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.

APPROACH

FEATURE SELECTION

Engagement score is calculated based on users' interaction with the content.

Engagement with video or any content is a joint product of both User and the content.

For example:- Certain users engage more with the content in comparison to other users. Also certain content get more engagement in comparison to other content.

Example: - Controversial videos attract greater attention.

So lets try to figure out key determinant of engagement score:

USER BEHAVIOUR

One can try to predict the user behaviour with factors like age, gender, profession etc.

A person is however more than a member of group/demographic. Group behaviour can't capture individuality.

What's the best predictor of an individual's behaviour?

Past actions of individuals. We see individuals as repeating patterns.

In this dataset the training set and test set comprise of same users and videos.

So, mapping the user behaviour from training set to test set becomes easy, it can be done with the help of user id.

The past user behaviour (or past engagement score) is the most important determinant, or feature.

Also, gender, age, profession are features which we used to refer to user.

So if one can directly refer to user using user id other feature become less important or even redundant.

Next Key determinant is Content of video: -

One can try to figure out whether the content of video is engaging by categorizing it with help of it's video category or views.

However, if one has video id and it directly refers to the video. So depending upon how engaging a video is in training set I try to predict it's engagement score on test set.

TARGETED ENCODING

User ID, Video Id, Video category are categorical data data. So we need to encode it.

However, there are >27000 users, >175 videos, and >47 categories, so categorical encoding would lead to large number of features.

I use targeted encoding. We replace categorical variable with mean/median of target variable. These captures mean of past behaviour.

I also encode other categorical features such as gender, profession etc using categorical features.

I scaled(standard) all variables except categorically encoded variables.

User_ID and Video_id together account for most of the variance in engagement score. The rest of the features only marginally improve r_2 score. However, since the ranking is based upon r_2 score more features are included to improve r_2 score. What is measured is improved.

MODEL SELECTION

After Going through a number of regression models from linear to neural networks.

Three models which gave best performance were shortlisted. The three are:-

1)Linear regression

2)Neural network regressor

3)Random Forest regressor

Linear regressor performed best on test data and was most consistent. It showed acceptable bias and comparatively low variance.

Neural Network model showed low variance and acceptable bias, in fact bias was lower than linear regression for training data. However, it performance was below that of linear regression on testing data, albeit difference was insignificant.

Random forest regressor worked best on training data but was over fitted and performance degraded for testing data.

Why Linear regression?

The reason behind better performance of Linear regression over other algorithms can be attributed to the fact that dataset is comparatively small and spread over too many categories(users,video_id)

This results in algorithms like Neural Networks failing to converge to optimum point, Algorithms like random forest overfit the data whereas simple algorithm like Linear regression performs best for dataset.

Linear regression did the best job and gave the best result.

Shortcoming of this approach: -

This approach would have worked better if we had more data points per user and per video. The reason being that human behaviour regresses to it's mean , so when we try to predict human behaviour we actually try to predict mean of human behaviour.

Due to less datapoints per user, this approach overfits for algorithms like neural net or random forest and thus Linear regression performs better for this hackathon.

Learning

The best approach must be decided taking availability of data into account.

Simple algorithms are more suitable when information is less, as they are more robust.

What is measured is improved, so one should widen what one measures.