European Football Team Rank Predictions

Jacob Hodges, Sam Ridderhoff, Anandhan Manoharan, Aidan Sherlock

December 6, 2023

1 Executive Summary

At the absolute top level, teams are in constant competition to finish above one another in their respective league. All teams are constructed differently. Some teams are very young and like to attack, other are more experienced and prefer to sit back and defend. This study aims to address what elements of a team are most correlated to their success. In order to dive into this, we use model building with Ordinal Logistic Regression. Our most successful model used Goals Scored, Goals Against, a team's Average Age, their Possession Rate, and their Annual Wages on a log scale to give log-odds of finishing in different positions on the table. Based on our model, we think the best way to increase a team's rank is by spending more money on defenders as our Annual Wages and Goals Against Coefficients were highest in magnitude. Based on our model, teams should focus more on being solid and not letting in goals then playing a heavy attacking setup with high press and free roaming attackers. For further research, we encourage incorporating more seasons of data to produce a more trained model. We are also interested in if results vary based on different leagues, as our model combines the top five European Leagues.

2 Problem Context

Every year, teams in their respective leagues fight to be high in the standings (more commonly known as the table in soccer), but how do we know which teams will finish the highest on the table? It is easy for us to say that teams with better players, coaches, more fans, etc. will finish higher in the table, but is there a more discernible way of saying this? Clearly teams that score a lot of goals are more likely to finish higher than a team that struggles to score. Conversely, conceding more goals will reduce the success of a team. We know that teams with better players are likely to have more success in the league, but how do you get these good players? The quickest way to do this is to go out and buy them from other teams. This line of reasoning tells us that teams who spend more money will probably end up doing better in their league. Another thing we can consider: does a team need older players with more experience or younger, more athletic players to produce wins? Do different leagues lead to different models to predict success? Is the amount of time in a game a team has

the ball factor into wins? All of these ideas prompt us to our research question: Can we predict a team's final league position based on the number of goals they score/concede, how much money they spend, where they play, their average age, and their possession rate? If we are able to create a model that helps predict a team's position, this gives us insight as to which teams we expect to win and which we do not. We will also understand the influence of several factors into a team's success and their relative importance against each other. Should we concentrate our efforts into scoring many goals or defending the other team from doing so? Should we focus on spending money towards acquiring our players or rather focus on our own player development? Answering questions like these will become feasible by diving into this research project.

The data that we used was collected from the website, Football Statistics and History (FBref) [1]. The explanatory variables we used were goal difference which is the number of goals scored minus the number of goals given up by a team, the number of expected goals defined as the probability of a shot resulting in a goal, the amount of money a team spent measured by the weekly wages and annual wages paid out by the team, the average age of a team, and the teams average possession over a season. The response variable we measured is the league position of a team at the end of the regular season. For most of the leagues we are looking at there are 20 teams, so teams can earn a rank of 1 through 20, except for the German league, Bundesliga, which has 18 teams, so the teams can earn a rank of 1 through 18. In total we have collected three of the past seasons for the five major European leagues, the English, Spanish, Italian, French, and German leagues resulting in 294 observations. We didn't find any clear outliers present in the data.

3 Methods and Data Analysis

To answer our research question, we utilized an ordinal logistic regression model. Ordinal logistic regression is a method that is used to analyze the relationship between a set of factors, our explanatory variables, and an outcome that is naturally ordered or ranked in its categories. Our response variable, a team's final position in the league standings, is categorized as ordinal data since the final ranking of a team is naturally ordered. Doing ordinal logistic regression allowed us to analyze the probability of a given team finishing in any given position in the table. With the probability of teams finishing in any given position, we will be able to further explore what are the key factors that differentiate teams from each other and why our model gives the expected probabilities.

For model 1 we decided to analyze Wins, Losses, Expected Goals Against, Expected Goal Difference, Goals For, Possession, and Annual Wages. Our model 2 we analyzed Goals For, Goals Against, Age, Annual Wages, and Possession. The assumptions for an ordinal logistic regression are:

1. Making sure our dependent variables are ordered.
2. One or more of our independent variables are either continuous categorical, or ordinal.
3. No multicollinearity
4. Proportional odds

Log-Odds Equation for Model 1

$$\begin{aligned}\log(\frac{P_i}{1-P_i}) = a_i - 0.84114x_{Wins} + 0.58123x_{Losses} + 0.14093x_{ExpectedGoalsAllowed} \\ +0.09162x_{ExpectedGoalDifference} - 0.05205x_{GoalsScored} - 0.06416x_{Possession} \\ -0.16939x_{\log(AnnualWages)}\end{aligned} \tag{1}$$

For our first model, the first assumption is satisfied because our dependent variables are ordered as they represent the standings in a league table. Our second assumption is also satisfied as we found that our independent variables are either continuous, categorical, or ordinal. More specifically, Possession, Annual Wages, and Expected Goal Difference are continuous variables and the wins and losses of a team are categorical variables. When testing for our third assumption of multicollinearity, we ran the variance inflation factor and found that the expected goal difference variable had a variance inflation factor of 4.12. Based on the variance inflation factor rule of thumb, we can conclude that this model passes the multicollinearity assumption. The final assumption tests for proportional odds and to test for this we conducted the Brant Test and found that the Wins, Expected Goals Allowed, and Possession variables violated our proportional odds assumption.

Log-Odds Equation for Model 2

$$\begin{aligned}\log(\frac{P_i}{1-P_i}) = a_i - 0.188x_{GoalsScored} + 0.2034x_{GoalsAgainst} - 0.1056x_{Age} \\ -0.04297x_{Possession} - 0.2415x_{\log(AnnualWages)}\end{aligned} \tag{2}$$

For the second model, both of the first two assumptions are satisfied since the model uses the same response variable and also uses continuous variables. For the assumption of multicollinearity, looking at a scatter plot matrix of the model, it looked like there could be multicollinearity between certain variables, specifically Goals For and Possession, however when looking at the variance inflation factors, there does not appear to be multicollinearity as all the variance inflation factors are around one, so that assumption holds. For the last assumption of proportional odds, we conducted the Brant Test and found that Goals For, Goals Against, and Age violated the assumption of proportional odds.

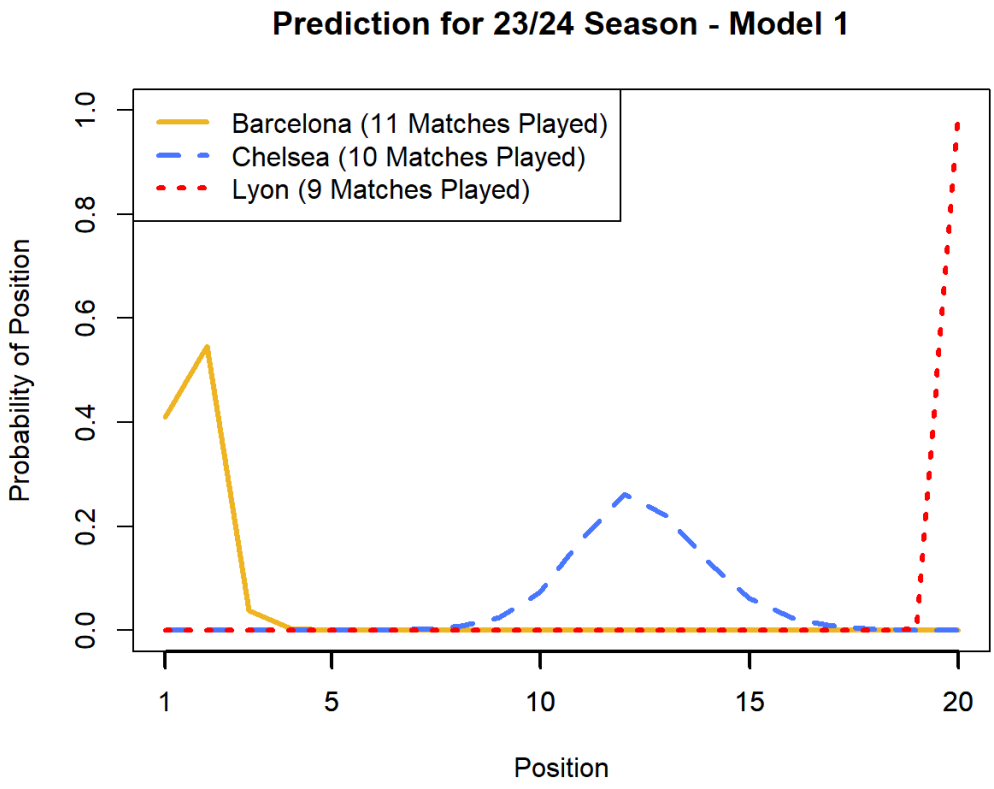

Figure 1: This is a prediction based on model 1 for 3 teams based on games played so far this season.

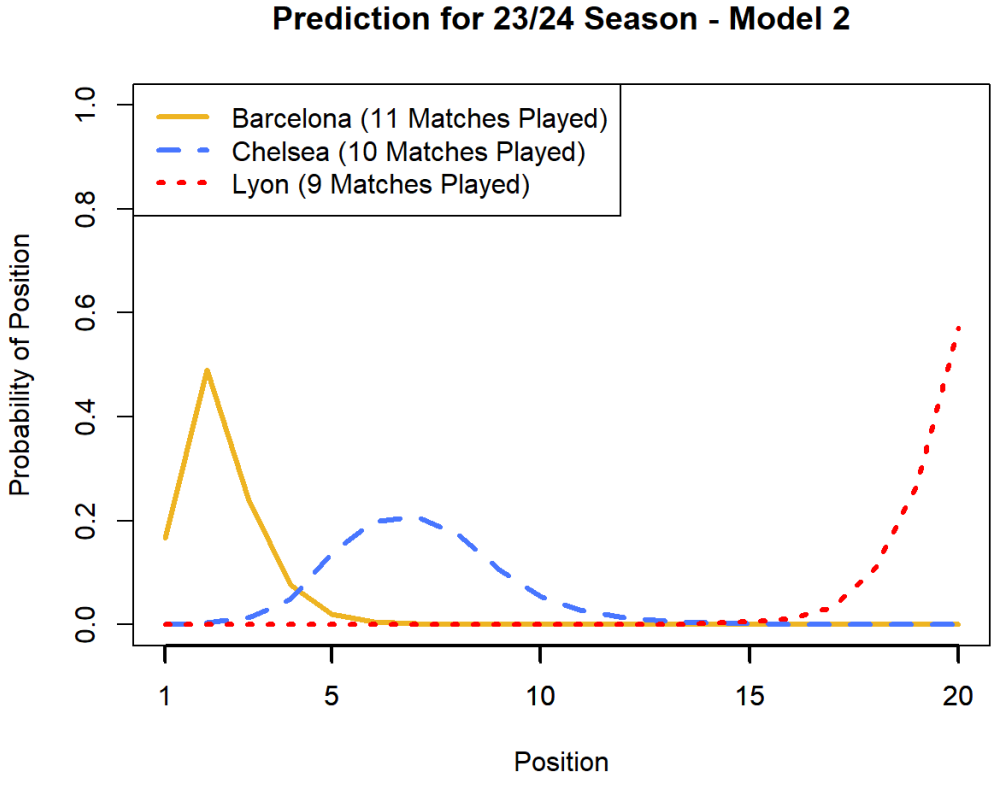

Figure 2: This is a prediction based on model 2 for 3 teams based on games played so far this season.

In these images, we are using our model to predict the final rank of a team in the current season using both of our models. We chose three teams from three different leagues, Barcelona from Spain, Chelsea from England, and Lyon from France. We also had to adjust the statistics to reflect what they would look like having played a full season. In the first model, since it incorporates wins and Lyon currently has no wins through the first nine games, the probability of them ending up in last is very high. The placement of Barcelona is similar in both of the two models, while the placement of Chelsea varies greatly between the two. This could be due to the inclusion of wins and losses in the first model, which lowers Chelsea's projected rank, even though they may have high annual wages and decent stats, their mediocre wins and losses record causes their placement to be much lower in the first model.

Squad	Rank
Arsenal	1
Manchester City	2
Newcastle Utd	3
Liverpool	4
Aston Villa	5
Tottenham	6
Brighton	7
Chelsea	8
Manchester Utd	9
Brentford	10

Figure 3: Top 10 of the English Premier League using Model 2

In the previous table, we used the second model to rank the teams in the premier league in the current season based on their stats thus far. With our model, it predicts that Arsenal will stay in first place with Manchester City in second, and Newcastle United will jump up from sixth, currently, to third at the end of the season. We won't know how accurate these predictions are until the end of the season, however, based on our knowledge of the English league right now, none of the predictions are extremely controversial. It will be interesting to see at the end of this season, how close our predictions will be to the actual rankings.

4 Conclusions

Through our research, we identified which facets of the game have the most effect on a team's position in the table, come the end of the season. By observing the coefficients in our models, the number of wins and losses were of great importance in Model 1, with coefficients of -0.84 and 0.58 respectively. Meanwhile, the amount of possession mattered little, with a coefficient of -0.06. However, Model 2, which excluded wins and losses, saw Goals Scored, Goals Against, and Annual Wages become more important. Although the most important variable in this second model was Annual Wages, Goals Against was second over Goals Scored, Age, and Possession, in which Goals Against has a coefficient of 0.20. Although the amount of goals scored is how most teams are compared and seems to be attacking players are often regarded as the most important, limiting the amount of goals you concede has the largest positive impact on a team's position. This follows the famous Manchester United manager Sir Alex Ferguson's phrase, "Offense wins you games, but defense wins you titles".

5 References

[1] Football statistics and history. FBref.com. (n.d.). https://fbref.com/en/