Engagement Score Prediction for Online Video Platform

AV COMPETITION

Problem Statement Objective

We have a dataset of an online video platform. The engagement score is assigned to the video based on the user's interactions. The engagement score defines how engaging the content of the video is. The objective is to predict the engagement score between the user and the video of different genres. The engagement score is the dependent variable here present in the training dataset, ranging from 0 to 5

Dataset

List of the features in the training dataset:

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 category
engagement_score	Engagement score of the video for a user

Approach for Feature Engineering

1:
   * filtering the data to the interquartile range.
   * creating dummies from `professional` feature
   * converting `gender` to label encode format
2:
   * creating dummies from `professional` feature
   * converting `gender` to label encode format
3:
   * merging `gender` and `profession` features and using mean encode
4:
   * merging `gender` and `profession` features and getting their dummies as feature
5:
   * creating dummies of `category_id` and `video_id`
   * using mean encode on `user_id`

Out of which Step 1’s performance was best.

Approach for Model Selection

- ➔ After diving training dataset in train – validate datasets, trained 3 models: Linear Regression, Random Forest, and XGBoost.
- ➔ R2 score on the train – validate dataset was best on XGBoost.
- ➔ After hyperparameter tuning, trained the XGBoost model and got the predictions.