Predicting English Premier League match results using Bayesian Hierarchical Model

Sung Min Ha and Kaushik Dutta

Introduction :

The statistical modelling of sports data has become increasingly popular in the last few years after the advent of machine learning, which gives the ability to construct predictive models with unprecedented accuracy. For ages the betting companies have employed statistical intuition based on the team's performance to calculate odds for predicting football results. But the biggest challenge in accurately predicting the results of a football match are the uncertainty pertaining to the results and thus it cannot be realized by a linear model. The football results prediction is based on complex non linear parameters encompassing a team's performance over a varying timeframe and depending on a wide range of factors, ranging from player performance statistics to home field advantage. To accommodate the non-linearity of the data, we propose a Bayesian Hierarchical Model that can be employed to predict the ranking of the English Premier League season alongwith a prediction of results and scores of a particular match in the season. A major reason behind choosing English Premier League is because it is highly competitive in nature and there is a high incidence of upsets (where weak teams outscore strong teams). The world was awed when the football club Leicester City won the championship in 2016, against all odds, but this demonstrates the highly unpredictable nature of the game, and therefore the difficulty of the problem we are trying to tackle.

Dataset and Preprocessing :

For this project we considered the English Premier League 2017-2018 season data. The English Premier League is the first division football league of the UK contested by 20 teams where each team plays against each other twice (once as home and another as away team). A total of 380 matches are played in a season and the team with the highest point is declared winner. The top four teams of the league table goes on to play in the UEFA Champions League. The next three teams qualify for the Europa League and the bottom three teams are relegated to the Second Division league. We obtained our dataset from http://www.football-data.co.uk/data.php. The dataset has match statistics for all the matches in the season in terms of goals scored, corners, shots, fouls, red and yellow cards for both home and away teams. For preprocessing we trimmed the dataset to only include features which directly or indirectly affects the final match results and arranged the matches in a longitudinal fashion according to their dates.

Feature Selection and Feature Engineering:

The quality of results in the prediction task is directly associated to the quality of the feature set used for modelling the system hence choosing the right features is of paramount importance. We divided the available features into two categories : 1) Attack and Defence Parameters consisting of the feature like goals scored and full/half time results which directly affect the match results and 2) Indirect Features consisting of the feature like corners, fouls, shots etc which provides us with mathematical insights regarding team performance and influence the match result in an indirect way.

When designing our model's features for the prediction of football results, we wanted to introduce features that could provide us with useful quantifiable insights for judging the recent performance of a team. To this end we engineered two unique features called *Form* and *Streak.* These features enhance the predictive power of the model by incorporating the fact that recent performance of the team has the ability to influence the current game. The mathematical formulation for the engineered features are given as :

Streak *:* This feature encapsulates the recent improving/declining trend in the performance of a team. The Streak value for a team is computed by assigning a score to each match result and taking the mean of the previous *k* scores, where *k* is a hyper-parameter. We also included a temporal dimension to the Streak feature by placing time-dependent weights on the scores of the previous games of a team, obtaining a feature that we refer to as the Weighted Streak (giving greater weights for recent games, and decreasing gradually for non-recent games).

$$\delta_j = \sum_{p=j-k}^{j-1} \frac{2(p-(j-k-1)res_p)}{3k(k+1)}$$

where, δ_j is the Weighted Streak of a team in the j^{th} match and the res_p is point each team gets as per the result of the match (0-Loss, 1-Draw , 3-Win)

Form: This feature encapsulates the recent performance of the team relative to its opponents. The *Form* value of each team is initialized to one at the beginning of each season and then updated after each match according to the result of the match i.e. win,draw,loss. Our mathematical formulation of *Form* ensures that a greater coefficient update is provided if a weak team triumphs over a strong team, and vice-versa. In the case of a draw, the Form of a weak team increases while that of a strong team decreases.

When a team α beats β we can write the *Form* (ξ) equation for the teams as:

$$\xi_j^{\alpha} = \xi_{j-1}^{\alpha} + \gamma\xi_{j-1}^{\beta}$$

$$\xi_j^{\beta} = \xi_{j-1}^{\beta} - \gamma\xi_{j-1}^{\beta}$$

where γ is the stealing fraction signifying the weight to be added/subtracted in case of win/loss

Model Description:

The Poisson Distribution is widely acceptable as a modelling approach for the distribution of the number of goals involving two competing teams. In this problem we are assuming two conditionally independent Poisson variables for the number of goals scored, correlation is taken into account, since the observable variables are mixed at an upper level. Moreover, as we are framed in a Bayesian context, prediction of a new game under the model is naturally accommodated by means of the posterior predictive distribution. As discussed the league is played by $N = 20$ teams and the number of goals scored by the home and the away team in g^{th} game (g = 1,2,3,…,380) of the season given by $\mathbf{y_{g1}}$ and $\mathbf{y_{g2}}$. The observed count vector is $\mathbf{y} = (\mathbf{y_{g1}}, \mathbf{y_{g2}})$, which are independent Poisson random variables given by Eqn. 1. The parameters $\boldsymbol{\theta} = (\boldsymbol{\theta_{g1}}, \boldsymbol{\theta_{g2}})$ gives the scoring intensity in the g^{th} game of the season and j = 1 for home team and j = 2 for away team.

$$y_{gj} | \theta_{gj} = Poisson(\theta_{gj}) \tag{1}$$

We model the scoring intensity parameters as a log linear random effect model based on the work of Karlis and Ntzoufras (2003) and is given by Eqn. 2 and Eqn. 3for the home and away.

$$log\, \theta_{g1} = home + att_{hg} + def_{ag} + intercept \tag{2}$$

$$log\, \theta_{g2} = att_{ag} + def_{hg} + intercept \tag{3}$$

The *home* parameter represents the advantage of the home team playing in favorable conditions and is designed to be a constant distribution for all teams across the season. The other two parameters are the *att* and *def* i.e. the attack and defense ability of the team. The goaling intensity of a team is calculated by the addition of its attacking ability and the opponents defensive ability. The attack and defense ability parameters are computed by the cumulative effect of the following parameters i.e. number of goals, shots and shots on target, corners, fouls, red and yellow cards, form and streak for both home and away team.

$$att_g = att_{goals} + att_{shots} + att_{shots-target} + att_{corners} - att_{fouls} - att_{yellow-card} - att_{red-card} + att_{streak} + att_{form}$$

$$def_g = def_{goals} + def_{shots} + def_{shots-target} + def_{corners} - def_{fouls} - def_{yellow-card} - def_{red-card} + def_{streak} + def_{form}$$

The attacking and defensive parameters are positively affected by goals, shots, corners, streak and form while they are negatively impacted by fouls and red/yellow cards (hence they are subtracted from the net attack and defense ability of the team). Now that we have the model parameters ready we have to specify some suitable prior distributions for the random parameters.

The *home* parameter is modelled as a fixed effect across season for all teams, hence it is assumed as flat standard prior Normal Distribution given in Eqn. 4. The attack and defence ability are modelled separately for each parameters for each teams T = (1,2,....,20) with a normal distribution having mean μ_{att} and variance τ_{att} and given in Eqn. 5.

$$home \sim Normal\,(0, 0.0001) \tag{4}$$

$$att_t \sim Normal\,(\mu_{att}, \tau_{att}) \qquad def_t \sim Normal\,(\mu_{att}, \tau_{att}) \tag{5}$$

As suggested by the works of Karlis and Ntzoufras (2003), we need to impose some identifiability constraints on the team-specific parameters. Hence, we use sum-to-zero constraints for all the parameters used to compute attacking and defensive ability i.e. $\sum_{t=1}^{T} att_t = \sum_{t=1}^{T} def_t = 0$. Finally, the hyper-priors of the attack and defense effects are modelled independently using again flat prior distributions given by Eqn. 6 and Eqn.7.

$$\mu_{att} \sim Normal\,(0, 0.0001) \qquad \mu_{def} \sim Normal\,(0, 0.0001) \tag{6}$$

$$\tau_{att} \sim Gamma\,(0.1, 0.1) \qquad \tau_{def} \sim Gamma\,(0.1, 0.1) \tag{7}$$

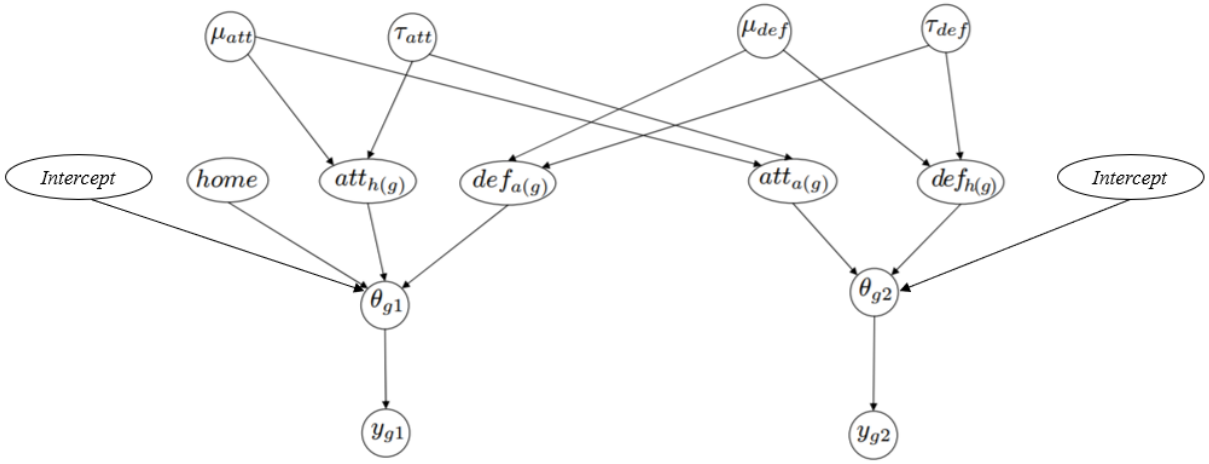


Figure 1. The Representation of the Hierarchical Model

The inherent hierarchical nature of the model kind of implies that there is an effect on the observed variable vector ***y*** due to the latent unobservable hyperparameters $\eta = (\mu_{att}, \mu_{def}, \tau_{att}, \tau_{def})$. In fact, the components of η represent a latent structure that we assume to be common for all the games played in a season and that determine the average scoring rate. Each game contributes to the estimation of these parameters, which in turn generate the main effects that explain the variations in the parameters ***θ*** and therefore implying a form of correlation on the observed counts ***y.***

Results :

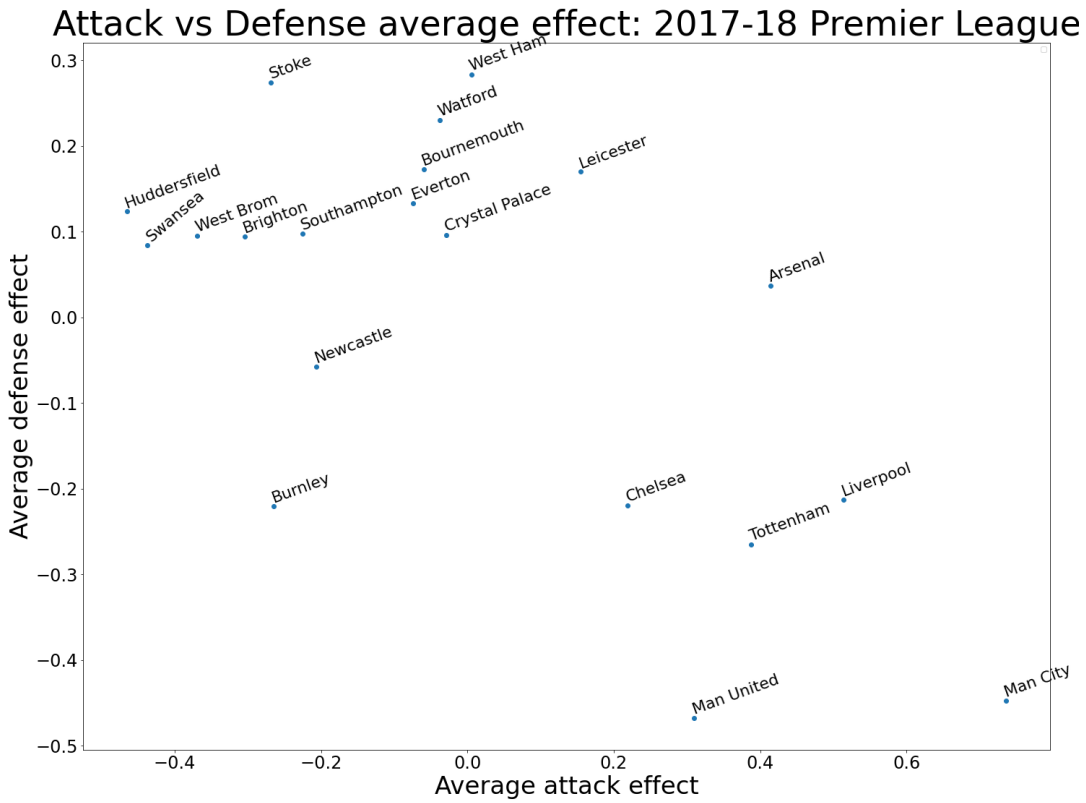


Figure 2. Attacking Ability VS Defensive Ability average effect for each team. Higher performing teams will appear at bottom right, and the lower performing teams will appear at top left.

According to the Bayesian approach the objective of our modelling is two-fold: first we wish to estimate the values of the attack and defence abilities of each team through which we can explain the scoring rate. This task is accomplished by the entering evidence observed in the goal scored vector (y) i.e. the likelihood and updating the prior distributions by a Markov Chain Monte Carlo (MCMC) procedure. Table 1 summarizes statistics for the posterior distributions of the coefficients for the log-linear model describing the scoring intensity. The average attacking and defensive ability for each team is depicted in Fig. 2. We observe that the top four teams of that season (Man City, Liverpool, Man United, Tottenham) have higher posterior value for *att* and lower posterior value for *def* parameter. Given the *att* and *def* parameters are zero-centered, positive *att* parameters would positively affect the scoring ability of the team, while negative *def* would negatively affect the scoring ability of the opposing team. Man City, the league champion

for that year, has the highest posterior for attack and the lowest posterior for defence as shown in Table 1. We also observe that for the bottom three teams (i.e. Swansea, West Brom and Stoke) have higher defensive posterior and lower attacking posterior suggesting they have more tendency to accede goals than scoring them. We also assumed that there is an inherent advantage for a team playing at home conditions and it is validated by the positive posterior mean for the home parameter of 0.284 (95% CI, 0.169-0.401). We also considered an intercept parameter to give a uniform weightage to all the teams and also for the evaluation of away team scoring propensity. It also has a positive posterior mean of 0.037 (95% CI, -0.063 - 0.136) signifying the away team factor also has some effect though not as dominant as the home team factor (Table 2). Furthermore, using equation 2 and 3, where $att_{h(g)}$ and $def_{a(g)}$ are zero-centered and $home$ and $intercept$ are constant values over a season and across all teams, we note that on average, the difference between home team score and away team score is predicted as $e^{home+intercept+0} - e^{intercept+0} \approx 0.3408$. With the observed data, we observed the average home score of 1.53 and away score of 1.15 for the season, resulting in an observed difference in scores of 0.38 goals. We note from the similar value of predicted difference of 0.3408 to actual difference of 0.38 in home and away scores that the model accurately estimated the home team advantage.

Team	Attack				Defense			
	mean	2.50%	median	97.50%	mean	2.50%	median	97.50%
Arsenal	0.16588	-0.2695	0.1714	0.50094	-0.0359	-0.275	-0.0556	0.21784
Brighton	-0.0884	-0.4394	-0.0954	0.35683	0.18618	-0.2253	0.1795	0.53247
Chelsea	0.01775	-0.4217	0.02633	0.38894	-0.1057	-0.5569	-0.1002	0.26632
Crystal Palace	0.16021	-0.0491	0.15133	0.39569	0.0497	-0.2823	0.01903	0.38524
Everton	-0.0762	-0.521	-0.025	0.23009	0.02192	-0.3693	0.01814	0.42418
Southampton	-0.0936	-0.533	-0.0679	0.26957	-0.0109	-0.2643	-0.0197	0.30495
Watford	0.13321	-0.3201	0.14177	0.53763	0.07882	-0.2791	0.06812	0.47623
West Brom	-0.1677	-0.5749	-0.1472	0.14236	0.04829	-0.2415	0.06221	0.35666
Man United	0.1574	-0.2284	0.15023	0.58351	-0.2089	-0.6698	-0.2023	0.1697
Newcastle	-0.1536	-0.5521	-0.1565	0.23599	0.00033	-0.233	0.013	0.2253
Bournemouth	-0.1121	-0.3705	-0.1058	0.1163	0.07689	-0.242	0.06963	0.4748
Burnley	-0.1398	-0.5708	-0.1275	0.29761	-0.066	-0.4522	-0.0589	0.25765
Leicester	0.10131	-0.1865	0.09265	0.43796	0.07916	-0.2879	0.07243	0.47187
Liverpool	0.17875	-0.2033	0.15436	0.57924	-0.1017	-0.482	-0.0999	0.25878
Stoke	-0.1221	-0.4594	-0.1061	0.25974	0.14927	-0.2038	0.1518	0.50423
Swansea	-0.178	-0.6732	-0.1461	0.31462	-0.0371	-0.4241	-0.0127	0.25693
Huddersfield	-0.1789	-0.6387	-0.1614	0.31492	0.07061	-0.2472	0.07273	0.40095
Tottenham	0.12312	-0.2371	0.11147	0.59363	-0.1254	-0.5289	-0.1179	0.17884
Man City	0.30424	-0.0658	0.29461	0.71546	-0.1742	-0.6236	-0.1681	0.24899
West Ham	-0.0315	-0.1341	-0.0292	0.07753	0.10457	-0.254	0.10108	0.46063

Table 1. Statistical Estimates on Attack and Defense Effects of the Model

	mean	2.50%	median	97.50%
Home	0.284818805	0.169524643	0.28921925	0.40125784
Intercept	0.03767081	-0.063000434	0.039115445	0.136080733

Table 2. Statistical Estimates on Home and Intercept Effects of the Model

The second and more interesting objective of the model is the prediction of the match results. We can use the results derived in the implied posterior distributions for the vector $\boldsymbol{\theta}$ to predict a future occurrence of a similar game. To that end, we produced a vector of 1000 replications for the posterior predictive distribution of $\boldsymbol{y}$ that we used for the purpose of model checking. From the 1000 model simulations, we took the median and the 95% CI for each team as shown in Fig. 3. In addition we observed per match predictive points against observed per team in Fig. 4, where for most of the teams the Bayesian model produced a good fit to the observed points. There are some disparities in prediction as we can observe from Fig 3., Supplementary Fig. 1, and Supplementary Fig. 2 for the higher and lower performing team due to overshrinkage. Under this effect, all parameters have the tendency to be pulled towards their group mean, hence top performing teams have lower predicted goals/points than the observed whereas bottom performing teams have higher predicted goals/points than the observed. Contextually, the top performing and bottom performing teams may have large deviations in their performances from the group mean, such that the current model where the attack and defense effects are drawn from a single common distribution may not be sufficient to capture all the variance in teams' performances. Fig 2. also supports this notion, as we note the top few teams are sparsely distributed on the bottom right and most of the average teams are clustered tightly near the center of the scatterplot. Table 3 depicts the league table signifying goals scored, goals conceded and total points for both the actual season 2017-18 and for the Bayesian Predictive model. The Bayesian Model successfully identifies the champion for the season i.e. Man City with a predictive point of 91 as compared to the actual point = 100. The model also successfully identifies the top 4 teams of the season (which proceeds to the Champions League) when compared to the actual data. The model successfully identifies two out of the next three teams (Europa League). For the bottom three teams which are relegated, the model identifies two (Swansea and Stoke) correctly and misses out on West Brom mainly due to the overshrinkage issue which incentivizes West Brom to have higher points.

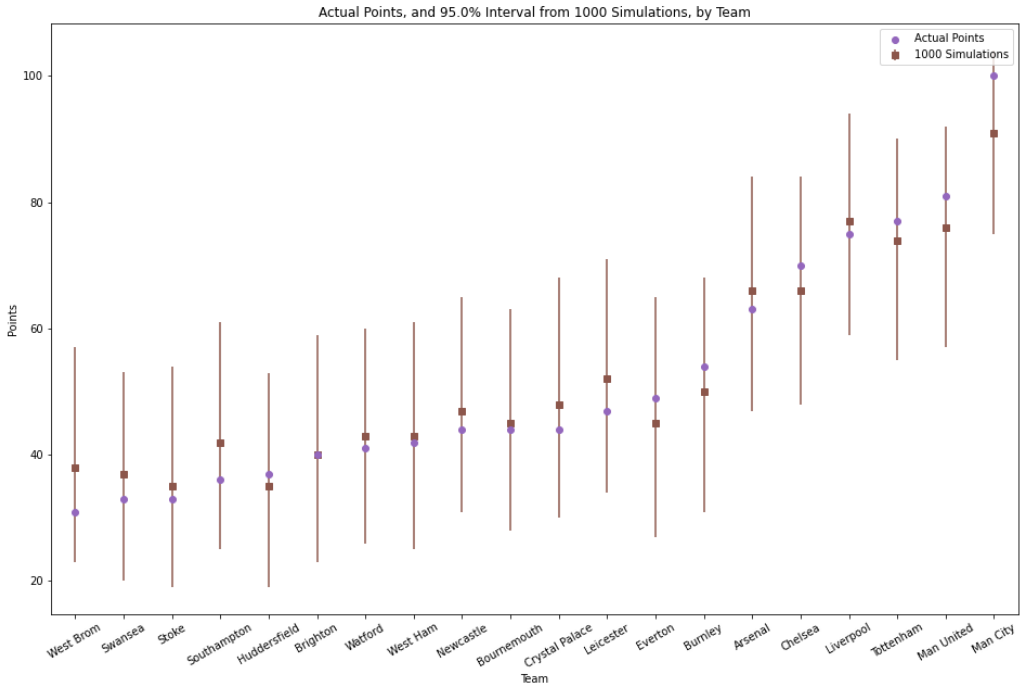


Figure 3. Actual Points of the Season VS Median Points from the Simulated Seasons along with the 95% CI for each team

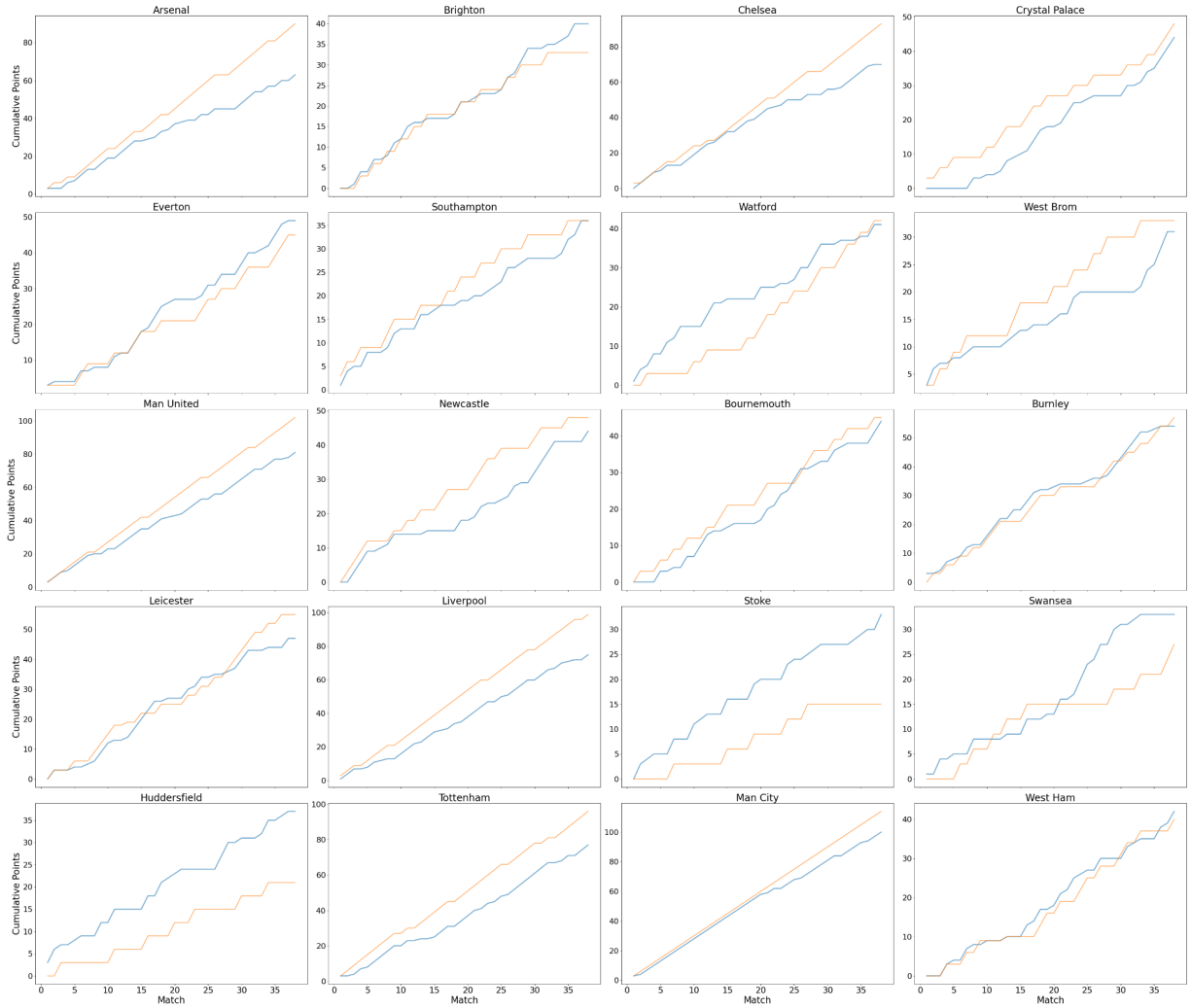


Figure 4. Posterior predictive validation of the hierarchical model: the blue line represents the actual cumulative points acquired by each teams through the season, while the orange line represents the cumulative points acquired by each teams through the season as predicted by the Bayesian Hierarchical Model

Team	Observed			Bayesian Model		
	Points	Goals Scored	Goals Conceded	Points	Goals Scored	Goals Conceded
Huddersfield	37	28	58	35	30	56
Swansea	33	28	56	37	31	54
West Brom	31	31	56	38	32	55
Brighton	40	34	54	40	34	54
Stoke	33	35	68	35	36	66
Burnley	54	36	39	50	37	39
Southampton	36	37	56	42	38	54
Newcastle	44	39	47	47	39	47

Everton	49	44	58	45	44	56
Watford	41	44	64	43	45	62
Bournemouth	44	45	61	45	44	58
Crystal Palace	44	45	55	48	46	54
West Ham	42	48	68	43	47	65
Leicester	47	56	60	52	55	58
Chelsea	70	62	38	66	60	39
Man United	81	68	28	76	66	30
Arsenal	63	74	51	66	72	50
Tottenham	77	74	36	74	71	37
Liverpool	75	84	38	77	80	38
Man City	100	106	27	91	101	30

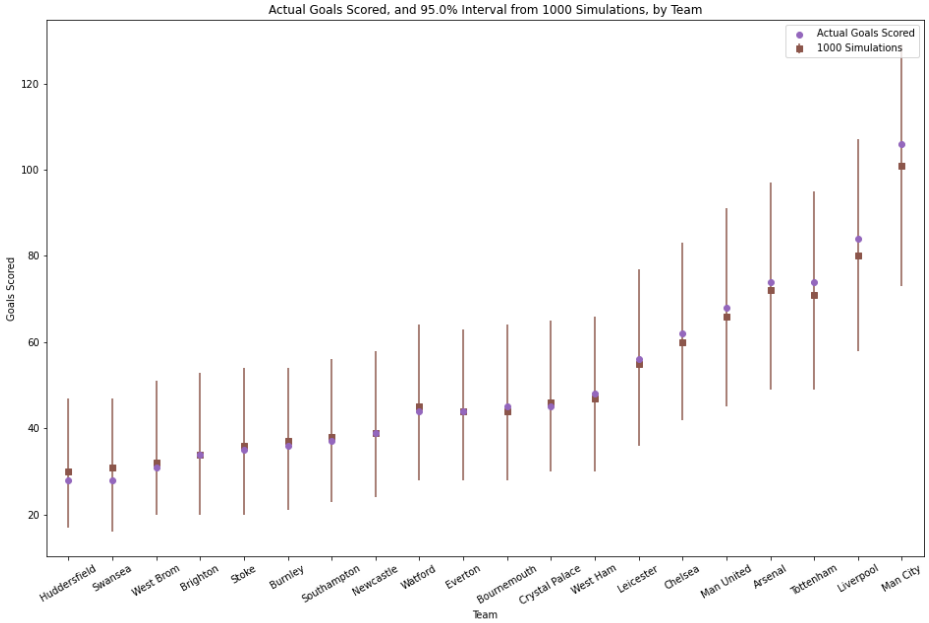
Table 3. Posterior Model Predictions of the Points and Scores against Observed Values

Conclusion and Future Directions:

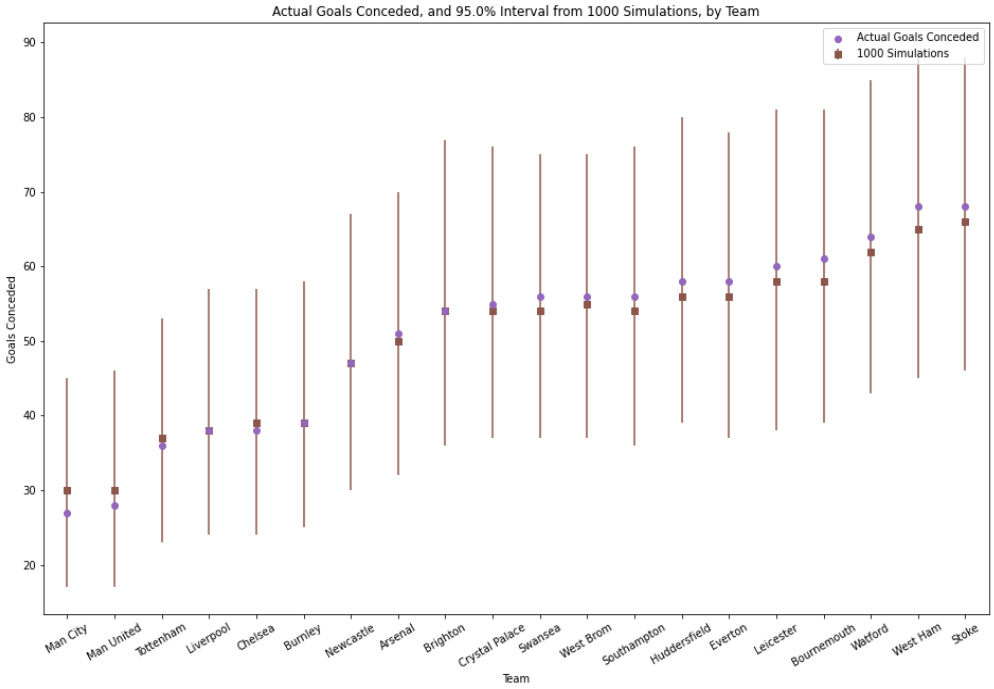
The designed Bayesian Hierarchical Model successfully simulated and predicted results and statistics of individual matches in a season. The model validated the fact that there is an undue advantage which the teams enjoy consistently while playing at home. Our model also incorporated multiple indirect features and engineered features to account for the perpetual performance of a team through the season. The overall predictions from the Bayesian model, as showcased on Fig. 2, 3, and 4, accurately matched the observed data with relatively small errors. As for future directions, there is the systematic pattern of overshrinkage observed that needs to be addressed to improve the performance of the model. With such large variance in the performance of the teams, particularly with the top and bottom few, we detected a consistent issue of overshrinkage for both prediction of the points and the goals scored/conceded. This overshrinkage issue can be addressed by utilising mixture model prior instead of a flat distribution prior to better capture the wide difference in the teams' attack and defense effects. Finally, we also hope to extend this model to calculate RPS score that is used by the betting organisations to guess match results.

References :

1. Karlis D, Ntzoufras I. Analysis of sports data by using bivariate Poisson models. Journal of the Royal Statistical Society: Series D (The Statistician). 2003 Oct;52(3):381-93.
2. Baio G, Blangiardo M. Bayesian hierarchical model for the prediction of football results. Journal of Applied Statistics. 2010 Feb 1;37(2):253-64.
3. Aitchison J, Ho CH. The multivariate Poisson-log normal distribution. Biometrika. 1989 Dec 1;76(4):643-53.
4. Karlis D, Ntzoufras I. On modelling soccer data. Student. 2000;3(4):229-44.
5. Baboota R, Kaur H. Predictive analysis and modelling football results using machine learning approach for English Premier League. International Journal of Forecasting. 2019 Apr 1;35(2):741-55.
6. Diniz MA, Izbicki R, Lopes D, Salasar LE. Comparing probabilistic predictive models applied to football. Journal of the Operational Research Society. 2019 May 4;70(5):770-82.
7. Sadeghkhani A, Ahmed SE. A Bayesian Approach to Predict the Number of Goals in Hockey. Stats. 2019 Jun;2(2):228-38.



Supplementary Figure 1. Actual Goals Scored of the Season VS Median Goals Scored from the Simulated Seasons along with the 95% CI for each team



Supplementary Figure 2. Actual Goals Conceded of the Season VS Median Goals Conceded from the Simulated Seasons along with the 95% CI for each team