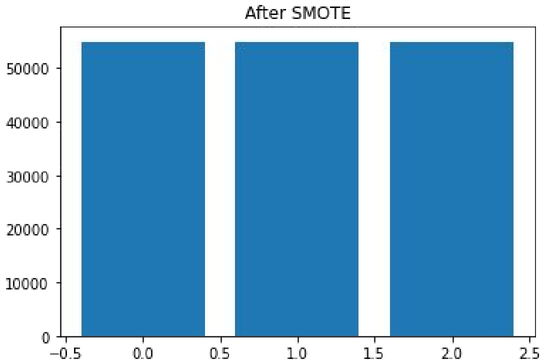
Created baseline models with undersampled data and it was observed that they underperformed primarily due to **loss of information.**

Handling Imbalanced data

2. Oversampling Technique:

- Technique used was **SMOTE**
- Created balanced data with **54941** records for each class.

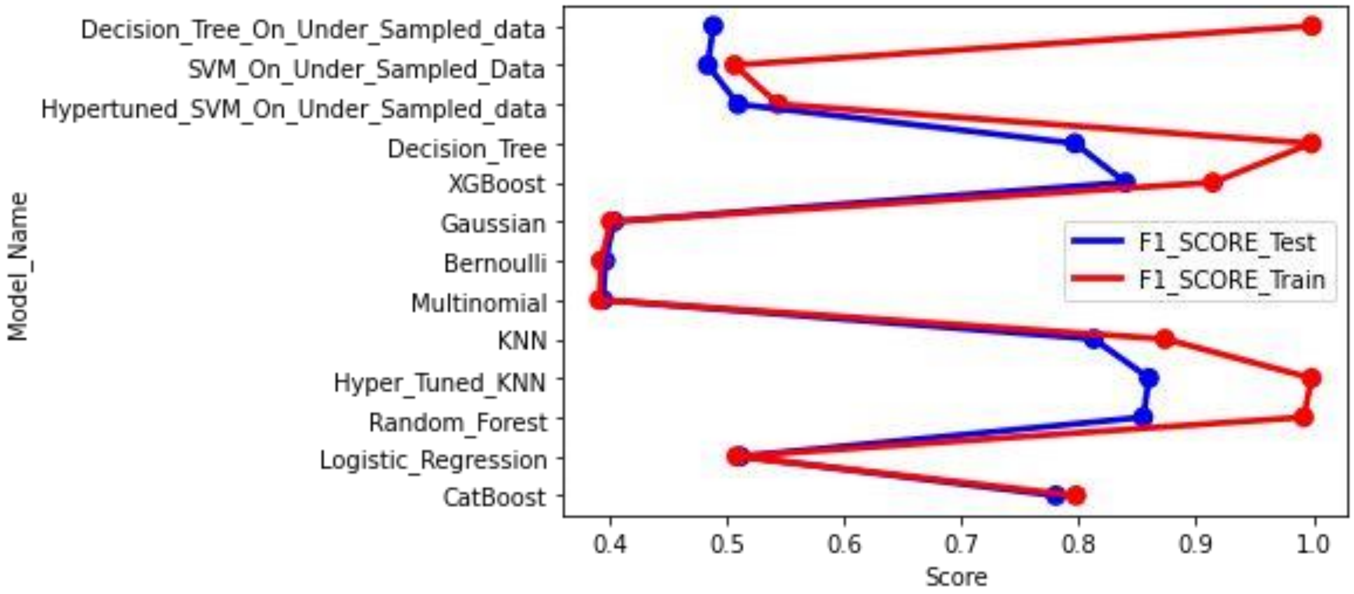




Different Models

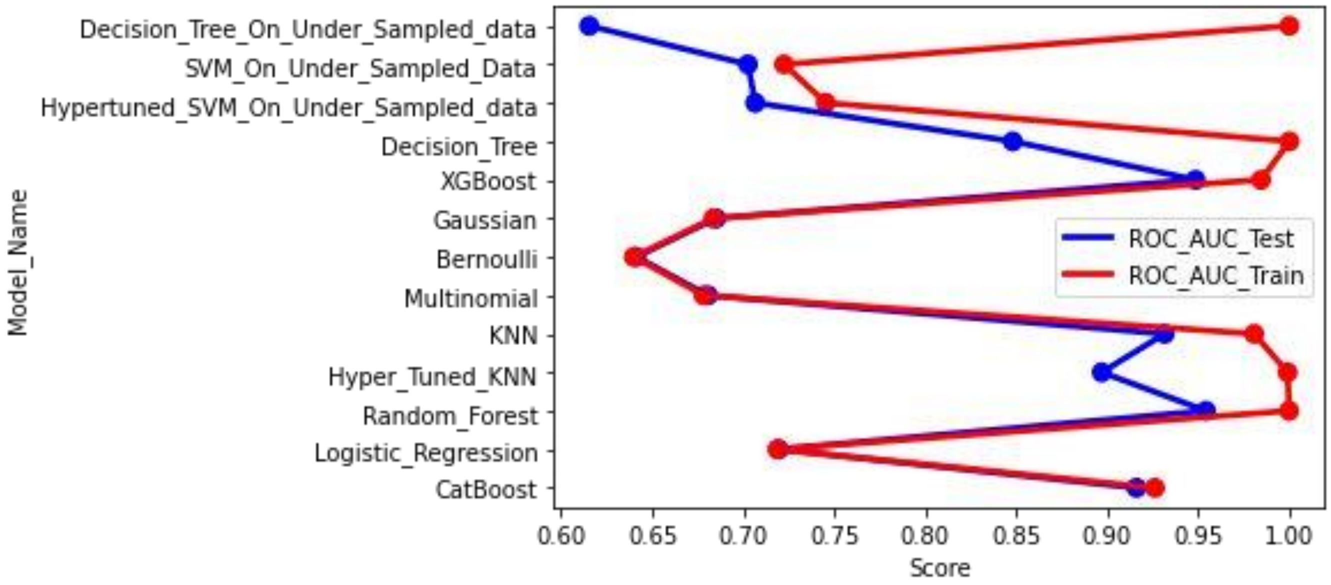
Evaluation Metrics:

1. **F1_Score**



Different Models

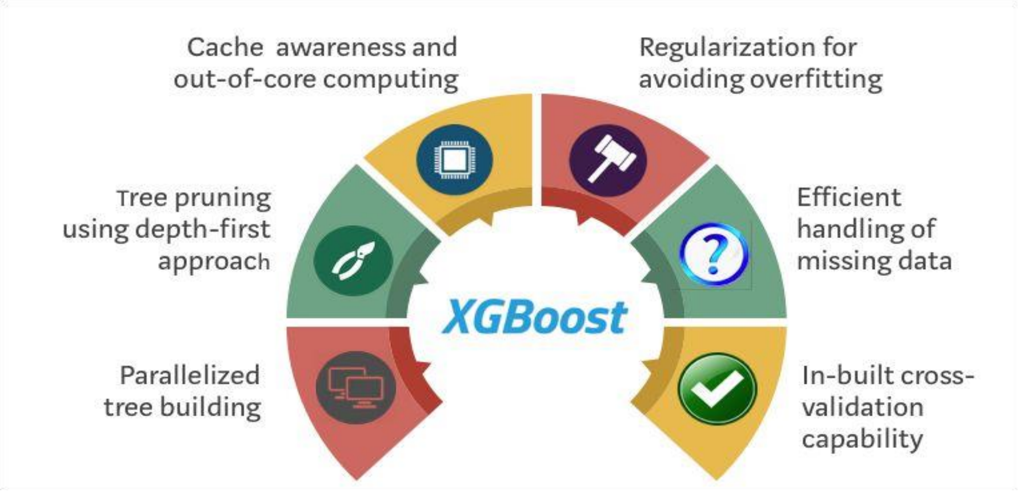
2. ROC_AUC_Score



Winner Model

XGBoost:

- Robust to outliers.
- Supports regularization.
- Works well on small to medium dataset.
- F1 score for train & test set were 89% & 81% respectively



Conclusion

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

- It was 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 location or the time frame the emails were sent on.

- 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 !

- For modelling, it was observed that for imbalance handling Oversampling i.e. SMOTE worked way better than undersampling as the latter resulted in a lot of loss of information.

- Based on the metrics, XGBoost Classifier worked the best giving a train score of 89% and test score of 81% for F1 score.

Challenges

- Choosing the appropriate technique to handle the imbalance in data was quite challenging as it was a tradeoff b/w information loss vs risk of overfitting.

- Overfitting was another major challenge during the modelling process.

- Understanding what features are most important and what features to avoid was a difficult task.

- Decision making on missing value imputations and outlier treatment was quite challenging as well.

Thank You
Q & A