Engagement score prediction

i. A brief on the approach used to solve the problem.

1. Importing libraries: Required libraries and dataset was imported.
2. Null values checking: Null value evaluation was done to see if there were any null values existed in dataset and none were existing.
3. Outliers checking: Also, outliers were checked if existed and existing outliers could be left as it is since it was not harming performance of dataset.
4. Object column: One hot encoding was done for object column to convert it to numerical column.
5. Approach 1: In the first approach, the numerical columns were run as it is without carrying any feature engineering process.
6. Heatmap: To find correlation between variables, heatmap was plotted.
7. Train test split: Train and validation set was split from train dataset to carryout cross validation using Random search cv.
8. Algorithms: Linear Regression, Decision Tree regressor and Random Forest regressor were carried out to predict the target variable.
9. Test prediction: Once the ML algorithm which gave better r2score was considered, the same was used to make test dataset prediction.

ii. Which Data-preprocessing / Feature Engineering ideas really worked? How did you discover them?

1. Approach 2: Approach 2 was carried out to perform feature engineering.

2. Encoding: Since the numerical variables in columns like category_id, video_id and age were not making much sense to run as it is, binning of those variables was done.
3. Frequency encoding: Frequency encoding was carried out to give precedence to variables in binned columns with respect to target variable. Other encoding methods like one hot encoding or dummy encoding could have been done but I wanted to use the precedence existing in binned column.
4. Hyperparameter tuning: HPT was carried out to find out the best estimators which could give better r2score. Random forest with HPT gave better r2score.
5. Scaling: Power transformer method was used to scale as well as remove skewness and convert dataset to normal distribution.

iii. What does your final model look like? How did you reach it?

1. Train prediction: Prediction were run on train dataset by splitting it as train and validation set to run cross validation to get r2score.
2. Test prediction: The model which gave better train prediction was used to run test prediction with best estimator. The machine learning algorithm which gave better r2score was random forest with hyperparameter tuning.