**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.

|  |  |
| --- | --- |
| 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* |

**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 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:**

Based on the insights from EDA, the following feature selection/Rejection is performed on datasets:

* 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)
* “is\_price” 🡪 From data feature vs target plot, the relationship of feature “is\_price” with target variable is very less. Hence dropping the feature.
* “day\_of\_week” 🡪 Since there is multicollinearity between “is\_weekend” and “day\_of\_week”, dropping day\_of\_week feature.

**Outliers:**

* For outliers have tried following techiniques:
  + Drop outliers: Dropping outlier rows from the dataset increased the cross validation score, but the R2 Score has reduced for test data.
  + Replace outlier: Replaced outlier with upper boundary vaule, which helped in increasing the R2 score. Hence handled outliers with upper boundary value.

**Feature Transformation:**

* Have followed Feature transformation techinique like Square, square-root, logarithmic, exponential, cube etc on the features which don’t have a linear relationship with target variable. But it increased the Cross validation score, but not the R2 score. Hence no Feature transformation technique is performed.

**Feature Scaling:**

Since models like KNNRegressor is used for the problem statement, have applied StandardScaler for converting all the features to same scale range.

**Best model for the problem statement:**

**Model Selection:**

* From Exploratory data analysis, we see that the most of the features don’t have linear relationship with the target variable, even after applying the transformation techiniques.
* Since most of the features are non-linear, have selected models like KNNRegressor, Random Forest Regressor, XGBoost Regressor, Gradient Boosting. Out of all the models, the XGBoost has better Cross validation score.
* Hence selected the model, did hyper parameter tuning for predicting the output with high accuracy.