**Problem Statement**

Your task at hand is to build a machine learning-based approach to predict the CTR of an email campaign.

**Data Description**

Train and Test set contains different sets of email campaigns containing information about the email campaign. Train set includes the target variable *click\_rate* and you need to predict the *click\_rate* of an email campaign in the test set.

|  |  |
| --- | --- |
| **Variable** | **Description** |
|  |  |
| campaign\_id | Unique identifier of a campaign |
| sender | Sender of an e-mail |
| subject\_len | No. of characters in a subject |
| body\_len | No. of characters in an email body |
| mean\_paragraph\_len | Average no. of characters in paragraph of an email |
| day\_of\_week | Day on which email is sent |
| is\_weekend | Boolean flag indicating if an email is sent on weekend or not |
| times\_of\_day | Times of day when email is sent: Morning, Noon, Evening |
| category | Category of the product an email is related to |
| product | Type of the product an email is related to |
| no\_of\_CTA | No. of Call To Actions in an email |
| mean\_CTA\_len | Average no. of characters in a CTA |
| is\_image | No. of images in an email |
| is\_personalised | Boolean flag indicating if an email is personalized to the user or not |
| is\_quote | No. of quotes in an email |
| is\_timer | Boolean flag indicating if an email contains a timer or not |
| is\_emoticons | No. of emoticons in an email |
| is\_discount | Boolean flag indicating if an email contains a discount or not |
| is\_price | Boolean flag indicating if an email contains price or not |
| is\_urgency | Boolean flag indicating if an email contains urgency or not |
| target\_audience | Cluster label of the target audience |
| *click\_rate (Target Variable)* | *Click rate of an email campaign* |

**Data Validation**

In this step, we perform different sets of validation on the given set of training and testing files.

1. Name Validation- we check length of training data and test data.
2. Number of Columns - We validate the number of columns present in the test and train data files.
3. Name of Columns - The name of the columns is validated and should be the same in the test and train data file.
4. The datatype of columns – we validate the data type of given test and train data.
5. Null values in columns - If any of the columns in a file have all the values as NULL or missing, we discard but the data haven’t and missing value.
6. we are using label encoding for (“times\_of\_day” )column to convert it into numeric values.
   1. Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.

**Model Training**

1) Data Export from local - The data in a stored local is exported as a CSV file to be used for model training.

2) Data Preprocessing

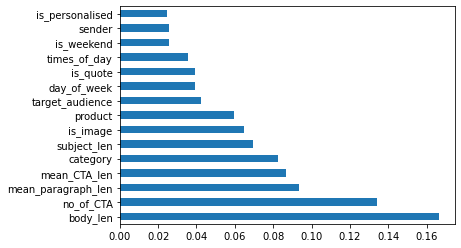
a) Check for null values in the columns.

b) Check if any column has zero standard deviation, remove such columns as they don't give any information during model training.

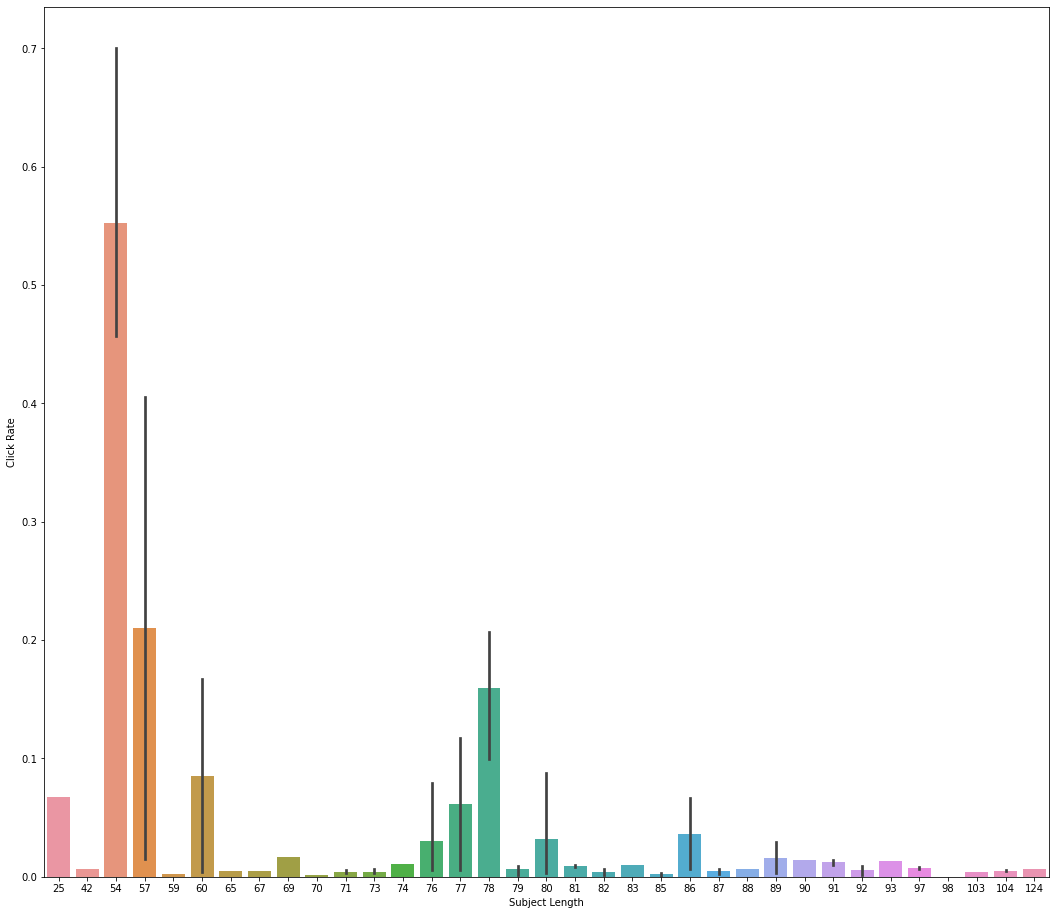
3) Model Selection - After creating base “Random Forest” and “XG Boost”, we find “XG Boost” is the best model. We are using two algorithms, “Random Forest” and “XG Boost”. Both the algorithms are passed with the best parameters derived from Grid Search. We calculate the r2 scores for both models and select the model with the best score. After that, we merge output of both algorithm and find mean of it which gives us best r2 score.

**Feature Selection**

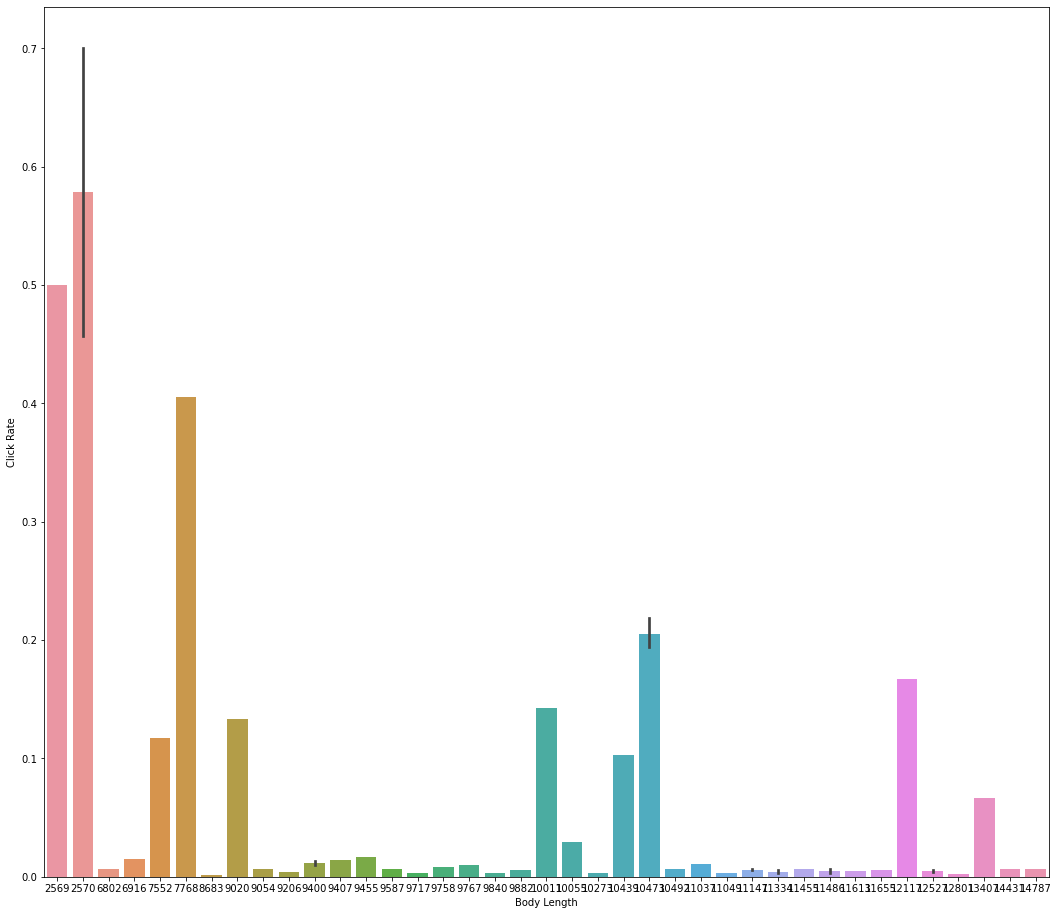
To perform feature selection using the “Extra Trees Regressor”, during the construction of the forest, for each feature, the normalized total reduction in the mathematical criteria used in the decision of feature of split (Gini Index if the Gini Index is used in the construction of the forest) is computed. This value is called the Gini Importance of the feature. To perform feature selection, each feature is ordered in descending order according to the Gini Importance of each feature and the user selects the top k features according to his/her choice.



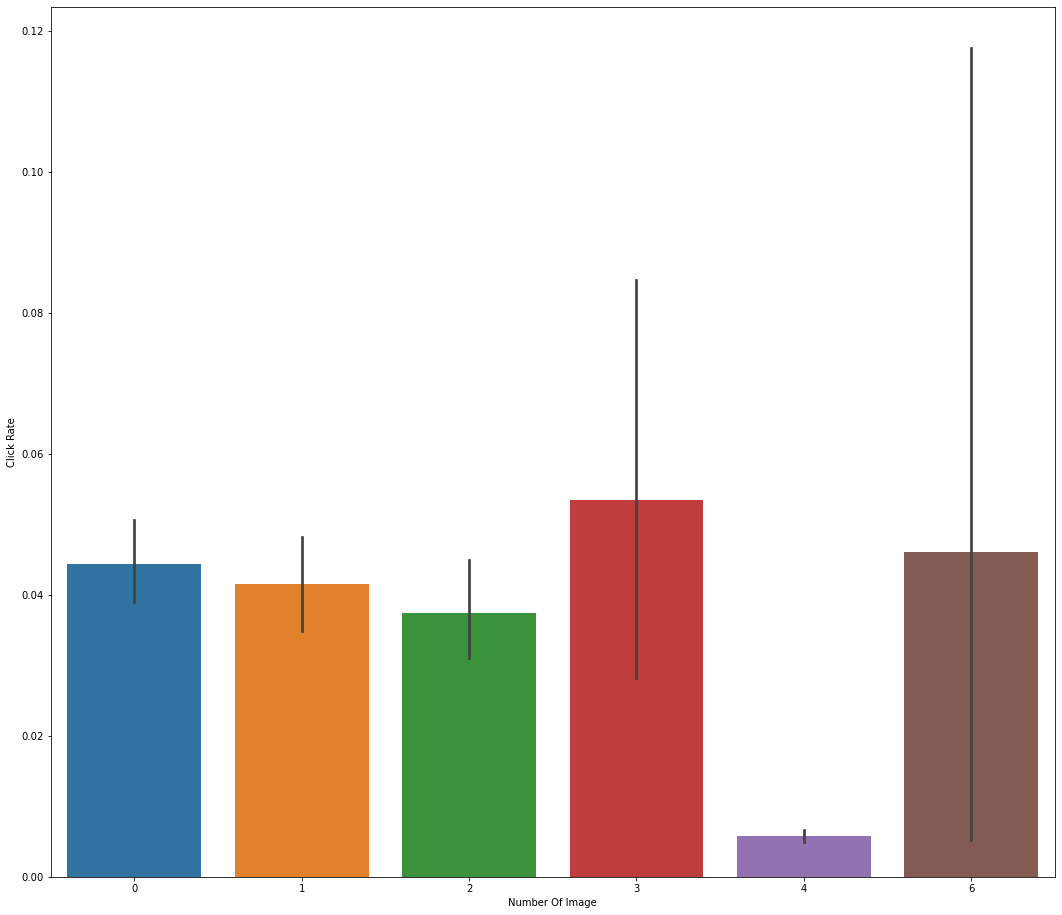
**Now let’s see the project graphical representations.**



According to this graph, the length of subject should be 54 to 57 then only the click rate will be high.



According to this graph, the length of body should be 2569 to 2570 then only the click rate will be high.



According to this graph, the image should be present in your email template.