Capstone Project

Email Campaign Effectiveness Prediction

Midhun R

Points for Discussion

- Business Task
- Data Summary
- Data Cleaning
- Exploratory Data Analysis
- Feature Engineering
- Modelling
- Conclusion

Business Task

A dataset containing 68353 records across 12 features has been given with information regarding the characteristics of e-mail marketing campaign.

The main objective is to understand the existing data so that a machine learning model can be built to predict the e-mail status to improve campaigns.

This is undertaken as an individual project.

Data Summary

- Number of records (rows): 68353
- Number of features (columns): 12
- Out of 68353 rows, none of them are duplicate rows.
- Out of 12 columns, 4 columns have missing values and 3 columns require conversion of data type.
- These irregularities will be handled later during data cleaning step.

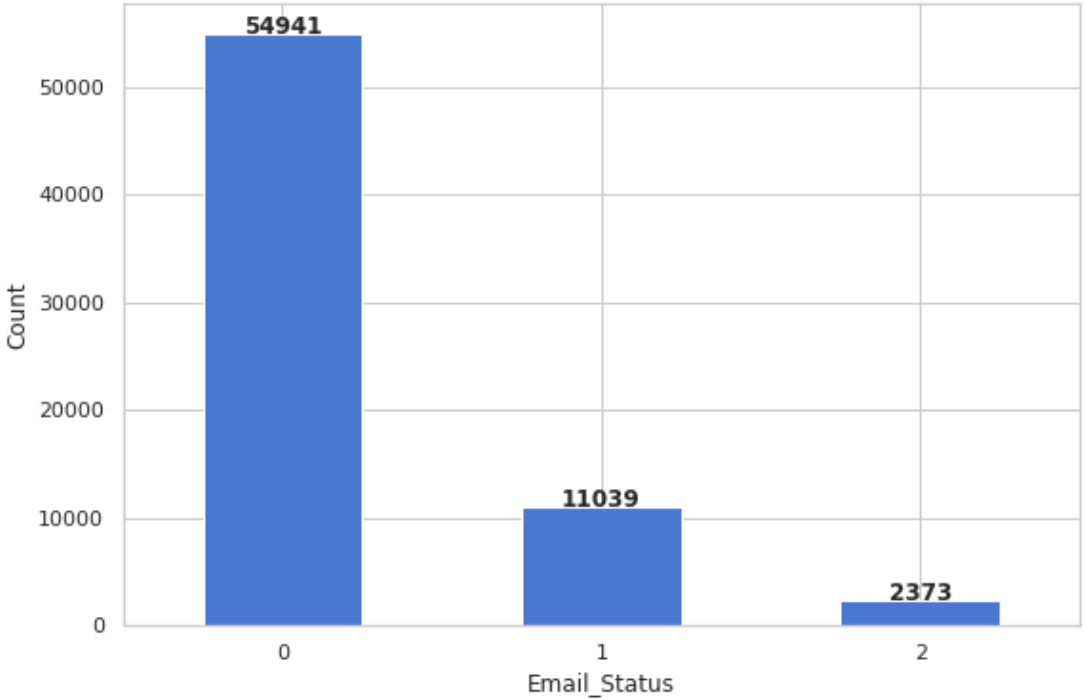
Data Summary (Contd.)

1. Email_ID: E-mail ID of recipients.
2. Email_Type: Differentiates between 2 different e-mail types: 1 and 2.
3. Subject_Hotness_Score: Measures the strength and effectiveness of mail subject.
4. Email_Source_Type: Differentiates between 2 different e-mail source types: 1 and 2.
5. Customer_Location: Differentiates between 7 different e-mail customer locations: A, B, C, D, E, F and G.
6. Email_Campaign_Type: Differentiates between 3 different e-mail campaign types: 1, 2 and 3.
7. Total_Past_Communications: Number of previous communications from the same source.
8. Time_Email_sent_Category: Differentiates between 3 different time of day category: 1, 2 and 3.
9. Word_Count: Number of words in the mail.
10. Total_Links: Number of links in the mail.
11. Total_Images: Number of images in the mail.
12. Email_Status: Differentiates between 3 different e-mail statuses: 1, 2 and 3, representing ignored, read & acknowledged, respectively. This is our target variable.

Data Cleaning

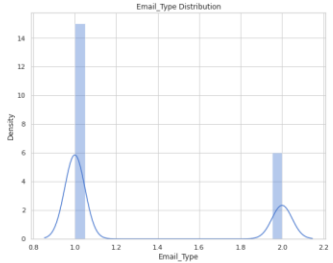
- About 17% of data in ‘Customer_Location’ has missing values. Since a large number of observations have missing values in it and it was difficult to find a value to impute in relation to other features, it was skipped.
- About 10% of data in ‘Total_Past_Communications’ has missing values; they were replaced with its mean.
- About 3% of data in ‘Total_Links’ has missing values; they were replaced with its median.
- About 3% rows in ‘Total_Images’ has missing values; they were replaced with its mode.
- Datatype of ‘Total_Past_Communications’, ‘Total_Links’ and ‘Total_Images’ were converted to int datatype for more convenience.

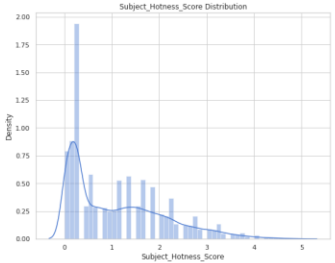
Exploratory Data Analysis

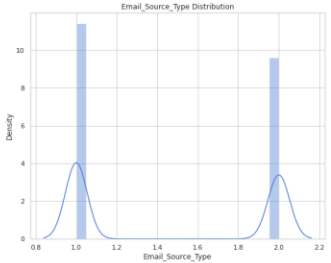


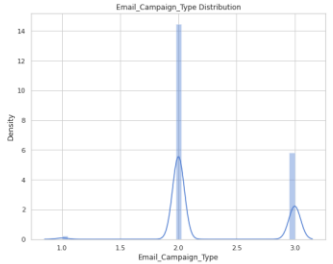
The dataset is highly imbalanced and ‘Email_Status’ 0 is the majority class and the rest of them are minority classes.

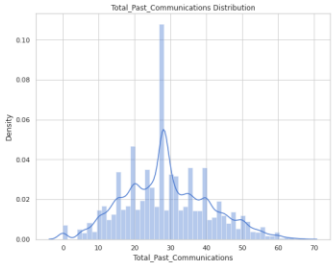
Exploratory Data Analysis (Contd.)

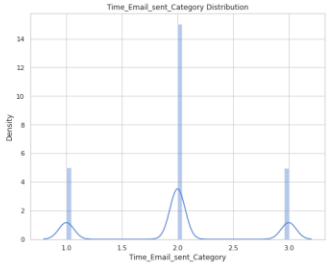


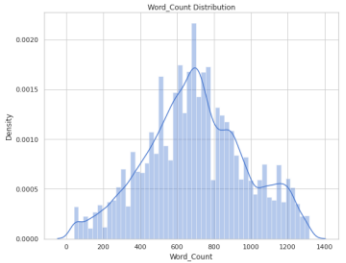


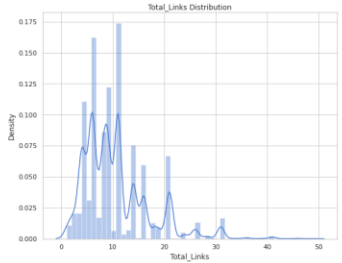


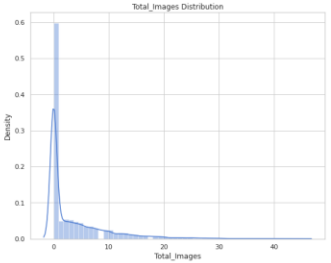


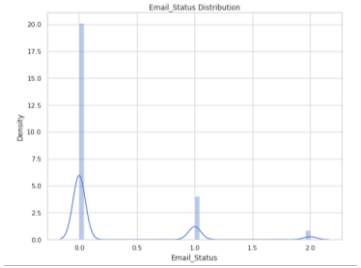






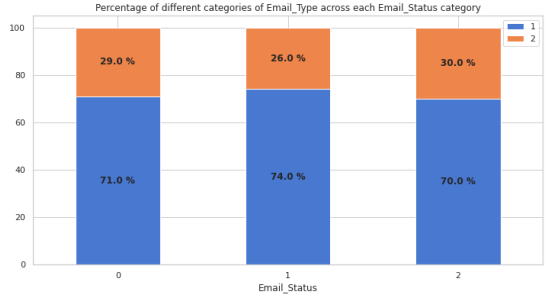


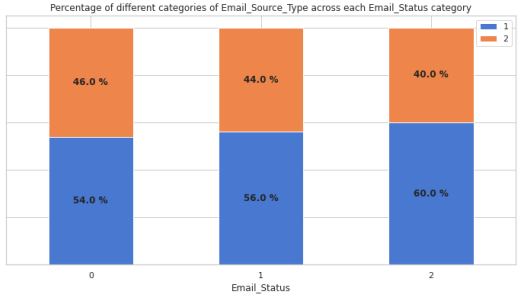


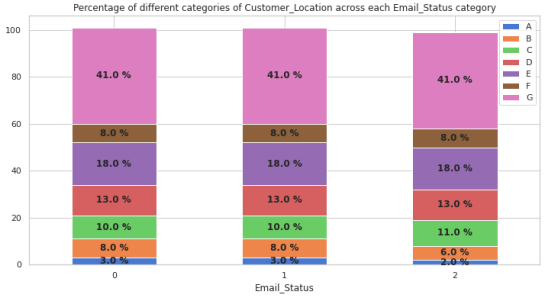


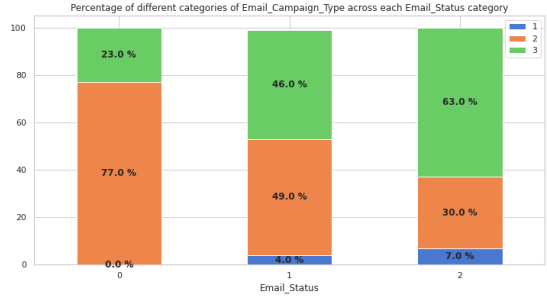
'Subject_Hotness_Score', 'Total_Links' and 'Total_Images' are positively skewed while 'Total_Past_Communications' and 'Word_Count' are almost normally distributed.

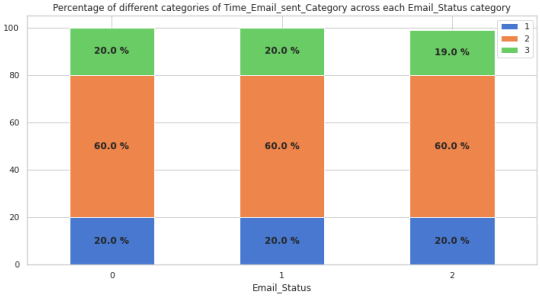
Exploratory Data Analysis (Contd.)





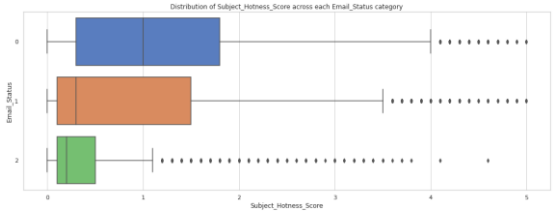


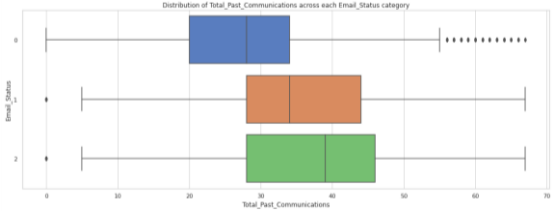


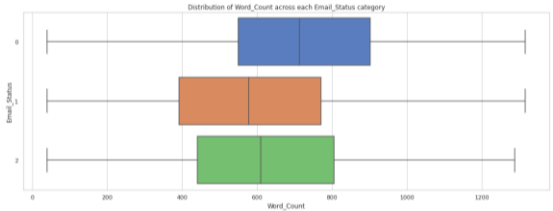


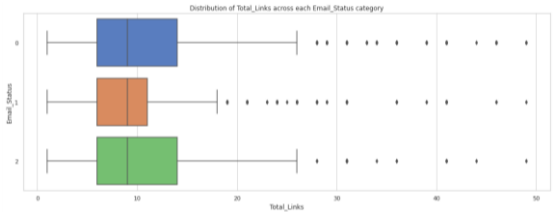
All categories of a feature have same distribution of e-mails across categories of ‘Email_Status'. ‘Email_Campaign_Type’ is the only feature which does not follow this trend. So it has the most impact on the target feature. If ‘Email_Campaign_Type’ is 1, then the mail has 66% chance of getting read and 23% chance of getting acknowledged.

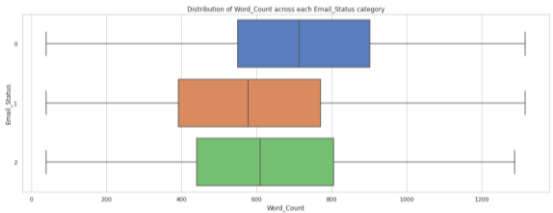
Exploratory Data Analysis (Contd.)











As the ‘Subject_Hotness_Score’ increases, probability of mails getting acknowledged decreases.

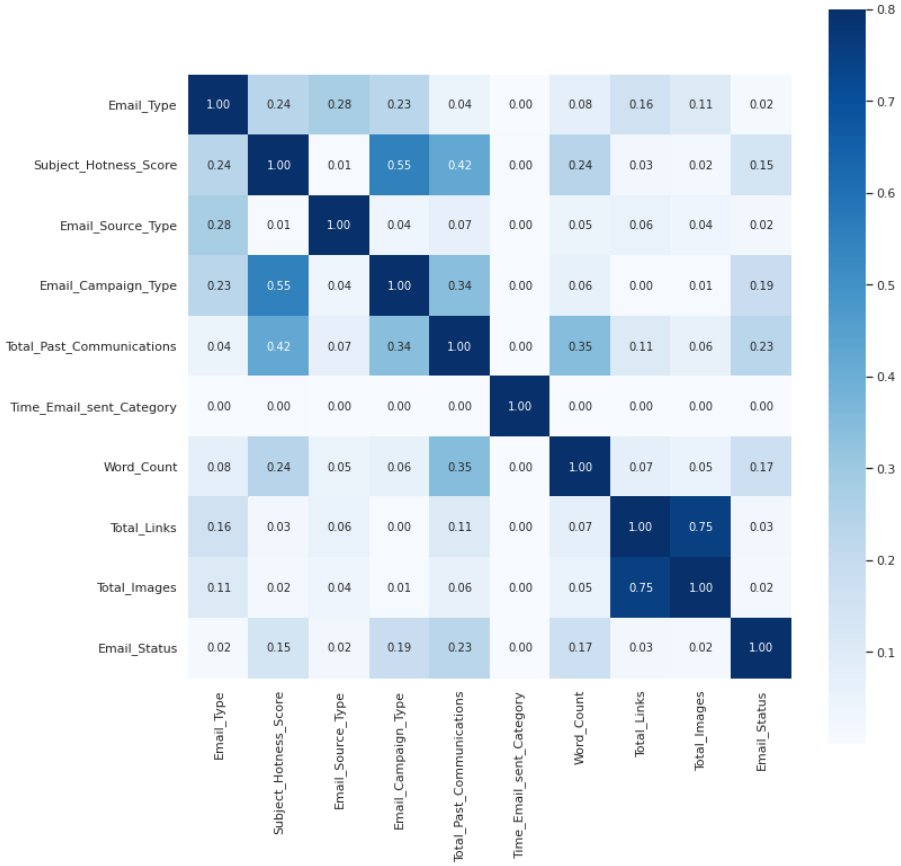
As the number of ‘Total_Past_Communications’ increases, probability of mails getting acknowledged or read also increases.

As the ‘Word_Count’ increases, probability of mails getting ignored also increases.

The ‘Total_Links’ have similar distribution across each ‘Email_Status’ category but read mails have slightly less variance than others.

The ‘Total_Images’ have similar distribution across each ‘Email_Status’ category but acknowledged mails have slightly more variance than others.

Exploratory Data Analysis (Contd.)



- ‘Time_Email_sent_Category’ has no correlation with Email_Status or any of the other independant features.
- Multicollinearity can be observed between ‘Subject_Hotness_Score’, ‘Email_Campaign_Type’ & ‘Total_Past_Communications’ and ‘Total_Links’ & ‘Total_Images.’

Feature Engineering

- ‘Email_ID’ is dropped since it does not affect the status of e-mail.
- ‘Customer_Location’ is also dropped since it has a lot of missing values which cannot be easily imputed and also it does not have much impact on the target variable as all the locations have same probability of mails getting ignored, read and acknowledged.
- ‘Time_Email_sent_Category’ is also dropped as it is already established in EDA that it has no correlation to any features and therefore, it doesn't affect the mail status.

Feature Engineering (Contd.)

- The variance inflation factor (VIF) of all numerical features is calculated in order to remove highly correlated features.
- Features having VIF greater than 5 should be eliminated.

	Feature	VIF
0	Subject_Hotness_Score	1.803914
1	Total_Past_Communications	3.911830
2	Word_Count	4.047726
3	Total_Links	8.581007
4	Total_Images	3.162623

- The only feature with VIF greater than 5 is ‘Total_Links’.
- But it cannot be removed from our dataset because it is an important feature. So, it was combined with ‘Total_Images’ since they have a good correlation.
- Then VIF of all remaining numerical features were calculated.

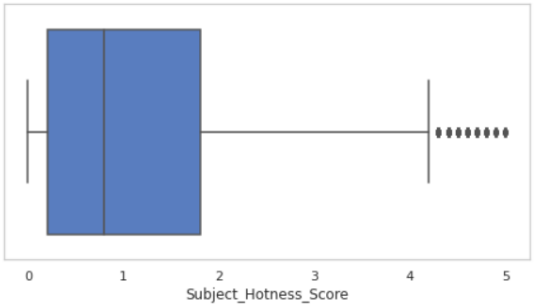
	Feature	VIF
0	Subject_Hotness_Score	1.733962
1	Total_Past_Communications	3.417183
2	Word_Count	3.678383
3	Total_Links_Images	2.613952

- Now, all features have VIF less than 5.

Feature Engineering (Contd.)

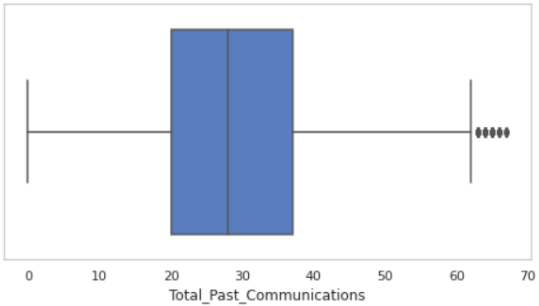
Outliers in ‘Subject_Hotness_Score’

Outliers in Subject_Hotness_Score : 247 (0.36%)



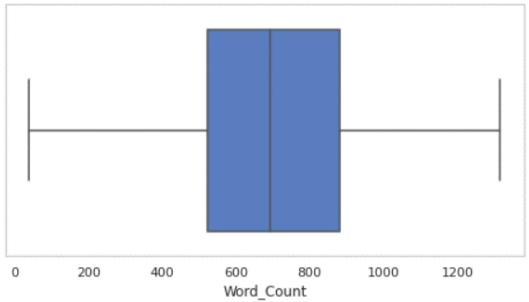
Outliers in ‘Total_Past_Communications’

Outliers in Total_Past_Communications : 136 (0.2%)



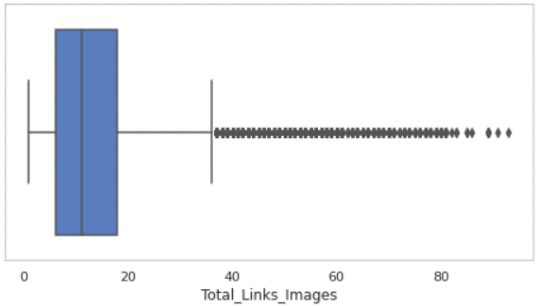
Outliers in ‘Word_Count’

Outliers in Word_Count : 0 (0.0%)



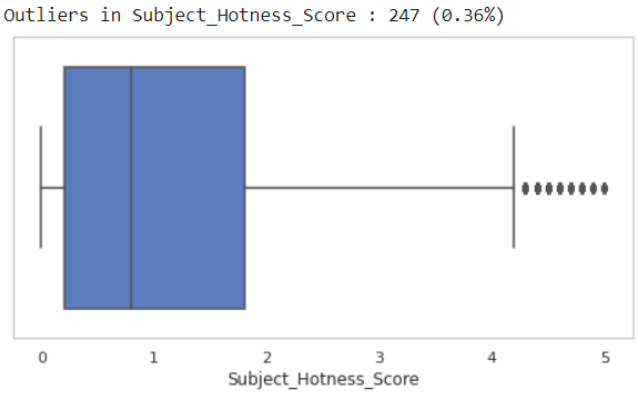
Outliers in ‘Total_Links_Images’

Outliers in Total_Links_Images : 3594 (5.31%)

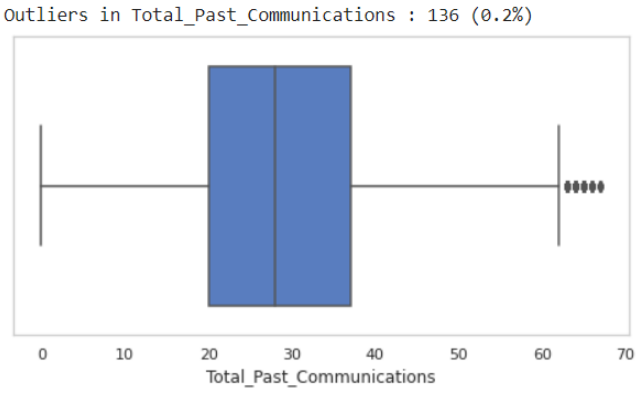


Feature Engineering (Contd.)

Outliers in ‘Total_Links_Images’ (Majority Class)

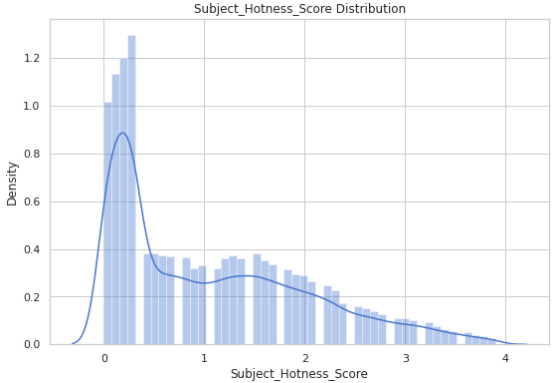


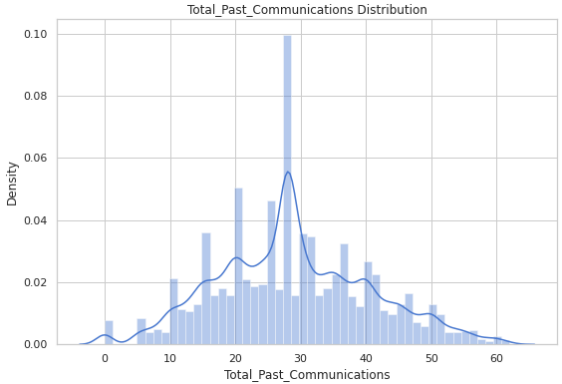
Outliers in ‘Total_Links_Images’ (Minority Classes)

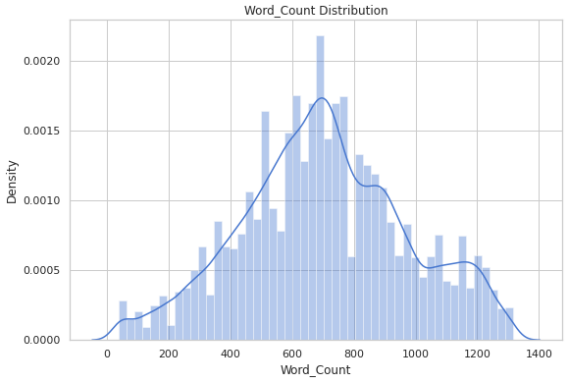


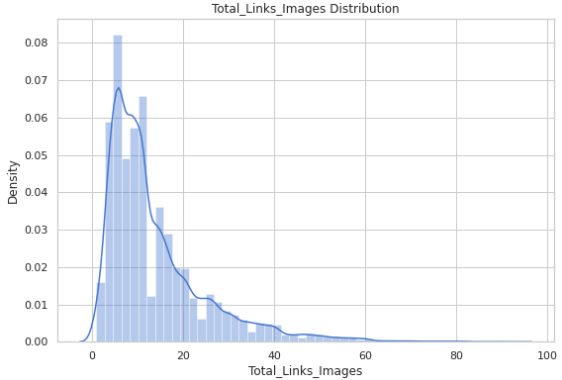
The percentage of outliers in minority classes is above 5. So, it is better to not remove them.

Feature Engineering (Contd.)





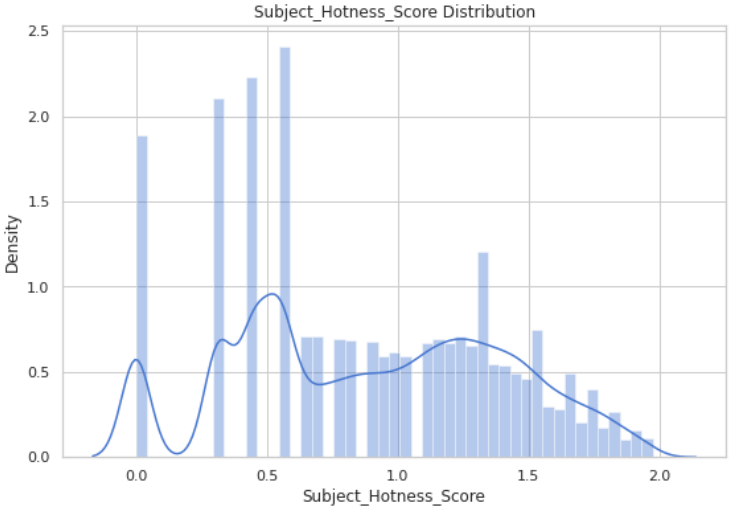


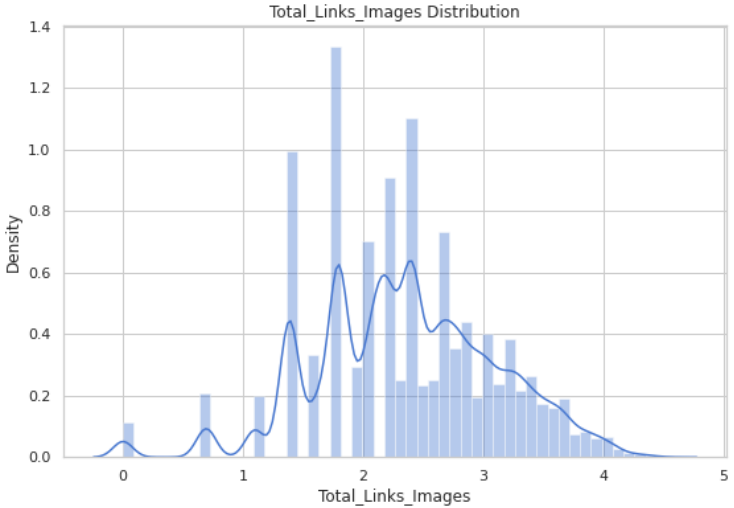


'Subject_Hotness_Score' and 'Total_Links_Images' are positively skewed. So, they must be transformed to normal distribution.

'Subject_Hotness_Score' has zero values while 'Total_Links_Images' has only positive values.

Feature Engineering (Contd.)





‘Subject_Hotness_Score’ is square root transformed and ‘Total_Links_Images’ is log transformed.

Feature Engineering (Contd.)

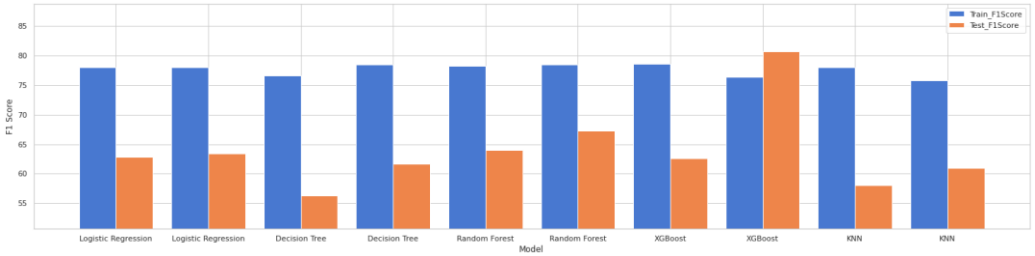
- ‘Email_Type’, ‘Email_Source_Type’ and ‘Email_Campaign_Type’ were encoded using one hot encoder.
- To overcome dummy variable trap, one resultant feature from each encoded feature must be removed. There are only two unique values in ‘Email_Type’ and ‘Email_Source_Type’, so removal of any one encoded feature from each of them would be sufficient. But in the case of ‘Email_Campaign_Type’, since there are more than two unique values, correlation matrix was used to decide which feature to remove.
- ‘Email_Campaign_Type_1’ was removed along with ‘Email_Type_2’ and ‘Email_Source_Type_2’.

Modelling

- Input and target data were separated, and both were split into training and test data with 25% test data.
- Training and test data of independent features were scaled using standardization.
- Two different techniques are used to balance the training data: Random Undersampling and SMOTE.
- Model training was done with these data using 5 different algorithms:
  1. Logistic classification
  2. Decision tree classification
  3. Random forest classification
  4. XGBoost classification
  5. KNN classification
- Models were trained with each algorithm twice, first with under-sampled data and second with over-sampled data.

Modelling (Contd.)

	Model	Sampling	Train_Accuracy	Test_Accuracy	Train_Precision	Test_Precision	Train_Recall	Test_Recall	Train_F1Score	Test_F1Score	Train_ROC_AUC	Test_ROC_AUC
0	Logistic Regression	RandomUnderSampling	53.389991	51.740423	53.389991	51.464354	0.717564	62.830969	78.074190	62.830969	68.572430	0.764563
1	Logistic Regression	SMOTE	53.662626	52.065803	53.662626	51.611125	0.718550	63.475177	78.060268	63.475177	69.018246	0.766232
2	Decision Tree	RandomUnderSampling	55.184136	57.747024	55.184136	52.096824	0.726046	56.294326	76.617504	56.294326	63.570629	0.705820
3	Decision Tree	SMOTE	57.574717	57.297754	57.574717	55.850958	0.757533	61.660757	78.556888	61.660757	68.128705	0.746796
4	Random Forest	RandomUnderSampling	55.901794	54.836561	55.901794	53.604339	0.747446	64.036643	78.228623	64.036643	69.452663	0.769335
5	Random Forest	SMOTE	60.795668	60.022655	60.795668	58.971044	0.797416	67.328605	78.478826	67.328605	71.760872	0.775618
6	XGBoost	RandomUnderSampling	60.132200	59.729420	60.132200	59.109163	0.795191	62.606383	78.610004	62.606383	68.558701	0.769354
7	XGBoost	SMOTE	87.754985	88.010946	87.754985	87.463419	0.966989	80.691489	76.355202	80.691489	77.814953	0.774185
8	KNN	RandomUnderSampling	99.848914	99.849277	99.848914	99.848887	0.999997	58.043735	78.044672	58.043735	65.033820	0.737761
9	KNN	SMOTE	99.865044	99.865363	99.865044	99.865065	0.999998	60.992908	75.758963	60.992908	66.467859	0.675422



- Evaluation metrics like Accuracy, Precision, Recall, F1 Score and ROC-AUC were calculated for each model.
- F1 score was used to compare different models and find out which one is better. Higher the F1 score, better the model.
- The model built using XGBoost algorithm with SMOTE dataset has the highest F1 score, followed by the one using random forest with SMOTE dataset.

Conclusion

EDA Conclusions

- No e-mails of campaign type 1 got ignored.
- If campaign type is 1, then the mail has 66% chance of getting read and 23% chance of getting acknowledged.
- Customer location or time of day does not affect the status of e-mail.
- As the number of previous communication increases, the chances of the e-mail being read or acknowledged also increases.
- E-mails tend to get ignored when word count is greater than 800.

Modelling Conclusions

- Oversampled data seems to be better than undersampled data. This can be due to the fact that undersampling causes loss of information.
- The model built using XGBoost algorithm with SMOTE dataset performed better than the other models. It should be preferred for predicting mail statuses.
- If model interpretability is more important than accuracy, model built using logistic regression algorithm and SMOTE dataset should be chosen over the one using XGBoost algorithm. It is the best performer among the white box models.