1. Data Preprocessing

As a part of data preprocessing the removal of column row_id enhanced the model performance.

For the encoding I tried Label_encoder and one hot encoding where the one hot encoding worked quite well.

The encoding was done on the categorical features which weren't removed from the data to improve the performance.

2. EDA

As a part of EDA we tried to figure out the duplicate values, null values and outliers but the data was cleaned enough that we didn't get any.

For the outlier detection we tried using inter quartile range formulas but it didn't improve the performance in any ways and practically considering those outliers is not feasible too by referencing the real life consumer

For correlation heatmap we had these features and weren't highly correlated still they were carrying some importance with them, so not so many columns were removed apart from row_id.

3. Deciding the model

We trained out model on

a. Linear Regression/ Polynomial Features.
b. GradientBoostingRegressor
c. XGBoostRegressor

GridSearchCV was used to find the optimal hyperparameter for all of the models and XGBoostRegressor performed the best among all with the best r2_score.