Engagement of video

Project Report

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.

Submitted by-

Abhay Surma

ACKNOWLEDGEMENT

First of all I would like to thank you Vidhya analytics For Providing me this dataset and give me a change to work on it.

I have given my all efforts while doing this project.

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

HARDWARE AND SOFTWARE REQUIREMENT

- For hardware I have used my laptop for doing EDA and training the model
- For Software I have used jupyter notebook
- And for rest of EDA I have used some library like
- Pandas
- Numpy
- Seaborn
- Matplotlib

Encoding

Applying get dummies (it is like one hot encoding) which changes column which are object type to integer (0,1)

DATA ANALYSIS

The data is having 89197 rows and 13 columns (After encoding)

In this dataset there was no missing values

data where we are having 2 columns that are only object type.

For checking outliers – box plot is plotted

For checking distribution – displot is plotted

For checking corelation – bar plot is plotted

For checking multicollinearity – heat plot is plotted

After this we done,

Data Normalization -

Data normalization is essential part and I have done it by -

power transformer

Log transformer

But best result is coming by not applying it and also removed outliers but results are not getting fine so removed it

Data Scaling -

Data scaling is done through -

Standard scaler

Min-max scaler

We are getting good results by applying Min-max scaler

MODEL TRAINING -

First, we have applied linear regression from linear model to select best random state i.e., 992.

Then we have train test split and taken test size as 0.10 as we already have test to be tested so, giving good chance for training data to be trained.

Linear regression -

- r2score – 27.086
- RMSE – 0.742
- Cross – validation score – 27.21

Then we applied we few other algorithms from regularization i.e., Ridge

Ridge is giving same result as linear regression

lasso is giving bad results removed from file

Decision tree -

Performing bad as the model so removed

Random forest -

- R2score - 36.16120122956173
- MSE - 0.5206511995515695
- RMSE: 0.6941604285881708

Random forest is giving as the best result in all what we also applied

Gradient boosting after this and Xtreme boosting but results of all the model are not better than Random Forest

Gradient Boosting

- R2score - 34.34286878769742
- MSE - 0.5375233145427317
- RMSE: 0.7039769684333769

Hyper tuning the random forest model and Gradient boosting model – by Gridsearchcv

R2score increases little from hypertuning so we selected the random forest model

CONCLUSION -

r2score is 36.72

Also applied PCA and done some feature engineering but couldn't improve result by that so removed from final submission.

THANK YOU