**Introduction:**

Our objective is to predict the engagement score of the video on the user level

**Approach:**

We checked the distribution of target variable Engagement score . It s not exactly normally distributed. It has some modes as most of variables are between 2 and 4

We tried log, boxcox, squared transformation but that does not help much

From the heatmap we could see that category id and video id are correlated. And video id is of not much use as we are checking target on user level. Video\_id did not help to improve r2 score as well. So we removed video\_id

Here we are predicting engagement score on user level. So user id is the important feature Linear regression won’t work well as the target is not uniformly distributed. So we go for tree based models

In Boosting algorithms the user id is not picked as primary variable when we checked feature importance so we got relatively low r2 score over there. WE have used RandomForestRegressor which gave considerably good score

We tried adding new features like mean views per user id,max views per user id etc but it did not help in improving r2 score

As the data is on user level and train and test both have same user we concatenated and took count of users in category and mean users in category which helped us to improve r2 score very little.

We have added new feature - engagement per category by dividing views/followers as they both are given in category level

We tried tuning randomforest hyperparameters using stratified kfold cross validation and we are able to get final score of 0.4167 in public leaderboard

There is still much scope for improvement. As we know boosting algorithms generally perform well. We should adjust hyperparameters accordingly so that it picks information based on user level

And in randomforest also we could do some more tuning to get good score

I tried many approaches like removing outliers from age column, removing category\_id, views as it is correlated with followers and new col categ-enga but it did not help in improving r2 score