Approach for finding the best model to predict engagement score of users for a video

1. **Brief:**
2. Data Inspection:
3. Finding if any attributes have missing/null values.
4. Finding number of categorical and numerical columns.
5. EDA:
6. Checking unique values in columns to see if they are discrete or continuous.
7. Checking unique values to find irrelevant attributes that we can drop.
8. Univariate analysis: Checking distribution of each attribute to find any anomalies/outliers.
9. Bivariate analysis: Checking the data w.r.t other attributes to dig more information.
10. Data Wrangling:
11. Removed data points having engagement score less than 1 to remove outliers.
12. Transformed target/label to make it normally distributed.
13. Dropped irrelevant columns row\_id and user\_id.
14. Label Encoding for categorical columns gender and profession.
15. Model Building:
16. Using KFold cross validation to find best split for train and validation data from train data set.
17. Building model using XGBoost algorithm using best split found using KFold.
18. Predicting the engagement score on the test data set.
19. Storing the predicted output in submission.csv file.
20. **Which Data-pre-processing / Feature Engineering ideas really worked? How did you discover them?**
21. Dropping irrelevant columns.
22. Removing data points which was outlier as per engagement score.
23. **What does your final model look like? How did you reach it?**
24. Explored various models like Linear Regression, Decision Tree, Random Forest, XGBoost, Light GBM.
25. Doing multiple rounds of hyperparameter tuning and then using K-Fold cross validation I ended up using XGBoost as it was giving be best score of all.