**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.

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

**Objective:**

Predict the engagement score of the video on the user level.

**Approach:**

To get the base model, I simply fit xgb model on train data and predict on unseen data & score was .20, So I need to improve from here.

1. I checked the unique users in train and test data.
   1. How many users were in test data but no on train data & all users were present in train data.
   2. So, I filtered the train data only with all user’s which are in test data. Now I have same users in both (train & test data)
   3. Fit the XGB model. (Tuned Hyperparameter)
   4. Score is improved to .27
2. I merged both (train & test) data set & create lag variable of engagement\_score
   1. User wise
   2. User-category wise
   3. Fit the XGB model.
   4. Score is improved to .37
3. Grab the mean/std deviation/min/max of engagement\_score for each user & create features using them.
   1. Fit the XGB model.
   2. Score is improved to .40.

**What I tried but did not work:**

1. Ensemble of two models (Cat + XGB OR Cat + LGBM or LGBM + XGB)
2. Create cluster using age/followers/like and created variable & used in model.
3. Which category video user uploaded the model & used in model.
4. Grab mean ofengagement\_score user wise from training data & use as feature in model.