Soccer Player and Team Rating Model

Motivation – Rating player contributions on a uniform scale for a team sport is always challenging considering the different roles that each player can perform in a team setup. We focus on developing an objective system to rate player performance in a soccer game and in the process, look at a novel way to assess relative team strengths in a Head-to-Head clash.

Abstract – We look at soccer performances in terms of the points every contribution is expected to add to the total team points. To arrive at our model, we need to consider the variable which affects the number of points that a team earns from a given game; Goals. We need to model player contributions in terms of the change it brings about to the expected goals that a team scores. The validity of this type of rating system is checked and the rating system is modified so that it satisfies the benchmarks of a good player rating system. Next, we look at ways this individual rating can be used to model the team's overall strength. We restrict our rating system to the 2016/17 English Premier League season and try to draw inferences based on this data and the conclusions our model gives.

Data – The data we obtained was scraped from whoscored.com. The data includes game level stats for different players and aggregate stats for the entire team during every game of the 2016-17 season.

Model – For modelling goals in a given game, we consider the Poisson model because the number of goals by a given team in a soccer match is less. Hence, the final score-line is a form of double-Poisson model which is just two independent Poisson distributions multiplied together. The expected number of goals for a team is just the expected shots the team takes multiplied by the probability that a shot goes inside the goal. Player contributions, thus are measured in terms of the expected shots and the probability that a shot hit goes inside the goal. Expected shots is assumed to be a linear combination of in game player stats like, contests won, accurate passes given, aerials won, tackles made by the opposition, and yellow and red cards the opposition gets. The linear regression results are available in the Appendix. Expected points added by a contribution is given by taking derivative of the expected points with respect to the contribution multiplied by the number of contributions. Player rating is the sum over all the contributions' expected points added.

The rating obtained by the above model has a desirable property of ratings sum for all players approximately equals the total points. But, the rating thus obtained has very large variance from game to game. We expect the ratings to be a good predictor of skill and hence not be prone to unexpected game performances. Also, the ratings we have obtained doesn't give any weight to the actual game result which is another desirable property, in the sense that, player performances should be incentivized for positive team results. Other than this, goals, assists and clean sheets have no representation in our rating. We call our rating obtained previously as the pure rating, and to include the desirable properties mentioned above we take a weighted sum of our pure rating with that of the specific indexes mentioned above. One key property that we wish our ratings should retain is that of the sum being equal to the expected points, and hence we make sure all our indexes individually sum to expected points and the sum of all weights is equal to 1.

Model Test – One way to test whether our model does makes sense is to compile a top player list from our ratings and compare it with lists from other sources. Easiest way to create such a list is to rank players by their average ratings during the season. The downside of this method is it overlooks the

possibility of our model might be biased towards one type of player. Therefore, we need to find a way to reflect each player's strength based on their performance among similar players in the league. There are six major roles in a soccer game, which includes goalkeeper (GK), defender (D), defensive midfielder(DM), midfielder (M), attacking midfielder (AM) and forward (FW). To ensure we have enough sample size, we categorized these roles into four categories. Then we computed average rating and standard deviation for each of the four categories. Lastly, we gave each player a z-score based on his most played position during the season. We then repeated the same procedure on the players rating data from Whoscored and produced two lists, one average based ranking and one z-score based ranking. The are some noticeable differences when we compare our lists to the lists from Whoscored. For average based list, our top 20 players include 2 goalkeepers where whoscored's list has none. Whoscored's rating system gives offensive position such as forward a lot of credit and gives goalkeeper not enough credit. Our rating model is more unbiased.

Inferences – Since we've played around the idea of individual player's position, we also took a deeper look at each team's formations. In soccer, the formation describes how the players in a team are positioned on the field. Different formations can be used depending on whether a team wishes to play more attacking or defensive soccer. During 2016-2017 premier league season, 380 matches had been played and we saw 15 different formations. The most frequent used formation is 4-2-3-1 which is to be expected. 4-2-3-1 formation has become almost the standard in modern soccer football because of its balance and flexibility. The second and third popular formation are 4-3-3 and 4-4-2. Looking at each team's most used formation, we found some teams such as Liverpool and Leicester were very comfortable to play one formation most of the time for the entire season. Meanwhile, teams like Hull, Watford and Sunderland tried multiple positions during the season. These three teams were the bottom four at the end of 2016-17 season. It seems like the lack of identity in formations lead to their lackluster performances. By tracking the outcome of each match, we created a matchup table for all formations. 4-4-2 formation was the third most used formation and its matchup records were not great, falling short against almost all matchups. 3-4-3 was played the most by Chelsea that season. It had a winning record against its matchups and helped Chelsea won the championship. Therefore, we have to ask ourselves is there a real advantage to play 3-4-3 in premier league or it's simply because Chelsea's roster during that season was stacked. In what way would formation affect a team's performance? Due to lack of data, we could not derive more meaningful insights regarding this question.

Team Rating – The model for calculating expected number of shots takes into account both the relative strength for attacking team (aerials won, accurate passes, etc.) as well as defending team (opposition tackles). This gives an intuition for developing team model as an aggregate for all the players in a team as it takes into account defense vs. offense of both the teams involved. Moreover, the property of ratings sum equal to expected points makes team rating points in a match a good predictor of the game result. For building the expected number of goals a team scores against other team, we need the total expected shots that each team has which will be defined by the total contributions a team makes in a game. The total contributions for a team is equal to sum over all players, the average player contributions when he starts multiplied by the probability that the player starts plus the average contributions a player makes when he is substituted multiplied by the probability that the player is a substitute in a game. This method requires a team roster prediction where probability of every player starting and playing as a substitute is given. For our team rating model, we used historical player probability in each case. This suggests the model performance could have been improved if better predictions for team roster were available.

Findings – The team rating model was used to make game result predictions for every game of the season 2016-17. The model accurately predicted the result of 61.3% of the games. Contrast this with the home win rate of 49.7%, which means that if a model predicted home win every game, then the model would be correct 49.7% of times. Our model gives a considerable improvement to this. Also, betting data from bet365.com were obtained and implied probabilities from the betting data compared with that of those predicted by our model. In particular, if we were to bet on the winner predicted by our model in every game of the season, we would make a profit of 13% over the 380 games of the season.

Further Improvements – The model relies heavily on the calculation of expected shots and because the expected shots are assumed to be linearly related to contributions, this assumption needs to be separately verified. Though, we tried plotting scatter plots between the dependent (expected shots) and independent variables (contributions) we have been inconclusive in determining if the linear model is correct fit for this. Some improvement can be obtained if we use log linear model for shots as even shots data appear to be Poisson distributed. Also, if we have data related to key passes, crosses, long pass, etc. this data is shown to be more strongly correlated to the expected shots and we may have better results from our model. Contributions can also be exponentially weighted such that recent performances carry more weight in the rating model which might give robust predictions for the game.

Appendix –

Linear Regression Results –

	coef	std err	t	P>|t|
const	2.3668	0.919	2.576	0.010
Won Contest	0.1556	0.042	3.701	0.000
Accurate Pass	0.0241	0.001	16.748	0.000
Tackles(Opp.)	-0.0943	0.034	-2.780	0.006
Aerial Won	0.0895	0.025	3.593	0.000
Yellow(Opp.)	0.2471	0.130	1.907	0.057
Red(Opp.)	1.5112	0.480	3.151	0.002

Expected Number of points/Game = 3*P(Wins) + 1*P(Tie)

Player Rating -- $\sum contribution_i * Partial\_derivative\ (E[points])_i$

Indexes

- Index 1 – Model Rating
- Index 2 – Game result*Touch_Factor
- Index 3 – 1.37*Game_Time
- Index 4 – 1.02*Goals
- Index 5 – 1.02*Assists
- Index 6 – 2.73*Clean_Sheets

Weights --

37%*Index1 + 25%*Index2 + 15%*Index3 + 5%*Index4 + 5%*Index5 + 13%*Index6

Directory Structure –

.csv Files –

- E0.csv – file with betting data from various betting sites
- playerContributions.csv – All Players' contributions in different matches and their rating obtained from the model.
- playerNameId.csv – All Players' name and their corresponding Id
- playerPositions.csv – data regarding the positions that every player has played during the season.
- playerRatings.csv – Player ratings in all matches
- playerStartsMeanContributions.csv – average player contributions when starting a game
- playerSubsMeanContributions.csv – average player contributions when substituting in a game
- scoreFile.csv – file with all the full time and half time scores for the season.
- teamMeanContributions.csv – Average contributions for teams
- teamRoster.csv – Roster predictions for all team
- winProbabilities.csv – Model predicted win Probabilities and implied probabilities.
- Formations.xlsx – excel file with all the calculations for the formation analyses carried out.
- TopPlayerComparison.xlsx – excel file with z-score and average score comparison

Python Files –

- playerRating.py – Main python file which calls the latter files for calculating player ratings and generating all player data related csv files.
- winProbabilities.py – Main python file which calculates team win probabilities for all games and generates csv files related to it.
- Index.py – python file which calculates the team Rating.
- Regression.py – Python file which computes the linear regression for expected shots
- teamRating.py – Reads data from a csv file and computes score in readable format.

Raw Data –

- Inside folder xml_stats/season16-17
- season_match_stats.json – detailed match stats for every game of the season
- season_stats.json – match stats for all games of the season

folder /epl_data –

- contains the code for scraping whoscored.com

References –

Doi.org. (2017). *On the Development of a Soccer Player Performance Rating System for the English Premier League | Interfaces*. [online] Available at: https://doi.org/10.1287/inte.1110.0589 [Accessed 14 Dec. 2017].