

# Adjusted Plus Minus for the English Premier League

## Abstract

I set out to create an adjusted plus minus for English Premier League (EPL) soccer player's that would reduce multicollinearity in evaluating the top players in a given season. Often a table of the player's with the 10 best plus minus' will be almost entirely comprised of players from the team that won the league that season. Some of these players deserve their spot at the top, but many are being dragged up by the coat tails of multicollinearity. I found that my adjusted plus minus helps dramatically decorrelate players from their team.

## Introduction

Plus minus is a common way of measuring a player's importance to their team and their individual impact in a game. Sports like basketball and hockey have been keeping track of player plus minus since the late 1960's. A player's plus Minus can be used in many ways, from making coaching decisions, to sports betting, to friendly MVP arguments. So if plus minus has been prominent in other sports for 60 years now, why has it not become a part of statistical analysis in soccer? Well lets start with what statistics are prominent in soccer: number of touches, tackles won, defenders beaten, passes made, shots on goal, goals, assists, fouls, and many more. There is an exhaustive list of stats in soccer, but the only simple one number summary I found was FIFA ratings, which are calculated from the opinions of some 9000 scouts, coaches, and players.

The comparisson I am about to make is true for many sports, but for simplicity I will make it for basketball. Basketball is high scoring, like really high scoring, in 2010 the average score of a basketball game was 100 points. The premier league (the top divison in England, and one of most popular and competitive leagues in Europe) in 2017 had an average goals per game of about 2.5. On any given day a basketball game is likely to score 40 times as many points as a soccer game. This means there is much less plus minus data in soccer than many other sports. Further, subsitutions are limited to three per game in the premier league. There is no substitution limit in basketball, and the average subsitutions per game is somewhere between 15 and 20. More subsitutions means there are more times during a game where different players from the same team are playing together, which helps to decorollate the data. Lastly, there are 38 games in a premier league season while there are 82 games per season in the NBA. More points, subsitutions, and games means basketball has much more higher quality data.

When we calculate a player's plus minus, and further their adjusted plus minus, there is likely to be much less collinearity between basketball players and their teams, because it is easier to isolate player performance. As you will soon see, the opposite is true in soccer. I hope to use player and match data from ESPN and FIFA over four (but really more like three and a half) seasons to create an adjusted plus minus for premier league players that reduces multicollinearity. FIFA ratings are a one number summary of a player's skill and they are aggregated from about 9000 coaches, professional scouts, players, and season ticket holders.

## Methods

ESPN has data on every EPL game for every recent season. I scraped their website to get information on who was on the field, if and when they were subsituted, and which team scored while they were on the field. I also scraped soFifa.com for player names, team, and FIFA ratings. This website has data on the top 600 players from each league for every day the FIFA ratings were updated. FIFA updates player ratings periodically throughout the season, so I decided to collect the data from the beginning of the season.

Table 1: Top Players by Plus Minus for 2016/17

	Player	PlusMinus	AdjustedPlusMinus	Team
87	Victor Moses	106	0.188956	Chelsea
219	Marcos Alonso	97	-0.058831	Chelsea
78	N Golo Kante	91	-0.022945	Chelsea
75	Thibaut Courtois	86	0.024667	Chelsea
77	Gary Cahill	84	0.013352	Chelsea
79	Cesar Azpilicueta	84	0.006228	Chelsea
85	Diego Costa	84	0.104601	Chelsea
172	David Luiz	80	-0.010754	Chelsea
81	Nemanja Matic	75	-0.123754	Chelsea
88	Pedro	75	0.050088	Chelsea

The ESPN data is a wide matrix where the first four columns are start, stop, home goal, and away goal. Start indicates the beginning of a period and stop indicates the end. Periods end when either a half ends (at 45 and 90 + extra time) or when a player is substituted out of the game. The rest of the columns are players, their rows filled with indicator variables defined as: 1 for being on the field playing for the home team, -1 for being on the field and playing for the away team, 0 for not being on the field.

I will use ridge regression for players on goal differential to calculate a penalized adjusted plus minus for every player. The resulting coefficients will be the adjusted plus minus, and I will use these coefficients to rank players. I will also use the adjusted plus minus to predict goal differential and plot the RMSE for each season.

## Results

With the ESPN data set I could calculate the score at any time in the game, by noting when a goal was scored and incrementing the respective team's score. I then calculated the goal differential at any period as (home score - away score). With this I could calculate a player's plus minus for any period as their value for that period (1, -1, or 0) multiplied by the difference between home goal and away goal. To calculate a player's plus minus over a particular game I simply summed their plus minus for every period in the game, and to get their plus minus for the entire season I did the same thing for every period of every game. This may seem alarming at first but if a player wasn't playing in a game (either on the bench or not on either team) their indicator variable would be 0 so any goals scored would not contribute to their plus minus. So I summed these multiplications for every player column to calculate each player's season plus minus. I did this for each season in my data set (2016/17 - 2019/20) and made a table of the top 10 players by plus minus for each season. These tables make the multicollinearity problem very clear. For reference, a win is three points, a draw is one point and a loss is zero points.

For the 2016/17 season when Chelsea won the title by 7 points (Table 1), a large margin but not an alarming one, every single player in the top 10 of plus minus was a Chelsea player. In a season with many other superstars like: Sadio Mane, Alisson, Marcos Alonso, and Salah to name a few.

The same is true for the 2017/18 season when Manchester City won the title by 19 points over second place Manchester United. The only player in the top 10 who did not play for Manchester City was Paul Pogba, a generational talent who was Manchester United's star player. The only table that is different is for the 2018/19 season, the top 5 player's all play for Liverpool and the next 5 all play for Manchester City.

The 2018/19 season was entirely dominated by these two teams, the third place team was 25 points behind second place, 9 wins behind Liverpool and 11 wins behind Manchester City. This was a record breaking season where the competition was completely outclassed and it was a clear two dog race from the beginning.

Finally, the unfinished 2019/20 season has been dominated by Liverpool, 29 games in and Liverpool is sitting comfortably 25 points above second place Manchester City. Although the season is only 3/4ths the way

Table 2: Top Players by Adjusted Plus Minus for 2016/17

	Player	PlusMinus	AdjustedPlusMinus	Team
307	Andre Gray	6	0.872721	Burnley
5	Sam Byram	11	0.697236	West Ham United
149	Fernando	10	0.635637	Manchester City
332	Jesse Lingard	2	0.633663	Manchester United
306	Ashley Barnes	-6	0.534991	Burnley
288	Tom Davies	1	0.521766	Everton
279	Wilfried Bony	10	0.483763	Manchester City
206	Son Heung	20	0.435440	Tottenham Hotspur
318	Jarrod Bowen	0	0.384752	Hull City
65	Alberto Moreno	6	0.372847	Liverpool

Table 3: Top Players by Plus Minus for 2017/18

	Player	PlusMinus	AdjustedPlusMinus	Team
100	Kevin De Bruyne	29	0.001533	Manchester City
103	Kyle Walker	28	0.001548	Manchester City
95	Ederson	27	0.001344	Manchester City
97	Nicolas Otamendi	25	0.001265	Manchester City
108	Raheem Sterling	24	0.001540	Manchester City
99	Fernandinho	20	0.001335	Manchester City
102	Leroy Sane	20	0.001322	Manchester City
101	David Silva	19	0.001373	Manchester City
34	Paul Pogba	16	0.001097	Manchester United
105	Sergio Aguero	16	0.001683	Manchester City

Table 4: Top Players by Adjusted Plus Minus for 2017/18

	Player	PlusMinus	AdjustedPlusMinus	Team
332	Aymeric Laporte	7	0.001922	Manchester City
31	Daley Blind	6	0.001816	Manchester United
325	Oleksandr Zinchenko	11	0.001801	Manchester City
330	Sebastian Prodl	4	0.001711	Watford
289	Ragnar Klavan	9	0.001696	Liverpool
105	Sergio Aguero	16	0.001683	Manchester City
103	Kyle Walker	28	0.001548	Manchester City
96	John Stones	6	0.001541	Manchester City
108	Raheem Sterling	24	0.001540	Manchester City
100	Kevin De Bruyne	29	0.001533	Manchester City

Table 5: Top Players by Plus Minus for 2018/19

	Player	PlusMinus	AdjustedPlusMinus	Team
107	Alisson	28	0.001501	Liverpool
108	Virgil van Dijk	28	0.001501	Liverpool
110	Andy Robertson	28	0.001518	Liverpool
116	Sadio Mane	26	0.001654	Liverpool
117	Mohamed Salah	25	0.001480	Liverpool
66	Ederson	23	0.001474	Manchester City
75	Raheem Sterling	23	0.001686	Manchester City
70	Kyle Walker	21	0.001508	Manchester City
71	Bernardo Silva	21	0.001441	Manchester City
73	Ilkay Gundogan	21	0.001785	Manchester City

Table 6: Top Players by Adjusted Plus Minus for 2018/19

	Player	PlusMinus	AdjustedPlusMinus	Team
79	Kevin De Bruyne	12	0.002937	Manchester City
219	Rob Holding	6	0.002634	Arsenal
217	Son Heung	15	0.002493	Tottenham Hotspur
149	Lucas Moura	13	0.002378	Tottenham Hotspur
120	Daniel Sturridge	9	0.002115	Liverpool
113	Naby Keita	16	0.002094	Liverpool
147	Eric Dier	11	0.002061	Tottenham Hotspur
142	Mousa Dembele	5	0.002047	Tottenham Hotspur
151	Serge Aurier	4	0.002006	Tottenham Hotspur
330	Oleksandr Zinchenko	13	0.001904	Manchester City

Table 7: Top Players by Plus Minus for 2019/20

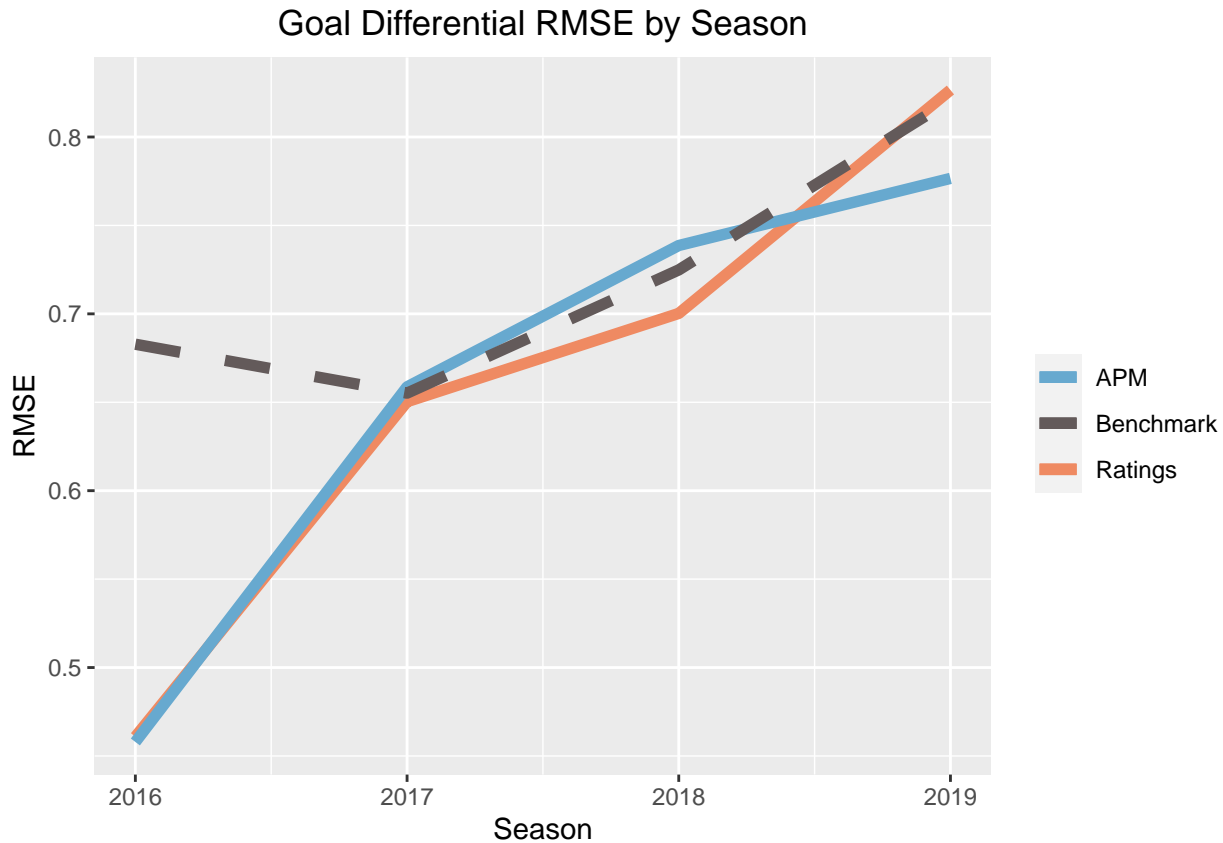
	Player	PlusMinus	AdjustedPlusMinus	Team
2	Virgil van Dijk	34	0.001503	Liverpool
4	Andy Robertson	34	0.001532	Liverpool
5	Trent Alexander	33	0.001504	Liverpool
7	Georginio Wijnaldum	33	0.001624	Liverpool
8	Jordan Henderson	30	0.001766	Liverpool
9	Roberto Firmino	30	0.001531	Liverpool
11	Mohamed Salah	28	0.001607	Liverpool
14	Sadio Mane	27	0.002022	Liverpool
1	Alisson	24	0.001813	Liverpool
3	Joe Gomez	23	0.001603	Liverpool

Table 8: Top Players by Adjusted Plus Minus for 2019/20

	Player	PlusMinus	AdjustedPlusMinus	Team
310	Dejan Lovren	5	0.002224	Liverpool
10	Divock Origi	14	0.002183	Liverpool
14	Sadio Mane	27	0.002022	Liverpool
368	Shkodran Mustafi	5	0.001895	Arsenal
1	Alisson	24	0.001813	Liverpool
8	Jordan Henderson	30	0.001766	Liverpool
168	Xherdan Shaqiri	3	0.001724	Liverpool
234	Marcos Alonso	6	0.001690	Chelsea
7	Georginio Wijnaldum	33	0.001624	Liverpool
11	Mohamed Salah	28	0.001607	Liverpool

through, Liverpool is within a win or two from winning the league.

The following are tables of the top 10 players for each season by adjusted plus minus. As expected the tables are somewhat decorrelated by team, and also include more of the traditionally good players from their respective season. The adjusted plus minus also helps to accomodate for playing time, a player who played excellent for part of the season before getting hurt likely would not make the top 10 in plus minus because they had less games played. This player however could crack the top 10 in adjusted plus minus.



This is a plot of RMSE for goal differential, calculated using ridge regression on an unused testing set for each season vs a benchmark that simply guesses the average goal differential. As expected the 2019/20 season has the largest RMSE, because the season is ongoing. While the 2016/17 season has lowest RMSE, because there is nearly three times as much match data for this season. In the 2016/17 season both models greatly outperform the benchmark, unfortunately the rest performed similarly to the benchmark.

## Conclusion

In conclusion traditional plus minus can be extremely misleading in soccer because of its high degree of multicollinearity. Players who had little impact on title winning teams will have large plus minus' simply because they played for a winning team. Because of this we must turn to another form of one number summary. Here I have calculated an adjusted plus minus using ridge regression to penalize player contributions. This APM (adjusted plus minus) succeeds in decorrelating the data, however the significance of player coefficients should be questioned. Running the same regression multiple times on different seeds will often lead to significant changes in a player's APM. In my experience the top 10 APM is relatively stable, as many of these players had distinct and important impact on their team's success. In addition the "super sub" (a player who is subbed on late in a game and makes a scoring impact) is highly weighted by the APM because they tend to have a large goal to time played ratio. The 2016/17 season has the most stable APMs and the best predictive ability, because the match data has nearly 1700 observations, compared to the 500-600 in the other seasons. In summary APM is likely a better measure of player importance than PM in soccer as it helps decorollate players from their team. APM can certainly seperate the good from the bad players, however small differences in APM should not be overstated, it is likely that a small change in APM is statistically insignificant, meaning that APM is a bad measure for deciding which players of a certain tier are better than others. If soccer players can be sorted into great, good, and bad tiers, APM does a decent job of placing player's into their respective tier but does a poor job of ranking them within that group.