Predicting Football Match Results of Spanish League using Bayesian Hierarchical Model

DATS6450 Bayesian Final Project - Team 2: Hao Ning, Xi Zhang

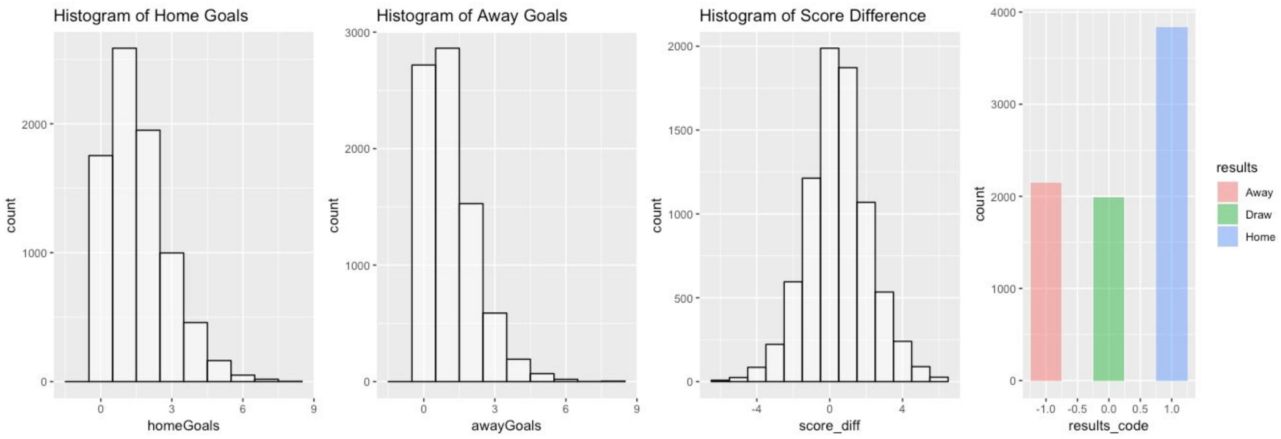
Introduction

Soccer(or internationally, football) is one of the most popular sports on our planet, with a well-developed industry worth more than $400 billion and billions of fans(estimated) around the world[1]. Predicting the matches results have always attracted many attentions and analyzing the game results proved to be very helpful with the business growth and player training[2]. There are many machine learning approaches for game prediction, however, we believe the Bayesian approach could be very helpful in this scenario (**given reliable historical data**). We will work on the data from the Spanish league, specifically the season 2015-2017.
Data link: https://www.kaggle.com/ricardomoya/football-matches-of-spanish-league

Preprocessing

After checking if there's any missing values, renaming columns, changing data types, we filter years after 1997 for EDA and 2015-2017 for modeling and prediction.

EDA



[1] Sawe, Benjamin Elisha. "The Most Popular Sports in the World." WorldAtlas, Apr. 5, 2018, worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html.

[2] Shahzeb, Farheen. "The Evolution and Future of Analytics in Sport". June 22, 2017, Proem Sports

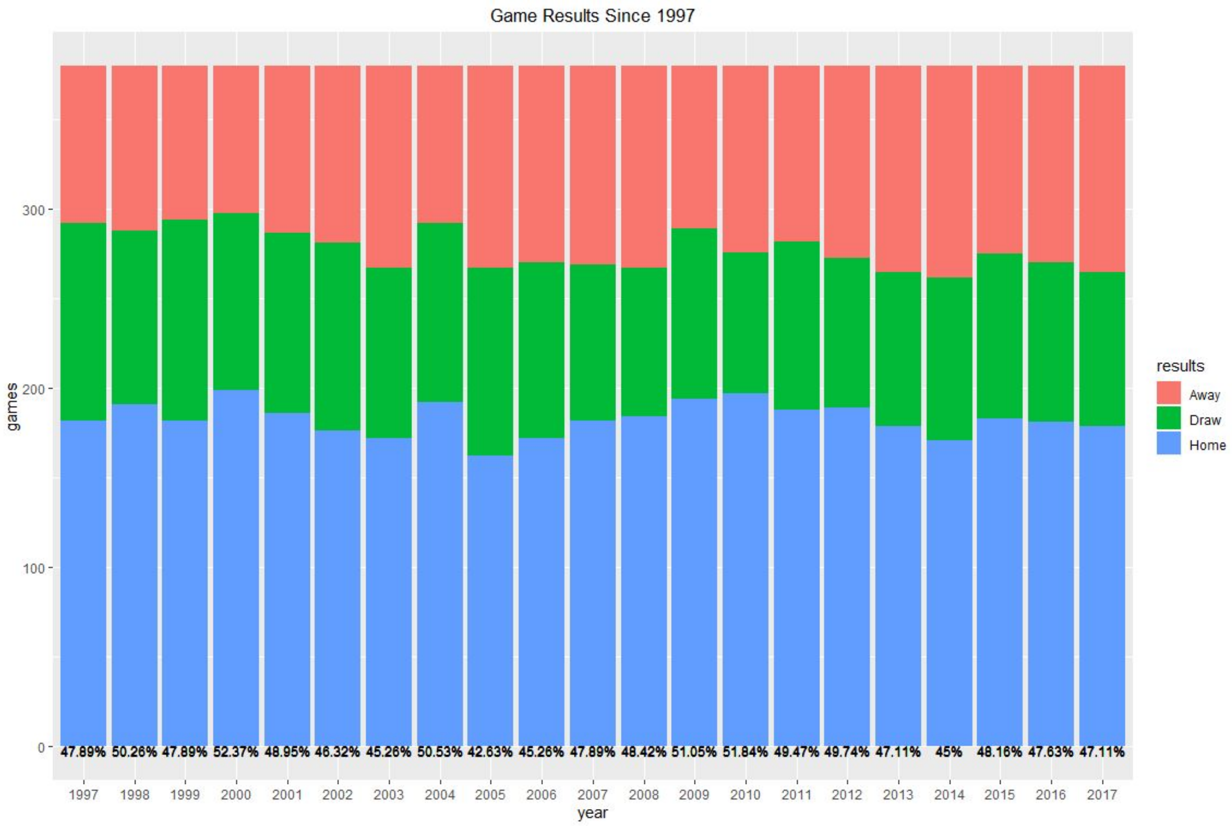


Figure Game Results since 1997, Comparison of home goals and away goals

From the figure shown above, we can see that home advantage is quite obvious, otherwise the proportion of home win/draw/away win would be very similar. Overall, we observed a ratio of home:draw:away = 2 : 1 : 1. This is a very important field knowledge that will help us to specify prior of our hierarchical model.

Home advantage really make sense for the following reasons:

- Better familiarity of the fields
- Better team shape, since no lodging or travel involved
- Ultimate support from fans
- Pressure on referees

Bayesian Hierarchical Model

Approach 1 - Using home win 0,1 for modelling and prediction

Our assumption is that all teams abilities will be different, otherwise we will likely to see each team take turns winning the championship every now and then.

Hierarchical Model:

We've already proved that team playing at home will have a higher winning probability, therefore, a home advantage factor will be added into our model. In this way, we can avoid overestimating the team ability when they are playing at home, they win not only because of their abilities, but also due to home advantage.

The hierarchical model is defined in this way:

- Home team win follows a binomial distribution, with parameter of home team winning probability
- Home team win probability is defined by the equation with parameter of team abilities (home and away) and home advantage factor
- Team abilities is e to the power of log(ability), which follows a normal distribution, with paratermeter of performance variation; home advantage follows a uniform distribution
- Performance variation follows a uniform distribution

The parameters and priors are specified here (bottom up):
performance variation ~ dunif(0,2)
home_advantage ~ dunif(1,1.5)
log(ability) ~ dnorm(0,1/$performance\_variation^2$)
ability[i] ~ exp(log(ability))
prob[i] = $\frac{ability(home[i]) * home\_advantage}{ability(home[i]) * home\_advantage + ability(away[i])}$
yi (home win) ~ dbin (prob[i], 1)

By implementing this model for years 2015-2017, we want to compare team abilities and find out the best team in the league. There are a total of 1140 games played and we are using 900 games for modelling. Below, is a boxplot of team abilities, we can see that Barcelona and Real Madrid are the top 2 teams.

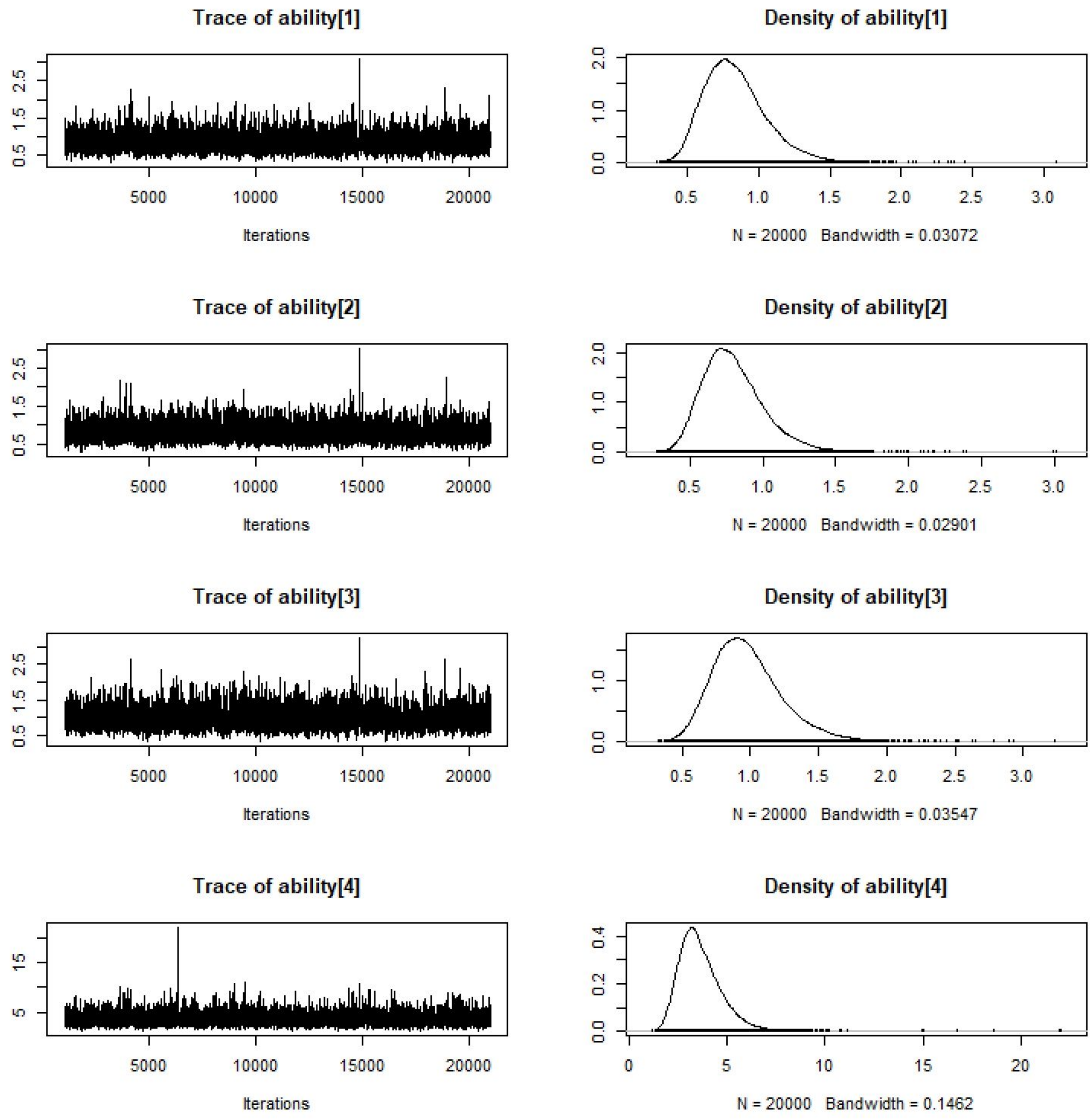
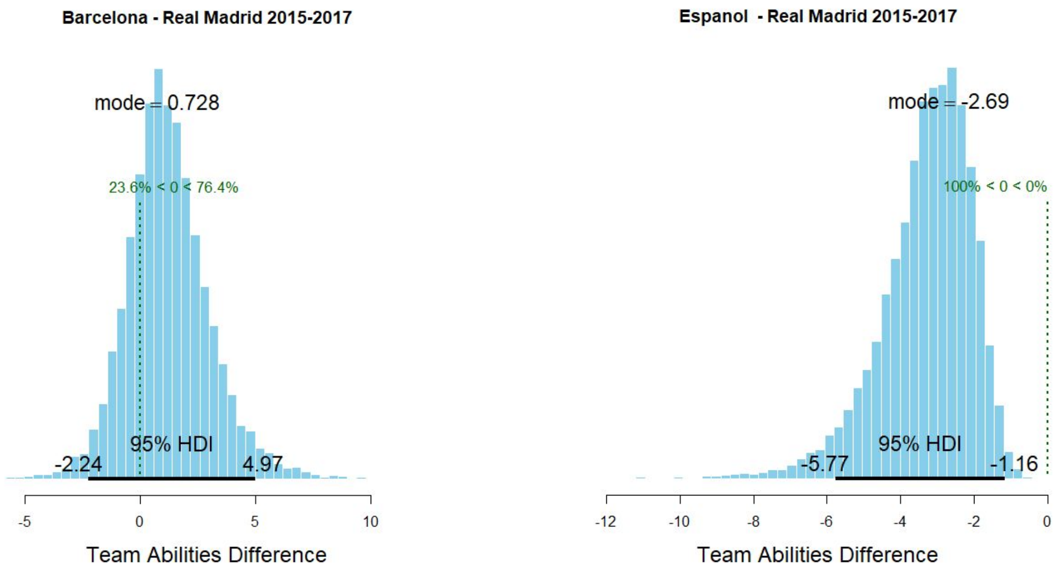


Figure Trace Plot of Team Abilities, only showing team 1 to 4(check code for all teams).

Now, it will be interesting if we compare the team abilities between good team v.s. good team, and normal team v.s. good team, using plotpost function, as shown below:



Plotpost function displaying the comparison between different teams

We can see that:
For Barcelona - Real Madrid:
Barcelona is the better team, because **mode>0**; however, this is **not significant**, because 0 is **within** the 95% HDI.
For Espanol - Real Madrid:
Real Madrid is the better team, because **mode<0**; and this is **significant**, since 0 is **not within** the 95% HDI.

Note: in our code, we can plot comparison of all teams pairs using nested for loops, we commend this part out since there are too many graph windows showing up.

Predictions

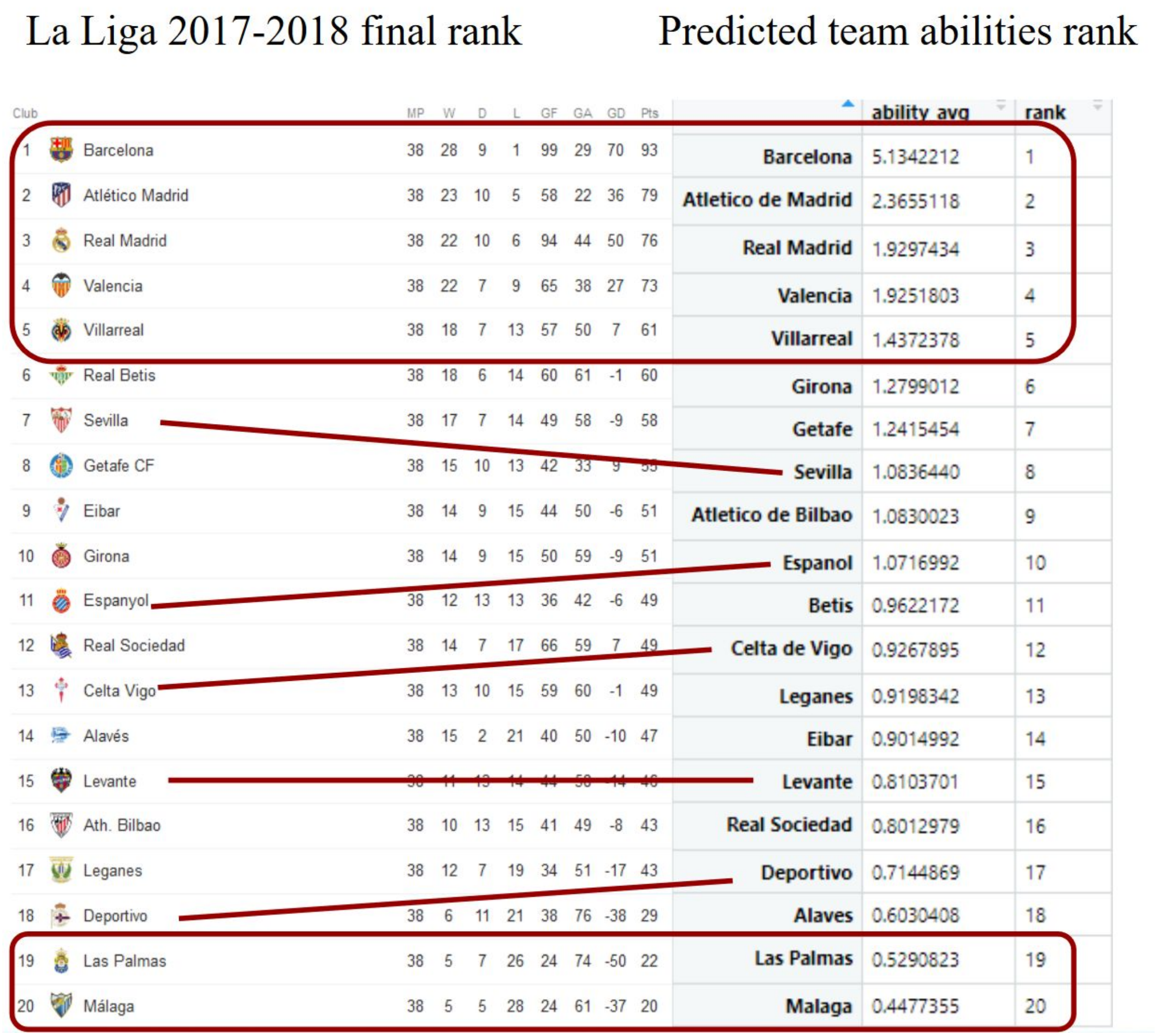
Since before the start of every new season, the team roster will likely change, we will be implementing the model for season 2017 for prediction. There are a total of 380 games, the first 200 games are used for modelling, then the rest for prediction.

We use a nested for loop to calculate winning probabilities for all pairs of home[i] vs. away team[j]. We are happy to see that the probability of home[i] vs away[i] is greater than 0.5 (although in reality this won't happen a team playing against itself), our model truly captured the home advantages!

Now, the critical part is how we use the winning probabilities to make predictions/bet. We made predictions with probability>0.6 and probability>0.7. That is to say, when the home team have a predicted winning probability greater than 0.6 or 0.7, we bet home win.

Here are the highlights of prediction results:
"P_pred >0.6, we bet the home team win, prediction accuracy : 70 %"
"P_pred >0.7, we bet the home team win, prediction accuracy : 82.61 %"
We can achieve a pretty decent overall accuracy!

Finally, we have another interesting findings by sorting the team abilities from the model and comparing final rank of the season, as shown below.



La Liga 2017-2018 final rank

Club		MP	W	D	L	GF	GA	GD	Pts
1	Barcelona	38	28	9	1	99	29	70	93
2	Atlético Madrid	38	23	10	5	58	22	36	79
3	Real Madrid	38	22	10	6	94	44	50	76
4	Valencia	38	22	7	9	65	38	27	73
5	Villarreal	38	18	7	13	57	50	7	61
6	Real Betis	38	18	6	14	60	61	-1	60
7	Sevilla	38	17	7	14	49	58	-9	58
8	Getafe CF	38	15	10	13	42	33	9	55
9	Eibar	38	14	9	15	44	50	-6	51
10	Girona	38	14	9	15	50	59	-9	51
11	Espanyol	38	12	13	13	36	42	-6	49
12	Real Sociedad	38	14	7	17	66	59	7	49
13	Celta Vigo	38	13	10	15	59	60	-1	49
14	Alavés	38	15	2	21	40	50	-10	47
15	Levante	38	11	13	14	44	58	-14	46
16	Ath. Bilbao	38	10	13	15	41	49	-8	43
17	Leganes	38	12	7	19	34	51	-17	43
18	Deportivo	38	6	11	21	38	76	-38	29
19	Las Palmas	38	5	7	26	24	74	-50	22
20	Málaga	38	5	5	28	24	61	-37	20

Predicted team abilities rank

	ability avg	rank
Barcelona	5.1342212	1
Atletico de Madrid	2.3655118	2
Real Madrid	1.9297434	3
Valencia	1.9251803	4
Villarreal	1.4372378	5
Girona	1.2799012	6
Getafe	1.2415454	7
Sevilla	1.0836440	8
Atletico de Bilbao	1.0830023	9
Espanol	1.0716992	10
Betis	0.9622172	11
Celta de Vigo	0.9267895	12
Leganes	0.9198342	13
Eibar	0.9014992	14
Levante	0.8103701	15
Real Sociedad	0.8012979	16
Deportivo	0.7144869	17
Alaves	0.6030408	18
Las Palmas	0.5290823	19
Malaga	0.4477355	20

Figure: La Liga 2017-2018 final rank (left) and predicted team abilities rank (right)

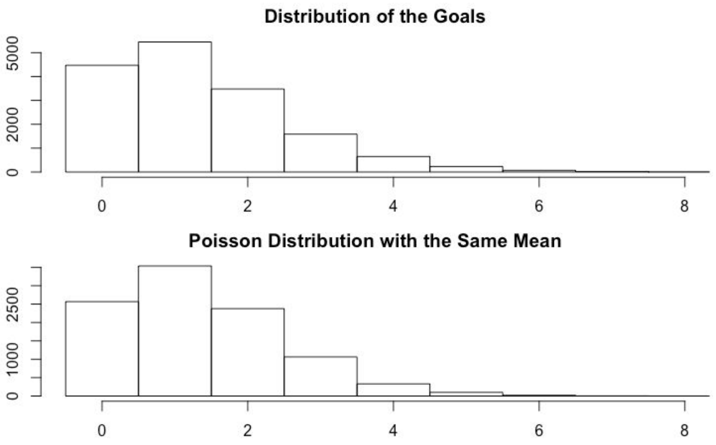
Top 5 teams and bot 2 teams matches exactly the same as predicted team abilities. For most of the other team rankings, they still kind of coincide with the predicted rank. This observation really make sense for the following reasons:

1. Strong teams are really dominating, thus easier to predict
2. There are many factors that could impact the game results, such as referee decisions, core player injury, weather conditions etc, especially for teams with very close abilities
3. Due to the data that we have at hand, we can't include other factors into our model, however, it will be very interesting when including all kinds of factors

Approach 1 complete.

Approach 2 - Using Hierarchical Model with Poisson Distribution for Predicting Goals

This approach is also based on hierarchical Bayesian modelling and Markov chain Monte Carlo but captured by the Poisson distribution. Because in the initial hypothesis, all football matches are about the same length, both teams have a lot of chances to score, and each team has the same probability of scoring a goal. The Poisson distribution is suitable for describing the probability distribution of the number of times a random event occurs in a unit time.



Because of the characteristics of poisson distribution, we have a lot of room for improvement.I'm going to show you three models and show you how to improve them step by step to make our predictions more accurate.

Of course, if the teams are all the same, the game is meaningless. We generally assume that the performance of the team has a skill variable. The skill of one team minus the other can predict the outcome of the game. Since the number of targets is Poisson-distributed, the skill of the team is naturally a logarithmic scale of the mean of the distribution. When facing the j team, the goal number distribution of team i is:

$$Goals \sim \mathrm{Poisson}(\lambda)$$

$$\log(\lambda) = \mathrm{baseline} + \mathrm{skill}_i - \mathrm{skill}_j$$

In this model, the baseline is the average number of goals scored when both teams are equally good:

$$HomeGoals_{i,j} \sim \mathrm{Poison}(\lambda_{\mathrm{home},i,j})$$

$$AwayGoals_{i,j} \sim \mathrm{Poison}(\lambda_{\mathrm{away},i,j})$$

$$\log(\lambda_{\mathrm{home},i,j}) = \mathrm{baseline} + \mathrm{skill}_i - \mathrm{skill}_j$$

$$\log(\lambda_{\mathrm{away},i,j}) = \mathrm{baseline} + \mathrm{skill}_j - \mathrm{skill}_i$$

With the addition of prior, we have a complete Bayesian model. The prior distribution of baseline and skills is set as follows:

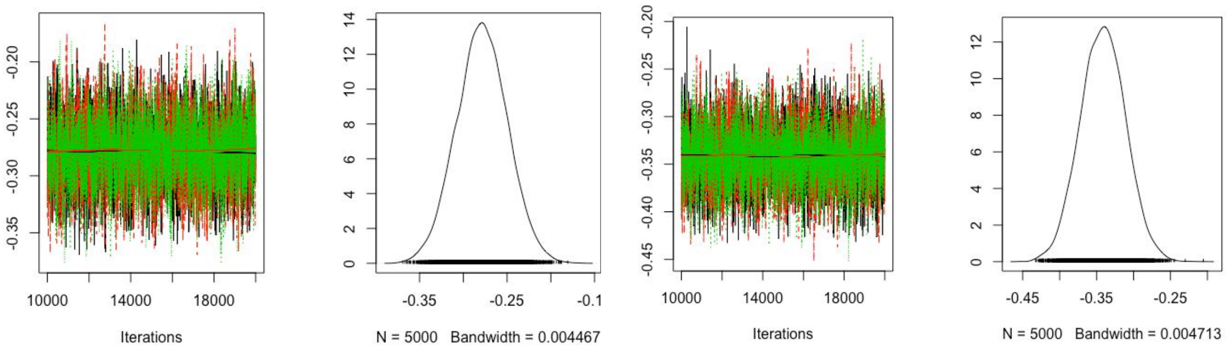
$$\mathrm{baseline} \sim \mathrm{Normal}(0, 4^2)$$

$$\mathrm{skill}_{1\ldots n} \sim \mathrm{Normal}(\mu_{\mathrm{teams}}, \sigma^2_{\mathrm{teams}})$$

$$\mu_{\mathrm{teams}} \sim \mathrm{Normal}(0, 4^2)$$

$$\sigma_{\mathrm{teams}} \sim \mathrm{Uniform}(0, 3)$$

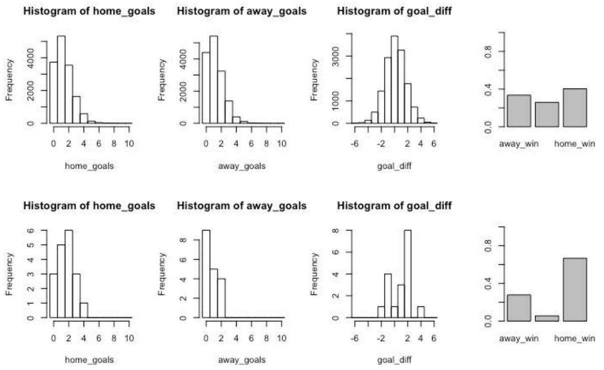
To compare the skill, we can look at the distribution of trajectory diagram and skill parameters for Valencia(left) and Sevilla(right):

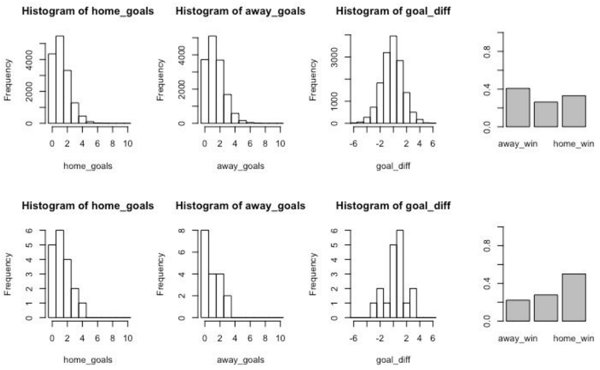


You can see that Valencia is a little bit better in the same base line.

To predict the goal of Valencia and Sevilla. We can check the following barplot for different home cases. The first row in the figure shows the simulation and the second row shows the historical data:

Valencia home to Sevilla away(left); Sevilla home to Valencia away(right)





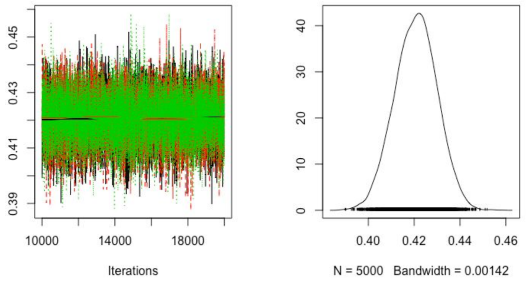
One thing our model didn't predict. Historical figures now show that Sevilla often beat Valencia at home. The current model does not take into account the advantage of the home team. We'll update the model to include the effect of home field advantage and look at the results.

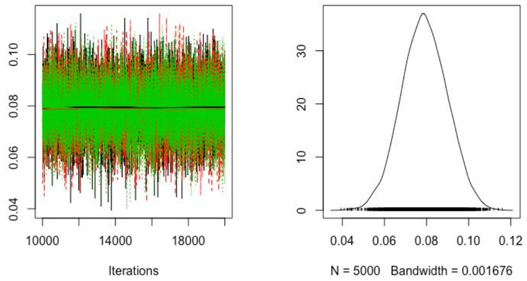
Updated Model with the Home Advantage

To add home field advantage, we implemented this change in the JAGS model by dividing the baseline into home_baseline and away_baseline:

The trace plots and distributions of home baseline and away baseline show home advantage does exist:

Home Baseline(left): Away Baseline(right):

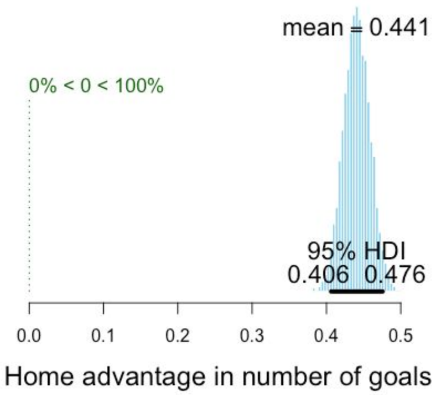




Through the following comparison of DIC, it also shows that model 2 is better than model 1:

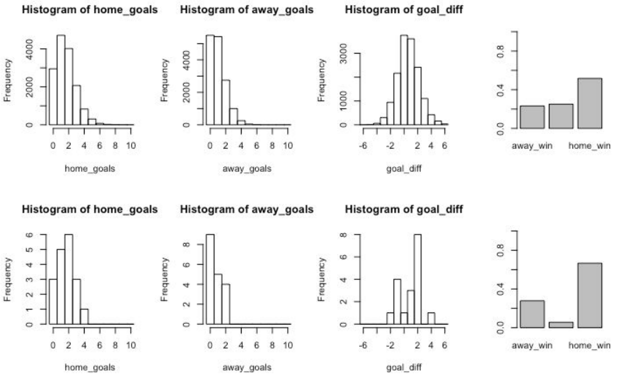
Difference: 593.5977
Sample standard error: 49.48685

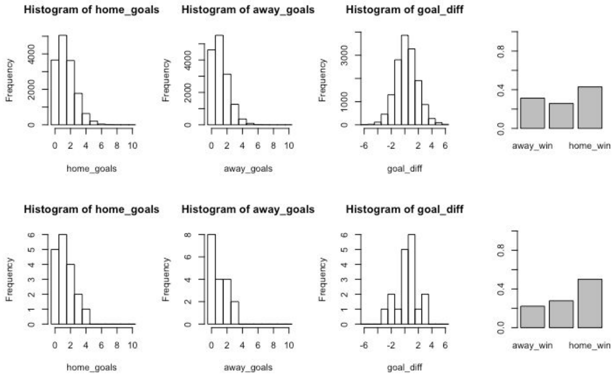
To compare the home baseline and away baseline, the home advantage is obvious. Home team most likely to gain 0.4 more goals.



Finally, we implement the new model to simulate and predict two teams:

Valencia home to Sevilla away(left); Sevilla home to Valencia away(right)



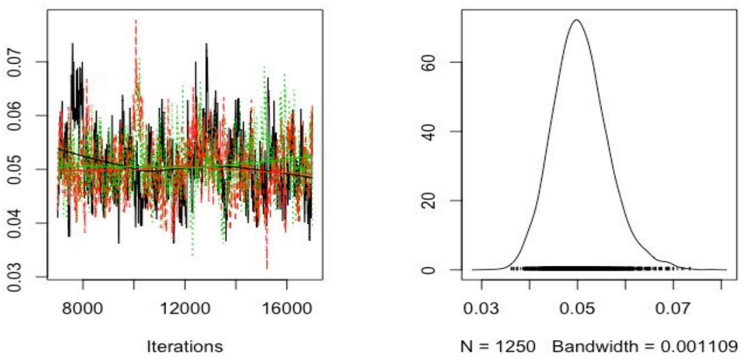


With the addition of home-field advantage, projections are reconciled with real historical data. Both teams have a better chance of winning at home.

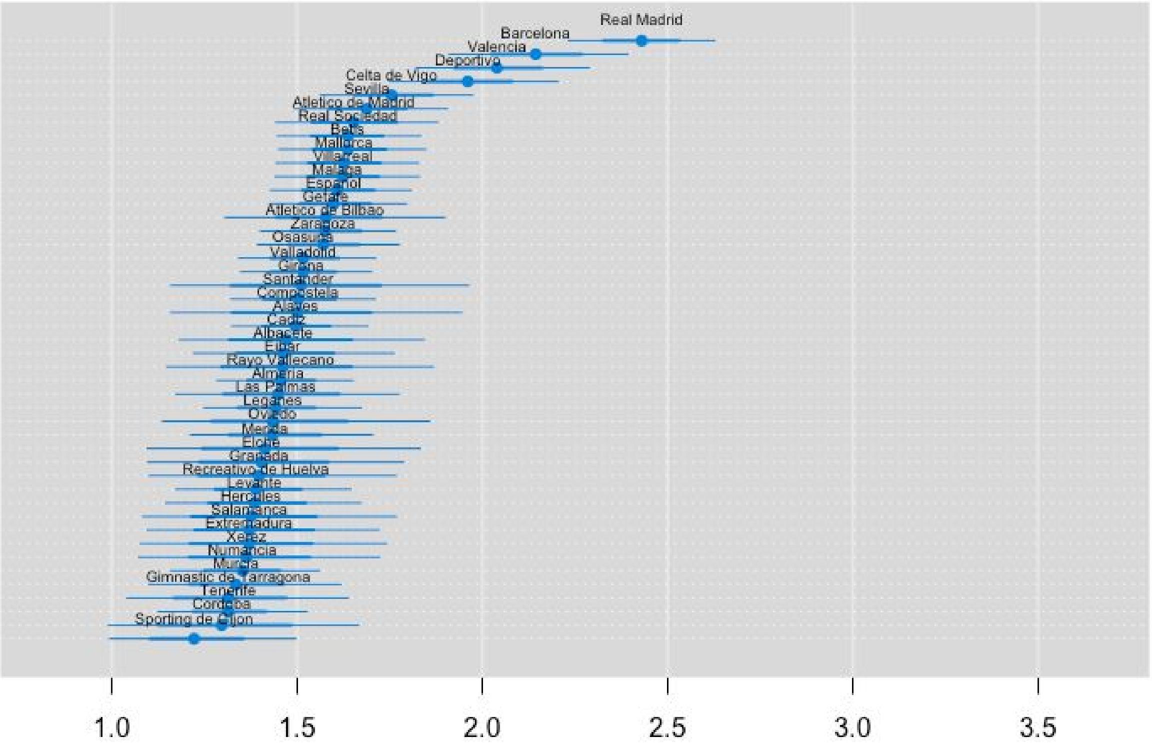
Updated Model with Skill Variability

The original assumption is that every team has the same skill level in every year, but it's not real. Team performs differently year to year. For fixing this part, we modify the model to include the year-to-year variability in team skills:

$$skill_{\ t+1} \sim \text{Normal}\ (skill_t,\ \sigma^2_{year})$$

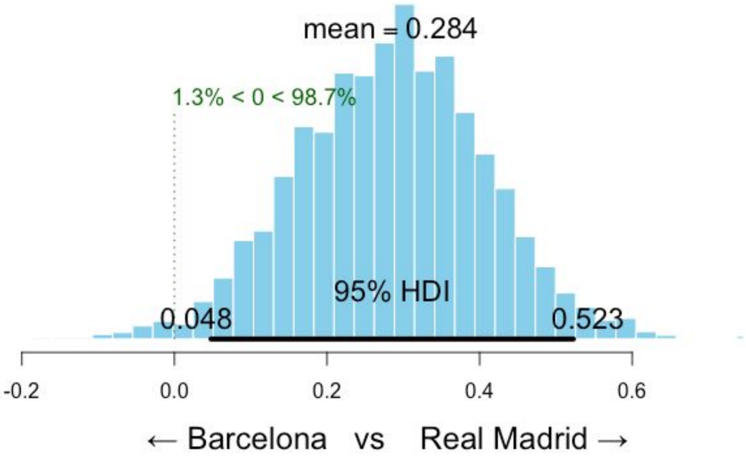


Ranking plot based on Model 3:



Check the credible difference between the top 2 teams:

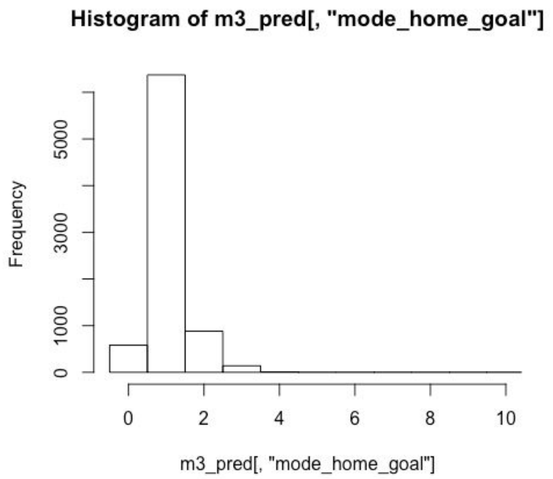
Team skills of Real Madrid - Team skills of Barcelona

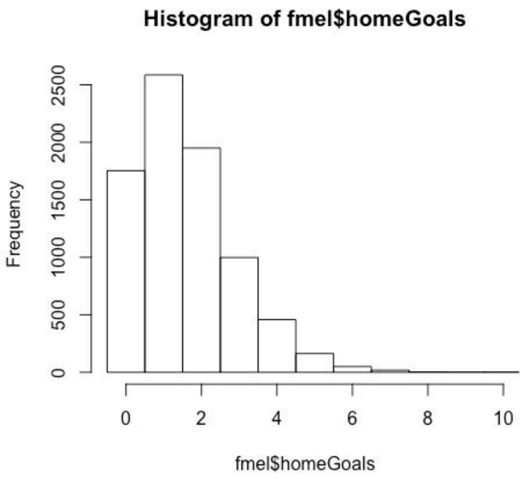


Since 0 is not in the range of HDI, the team skill of the Real Madrid is significantly better than the Barcelona.

Predictions

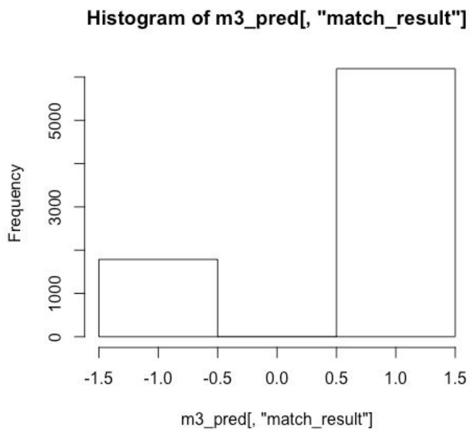
If we want to use our model to bet on the number of goals scored in a game, we will simulate the pattern of the number of goals scored, that is, the most likely number of goals scored.

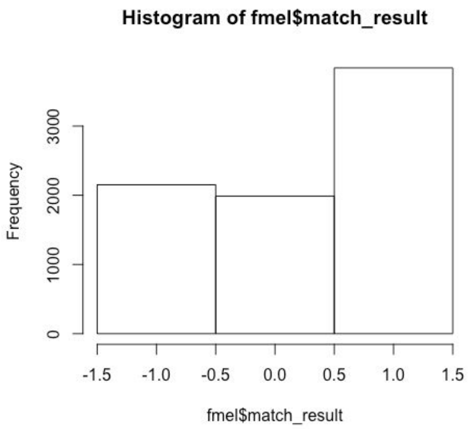




According to our distribution model of the most likely goals (left), 1 home goal has the highest probability, so it is the safest choice to bet on one goal.

We directly compare the predicted value of home goals with the actual value to obtain the prediction accuracy, then we get 0.3310777 with mean square error of 1.464646.





When the same model was used to predict the outcome, we got an accuracy of 0.5358396.

Results Analysis and Conclusions

Many factors could affect the results of football games, such as team roster change, referee decisions, the shape of the team, weather, injuries and so on. Therefore, building a comprehensive model that captures all the factors proved to be very complex and tedious.

Approach 1 that we have demonstrated could set things straight by only considering home team winning probability with parameters of team abilities, performance variations and home advantages.We have demonstrated a prediction accuracy up to 82.61 % (when betting with 0.7 home win probability), and find out that team ability rank coincide with the season final rank, especially the top teams and bottom team aligned exactly!

However, approach 1 has some limitations, since we can only predict whether the home team can win or not, game outcomes of draws and away team wins are not incorporated in the model. Also, the number of goals are not considered in the model, thus we missed some information that contributes to the ability of the team. For example, A:C with a score of 6:1, compared to B:C with a score of 2:1, in this case, we will predict that A and B are equally good.

Approach 2 we are using number of goals and score difference for modelling and predictions. The number of goals a team can score is the reflection of their abilities (ability is the parameter for number of goals a team can score). By calculating the score difference, we are able to predict all the outcomes: home win/draw/away win. However, the accuracy is not as good as approach 1. This is due to some teams are good at defence but bad at attacking, while other teams might be the opposite. Thus, only including score difference might not be accurate for results prediction.

Finally, game results prediction is not an easy field since many factors can impact the outcomes. We demonstrated 2 different approaches that can solve the problem in a relatively straightforward way. Including more factors will be very interesting and for future works to dive deeper!

Reference

Bååth, Rasmus. "Modeling Match Results in Soccer using a Hierarchical Bayesian Poisson Model." 2015, Sumsar.

Sawe, Benjamin Elisha. "The Most Popular Sports in the World." WorldAtlas, Apr. 5, 2018, worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html

Shahzeb, Farheen. "The Evolution and Future of Analytics in Sport". June 22, 2017, Proem Sports