English Premier League (EPL) - Predict Winner

Ashish Vinodkumar

11/19/2020

Summary:

The following report is looking to predict who is going to win the 2015-2016 english premier league season, along with identifying key features that influence a team's ability to win the league. The report details building a hierarchical linear regression model, to account for all necessary predictors and any potential interactions that would improve the prediction ability of the model. The main takeaway from this report will be to assess the key features, and prediction accuracy of the model.

Introduction:

English Premier League is the top soccer league in England which is contested by 20 teams each season. Each of these 20 teams play the other teams twice, once at their home stadium and once at their opponent's home stadium (usually referred to as 'away'). A win in a match between 2 teams results in 3 points, a draw results in 1 point, and a loss results in 0 points. As teams play each of the other teams in a round robin fashion, an official points table gets updated after each game to keep track of who is currently leading the points tally. At the end of the season, the team with the most points, is crowned champions of the english premier league for that given season. In this race to first place in the premier league, a team's attributes such as defensive style, whether they press high up the pitch or sit deep in their own half, passing style, whether they make long diagonal passes or short and quick passes, and chance creation, whether a team is creating enough chances to score goals in the final third of the pitch, and many more such predictors determine the quality of the team throughout the entire season. I aim to incorporate all of these predictors to answer the following questions:

- What are the key features that influence a team's ability to win an english premier league season?
- Assess the hierarchical linear regression model's prediction accuracy both in-sample and out-of-sample. For the out-of-sample prediction, the report details predicting the winner of the 2015-2016 EPL season and comparing the result against the expected/actual winner.

The article is organized into the following sections, Data and EDA, Model Selection, Results, and Conclusion. I aim to outline key predictors and trends in the EDA section, identify the best model in the Model Selection section, and answer the key research questions in the Results and Conclusion sections.

Data and EDA:

The dataset for this report was obtained from Kaggle's 'European Soccer Database'. This dataset contains data for all the top soccer leagues across Europe ranging from 2008 to 2016. Given the large dataset size, I spent a lot of time transforming the data by first identifying and filtering the dataset to only contain records pertaining to EPL. I then proceeded to replace team ids with the corresponding team's name, and derive win, draw, and loss for every team based on the number of goals scored by the home team compared to that of the away team. Having derived the outcome of every game, I proceeded to derive my response variable of "overall win percentage", by calculating the weighted average between total wins and total draws. As detailed in the Introduction section, a win is naturally worth 3 times as much as a draw. After deriving my response, overall win percentage, I proceeded to identify teams that appear at least 4 times across all the EPL seasons in my final dataset ranging from 2009-2015. This was done because the bottom 3 teams in the premier league get relegated to the lower division, and the top 3 teams get promoted from the lower division. As a result, 3 teams get replaced each season in the EPL and I wanted to ensure that when I create a hierarchy based on teams, there is enough data for each team to build an accurate model. Upon using my prior knowledge and intuition about the EPL and after

checking with Professor Akande, I have chosen 4, to be the cutoff for the least number of times a team must appear in the data ranging from 2009-2015, to be a part of my final dataset. Finally, I merged this dataset with all of the other season level attributes for each team, to create my final dataset. The final dataset had 88 records, where each record corresponded to a year/season and team between 2009-2015. Moreover, I decided to take a log transformation for the response variable, overall win percentage, because the original distribution (left) was skewed to the right and was not normally distributed.

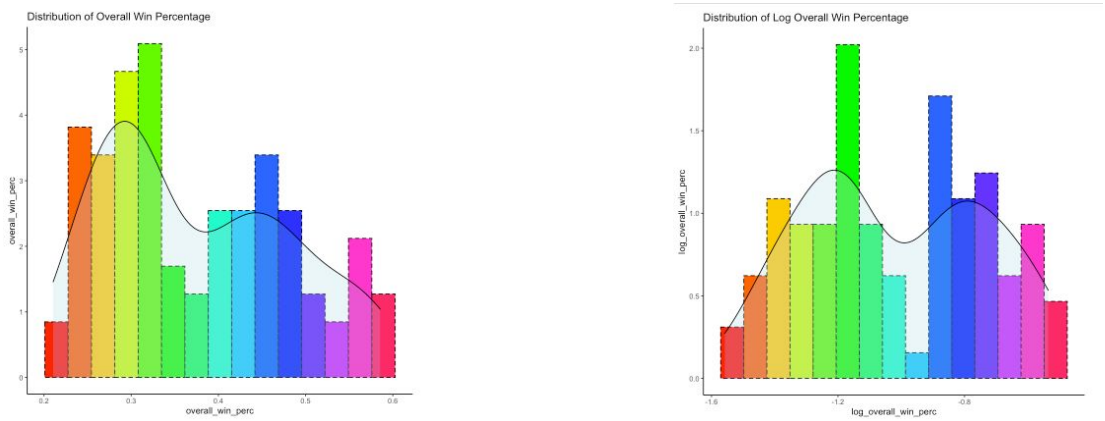

Figure 1. Distribution of response variable before and after log transformation.

I noticed that even after the transformation, there are two peaks in the distribution. I discovered that the two peaks are likely caused by teams either battling to stay away from relegation or battling to be the champion. As mentioned earlier in the introduction section, the bottom 3 teams get relegated to the lower division. I believe the fierce competition on either end of the points table-race to be league champions and race to prevent relegation, is responsible for causing the two peaks observed in the histogram. As a result, I was convinced that this is the nature of the dataset and no more transformation methods can fix the issue.

I also explored the effects of each predictor against the response variable, along with any potential interactions. The most notable EDA was the interaction between ChanceCreationCrossing and buildUpPlayPassing. ChanceCreationCrossing determines the number of times in the entire season that a given team creates a goal scoring chance by crossing the ball into their opponent's half. BuildUpPlayPassing signifies the style of play leading up to a chance, ranging from long diagonal passes, mixed balls (combination of long and short passes), and short passes.

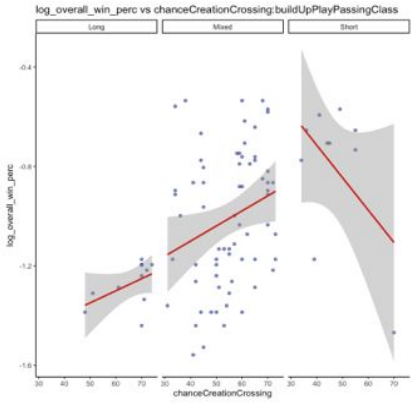

Figure 2. log(overall win percentage) vs ChanceCreationCrossing : BuildUpPlayPassing

As you can see in the above plot, for both long and mixed BuildUpPlayPassing levels, as the ChanceCreationCrossing increases, the overall win percentage also increases. However, when BuildUpPlayPassing is Short, and as the ChanceCreationCrossing increases, the overall win percentage decreases. I took note of this interesting interaction, to further analyze if this trend is statistically significant in the Model Selection section.

Model Selection:

I began model selection with a null model utilizing log(overall win percentage) as the response variable, 'team' as the random effect hierarchy level, and including the predictors BuildUpPlaySpeedClass, buildUpPlayDribblingClass, ChanceCreationCrossing, ChanceCreationShootingClass and defenceAggressionClass as part of my null model based on my EDA. I also created a full model with the same specifications as the null model, except now including all predictors in my dataset. I performed an anova test between the null model and my full model and noticed that the p-value of the full-model was significant at 0.00647, and was less than the alpha threshold of 0.05. I then proceeded to create 2 more hierarchical models that contained all the specifications as the Full-Model, but one of them contained (Model1) the interaction between ChanceCreationPassing and DefenceAggression, and the other contained (Model2) an interaction between ChanceCreationCrossing and BuildUpPlayPassing as observed during my EDA process. I performed an anova test between the Full Model and each of these 2 new models, Model1 and Model2. I noticed that the p-value of Model1 was significant compared to both the Full-Model and Model2. As a result, I proceeded with Model1 as my final model.

$$\begin{aligned}\text{Log}(Overall\_Win\_Perc_{ij}) \;=\; (\beta_0+\gamma_{0j}) + \beta_1 buildUpPlaySpeedClass_{ij} + \beta_2 buildUpPlayDribblingClass_{ij} + \\ \beta_3 buildUpPlayPassingClass_{ij} + \beta_4 buildUpPlayPositioningClass_{ij} + \beta_5 chanceCreationCrossingClass_{ij} + \\ \beta_6 chanceCreationShootingClass_{ij} + \beta_7 chanceCreationPositioningClass_{ij} + \beta_8 defencePressureClass_{ij} \\ +\beta_9 defenceAggressionClass_{ij} + \beta_{10} defenceAggressionClass{:}\,chanceCreationPassing_{ij}\end{aligned}$$

$$\begin{gathered} i = 1, \ldots n_j \,; \quad j = 1, \ldots J \\ \varepsilon_{ij} \sim N(0, \sigma^2) \\ \gamma_{0j} \sim N(0, \tau_0^2) \end{gathered}$$

After creating the final model I proceeded to assess the model assumptions:

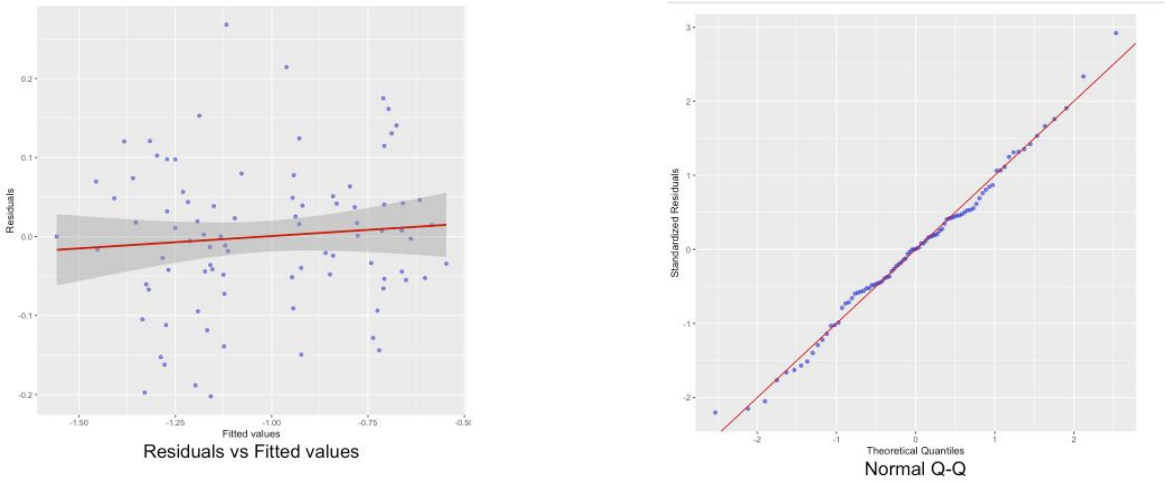

Figure 3. Plots of Model Assumptions Validation

I looked at the covariance matrix to check the model for multicollinearity and redundant variables. I was satisfied with the low covariance value and went on to check the model assumptions. I created graphs to check Linearity, Independence and

Equal Variance, and Normality. There was only one continuous predictor in the model, ChanceCreationPassing, and upon inspection of the graph, there was no visible underlying pattern. As a result, the check for linearity was successful. I then proceeded to plot the Residuals vs Fitted to check for Independence and Equal Variance, looking at the Figure 3 above (left), the data is randomly distributed across both sides of the line and there is no visible pattern in the data. As a result, the check for Independence and Equal Variance is successful. Looking at the Figure 3 above (right) for Normality, after taking the log transformation of the response, overall win percentage, the assumptions for normality considerably improved as the data points are very close to the 45 degree diagonal line in the plot.

Result:

According to the regression table on the right, my analysis sought to identify the key features that influence the overall win percentage for a team during an entire season of EPL. It is observed that BuildUpPlayPassing, DefencePressure, DefenceAggression, ChanceCreationPassing, and the interaction between DefenceAggressionClass and ChanceCreationPassing are all significant. One interesting observation is that surprisingly BuildUpPlaySpeed, BuildUpPlayDribbling, and ChanceCreationShooting are not significant. Based on my understanding of the game, I initially thought that the speed with which a team moves from defence to attack, how well they are able to dribble the ball, and how many shooting chances they create will influence the team's overall ability to win the EPL season. However, I was amazed to find out that quality of play detailed by the build up play passing between long, short and mixed passes, how well the team is able to press and showcase aggression when they lose the ball to win it back, and creation of goal scoring chances via passes, have a much higher influence on the game and entire season.

As a result, when the BuildUpPlayPassing is Mixed, the overall win percentage increases by 17.35%. Similarly when BuildUpPlayPassing is Short, the overall win percentage increases by 41.90%. When the BuildUpPlayPositioning is Organized, the overall win percentage increases by 13.88%. As noted above, when the DefensePressureClass is Medium, the overall win percentage decreases by 10.42%.

I further performed an in-sample prediction to validate the accuracy of the model between the actual winners between the 2009-2015 seasons, and the predicted winners based on my model. The model predicted the EPL season winner accurately for 2010/2011, 2012/2013, and 2013/2014. Therefore, it had a 50% accuracy for in-sample prediction. Please find the table below that details in-sample prediction validation:

	log_overall_win_perc		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	27.55	6.73 – 48.38	**0.010**
buildUpPlaySpeedClass [Fast]	0.05	-0.02 – 0.12	0.183
buildUpPlaySpeedClass [Slow]	-0.14	-0.35 – 0.07	0.194
buildUpPlayDribblingClass [Normal]	-0.04	-0.10 – 0.03	0.277
buildUpPlayPassingClass [Mixed]	0.16	0.05 – 0.26	**0.003**
buildUpPlayPassingClass [Short]	0.35	0.16 – 0.53	**<0.001**
buildUpPlayPositioningClass [Organised]	0.13	0.00 – 0.27	**0.049**
chanceCreationCrossingClass [Lots]	0.18	-0.01 – 0.38	0.066
chanceCreationCrossingClass [Normal]	0.16	-0.02 – 0.35	0.084
chanceCreationShootingClass [Lots]	-0.06	-0.22 – 0.10	0.480
chanceCreationShootingClass [Normal]	-0.09	-0.22 – 0.05	0.212
chanceCreationPositioningClass [Organised]	-0.08	-0.18 – 0.03	0.145
defencePressureClass [Medium]	-0.11	-0.22 – -0.00	**0.043**
defenceAggressionClass [Double]	-29.61	-50.43 – -8.79	**0.005**
defenceAggressionClass [Press]	-28.68	-49.50 – -7.85	**0.007**
chanceCreationPassing	-0.51	-0.87 – -0.14	**0.006**
defenceAggressionClass [Double] * chanceCreationPassing	0.52	0.16 – 0.88	**0.005**
defenceAggressionClass [Press] * chanceCreationPassing	0.50	0.14 – 0.87	**0.007**
Random Effects			
σ^2	0.01		
$\tau_{00\ team}$	0.05		
ICC	0.81		
$N_{\ team}$	16		
Observations	88		
Marginal R^2 / Conditional R^2	0.158 / 0.837		

Season	Actual Winner	Predicted Winner	log_overall_win_perc	y_pred
2009/2010	Chelsea	Manchester United	-0.5812293	-0.5467706
2010/2011	Manchester United	Manchester United	-0.6418539	-0.6390876
2011/2012	Manchester City	Manchester United	-0.5352442	-0.6967373
2012/2013	Manchester United	Manchester United	-0.5352442	-0.6758864
2013/2014	Manchester City	Manchester City	-0.5695332	-0.6157727
2014/2015	Chelsea	Manchester City	-0.6544327	-0.6019508

I similarly performed an out-of-sample prediction, to predict the winner of the 2015-2016 EPL season. Below is the result:

Season	Actual Winner	Place	Predicted Winner	log_overall_win_perc	y_pred
2015/2016	Leicester City	1st	Arsenal	-0.7612006	-0.6727761
2015/2016	Arsenal	2nd	Manchester City	-0.8342258	-0.7003402

I believe that the model prediction accuracy can be greatly improved by including all the seasons played in the premier league so far and not just between 2009-2015. This would allow the model to train on all teams' cyclical (success and failure) performance over the seasons, allowing the model to then make more accurate predictions. It would also be beneficial to explore the playing 11's overall per game rating for the entire season, and average it. I believe this would also improve the prediction ability of the model as certain players have an immense ability to single-handedly drive the game. These "star" players will consistently have higher overall ratings throughout the entire season, thus influencing the team's overall performance, and the team's overall win percentage. With the out of sample prediction, it is important to note that Leicester City were not a part of my training dataset. This was a deliberate choice as detailed in the Data section above, where only teams that appeared at least 4 times in the dataset across 2009-2015, will be retained in the final dataset. Given that the model predicted Arsenal to win the league, it is a fairly reasonable prediction considering Leicester City experienced a meteoric rise starting with the 2015-2016 season, where they went from a team that was constantly in the mid-table and relegation battle, to a team that was contending to be premier league champions. The 2015-2016 season was labelled as one of the most unpredictable seasons in the history of the english premier league.

Conclusion:

Using the hierarchical linear regression model, this report tried to explore and identify the key predictors that influence the overall win percentage for a team. The analysis identified BuildUpPlayPassing, DefencePressure, DefenceAggression, ChanceCreationPassing, and the interaction between DefenceAggressionClass and ChanceCreationPassing as significant predictors that influence a team's ability to win the premier league season. I also sought to perform an out of sample prediction with predicting the winner of the 2015-2016 season. My model predicted Arsenal to have the highest overall win percentage, and given the unpredictability of the 2015-2016 season and Leicester City's meteoric rise, the model predicted reasonably well by selecting Arsenal, the next best team as champions for the 2015-2016 season. However, I would argue that there are limitations to the model. Firstly, only 6 years of data ranging from 2009-2015 is not sufficient to make reliable predictions. Also, by retaining teams that only appear at least 4 times in the dataset, it eliminates teams like Leicester City. Also, it would be beneficial to include season-wide player ratings as certain 'star' players would consistently have a higher rating, thus positively impacting the team's overall performance. In all, there was insufficient data to further improve the model. To conclude, this analysis only serves as a preliminary insight and there remains room for improvement.