Prediction of Engagement Score

Overview

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

This is a regression problem.

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

We have got 3 files from https://www.analyticsvidhya.com/ - train.csv, test.csv and sample_submission.csv.

Training set -> train.csv contains the user and video information along with the engagement score, size of the dataset is around 4.93 MB. All the necessary features are described below –

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, size of the dataset is around 492 KB. All the necessary features are described below –

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

Approach

1. Import libraries

I have imported all the relevant libraries.

2. Data Inspection and Data Cleaning

I used train and test dataset provided by https://www.analyticsvidhya.com/ for training and testing purpose.

Approach I have taken for data inspections and cleaning are –

- Checked shape of the datasets
- Counted the datatypes of columns for both the datasets
- Checked the memory uses of the datasets and then reduced the memory using down casting function.
- Lastly checked any null value present in the datasets or not.

3. Exploratory Data Analysis

- At first checked total numbers of columns
- I have analysed all the features very closely, and tried with different plotting techniques as shown in the original notebook.
- I found these are the 3 most important features as mentioned **(age, gender, profession),** actually made very big factor in this case study.

4. Feature Engineering:

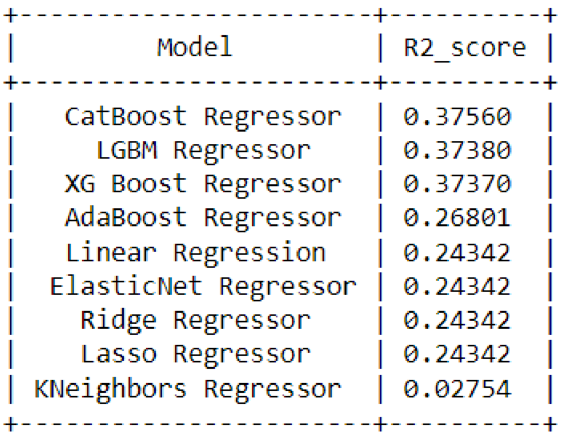
- I have used label encoding on object type i.e., Gender and Profession features.

- Rest of the all features are all numeric features, and not corelated with each other. So, I have kept all those features as same as provided in the dataset.

5. Modeling

- All though it is a regression problem, so I have used different regression algorithms as mentioned below, and tried to evaluate this the R-square score.

- **Although the R-square is quite low because of lack of information provided by the dataset,** Still based on the highest result of R-square score I found **CatBoost algorithm** worked well on the validation data.



```
+-----------------------+----------+
|         Model         | R2_score |
+-----------------------+----------+
|   CatBoost Regressor  | 0.37560  |
|     LGBM Regressor    | 0.37380  |
|   XG Boost Regressor  | 0.37370  |
|   AdaBoost Regressor  | 0.26801  |
|   Linear Regression   | 0.24342  |
|  ElasticNet Regressor | 0.24342  |
|    Ridge Regressor    | 0.24342  |
|    Lasso Regressor    | 0.24342  |
| KNeighbors Regressor  | 0.02754  |
+-----------------------+----------+
```

- So Final modeling I have used **CatBoost algorithm** to train the data, and finally tested with the test data provided by www.analyticsvidhya.com.
- Lastly stored all the result data into **my_submission.csv** for submission.

Thank You