Approach : Analytics Vidya - Football Hackathon

by ABIR CHAKRABORTY (mail2abirchakraborty@gmail.com)

Data Exploration:

1. The train data is not uniformly distributed , there is large number of observation rated as 10. Which need to be treated to prevent biasness towards higher rating.
2. There are plenty of Null and Zero columns which needs to be removed.
3. There are plenty of highly corelated columns which needs to be removed.
4. The outliers of the numeric column should be treated.

Approach :

The approach was divided in the following parts

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

Feature Engineering:

By adding weight and height and convert it in the categorical columns was getting feature importance one in my catboost model. However this was performing extremely poor on training data so in my final model i dropped it.

Final Model and Submission:

Even after taking the average of both the Catboost and LGBM models the minimum rating point was still 1.42 , which should be closer to zero. I tried with dropping few rows of training data with Rating 10 and got better score. which proves that the result is getting baised towards higher rating. Eventually i found the below method to reduce that biasness. I started with -1.4 (minimum of my prediction was -1.42) 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 step by step code is submitted in the notebook which scored 0.3251 in the public leaderboard.