Engagement Score Prediction (JOB-A-THON)

Data approach:

1. It is a regression problem.
2. Dropped the column ‘row_id’ as it’s just a serial number.
3. There are two categorical columns, gender, and profession, converted into numerical by one-hot encoding.
4. Done univariate, bivariate, and multivariate visual analysis.
5. Done Exploratory data analysis.

Observations:

1. The age of the users and category of the video both are right-skewed distributed or positively skewed which shows the mean, median, and mode are all different. Mode is the highest point of the histogram is nearby 20 for the age of the users and for the category of the video mode is 5.
2. The engagement score of the users is left-skewed distributed or negatively skewed and mode is around 4.

Model Selection approach:

1. As this is a regression problem so I have tried the **first model ‘linear regression**”.
   Used standard scaling on the data.
   Find the below score for the training data.
   RMSE: 0.736, MAE: 0.568 and R2 score: 0.273.
2. The Second model I have tried with decision tree and done hyperparameter tuning.
   It has improved the r2 score. The best parameters it has given me on the train data:
   max_depth=3,
   max_features='auto',
   min_weight_fraction_leaf=0.1.

3. The third model I have tried is random forest and r2 score is improved to 0.3088 on the train data.

Later on, I have tried XG Boosting and SVR (did not get a good score) and thought to try stacking meta-models by these models but I couldn’t apply as time restricts me.

Used eli5 for explaining the model:

The random forest model explains that male student users have more weight for the data and are more engagement scores.

Thank you for this opportunity Vidhya Analytics.

Sonya Naidu