Dataset

Our data comes from the European Soccer Database through Kaggle. The dataset is quite massive, containing over 25k matches, 10k players, 11 countries, and 8 seasons. More importantly, we have match results and specific attributes to each team. We also have player-specific information, and a mapping from players to teams. The ultimate goal would be to predict the outcome of any matchup. However, we choose to focus on specific causes on match outcomes, as well as the relative strength of teams throughout the years.

Exploratory Data Analysis

We performed quite a bit of EDA to understand more about the teams, as well as provoke certain questions. One larger question we had was what specific factors result in a team winning or not. Take a look at the count of Home vs. Away goals for a subset of Teams. Our dataset masked the teams by ID, so we present them here as well.

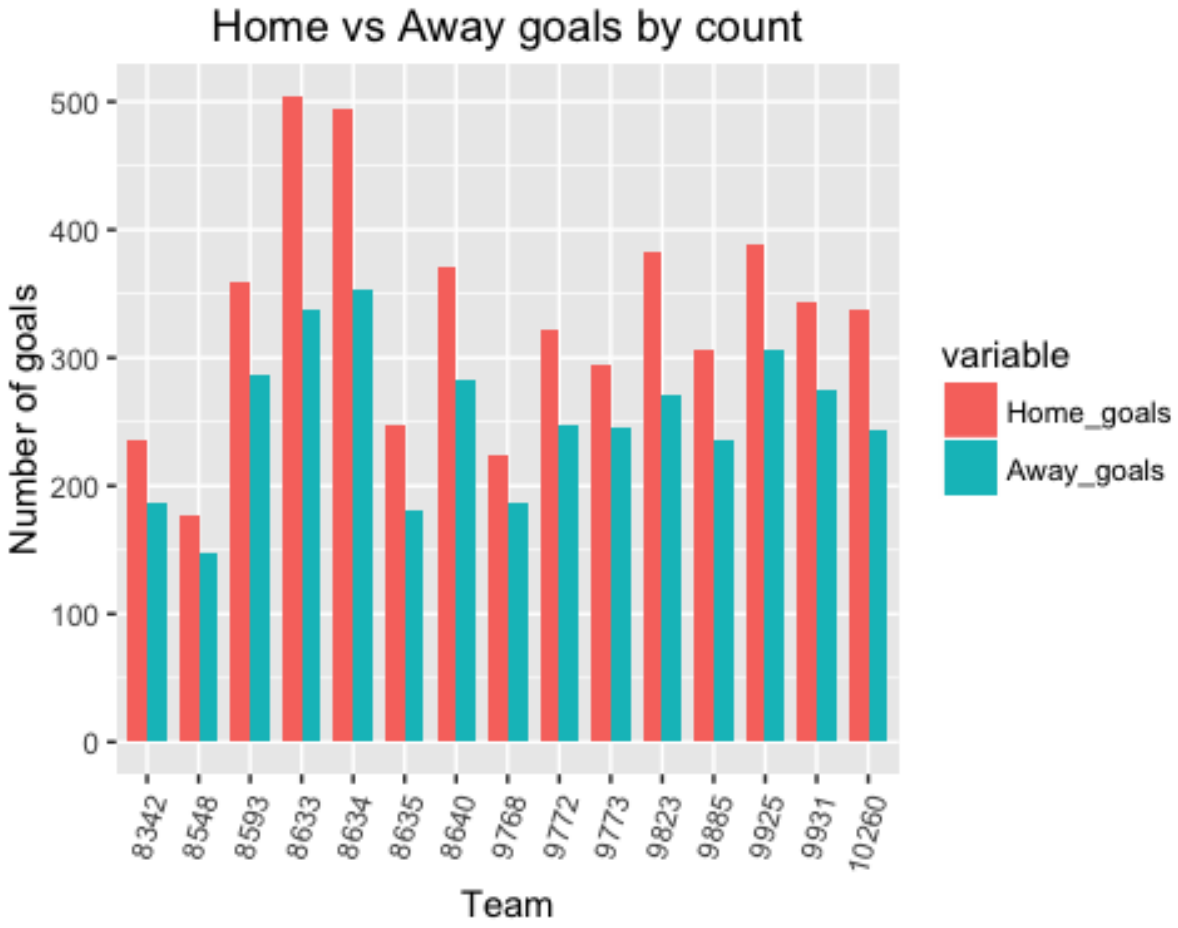

Figure 1: Count of Goals

Our analysis prompted a few questions that we hope subsequent modeling can answer:

- Which teams are the best and worst based on match play ?
- Does playing at home offer a significant advantage?

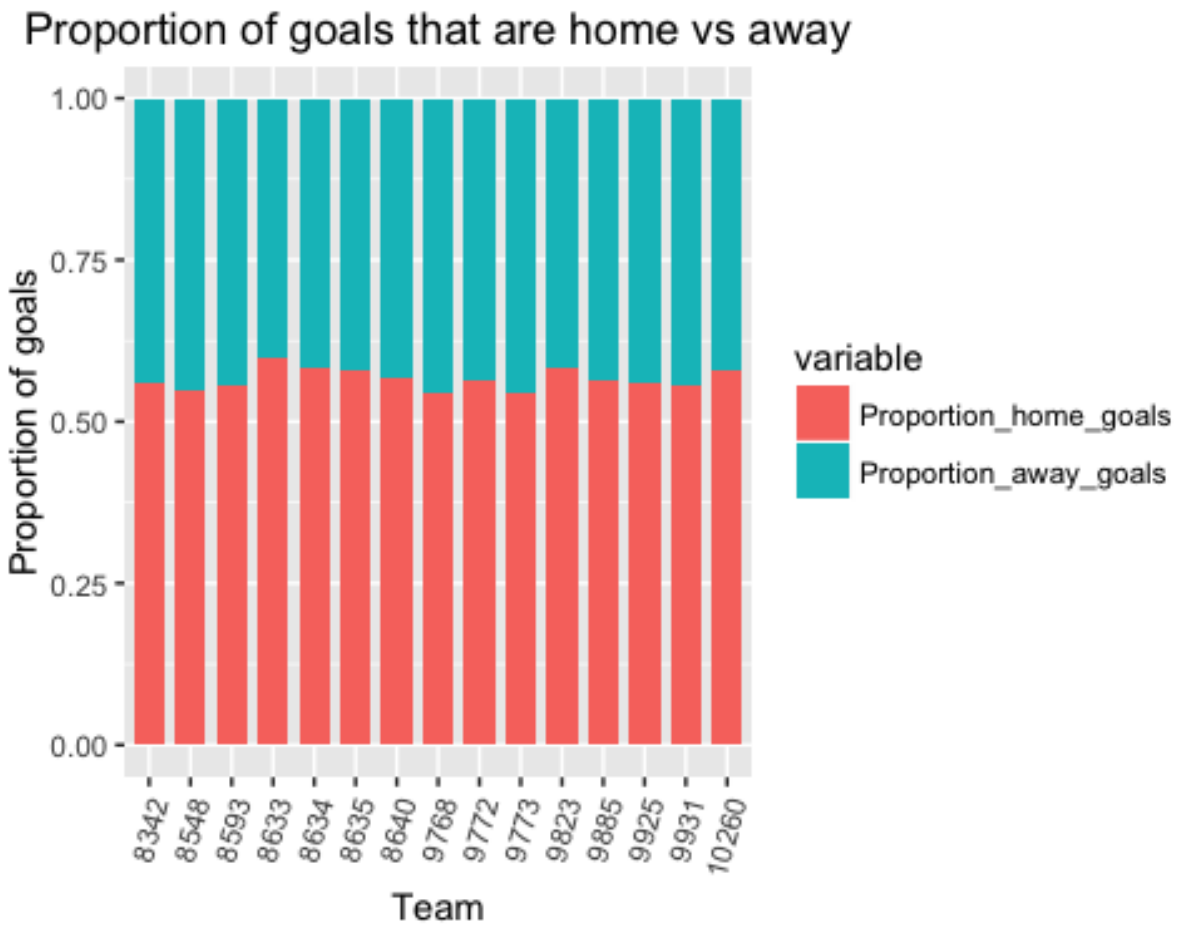

Figure 2: Proportion of Goals

- Which Team Attributes contribue most significantly to winning a match ?

We explore these questions in further detail in our modeling section.

Modeling

At its core, this dataset begs for a model that can predict the outcome of a competition. Furthermore, it would be ideal to estimate the probability of a pairwise comparison or matchup. As such, the canonical Bradley-Terry model is a well-known and ideal option for modeling outcomes. We also consider extensions to the basic model, studying Home Effect as well as Team-Specific Predictors.

PUT LITTLE EQUATION HERE

Standard Bradley-Terry Model

The standard Bradley Terry model works surprisingly well to rank the teams in terms of relative strength. The model can be expressed in the logit-linear form:

$$\text{logit}[P(i\text{beats}j)] = \beta_i - \beta_j$$

We assume independence in all contensts, and parameters are found by Maximum Likelihood Estimation. The top and bottom 5 teams are:

	Top Teams	Estimate	Bottom Teams	Estimate
1	9847.00	0.00	6631	0.00
2	9773.00	0.00	9839	0.00
3	8639.00	0.00	8526	0.00
4	8634.00	0.00	10218	0.00
5	0.00	0.00	8525	0.00

With a residual deviance of 10114 compared to a df of 6604, there is likely dispersion. As such, there is more variation in our data than our covariates can explain. Given these estimates, and the standard Bradley-Terry formula above, we can easily compute the Probability that of the outcome of a match:

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

A random matchup between 5 teams results in the following probabilities.

	Top Teams	Estimate	Bottom Teams
1	9847.00	6631	0.85
2	9773.00	9839	0.34
3	8639.00	8526	0.56
4	8634.00	10218	0.71
5	0.00	8525	0.15

So using the Standard Bradley-Terry Model, it is possible to predict the outcome of a game in a probabilistic manner.

Bradley-Terry Model with Contest-Specific Predictors

Given that our previous model was overdispersed, we would want to take into account more information if possible. Specficaly, the Standard model above does not take into account any contest-level predictors. One of our major questions after EDA was if Home field advantage has any effect on the competition. The Bradley-Terry model offers several extensions, one of which is the addition of Contest-Specific predictors. In our case, the only contest level predictor we could attain is whether the team was at home or not. Luckily, adding a contest level predictor is a natural extension to the standard Bradley-Terry model, as we only increase our degrees of freedom by 1.

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

After running the model, we find that our Home Team effect, $\alpha = 0.546508$, with a SE of 0.01668. This tells us that there is indeed an advantage to playing at home, as evidenced by the game outcomes. However, is adding an extra paremeter to our model meaningful? Does it improve the fit? We first noticed that there does not appear to be a lot of overdispersion in our model, since the Residual Deviance / Degrees of freedom = 1.17, which is fairly close to 1. Then we ran an Analysis of Deviance Test using a Chi-Squared approximation. We find a p-value of 0.034, which at a significance level of $\alpha = 0.05$, tell us that adding in Home-Field advantage is significant for determining the outcome of a match. Our updated best and worst teams, as well as random matchup probabilities are below.

	Top Teams	Estimate	Bottom Teams	Estimate
1	9847.00	0.00	6631	0.00
2	9773.00	0.00	9839	0.00
3	8639.00	0.00	8526	0.00
4	8634.00	0.00	10218	0.00
5	0.00	0.00	8525	0.00

	Top Teams	Estimate	Bottom Teams
1	9847.00	6631	0.85
2	9773.00	9839	0.34
3	8639.00	8526	0.56
4	8634.00	10218	0.71
5	0.00	8525	0.15

Bradley-Terry Model with Team-Specific Predictors

The final extension to our standard Bradley-Terry model is to incorporate Team-Specific predictors. Specifically, what attributes to a team contribute significantly to winning a match? It is natural to consider model simplification of the form:

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

in which ability of each team i is realted to explanatoy variables $x_{i1}, ..., x_{ip}$, through a linear predictor with coefficients β and independent errors U_i. We treated the teams as a random effect so we could focus on the covariates itself. FIFA provided a variety of Team attributes, both Categorical and Continuous. The top and bottom 5 predictors are:

These results are very interesting! We can see certain attributes for a team are significant in the team winning, vs. certain ones that correspond to losing. EXPLANATION OF A FEW.

	Predictors for Winning	Estimate	Predictors for Losing	Estimate
1	Creation Shooting Class (Lots)	0.00	Play Passing Class (Short)	0.00
2	Creation Passing Class (Safe)	0.00	Defence Pressure Class (High)	0.00
3	Defence Team Width (Wide)	0.00	Defence Pressure Class (Medium)	0.00
4	Creation Crossing Class (Normal)	0.00	Play Passing Class (Mixed)	0.00
5	Play Speed Class (Slow)	0.00	Creation Passing Class (Risky)	0.00

Accuracy of Model

Further Extensions

The Bradley-Terry models are widely used and many extensions beyond what we have explored have been created. One limitation of our current analysis is that we throw away all tied games. However, extensions exist which accommodate ties, and further improve our understanding of the teams relative ranking. Another limitation is that these models treat each game as Independent, which is quite a strong assumption. Teams get better throughout a season, and certain games may influence others. We also did not incorporate player effect within a team, and how those players move around different teams (as this is European Soccer).