THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Predicting Football Match Results of Spanish League using Bayesian Hierarchical Model

Group 2: Hao Ning, Xi Zhang

Introduction

- Soccer 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

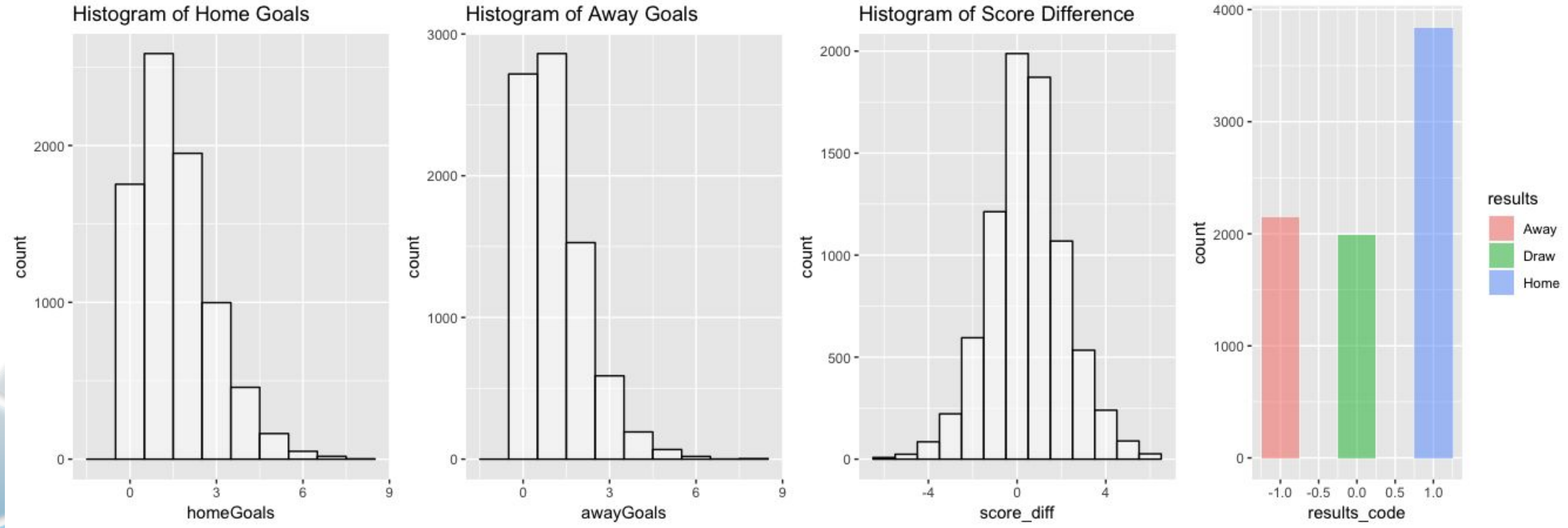
- Predicting the matches results have always attracted many attentions

- Bayesian approach could be very helpful in this scenario (**given reliable historical data**).

- We will work on the data from the **Spanish League** with 2 approaches
  - From Kaggle: https://www.kaggle.com/ricardomoya/football-matches-of-spanish-league

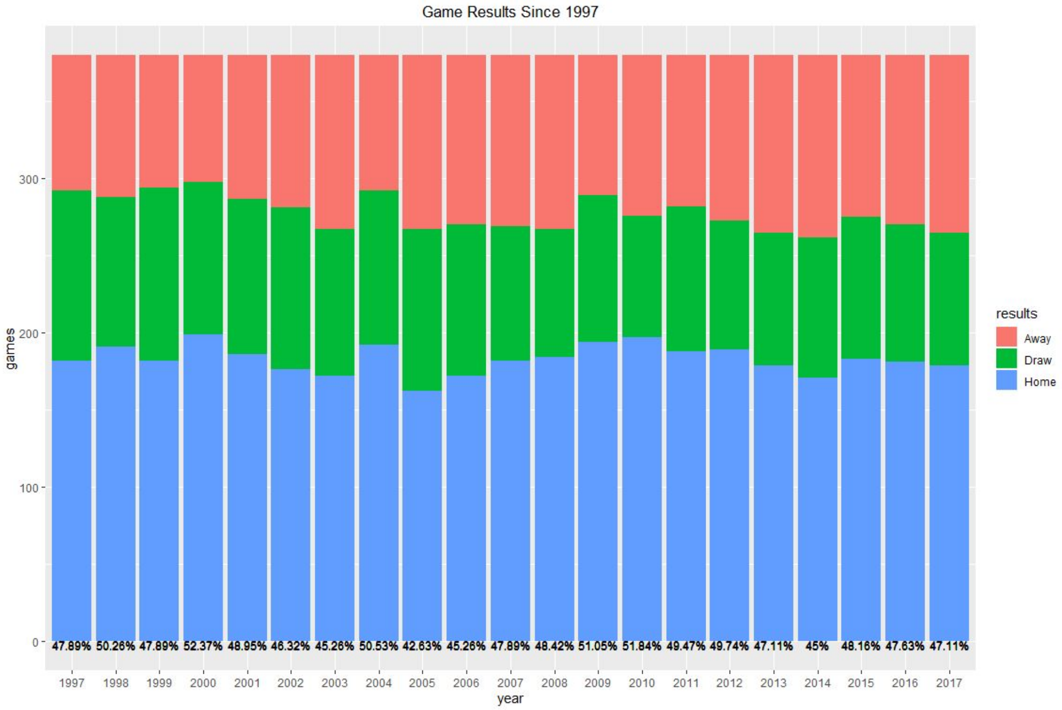
EDA - Home Advantage

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



Results Proportion

- If there's no home advantage, the proportion of home win/draw/away win would be the same
- Home win ~48%
- Home Advantage confirmed!



THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Approach 1 - Model Set-up

- 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

$$\text{performance variation} \sim \text{dunif}(0,2)$$

$$\text{home\_advantage} \sim \text{dunif}(1,1.5)$$

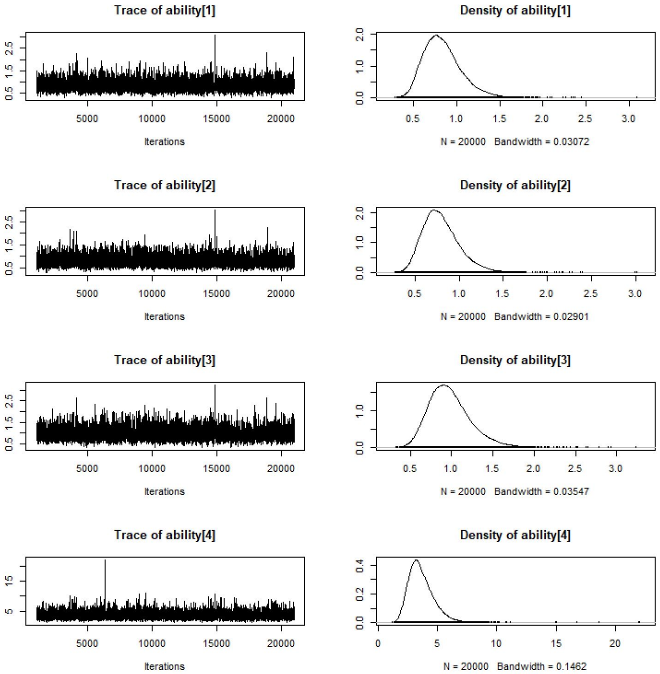
$$\log(\text{ability}) \sim \text{dnorm}(0, 1/\mathit{performance\_variation}^2)$$

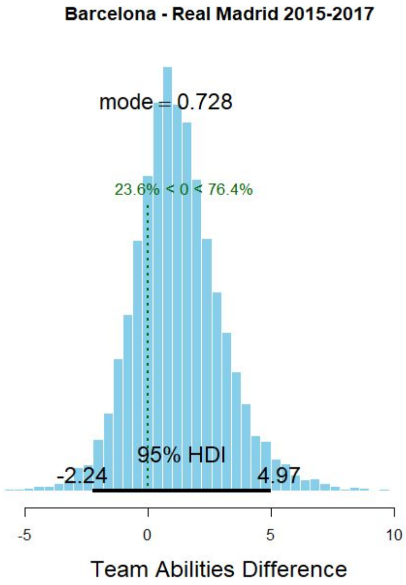
$$\text{ability}[i] \sim \exp(\log(\text{ability}))$$

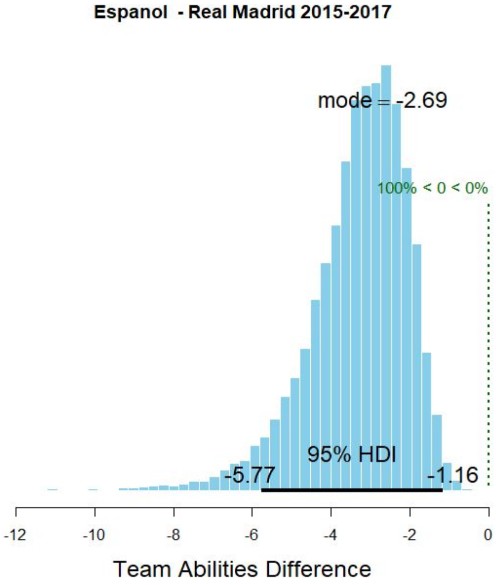
$$\text{prob}[i] = \frac{\mathit{ability}(\mathit{home}[i]) * \mathit{home\_advantage}}{\mathit{ability}(\mathit{home}[i]) * \mathit{home\_advantage} + \mathit{ability}(\mathit{away}[i])}$$

$$\text{yi (home win)} \sim \text{dbin}(\text{prob}[i], 1)$$

Approach 1 - Team Comparison







THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Approach 1 - Prediction

Season 2017, 380 games total

Modeling-set: 200

Predicting-set:180

How should we bet?

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 %

Interesting Findings

Rank alignment and mismatch

Reasons:

1. Strong teams are really dominating, thus easier to predict
2. Many factors that could impact the game results, such as referee decisions, core player injury, weather conditions, especially for teams with very close abilities
3. Limited data at hand, we can't include other factors into our model

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

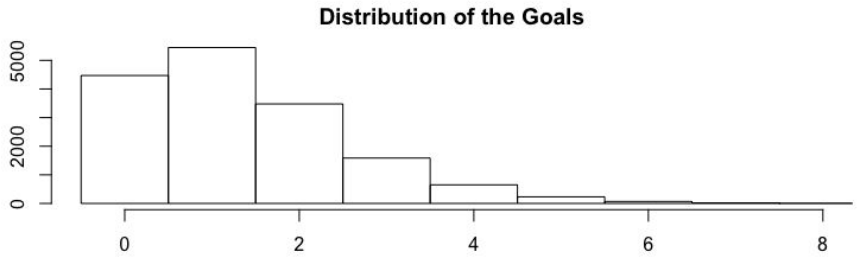
Predicted team abilities rank

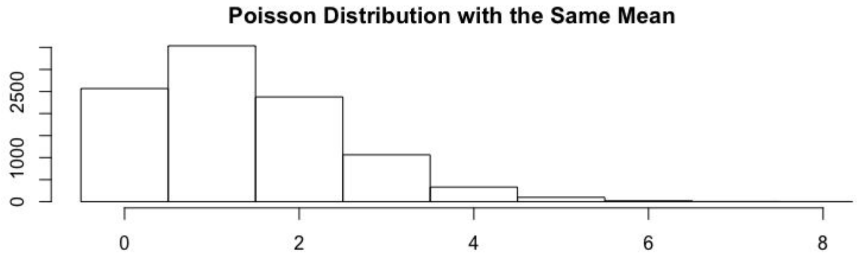
	ability avg	rank
Barcelona	5.1342212	1
Atletico de Madrid	2.3655118	2
Real Madrid	1.9297434	3
Valencia	1.9251803	4
Villarreal	1.4372378	5
Girona	1.2799012	6
Getafe	1.2415454	7
Sevilla	1.0836440	8
Atletico de Bilbao	1.0830023	9
Espanol	1.0716992	10
Betis	0.9622172	11
Celta de Vigo	0.9267895	12
Leganes	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

Approach 2 - Model Set-up

- In ideal poisson distribution, 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.

- In reality, if the teams are all the same, the game is meaningless, and the distribution of real data also prove it.





THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Approach 2 - Model 1 Set-up

The skill of one team minus another can predict the result of the game.

$$Goals \sim \text{Poisson}(\lambda)$$
$$\log(\lambda) = \text{baseline} + \text{skill}_i - \text{skill}_j$$

The baseline is assumed that both teams are equally good:

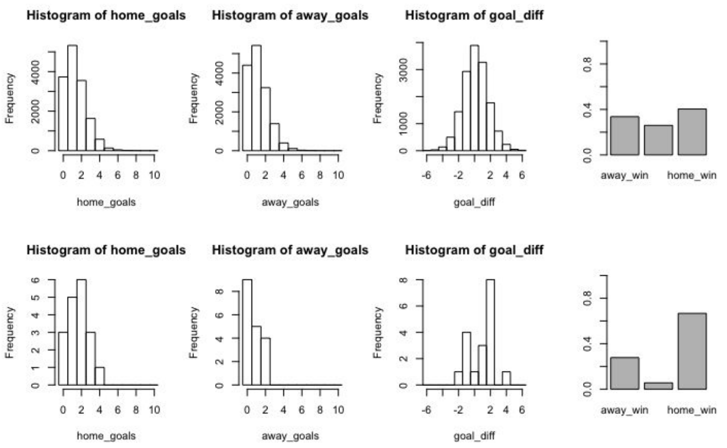
$$HomeGoals_{i,j} \sim \text{Poison}(\lambda_{\text{home},i,j})$$
$$AwayGoals_{i,j} \sim \text{Poison}(\lambda_{\text{away},i,j})$$
$$\log(\lambda_{\text{home},i,j}) = \text{baseline} + \text{skill}_i - \text{skill}_j$$
$$\log(\lambda_{\text{away},i,j}) = \text{baseline} + \text{skill}_j - \text{skill}_i$$

The prior distribution of baseline and skills is set as follows:

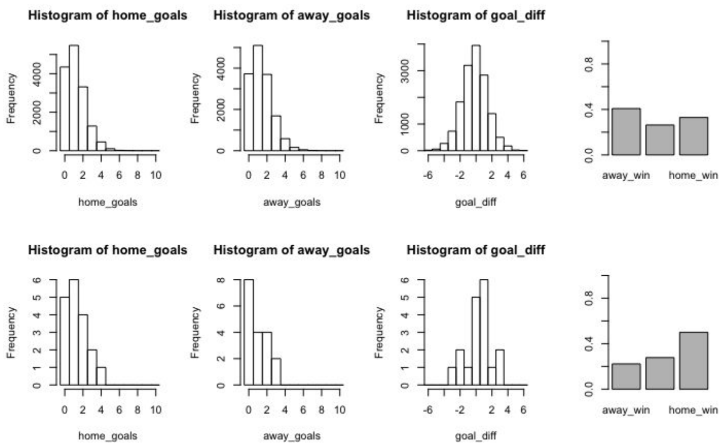
$$\text{baseline} \sim \text{Normal}(0, 4^2)$$
$$\text{skill}_{1\ldots n} \sim \text{Normal}(\mu_{\text{teams}}, \sigma^2_{\text{teams}})$$
$$\mu_{\text{teams}} \sim \text{Normal}(0, 4^2)$$
$$\sigma_{\text{teams}} \sim \text{Uniform}(0, 3)$$

Approach 2 - Model 1 Evaluation

Valencia home to Sevilla away:



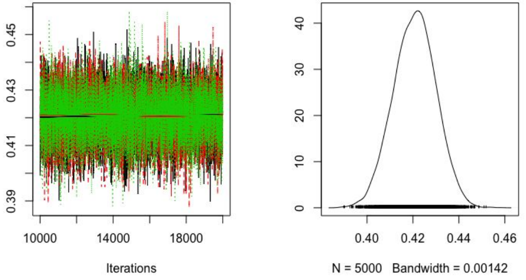
Sevilla home to Valencia away:



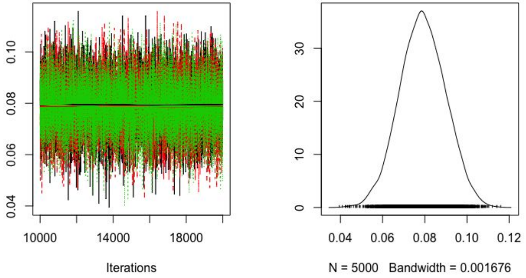
To predict the goal of Valencia and Sevilla. We can check the following barplot for different home cases. The first row in the figure shows the simulation and the second row shows the historical data.

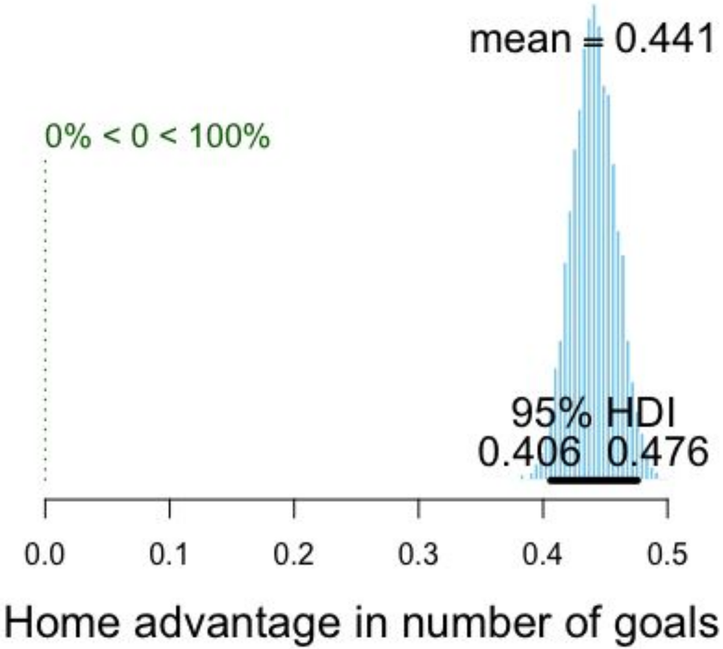
Approach 2 - Model Updating

Home Baseline



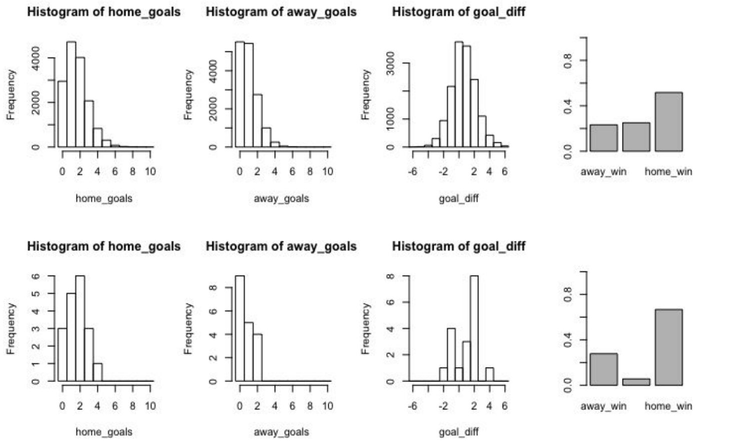
Away Baseline



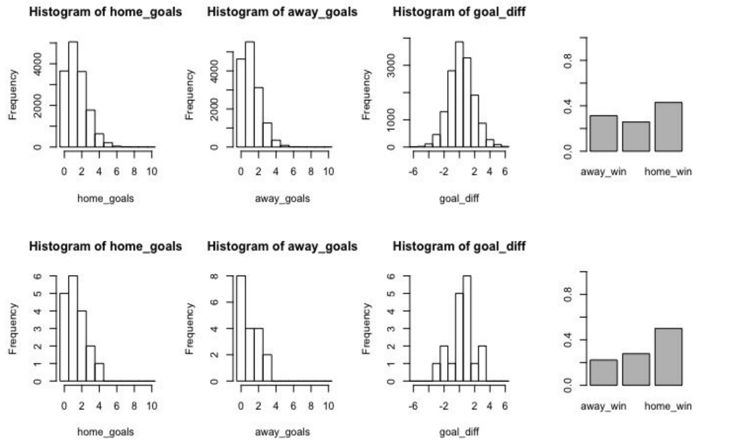


Approach 2 - Model 2 Evaluation

Valencia home to Sevilla away:



Sevilla home to Valencia away:

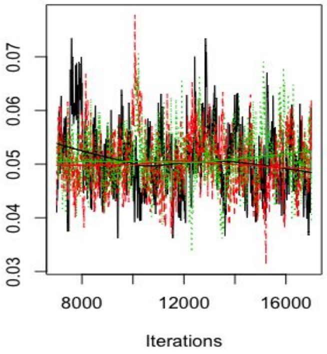


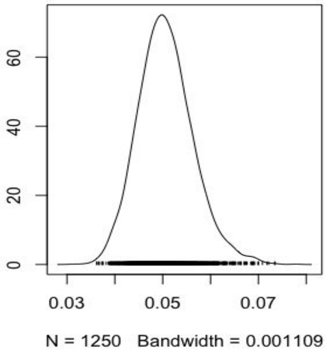
To predict the goal of Valencia and Sevilla. We can check the following barplot for different home cases. The first row in the figure shows the simulation and the second row shows the historical data.

Approach 2 - Model Updating

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:

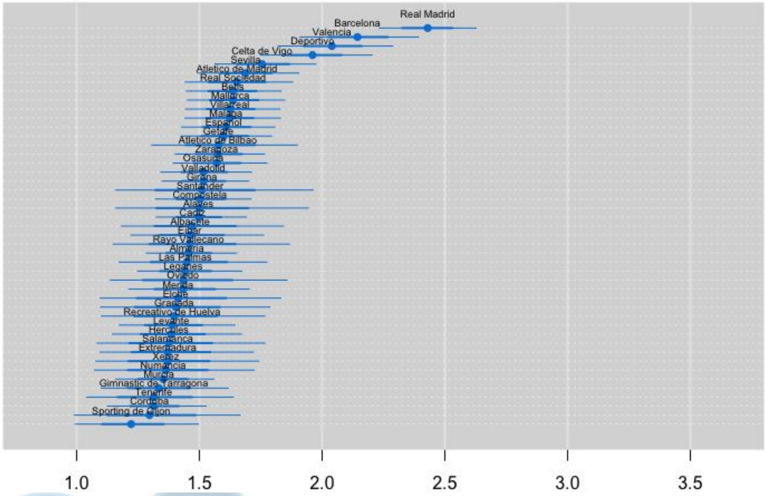
$$skill_{\,t+1} \sim \text{Normal}\,(skill_t,\, \sigma^2_{year})$$



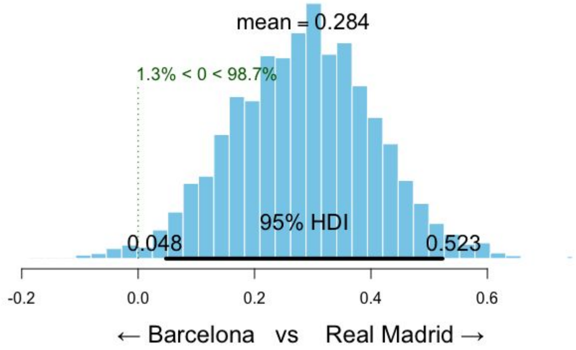


Approach 2 - Model 3 Evaluation

Ranking plot based on Model 3:



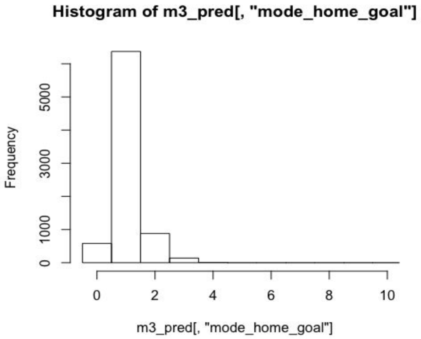
Team skills of Real Madrid - Team skills of Barcelona:

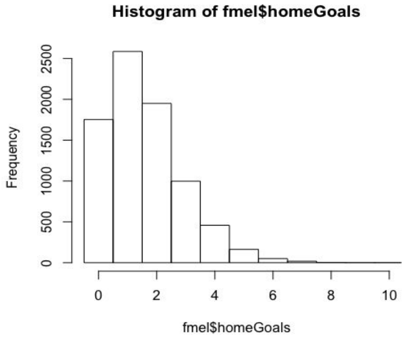


THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

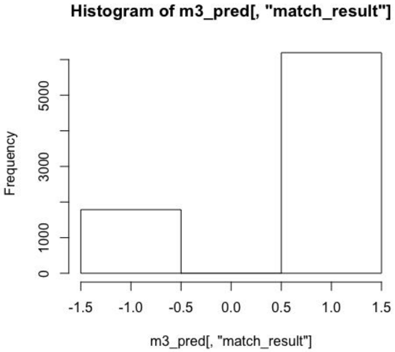
Approach 2 - Prediction

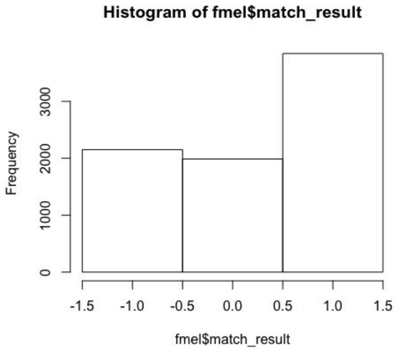
- 1 home goal has the highest probability.
- Prediction accuracy: 0.3310777.
- Mean square error : 1.464646.





- Prediction of win or lose.
- Prediction accuracy: 0.5358396.





THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Analysis of Two Approaches

Approach 1

- Good Accuracy
- Missed goal information
  - 6:1 home win - 2:1 home win
- Other outcomes are not included

Approach 2

- Able to include all game outcomes
- Normal accuracy
  - 97-17, there will be many roster changes
  - Some teams good at attacking, while some teams good at defending
  - Modelling goals rather than “ability to win”

Conclusion

Summary

- Many factors can affect the results of football games, game results prediction is hard
- Demonstrated 2 relatively straightforward approaches
  - Home win - Binomial
  - Goals - Poisson

Future Work

- Get more data, include other factors into model
- More complicated model, or combine the power of our 2 approaches

Reference

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

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

Q & A

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC