**Assessment 2: Engagement Score Prediction**

Can you predict the engagement score of the video?

**Problem Statement**

ABC is an online content sharing platform that enables users to create, upload and share the content in the form of videos. It includes videos from different genres like entertainment, education, sports, technology and so on. The maximum duration of video is 10 minutes.

Users can like, comment and share the videos on the platform.

Based on the user’s interaction with the videos, engagement score is assigned to the video with respect to each user. Engagement score defines how engaging the content of the video is.

Understanding the engagement score of the video improves the user’s interaction with the platform. It defines the type of content that is appealing to the user and engages the larger audience.

**Objective**

The main objective of the problem is to develop the machine learning approach to predict the engagement score of the video on the user level.

**Data Dictionary**

You are provided with 3 files - train.csv, test.csv and sample\_submission.csv

**Training set**

**train.csv** contains the user and video information along with the engagement score

|  |  |
| --- | --- |
| **Variable** | **Description** |
| row\_id | Unique identifier of the row |
| user\_id | Unique identifier of the user |
| category\_id | Category of the video |
| video\_id | Unique identifier of the video |
| age | Age of the user |
| gender | Gender of the user (Male and Female) |
| profession | Profession of the user (Student, Working Professional, Other) |
| followers | No. of users following a particular category |
| views | Total views of the videos present in the particular category |
| engagement\_score | Engagement score of the video for a user |

**Test set**

**test.csv** contains only the user and video information, and you have to predict the engagement score

|  |  |
| --- | --- |
| **Variable** | **Description** |
| row\_id | Unique identifier of the row |
| user\_id | Unique identifier of the user |
| category\_id | Category of the video |
| video\_id | Unique identifier of the video |
| age | Age of the user |
| gender | Gender of the user (Male and Female) |
| profession | Profession of the user (Student, Working Professional, Other) |
| followers | No. of users following a particular category |
| views | Total views of the videos present in the particular category |

**Submission File Format**

**sample\_submission.csv**contains only 2 variables - row\_id and engagement\_score

|  |  |
| --- | --- |
| **Variable** | **Description** |
| row\_id | Unique identifier of the row |
| engagement\_score | Engagement score of the video for a user |

**Evaluation metric**

The evaluation metric for this hackathon is [r2 score](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html).

**Guidelines for Final Submission**

Please ensure that your final submission includes the following:

1. Solution file containing the predictions for the row\_id in the test set (Format is given in sample\_submission.csv)
2. A zipped file containing code & approach (Note that both code and approach document are mandatory for shortlisting)
   1. Code: Clean code with comments on each part
   2. Approach: Please share your approach to solve the problem (doc/ppt/pdf format). It should cover the following topics:
      1. A brief on the approach used to solve the problem.
      2. Which Data-preprocessing / Feature Engineering ideas really worked? How did you discover them?
      3. What does your final model look like? How did you reach it?

**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.