Predicting Soccer Matches with Bradley-Terry Models and Extensions

Sanjay Hariharan and Ilan Man

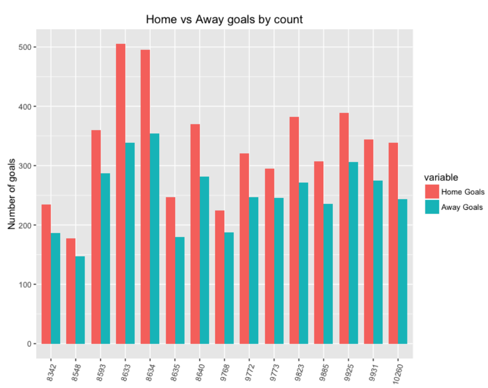
December 5, 2016

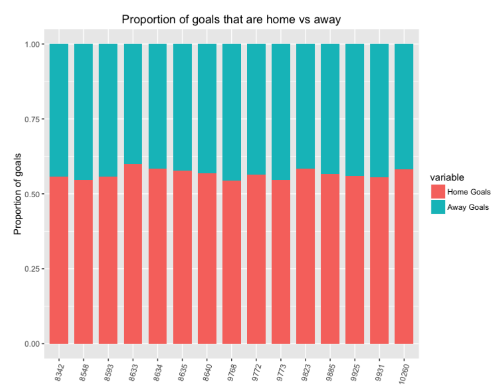
1 Motivation

Predicting the outcome of competitions is a well-studied and desirable subject, for both gamblers and sports fans alike. Given a Kaggle dataset of competing European Soccer teams, we set out to build a predictive model that can correctly predict the winner of head-to-head matchups. The dataset contains match results, team attributes, and player-specific information over eight years. We focus on our analysis on specific match outcomes, as well as contest and team-based attributes, in order to better understand the relative strength of teams throughout the years.

2 Exploratory Data Analysis

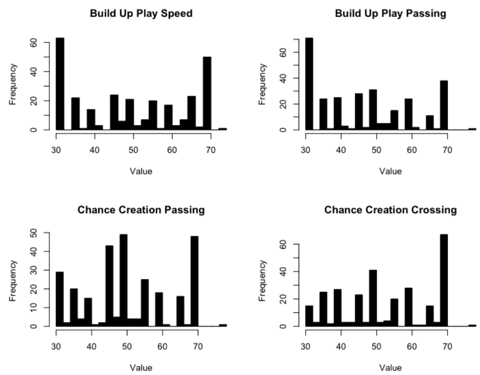
As our dataset is quite large, we perform a myriad of exploratory analyses, not only to better understand the variables, but also to provoke further questions to explore. For example, we plotted goals scored at home verses away for a random subset of teams.

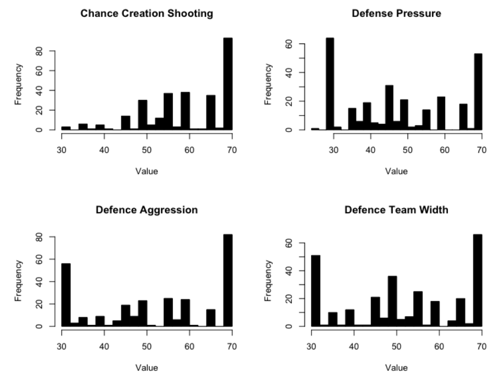




From the plots above, we can see that these teams average a greater number and proportion of goals scored at home than away. Bookies generally favor a home team at 53%, but it would be interesting to further explore this result using a model-based approach.

The database also contains attributes specific to each team. Specifically, FIFA provides a score on a variety of measures, including defense, passing, shooting, and build up speed.





We see that these values are quite spread out amongst the teams with no discernible pattern. It would be valuable to see which team attributes correspond most significantly with winning or losing a match.

Our analysis above prompted a few questions that we answer more in depth in our paper:

- What is the relative strength of one team compared to another ?
- Does playing at home offer a significant advantage ?
- Which team attributes contribute most significantly to winning a match ?

3 Modeling

Our main goal is to predict the outcome of any head-to-head soccer game. As such, the ability to estimate the probability of a pairwise matchup is desireable. The canonical Bradley-Terry model is flexible, interpretable, and provides us with the inference necessary to achieve our goal. We also consider extensions to the basic model, studying home team effect and team-specific predictors.

3.1 Standard Bradley-Terry Model

The standard Bradley-Terry model does not take into account any covariate effects. It is simply the head-to-head matchups along with the winner. The model can be expressed in the logit-linear form, for teams i and j:

$$\text{logit}[P(i > j)] = \lambda_i - \lambda_j$$

λ_i here is the coefficient estimate for team i. There are 2 distinct interpretations of the model output:

1. The λ_i parameters represent team i's relative strength, compared with other team's parameters
2. $\lambda_i - \lambda_j$ represents the log-odds of Team i winning over Team j

3.1.1 Results

We present the model outcome for the 3 teams with the best and worst winning percentage:

	Team Name	Win Percentage	Estimate	Standard Error
Best	FC Barcelona	0.769	2.792	0.299
	Real Madrid CF	0.750	2.385	0.280
	SL Benfica	0.745	3.079	0.497
Worst	Cordoba CF	0.079	-1.482	0.660
	SpVgg Greuther Furth	0.118	-1.586	0.577
	FC Dordrecht	0.118	-2.389	0.583

We can see that the teams with the better winning percentage have a higher coefficient result and are thus relatively stronger than the teams with the lower winning percentage and coefficient result. However, the two teams with the worst, but identical, winning percentage, have different values for their λ estimate. Their opponents could have been much different, or their could be latent variables affecting their overall performance.

Given these estimates, we can compute the probability of team i defeating team j:

$$P[i > j] = \frac{e^{\beta_i - \beta_j}}{1 + e^{\beta_i - \beta_j}}$$

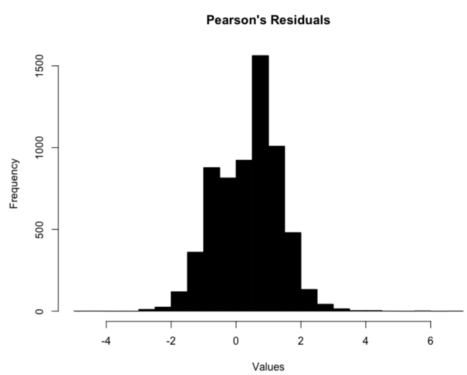
We select a random matchup between 5 Teams and present their probabilities below:

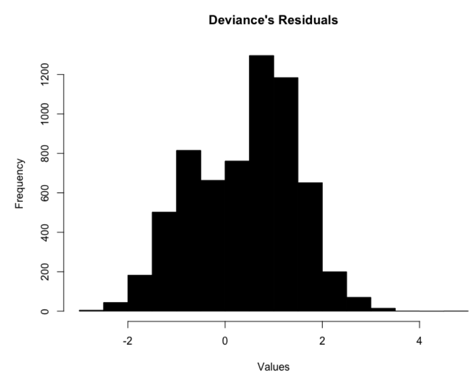
Team i	Team j	Win Percentage (i, j)	P[$i > j$]
SL Benfica	Standard de Liege	(0.75, 0.50)	0.861
Torino	LOSC Lille	(0.30, 0.48)	0.085
FC Nantes	SM Caen	(0.29, 0.29)	0.518
FC Dordrecht	FC Porto	(0.12, 0.74)	0.003
Celtic	Reading	(0.72, 0.16)	0.958

Here we see that the Probability of Team i beating Team j roughly follows their win percentage. For the two teams with very similar win percentages, FC Nantes and SM Caen, the model finds it difficult to conclude whether Team i will beat Team j, as their probability hovers around 0.50.

3.1.2 Model Adequacy

To check for Model Adequacy, we first analyzed the Pearson's and Deviance Residuals:





Both histograms look approximately normal, but the slight skewness to the right that is a little troublesome. We also checked for found that the sum of residuals squared divided by the degrees of freedom for Pearson's and Deviance residuals is 1.15 and 1.36, respectively. This result indicates that there may be over-dispersion in our model. One possible cause of over-dispersion is heterogeneity, or latent variables not specifically addressed by our model. We solve this problem in our next section by introducing a contest-specific predictor.

3.2 Bradley-Terry Model with Contest-Specific Predictors

The standard Bradley-Terry model above does not take any covariates into account. We offer a natural extension to the standard model, which adds an indicator if the team was playing at home or not. This addition results in the following logit representation:

$$\text{logit}[P(i\text{beats}j)] = \alpha[\mathbb{1}_{\text{Team i is home}}] + \beta_i - \beta_j$$

Here α is the 'Home Team Effect', or the increase in the Log-Odds of Team i winning if they are playing at home.

3.2.1 Results

After running the model, the coefficient estimate for α is 0.546, with a Standard Error of 0.016. This estimate means that playing at home increases the odds of winning by a factor of 1.726. This result is quite significant, since a team with 2:1 normal odds would see their odds increase to almost 7:2 if playing at home.
As above, we present the model outcome for the 5 teams with the best and worst winning percentage:

	Team Name	Win Percentage	Estimate	Standard Error
Best	FC Barcelona	0.769	2.852	0.307
	Real Madrid CF	0.750	2.447	0.287
	SL Benfica	0.745	3.162	0.510
Worst	Cordoba CF	0.079	-1.584	0.673
	SpVgg Greuther Furth	0.118	-1.674	0.588
	FC Dordrecht	0.118	-2.537	0.601

Our $\hat{\lambda}$ estimates have changed, but still follow the same relative magnitude as before. We note that the same random matchup from before, depending on which team is home or not, changes:

Team i	Team j	Win Percentage (i, j)	P[$i > j$]
SL Benfica	Standard de Liege	(0.75, 0.50)	0.914
Torino	LOSC Lille	(0.30, 0.48)	0.138
FC Nantes	SM Caen	(0.29, 0.29)	0.650
FC Dordrecht	FC Porto	(0.12, 0.74)	0.006
Celtic	Reading	(0.72, 0.16)	0.975

We see that the probability of winning significantly increases if we account for the 'Home Field Advantage'! This result corroborates our original analysis and answers our question of whether playing at home offers a significant advantage.

3.2.2 Model Comparison

The estimate for over-dispersion in our current model, that is, the sum of residuals squared divided by the degrees of freedom for Pearson's and Deviance residuals, is now 1.02 and 1.17, respectively. It is clear that adding the 'Home Team' contest-level predictor provides a better fit to the model than the standard one. This predictor may help explain variation in the data that only the Team's were not able to from before.

To test this finding, we perform an Analysis of Deviance test between our Standard Bradley-Terry model, and the current one with the added contest-level predictor. By adding the single predictor, the Residual Deviance decreases by 1125, and the Chi-Squared Test between the 2 models results in a p-value of 2.2e-16. This value indicates that adding the extra degree of freedom is significant, so we conclude that the larger model fits the data better.

3.3 Bradley-Terry Model with Team-Specific Predictors

The last question we seek an answers to is how important certain team-specific predictors are to the outcome of a head-to-head matchup. We consider an extension of the model of the form:

$$\lambda_i = \sum_{r=1}^{p} \beta_r x_{ir} + \epsilon_i$$

The ability of Team i is related to explanatory variables $x_{i1}, ..., x_{ip}$, through a linear predictor with coefficients β and independent errors ϵ_i. Our Coefficient and Standard Error estimates of these team-specific variables are the following:

	$\hat{\beta}$	Standard Error
Play Speed	-0.004	-0.0009
Play Passing	-0.001	0.001
Creation Passing	0.0007	0.001
Creation Crossing	0.002	-0.0009

	$\hat{\beta}$	Standard Error
Creation Shooting	-0.002	0.0012
Defense Pressure	0.013	0.001
Defense Aggression	0.004	0.0009
Defense Team Width	-0.013	0.001

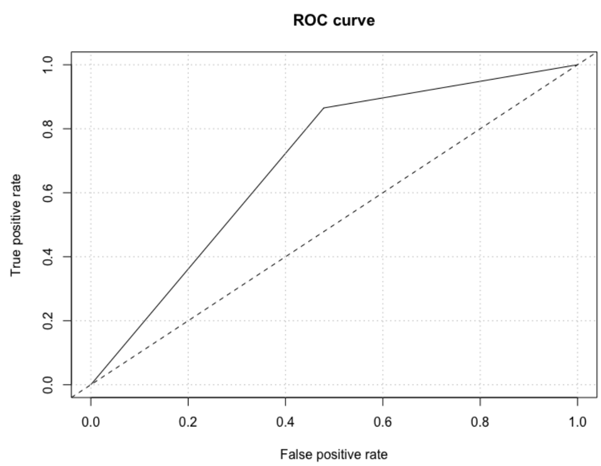
Though these values seem small, we note that these variables are continuous between 0 and 100. If the score for 'Defense Pressure' increases by 10 points for a particular team, the odds of that team beating any other team increases by a factor of 1.13. Furthermore, if a team's score for 'Play Speed' increased by 10 points, their odds of defeating another team actually decreases by a factor of 0.33.

Defense Pressure, Aggression, and Team Width, when introduced to the model, decrease Deviance the most compared to the other variables. As a result, these variables are the most significant in this model in affecting the contest. This result is not surprising, as the famous sports adage goes, *Defense Wins Championships.*

Overall, unfortunately, adding in team-level predictors does little to improve the overall model fit. An analysis of Deviance compared with the previous 'Home Advantage' model results in a large p-value, which reveals that adding these team-specific effects does little to improve the model fit. We believe this result is due to the fact that the FIFA score is too subjective, as well as the fact that players move around teams too much for the scores to have much meaning. The one constant in each match is that one team is at home and one is away, which contributes significantly to the result.

3.4 Accuracy

We assess the accuracy of our 'Home Advantage' model by training it on 70% of the data and predicting on the other 30%, building an ROC Curve and Confusion Matrix.



		Win	Loss	Reference
Prediction	Win	412	135	
	Loss	378	865	

We can see that the 'Home Advantage' model does a decent job of predicting the outcome for our test set. From our confusion matrix, our correct predictions outweigh our incorrect ones, but we still predict incorrectly about 25% of the time. Furthermore, our AUC is 0.693, which is quite good given that our only external predictor is Home Field Advantage.

4 Limitations and Further Extensions

The key assumptions are in itself the limitations of the Bradley-Terry model for our analysis. The model assumes independent contests, which is unrealistic in a sports setting. Teams change throughout the season, and certain games will most definitely influence others. Another issue is that the Standard Bradley-Terry model removes all 'ties'. Davidson (1970) extended the B-T model to accommodate ties, and given that 25% of the games in the dataset were ties, it would be very beneficial to explore this extension as to make full use of our dataset. Overall, however, the Bradley-Terry models are a robust, scalable, and interpretable method for modeling outcomes of competitions. They work particularly well for soccer matches, which have a clear win, loss, or tie, and their extensions add an interesting complexity.

5 References

[1] Agresti, Alan. Categorical Data Analysis. Hoboken, NJ: Wiley, 2013. Print.

[2] Borjigin, Khachin. "Players - FIFA 17 - Dec 1, 2016 - SoFIFA." Players - FIFA 17 - Dec 1, 2016 - SoFIFA. N.p., n.d. Web. 05 Dec. 2016.

[3] Bradley RA, Terry ME (1952). ?Rank Analysis of Incomplete Block Designs I: The Method of Paired Comparisons.? Biometrika, 39, 324?45

[4] Davidson, Roger R. ?On Extending the Bradley-Terry Model to Accommodate Ties in Paired Comparison Experiments.? Journal of the American Statistical Association, vol. 65, no. 329, 1970, pp. 317?328. www.jstor.org/stable/2283595.

[5] Kaggle: Your Home for Data Science. N.p., n.d. Web. 05 Dec. 2016.