Junyuan Gao

Research Proposal- Predicting final points of Premier League Teams using Linear Modeling techniques

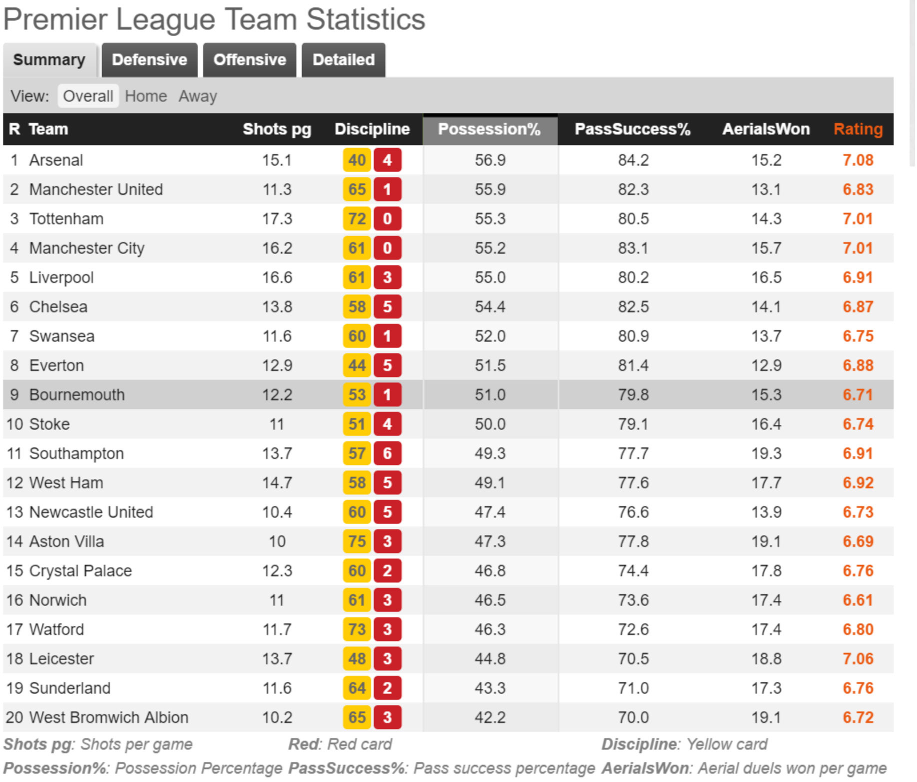
Motivation:

What makes a soccer team win the game? How could a team earn as much points as possible in a season to achieve the championship? These questions have puzzled professional soccer coach and fans for a long time. Professional coaches optimizes their teams' performance by designing various strategies and assigning tactics on different players. However as a statistics student with less professional knowledge and more statistical techniques, I would like to figure out how does factors from all aspects influence the final points within various models and data from the former seasons. Moreover, predicting final points after the 2016/17 match season within the current data and the best-fitting model will also be do-able.

Methodology:

1. Data collection

I'm going to collect historical performance data from https://www.whoscored.com/Statistics through the past 12 years, which records detailed performance in terms of different factors (such as possession rate, pass success rate, shots per game, etc.) All those data will be collected and classified into different types (offensive data, defensive data etc.)



Premier League Team Statistics

R	Team	Shots pg	Discipline		Possession%	PassSuccess%	AerialsWon	Rating
1	Arsenal	15.1	40	4	56.9	84.2	15.2	7.08
2	Manchester United	11.3	65	1	55.9	82.3	13.1	6.83
3	Tottenham	17.3	72	0	55.3	80.5	14.3	7.01
4	Manchester City	16.2	61	0	55.2	83.1	15.7	7.01
5	Liverpool	16.6	61	3	55.0	80.2	16.5	6.91
6	Chelsea	13.8	58	5	54.4	82.5	14.1	6.87
7	Swansea	11.6	60	1	52.0	80.9	13.7	6.75
8	Everton	12.9	44	5	51.5	81.4	12.9	6.88
9	Bournemouth	12.2	53	1	51.0	79.8	15.3	6.71
10	Stoke	11	51	4	50.0	79.1	16.4	6.74
11	Southampton	13.7	57	6	49.3	77.7	19.3	6.91
12	West Ham	14.7	58	5	49.1	77.6	17.7	6.92
13	Newcastle United	10.4	60	5	47.4	76.6	13.9	6.73
14	Aston Villa	10	75	3	47.3	77.8	19.1	6.69
15	Crystal Palace	12.3	60	2	46.8	74.4	17.8	6.76
16	Norwich	11	61	3	46.5	73.6	17.4	6.61
17	Watford	11.7	73	3	46.3	72.6	17.4	6.80
18	Leicester	13.7	48	3	44.8	70.5	18.8	7.06
19	Sunderland	11.6	64	2	43.3	71.0	17.3	6.76
20	West Bromwich Albion	10.2	65	3	42.2	70.0	19.1	6.72

Shots pg: Shots per game *Red*: Red card *Discipline*: Yellow card

Possession%: Possession Percentage *PassSuccess%*: Pass success percentage *AerialsWon*: Aerial duels won per game

2. Data analysis and tests of different models

After finishing step1 above, data will be divide into independent variables (goals per game, fouls per game, etc.) as predictors and dependent variable (i.e. final points after a season). After that, I will figure out correlations between dependent variables to see if there are any useless ones. Furthermore, various statistical models (OLS, logistic regression, etc.) will be tested and accuracy of models will be checked by comparing predictions and real-value. The best-fitting one will be selected to expect the 2016-17 Premier League result.