METHODOLOGY AND APPROACH: Analytics Vidya - Football Hackathon by **OLUBAYODE EBENEZER** (**olubayodeeben@yahoo.com**).

Data Exploration:

1. The train data is not uniformly distributed.
2. Converted Team, winner columns to numerical variable
3. I removed Outliers.
4. There are plenty of Null and Zero columns which needs to be removed. I dropped column with almost 90% zeros, and replace nan values with mean.
5. There are plenty of highly correlated columns which needs to be removed.
6. . The outliers of the numeric column were treated again.
7. I did some features engineering and features interaction so as to be able to get more features that can help in our model performances.

MODELLING:

I model two different regressor models which helps in getting a better result. The two baseline model used are **CATBOOST REGRESSOR and LGBM REGRESSOR**;

1. Building the first model by Catboost without encoding of the categorical variables.
2. Building the second model by LGBM with encoding of the categorical variables.
3. Take the average of the both result as they are expected to find patterns in different way.
4. Treatment of biasness towards higher ratings which is mainly due to presence of higher frequency of rating 10.

5. Final Model and Submission: Even after taking the average of both the Catboost and LGBM models the minimum rating point was still 6.96, which should be closer to zero. I guess if I tried with dropping few rows of training data with Rating 10 and it could as well give a better score but didn't do that, which can be used to proves that the result is getting biased towards higher rating. Eventually I tried to reduce that biasness in the rating columns by subtracting the rating with – 0.95 (which Is very close to zero) and gradually changed it with different section. Rating between 5 to 7.5 are treated more finely due to presence of higher number of predictions. Just like parameter tuning had to perform plenty of trial and errors to get what is working best. The submission file is called ***Football_final_sub.csv***

6. Also, I tried to do features importance for both **catboost and lgbm** and lgbm score doesn't give a better result as compared to when I did it without feature importance. The R squared was 0.44 without features selection and the one with features selection gave 0.32 R squared value. While Catboost almost gave same r squared i.e. R2 value of 0.37.

7. I then take the average of both the Catboost and LGBM models with features Importance and the minimum rating point was 5.39 which looks much better than the model without feature interaction that has a value of 6.96 minimum rating_num. The submission file is called ***Football_both_new_features_selected_sub.csv.*** Trying to reduce possible biasness in the rating

columns by subtracting the rating with – 0.95 (which Is very close to zero) and gradually changed it with different section. Rating between 5 to 7.5 are treated more finely due to presence of higher number of predictions. Just like parameter tuning had to perform plenty of trial and errors to get what is working best doesn't really affect the model with feature Importance. The submission file is called ***Football_final_sub_2.csv***

8. So step 7 gave me the best results with ***Football_both_new_features_selected_sub.csv***