Capstone Project

Email Campaign Effectiveness Prediction

Individual Project

Name : Aakash Ramnani

Content

- Problem Statement
- Data Summary
- Project Summary
- EDA (Exploratory Data Analysis)
- Hypothesis Testing
- Feature Engineering and Data Pre-processing
- Model Implementation
  1. Logistic Regression
  2. Decision Trees
  3. Random Forest
  4. KNN Classifier
  5. XGBoost Classifier
- Hyperparameter Tuning
- Model Performance and Evaluation
- Conclusion

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

- **Email Id -** It contains the email id's of the customers/individuals
- **Email Type -** There are two categories 1 and 2. We can think of them as marketing emails or important updates, notices like emails regarding the business.
- **Subject Hotness Score -** It is the email's subject's score on the basis of how good and effective the content is.
- **Email Source -** It represents the source of the email like sales and marketing or important admin mails related to the product.
- **Email Campaign Type -** The campaign type of the email.
- **Total Past Communications -** This column contains the total previous mails from the same source, the number of communications had.
- **Customer Location -** Contains demographical data of the customer, the location where the customer resides.
- **Time Email sent Category -** It has three categories 1,2 and 3; the time of the day when the email was sent, we can think of it as morning, evening and night time slots.

- **Word Count -** The number of words contained in the email.
- **Total links -** Number of links in the email.
- **Total Images -** Number of images in the email.
- **Email Status -** Our target variable which contains whether the mail was ignored, read, acknowledged by the reader.

Project Summary

- In this problem statement, we will be trying to create machine learning models that characterize and predict whether the mail is ignored, read or acknowledged by the reader. In addition to this, we will be trying to analyze and find all the features that are important for an email to not get ignored.
- The main steps of the project are:

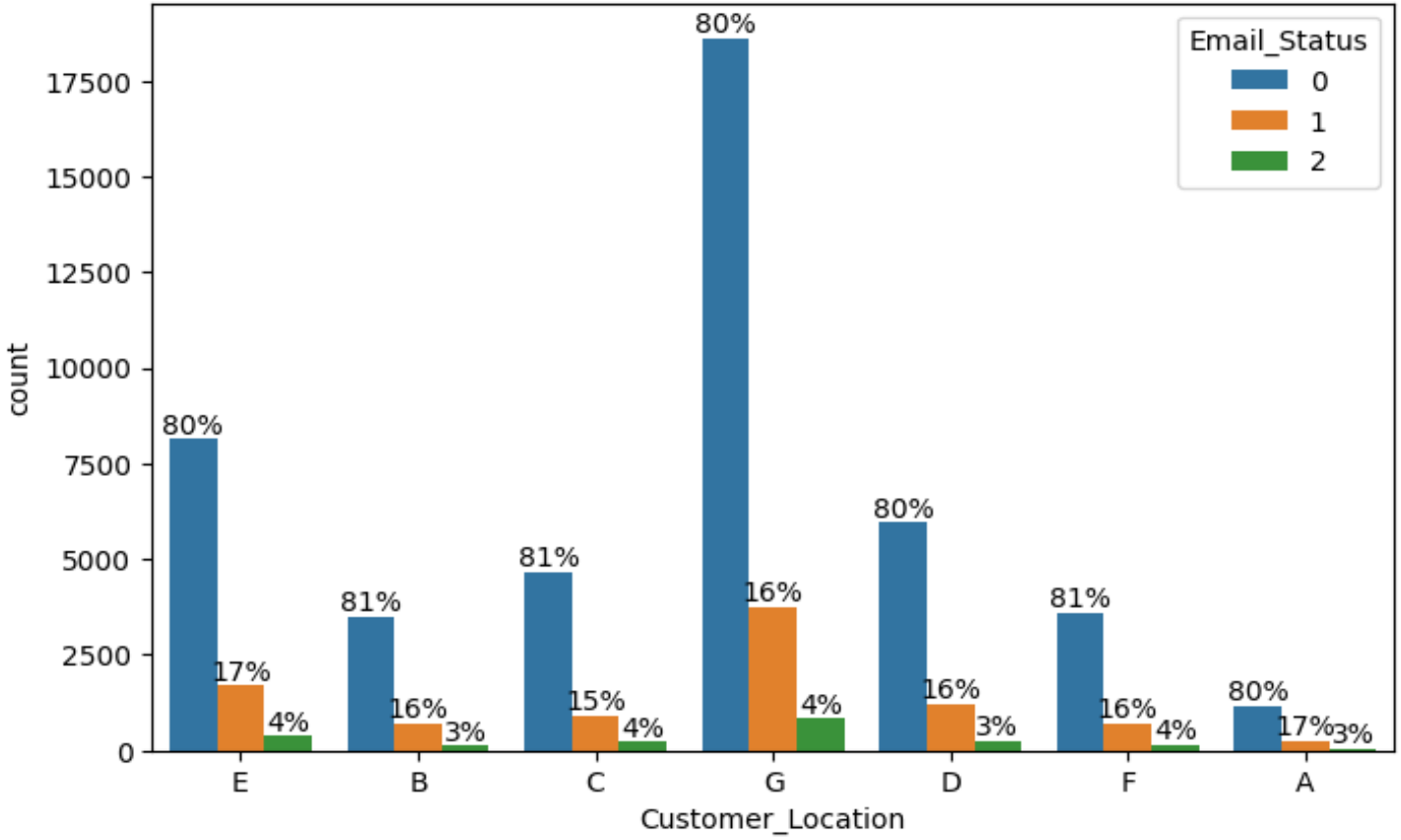
1. **Basic EDA(Exploratory Data Analysis):** I have performed basic EDA to understand the data and its characteristics. Also, have used Univariate - BI variate - and Multivarite analysis to Understanding the correlation between variables and to explore distribution and intercation of variables.
2. **Hypothesis Testing:** I have performed hypothesis testing to test the relationship between the variables. Also have used statistical tests such as t-test, ANOVA, f-test to compare the means or proportions of different groups or categories.
3. **Feature Engineering and Data Preprocessing:** To create and select the most relevant and informative features for the model. Have used methods such as correlation analysis and VIF(Variance Inflation Factor) for feature selection.
4. **ML Model Development and Evaluation:** In this project Have developed and evaluated different machine learning Models. I have used logistic regression, KNN Classifier, Decision Tree, Random Forest and XGBoost.
5. **Model Interpretation and Explanation:** Explained the model predictions using feature importance. It helps to analysis the effect of different factors on the output variable.

EDA(Exploratory Data Analysis)

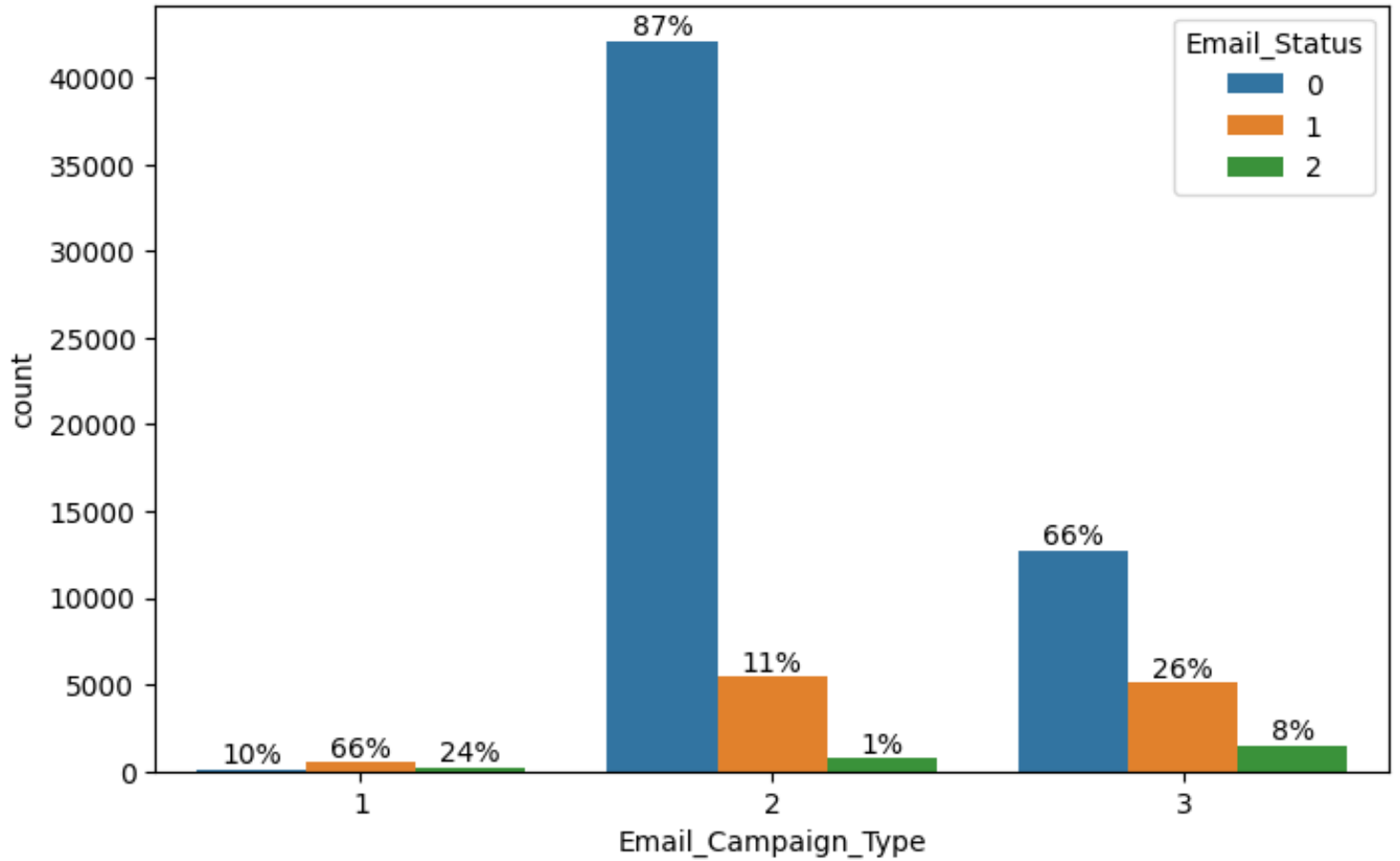
- We did a detailed exploratory data analysis of the categorical and continuous variable against their target variable.

Categorical Data

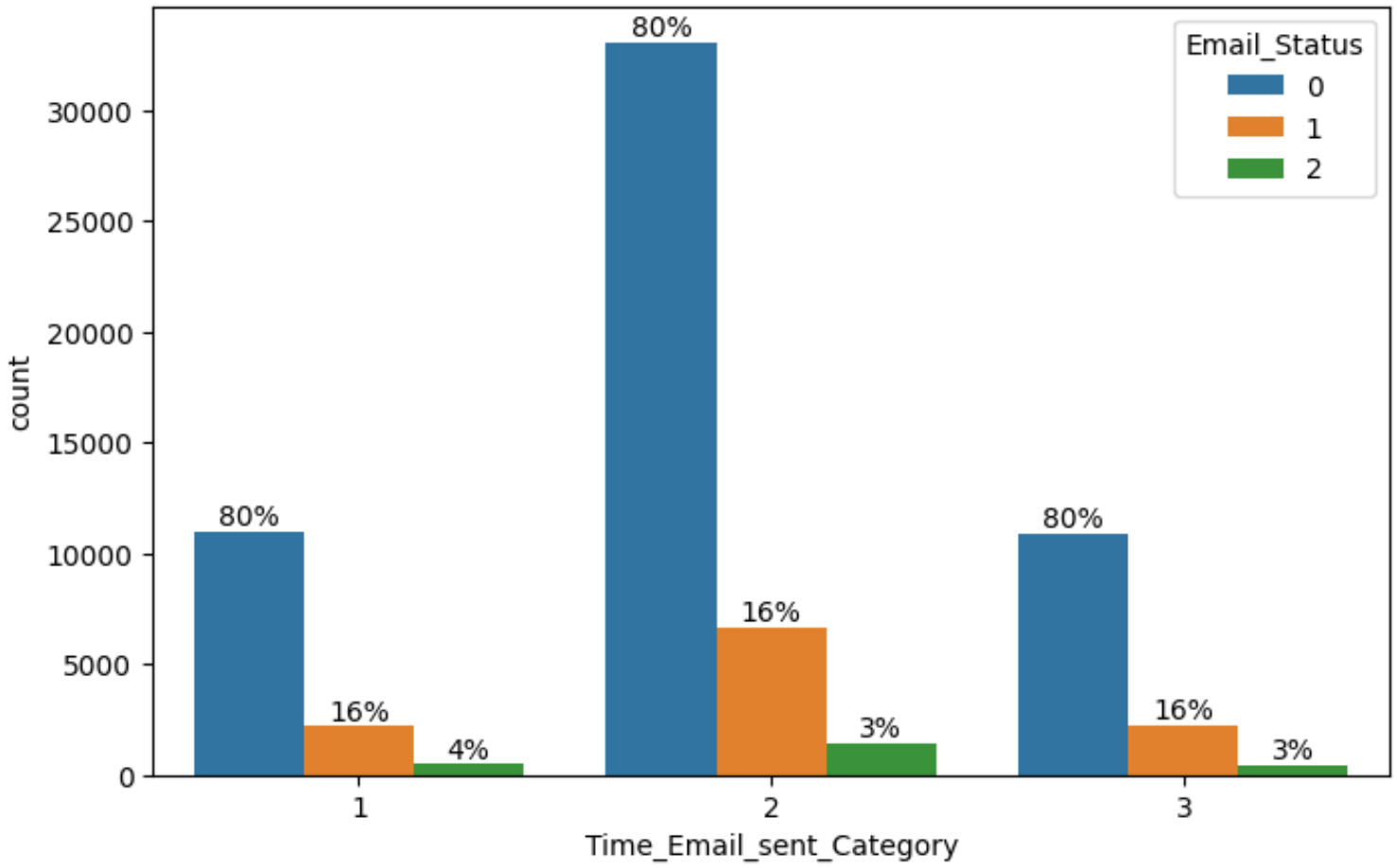
- In Customer location, irrespective of location, the percentage of emails being ignored, read and acknowledged is kind of similar. It does not exclusively influence our target variable.



- Emails sent through campaign type 1 are very less in number. However the rate at which the emails are read and acknowledged is better than campaign type 2 and campaign type 3.
- Campaign type 3 has good results, with less number of emails sent but comparatively more number of emails out of sent emails were read and acknowledged.

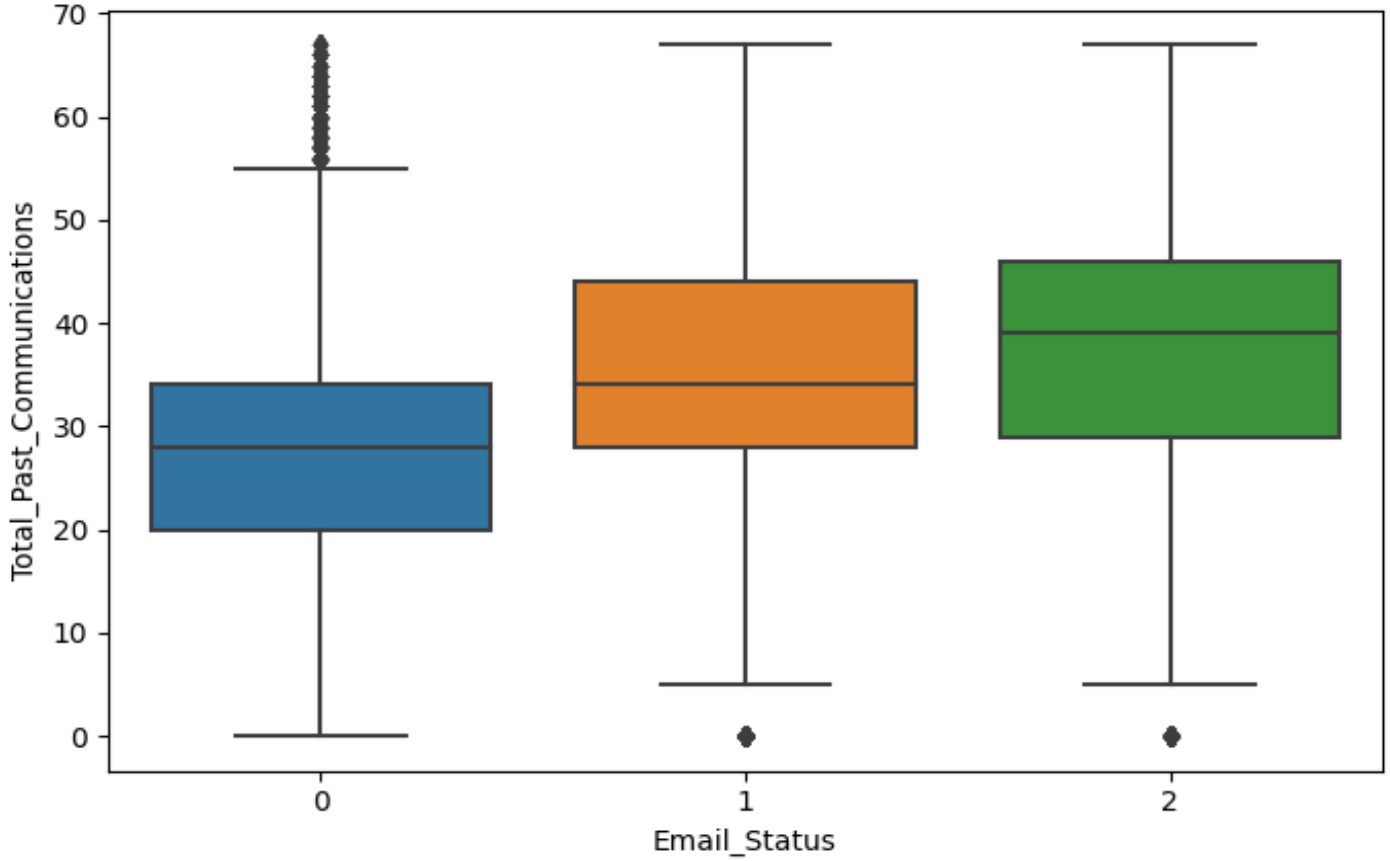


- Time email sent feature cannot be considered relevant in classifying the emails.
- It can also be seen that most number of emails are sent in Time Email Sent Category 2(afternoon). Less number of emails are sent in Morning and night.

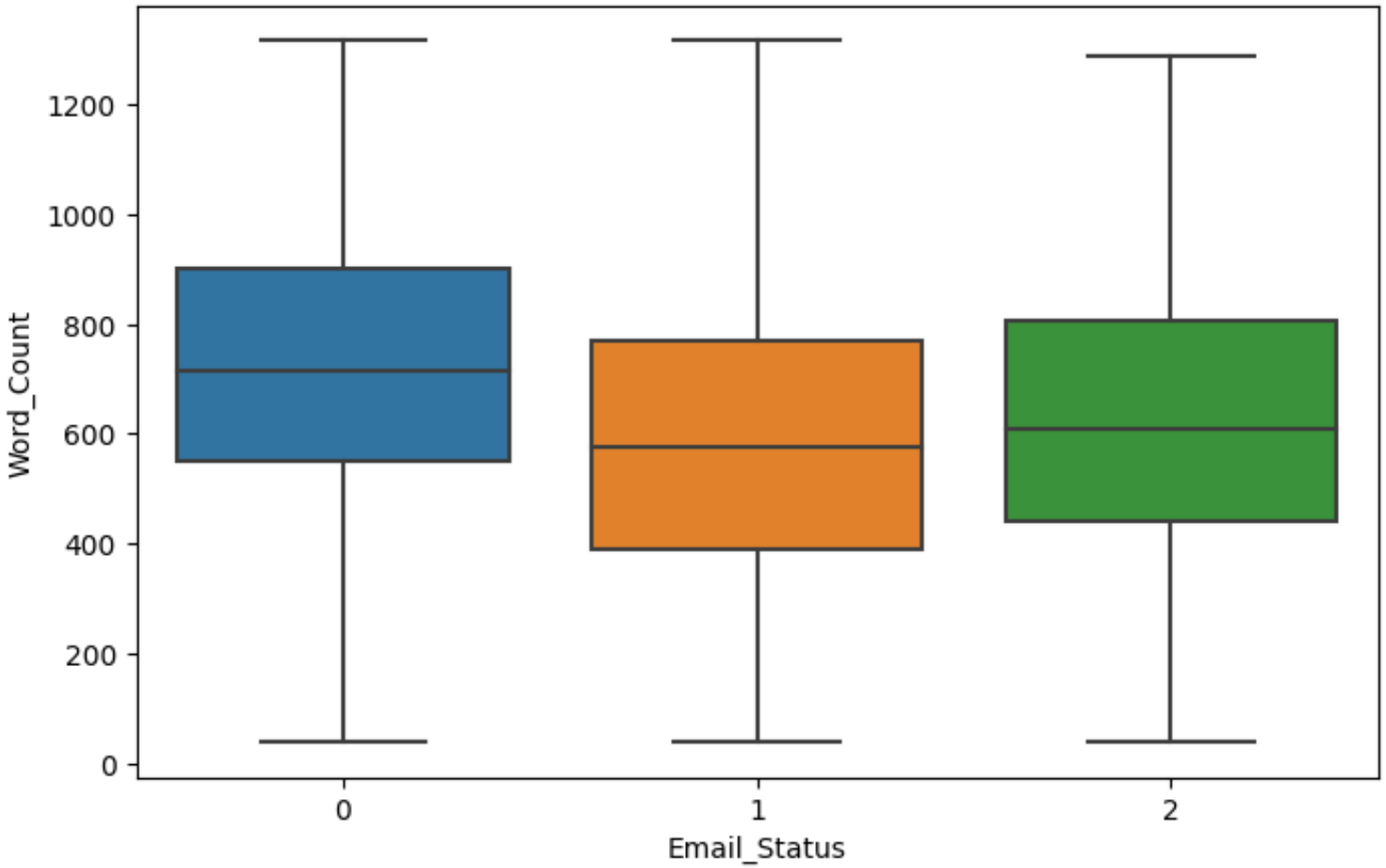


Continuous Data

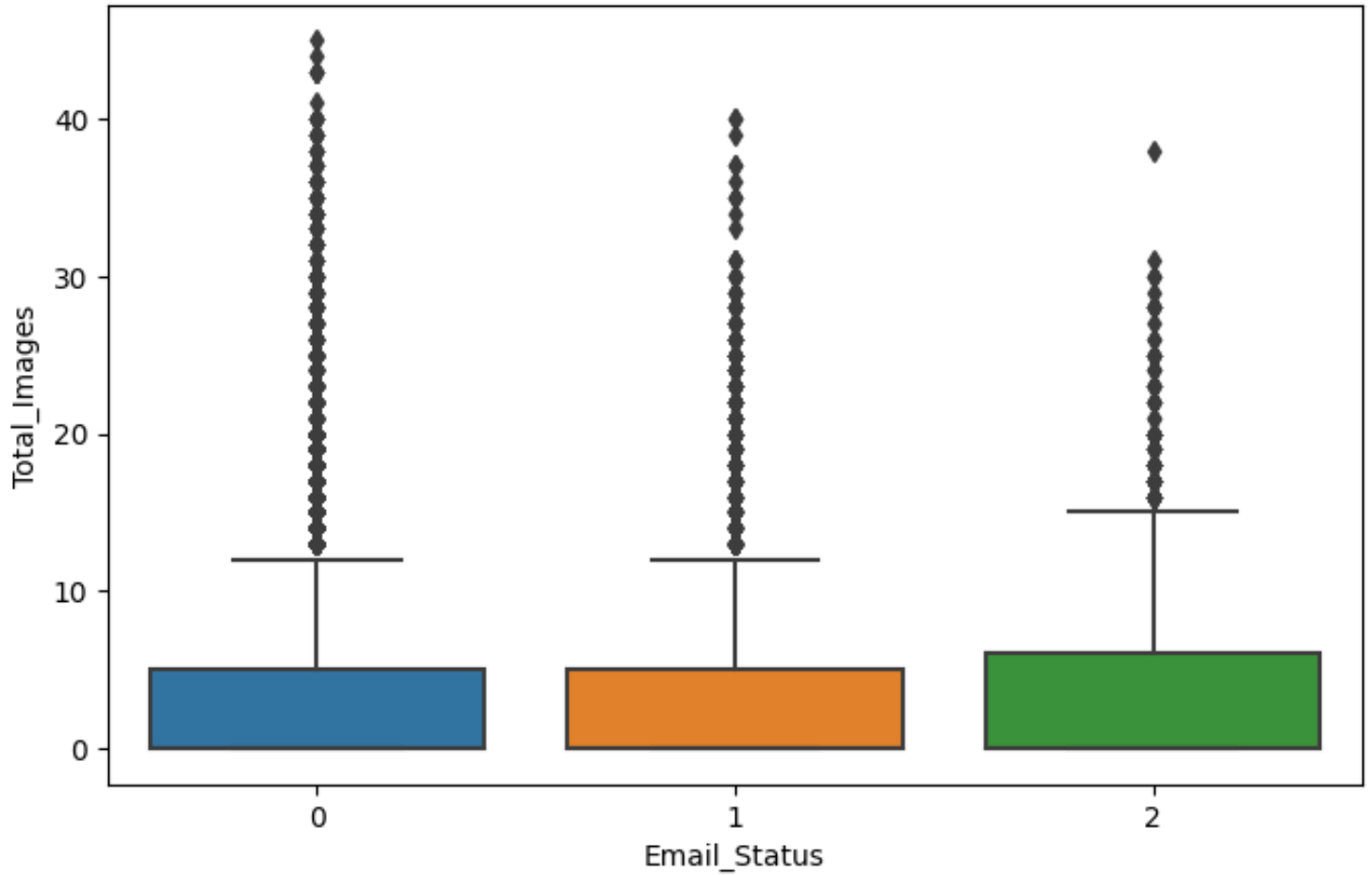
- From EDA it was very evident from the plot that more the number of previous interactions with customer, more the customer tends to read and acknowledge the mail. It is important to build a good rapport and connection with customer.



- E-mails with high number of words are ignored more than read or acknowledged.
- Lengthy E-mails should be avoided.

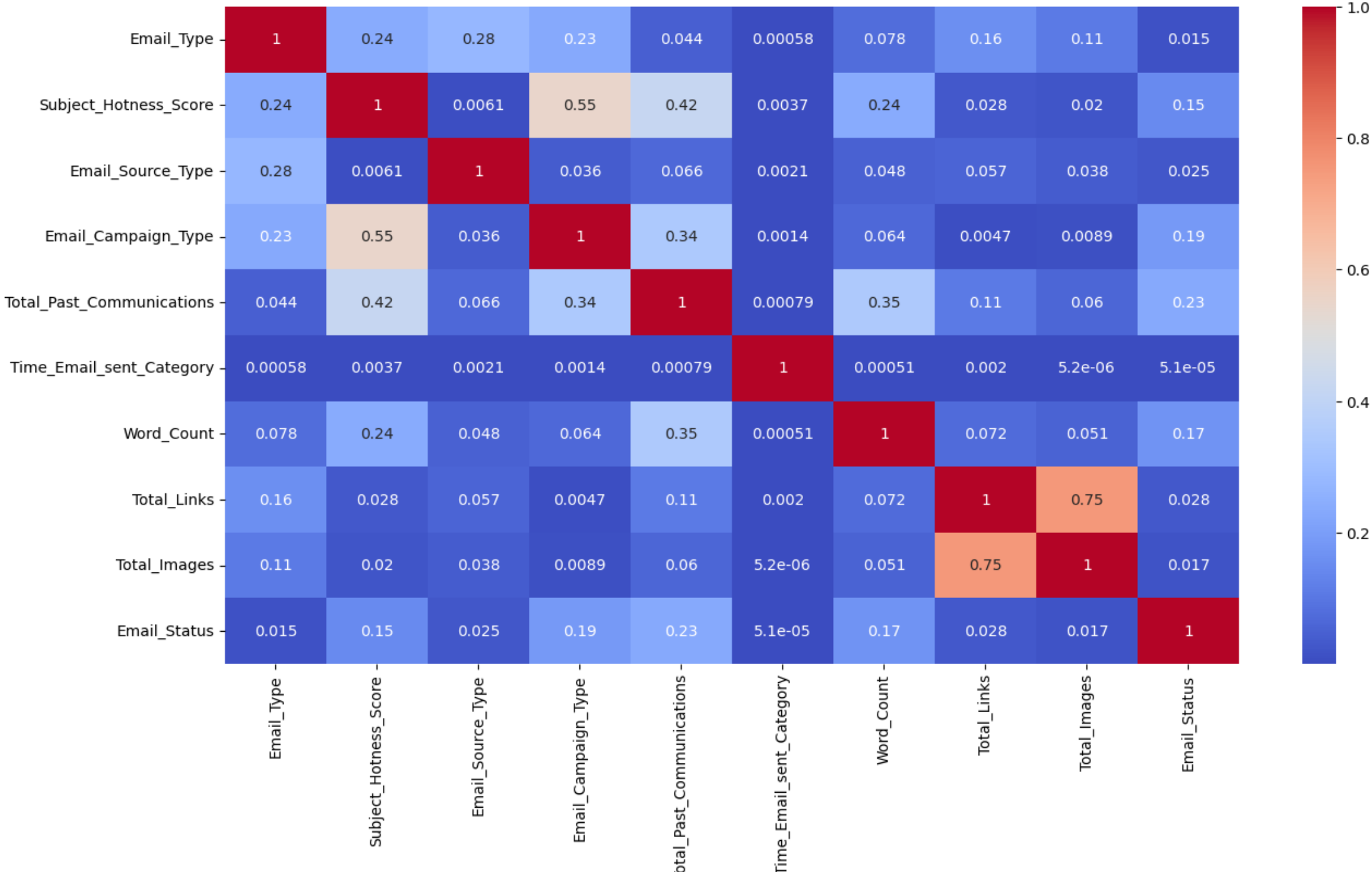


- High number of outliers in Total_Images value 0 suggest that there are more images in ignored emails.



Correlation Heatmap

- Correlation is a statistical term used to measure the degree to which two variables are related.
- Correlation coefficient being greater than 0 means the variable show positive correlation, for example if one variable increase the value of other will also increase.
- Correlation coefficient being less than 0 means the variables show negative correlation, for example if one variable increases the value of other will decrease.
- Zero correlation implies no linear relationship.
- Email Campaign type and total past communications show positive correlation with emails being read and acknowledged. Total Images and Total Links show highest positive correlation. Email Campaign type and Total Past Communications show positive correlation.

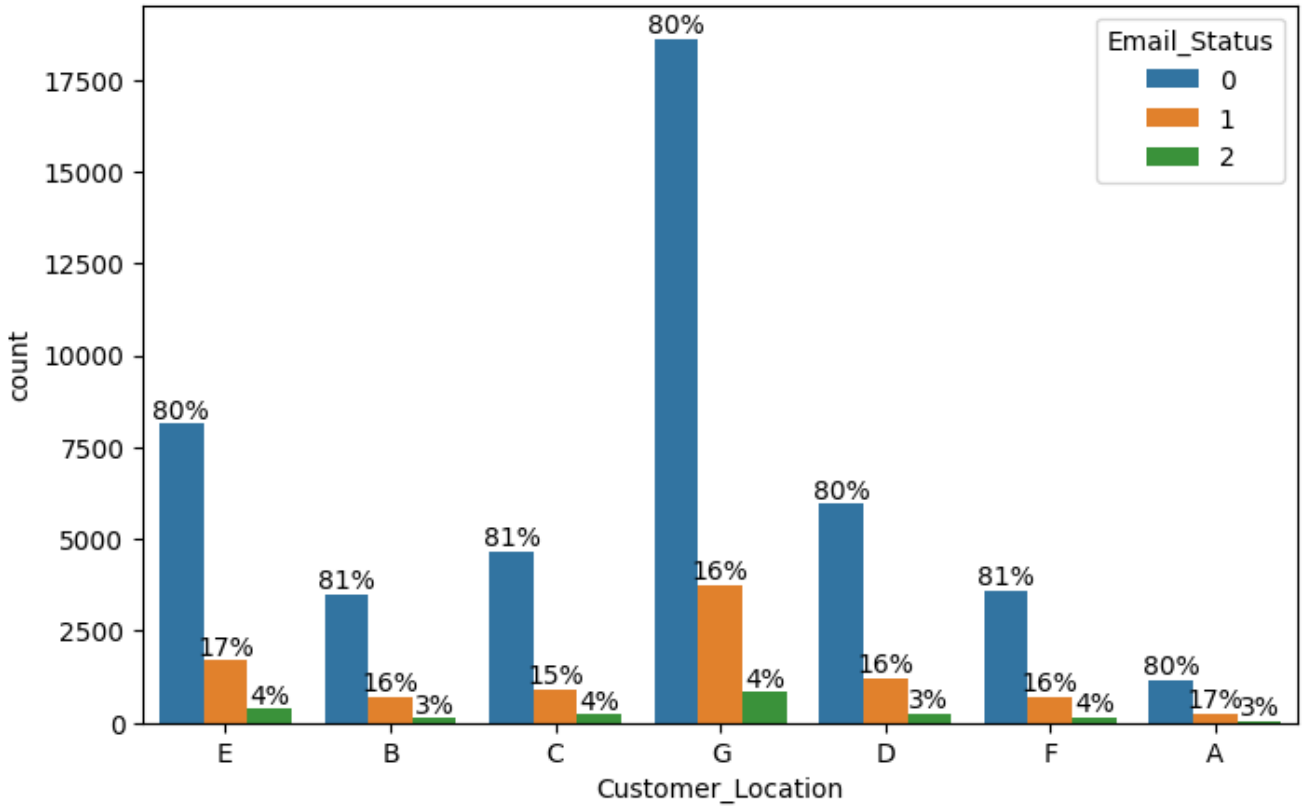

Email_Type 1 0.24 0.28 0.23 0.044 0.00058 0.078 0.16 0.11 0.015
Subject_Hotness_Score 0.24 1 0.0061 0.55 0.42 0.0037 0.24 0.028 0.02 0.15
Email_Source_Type 0.28 0.0061 1 0.036 0.066 0.0021 0.048 0.057 0.038 0.025
Email_Campaign_Type 0.23 0.55 0.036 1 0.34 0.0014 0.064 0.0047 0.0089 0.19
Total_Past_Communications 0.044 0.42 0.066 0.34 1 0.00079 0.35 0.11 0.06 0.23
Time_Email_sent_Category 0.00058 0.0037 0.0021 0.0014 0.00079 1 0.00051 0.002 5.2e-06 5.1e-05
Word_Count 0.078 0.24 0.048 0.064 0.35 0.00051 1 0.072 0.051 0.17
Total_Links 0.16 0.028 0.057 0.0047 0.11 0.002 0.072 1 0.75 0.028
Total_Images 0.11 0.02 0.038 0.0089 0.06 5.2e-06 0.051 0.75 1 0.017
Email_Status 0.015 0.15 0.025 0.19 0.23 5.1e-05 0.17 0.028 0.017 1
Email_Type
Subject_Hotness_Score
Email_Source_Type
Email_Campaign_Type
Total_Past_Communications
Time_Email_sent_Category
Word_Count
Total_Links
Total_Images
Email_Status
1.0
0.8
0.6
0.4
0.2

Hypothesis Testing

- Based on our EDA, we have defined these three hypothesis.

    - The proportion of ignored, read and acknowledged emails is almost same for all the customer location.
    - There is no correlation between Total_Links and Total_Images.
    - Total_Past_Communication has no effect on Email_Status.

Hypothesis - 1



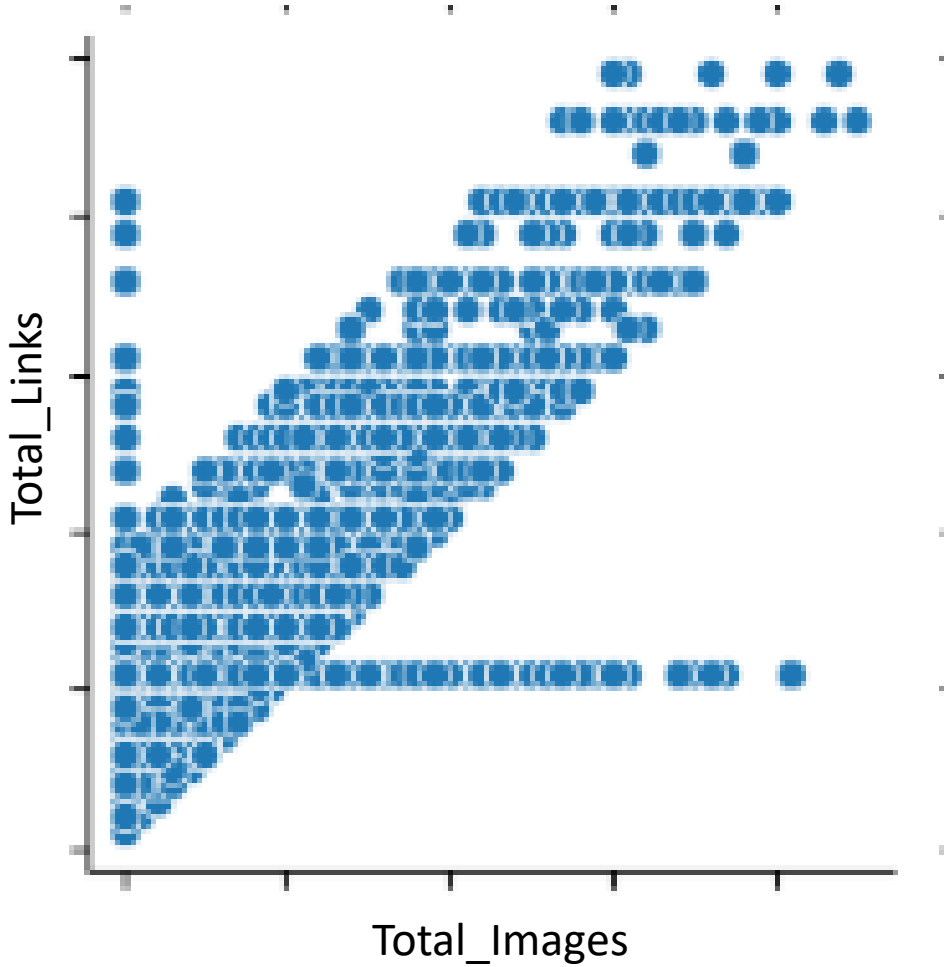
- **Null Hypothesis (H0) -** The proportion of ignored, read and acknowledge emails is almost same for all the customer locations.

- **Alternate Hypothesis (HA) -** Their is a huge difference in proportion of ignored, read and acknowledged emails for all the customer location.

- As we have two categorical features to compare we have used chi-square test with significance level 0.05 for these hypothesis.

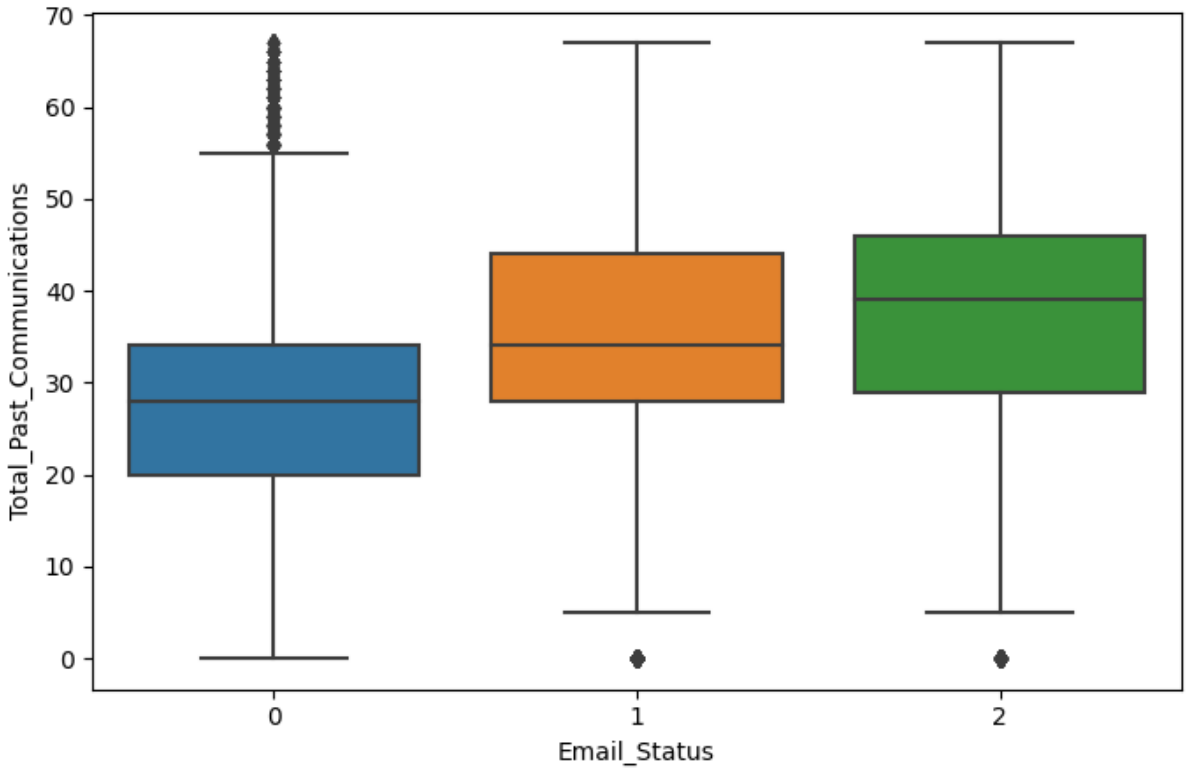
- After performing the test p-value we obtained is 0.52 which is greater than significance level 0.05, so we can not reject the null hypothesis.

Hypothesis - 2



- **Null Hypothesis (H0) -** There is no correlation between Total_Links and Total_Images.
- **Alternate Hypothesis (HA) -** Total_Links and Total_Images are highly correlated.
- As we have two quantitative variables and we need to find the correlation we have used pearson correlation test.
- After the test, Value for r (correlation coefficient) is 0.75 which is greater than 0.5 which indicates a high correlation between two variable.
- Value of p is 0 which less than significance level 0.05, so we will reject the null hypothesis.

Hypothesis - 3



- **Null Hypothesis (H0) -** Total_Past_Communication has no significant effect on Email_Status.
- **Alternate Hypothesis (HA) -** Total_Past_Communication has significant effect on Email_Status.
- We need to test for one categorical variable and one quantitative variable we have used one anova test with significance level 0.05.
- Value of p is 0 which less than significance level 0.05, so we will reject the null hypothesis.

Feature Engineering and Data Pre-Processing

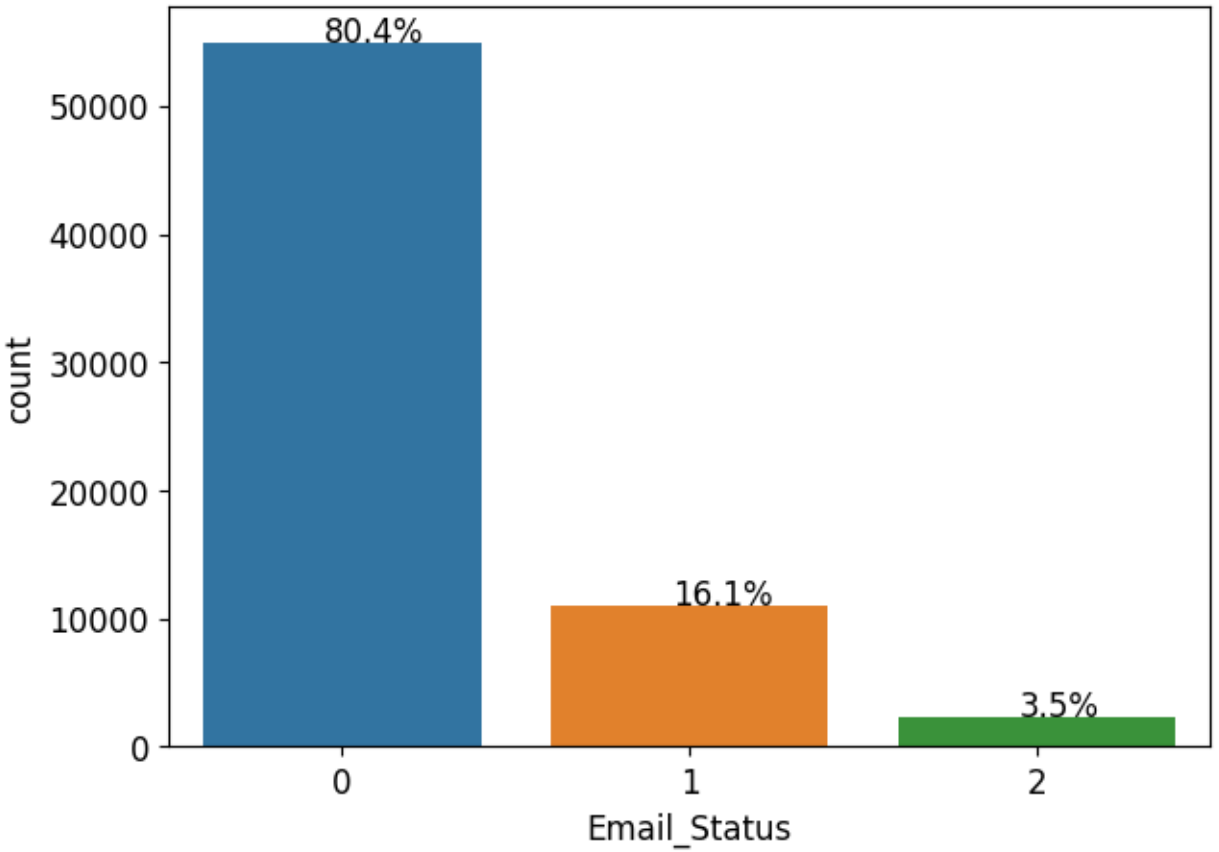
- **Handling Missing Values :**
  - There are missing values in Customer_Location, Total_Past_Communications, Total_Links and Total_Images. Max number of missing values are in Customer_Location.
  - As it can be seen that distribution for Total_Past_Communications is almost normal. So it will be right to impute the missing the values with mean.
  - It is clear from the plot that distribution for both Total_Links and Total_Images is right skewed. So it would be right to impute the missing values with mode.
  - As Customer_Location has no direct effect on output variable and there is no way we can guess customer location to impute the missing value so we dropped the missing values.

- **Handling Outliers :**
  - We have the number of outliers with respect to Email_Status.
  - Our output data is imbalanced, Email_Status value 1 and value 2 are in minority class.
  - For our model to predict coreclty and not be biased to one class, we need to take care that we are not deleting more than 5% of useful data related to minority class.
  - There are more than 5% outliers for minority class so we will not be deleting them.
  - Deleting the outliers for majority class.

- **Categorical Encoding :**
  - We have used one hot encoding as the categorical data was nominal and not ordinal.

- **Feature Selection :**
  - We have used VIF to check what all continuous features follow multicollinearity and select the feature accordingly.
  - We created new feature Total_links_images from Total_Images and Total_Links.
  - We do not need Email_ID and Customer Location Column for prediction, so we dropped them.

Handling Imbalanced Data



- In EDA we clearly saw that number of ignored emails are very high in compare to Read and Acknowledged emails.
- This imbalance in class can lead to biased classification towards ignored emails.
- We have used SMOTE technique to treat imbalanced data.
- SMOTE generates synthetic data for minority class.
- 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.
- SMOTE will save us from the problem of loss of information unlike other techniques like under sampling.

Model Implementation

- We have Tried implementing the following models on our dataset :

    - Logistic Regression
    - Decision Tree
    - Random Forest
    - KNN Classifier
    - XGBoost Classifier

Logistic Regression

- Logistic Regression is a classification algorithm that Predicts the probability of an outcome that can only have two values.
- Multinomial Logistic Regression is an extension of logistic regression that adds native support for multi-class classification problems.
- After evaluating we got the following results.

	Model_Name	Train_Accuracy	Train_Recall	Train_Precision	Train_F1score	Train_AUC	Test_Accuracy	Test_Recall	Test_Precision	Test_F1score	Test_AUC
0	Logistic Regression	0.536502	0.536502	0.521232	0.511283	0.72392	0.537942	0.537942	0.521233	0.51171	0.725308

- As our test set was highly imbalanced we can not trust accuracy with this, but seeing F1 score and AUC ROC we can see the results were not that good.

Decision Tree

- Decision Tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems.
- Decision Tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree.
- After evaluating we got the following results.

	Model_Name	Train_Accuracy	Train_Recall	Train_Precision	Train_F1score	Train_AUC	Test_Accuracy	Test_Recall	Test_Precision	Test_F1score	Test_AUC
0	Logistic Regression	0.536502	0.536502	0.521232	0.511283	0.72392	0.537942	0.537942	0.521233	0.511710	0.725308
1	Decision Tree	0.999439	0.999439	0.999440	0.999439	1.00000	0.780824	0.780824	0.780975	0.780715	0.835749

- Clearly Decision Tree model was overfitting. Showing great results on train data and not working well on test data.

Random Forest

- Random Forest is an ensemble learning method for classification and regression that operates by constructing a multitude of decision trees at training time.
- For Classification task the output of the random forest is the class selected by most trees.
- To prevent overfitting and get better results we also trained a hyperparameter tunned model.
- After evaluating we got the following result.

2	Random Forest	0.568954	0.568954	0.555191	0.548399	0.758494	0.569902	0.569902	0.555327	0.548952	0.758297
3	Random Forest tunned	0.892869	0.892869	0.892978	0.892028	0.979517	0.810884	0.810884	0.809475	0.808529	0.940415

- Hyperparameter tuned model is showing better results.

KNN Classifier

- The abbreviation KNN stands for “K-Nearest Neighbour”. It is a supervised machine learning algorithm. The algorithm can be used to solve both classification and regression problem statements.
- Nearest neighbor classification is a machine learning method that aims at labeling previously unseen query objects while distinguishing two or more destination classes.
- To get better results we also trained a hyperparameter tunned model.
- After evaluating we got the following result.

	Model_Name	Train_Accuracy	Train_Recall	Train_Precision	Train_F1score	Train_AUC	Test_Accuracy	Test_Recall	Test_Precision	Test_F1score	Test_AUC
4	KNN Classifier	0.875767	0.875767	0.883326	0.873308	0.980332	0.812379	0.812379	0.821034	0.807896	0.929979
5	KNN Classifier tunned	0.999439	0.999439	0.999439	0.999439	0.999579	0.862594	0.862594	0.867221	0.859958	0.896945

- Hyperparameter tunned model show the best results.

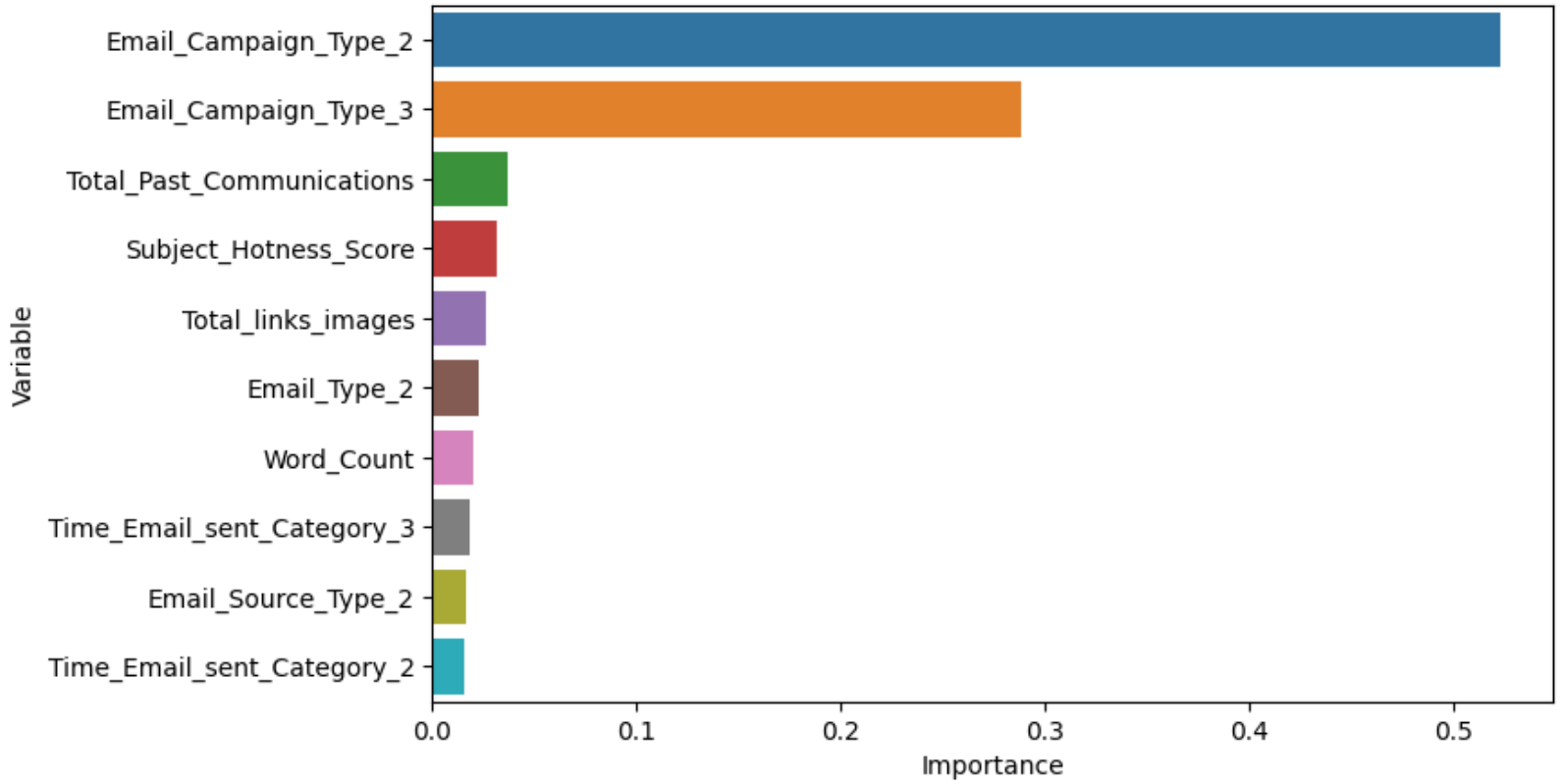
XGBoost Classifier

- XGBoost is a decision tree based ensemble machine learning algorithm that uses a gradient boosting framework.
- The two reasons to use XGBoost are also the two goals of the project.
    1. Execution Speed
    2. Model Performance
- After evaluation we got the following results.

	Model_Name	Train_Accuracy	Train_Recall	Train_Precision	Train_F1score	Train_AUC	Test_Accuracy	Test_Recall	Test_Precision	Test_F1score	Test_AUC
6	XGBoost	0.960095	0.960095	0.961723	0.959851	0.996812	0.860850	0.860850	0.861371	0.858131	0.955793

Feature Importance

- Feature with highest importance for XGBoost Classifier is shown by the bar plot.
- Email_Campaign_Type is most important feature followed by Total_Past_Communications for XGBoost Classifier model.

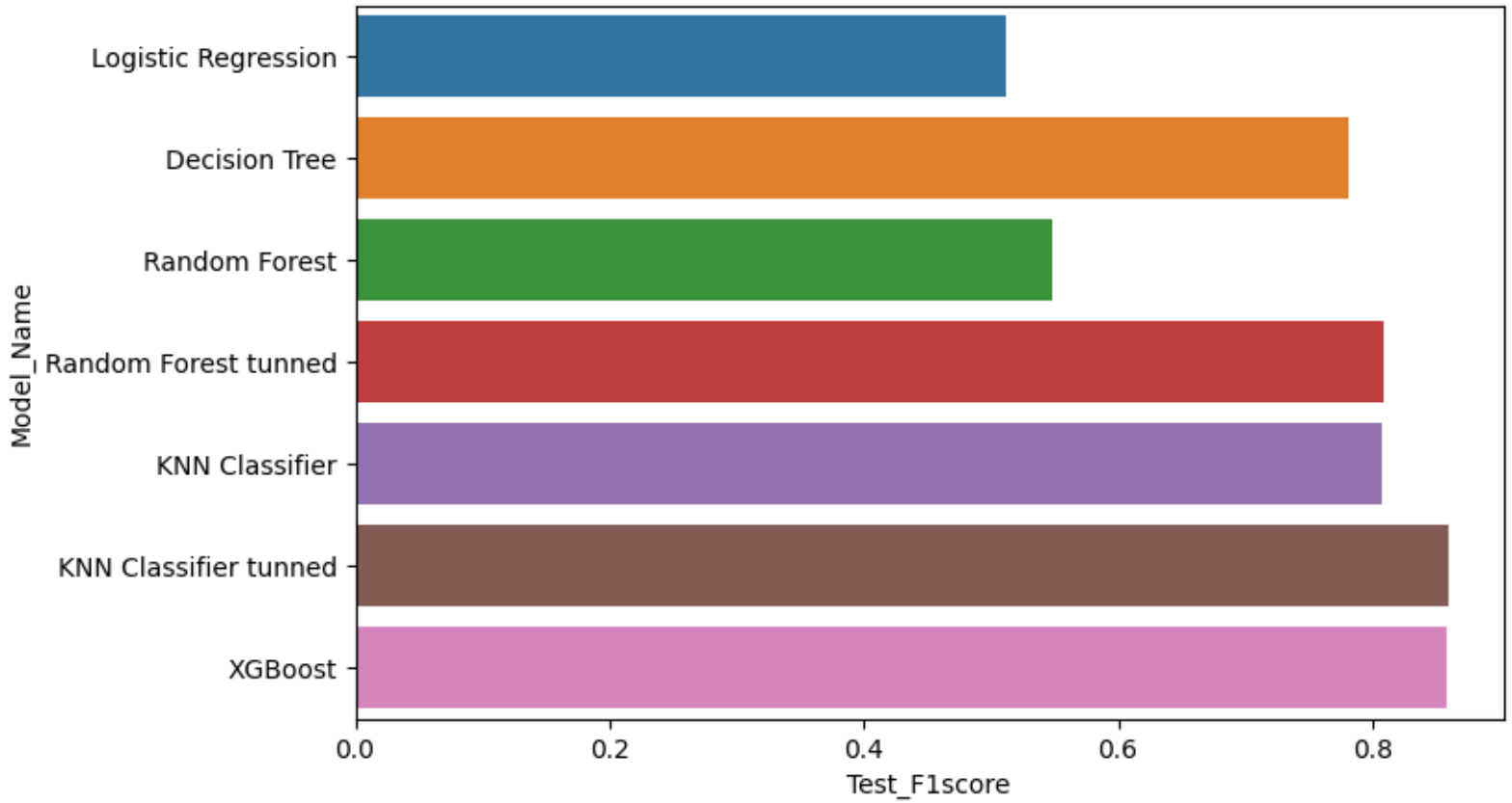


Model Performance and Evaluation

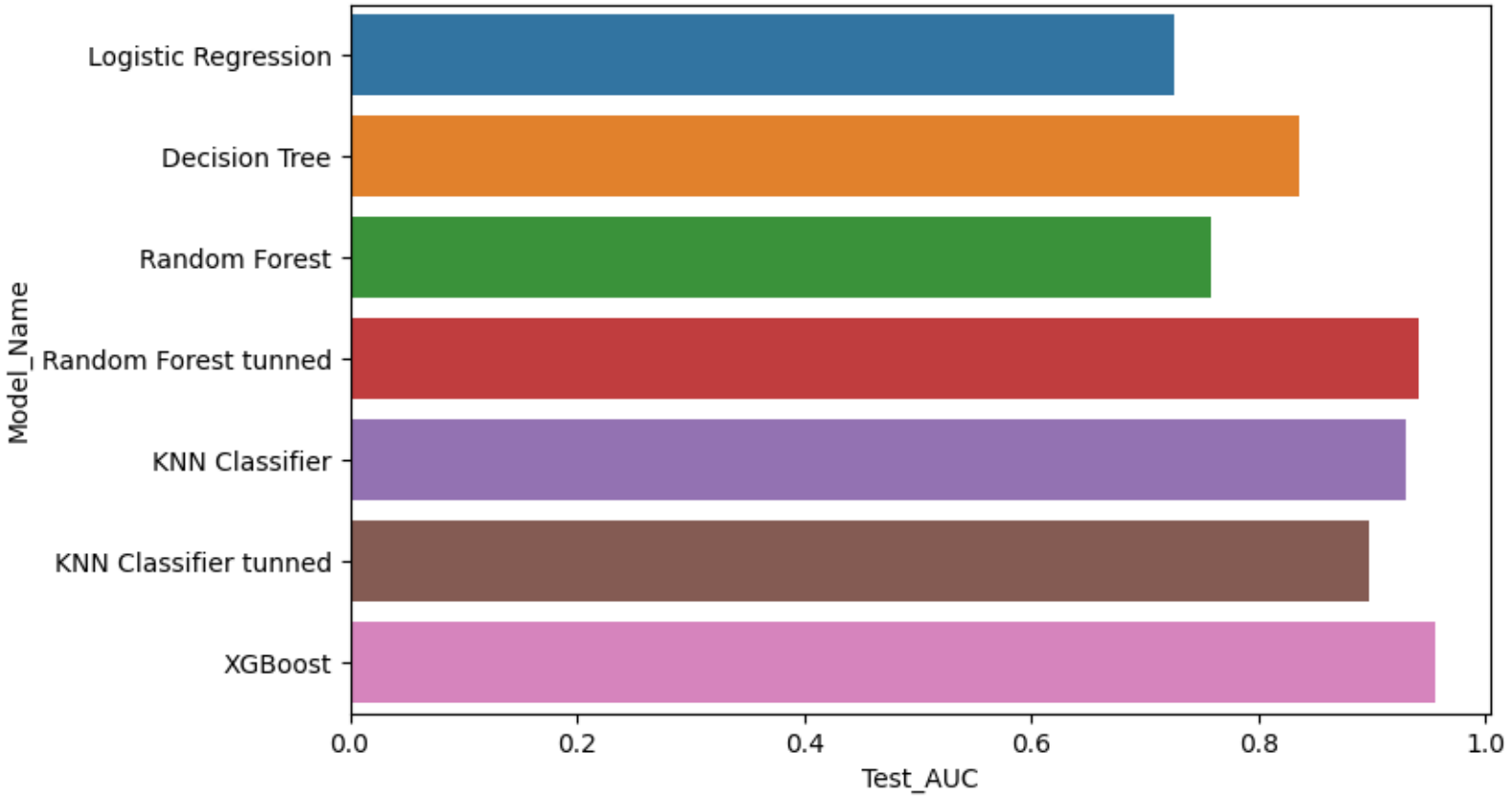
- Based on the metrics, XG Boost Classifier and KNN Classifier with best parameter works the best giving a train F1 score of 96% and 99% respectively and test F1 score of 86.17% and 86.12% respectively.

	Model_Name	Train_Accuracy	Train_Recall	Train_Precision	Train_F1score	Train_AUC	Test_Accuracy	Test_Recall	Test_Precision	Test_F1score	Test_AUC
0	Logistic Regression	0.536502	0.536502	0.521232	0.511283	0.723920	0.537942	0.537942	0.521233	0.511710	0.725308
1	Decision Tree	0.999439	0.999439	0.999440	0.999439	1.000000	0.780824	0.780824	0.780975	0.780715	0.835749
2	Random Forest	0.568954	0.568954	0.555191	0.548399	0.758494	0.569902	0.569902	0.555327	0.548952	0.758297
3	Random Forest tunned	0.892869	0.892869	0.892978	0.892028	0.979517	0.810884	0.810884	0.809475	0.808529	0.940415
4	KNN Classifier	0.875767	0.875767	0.883326	0.873308	0.980332	0.812379	0.812379	0.821034	0.807896	0.929979
5	KNN Classifier tunned	0.999439	0.999439	0.999439	0.999439	0.999579	0.862594	0.862594	0.867221	0.859958	0.896945
6	XGBoost	0.960095	0.960095	0.961723	0.959851	0.996812	0.860850	0.860850	0.861371	0.858131	0.955793

- Bar Plot shows the test F1 score for all the models.
- Its is clear from the plot that KNN Classifier and XGBoost Model show best results out all the other models.



- Bar Plot shows the test AUC score for all the models.
- Its is clear from the plot that KNN Classifier and XGBoost Model show best results out all the other models.



Conclusion

- **From EDA**
  - The percentage ratio of ignored, read and acknowledged emails are almost same for all the customer location. Customer Location does not exclusively influence the Email_Status. Both EDA and Chi-square hypothesis test showed the same result. So we should not consider location as a factor for people ignoring reading or acknowledging the emails.
  - If campaign type is 1, then there 66% chances of emails being read and 23% chances of emails being acknowledged.
  - Time_Email_Sent_category has no signifiacnt effect on Email_Status.
  - Analyzing total past communications by performing EDA and one way ANOVA test, it is evident that the more the number of previous emails, the more it leads to read and acknowledged emails. This is just about making connection with the customers.
  - The more the words in an email, the more tendency it has to get ignored. Too lengthy emails are getting ignored.
- **Modeling**
  - Based on the metrics, XG Boost Classifier and KNN Classifier with best parameter works the best, giving a train F1 score of 96% and 99% respectively and test F1 score of 86.17% and 86.12% respectively.