Documentation

1. **Approach**:

In the problem the first thing that caught my eye was dupicate column names, so I analyzed these columns and came to a conclusion that, I will use the other column to fill the missing values.

Then I looked for other features and tried to find the relationship between them, this methodology did help me by filling the missing values of columns like game\_season, home/away and lat/lng.

A single model is used for prediction of probabilities of ronaldo scoring a goal.

1. **Feature Generation/Data Pre-processing**:

I have formed two new features – home and away.

A new feature is made to capture the information about the match location whether it’s an home match or away match.

total\_time\_rem – It represents the total remaining time in seconds.

1. **Key Observations/Trends**:

From my manual interpretation i observed various relationship between the variables to fill the missing values.

I used match\_id to fill my various columns like home/away,lat/lng,game\_season.

Also, I assumed that remaining\_sec column can’t have a value more than 60, similary a match would not be longer than 90 mins, so i replaced the values in remaining\_sec and remaining\_min with null if they are greater than 60 and 90 respectively.

1. **Model Selection**:

I created a single model, I tuned the parameters intuitively, I wanted to use Grid Search CV for the best model parameters given by XGBoost classifier.

1. **Expected error for submission**:

Error may occur while reading the train and test file (‘yds\_train.csv’ and ‘yds\_test.csv’) if proper current directory is not specified.

1. **Ways to Improve**:

Due to time constraints, I couldn’t implement these ideas:

Firstly, the hyperparameters can be easily tuned using Grid Search CV, apart from that ensembling of different models can also help us improve the accuracy.