**Predict CTR of an Email Campaign**

Most organizations today rely on email campaigns for effective communication with users. Email communication is one of the popular ways to pitch products to users and build trustworthy relationships with them. The primary aim of our project is to increase the Click Through Rate (CTR).

**CTR =   No. of users who clicked on at least one of the CTA / No. of emails delivered**

CTR depends on multiple factors like design, content, personalization, etc.

We are expected to build a smart system to predict the CTR for email campaigns and therefore identify the critical factors that will help the marketing team to maximize the CTR.

**Approach:**

Below are the technique performed on the dataset:

* **Feature Engineering:**
  + Load & Summarize dataset
  + Handle Missing values, outliers, Duplicates
  + Encoding the Categorical Features
  + Dropping unwanted features which may lead to overfitting
* **Exploratory data analysis to understand more about data**
  + Discrete Feature vs Target
  + Continuous feature vs Target
  + Correlation between features
  + Dropping unwanted features that does not have relationship with the output feature
* **Data Processing:**
  + Train test split
  + Feature Scaling(Normalization/Standardization)
* **Building Machine Learning Model**
  + Select few Regression models for the datasets, train and test the model accuracy.
* **Create Submission file to Submit the results.**

**Data-Preprocessing/Feature Engineering:**

* campaign\_id --> Removed since it will lead to model overfitting
* “is\_timer" columns has only one class and is not going to affect the output class. Hence dropped the feature.
* times\_of\_day 🡪 Ordinal 🡪 Evening(2), Noon(1), Morning(0)
* **Boolean Value convertion:** As mentioned in the data description the features like ('is\_weekend', 'is\_image', 'is\_personalised', 'is\_quote', 'is\_emoticons','is\_discount', 'is\_price', 'is\_urgency') are Boolean values. But few boolean columns has values other than 0 and 1.

On visualization, we see the values greater than 1 are negligible comparing with Boolean values. So we are encoding these values into 0 and 1(data greater than 1 into 1)

* There is very less relationship for the feature: "is\_price" to the Target variable. So removing the feature: "is\_price"

**Train and Test Set**

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* |

**Submission File Format**  
**sample\_submission.csv**contains 2 variables - campaign id and click\_rate

|  |  |
| --- | --- |
| **Variable** | **Description** |
| campaign\_id | Unique Identifier of a campaign id |
| click\_rate (Target Variable) | Click rate of an email campaign |

**Evaluation Metrics:**  
  
The evaluation metric for this hackathon would be r2\_score.

**Public and Private Split**  
  
Test data is further divided into Public (40%) and Private (60%) data. Your initial responses will be checked and scored on the Public data. The final rankings would be based on your private score which will be published once the competition is over.