Georgia Institute of Technology

ISYE 6416: Computational Statistics

Predicting Premier League Table

Abhijeet Mavi
(GTID: 903518738)

Arjun Kudva
(GTID: 902759271)

Felipe Simon
(GTID: 903413707)

Li Wei Yap
(GTID: 903526115)

Instructor:
Dr. Yao Xie

April 25, 2020

Contents

List of Figures

List of Tables

Abstract

English Premier League is one of the most fiercely competitive league based tournaments in the world with 20 teams from the top division in English soccer contesting against each other to aggregate points. One of the most interesting elements of this league is its unpredictability. Each game has a lot of uncertain features guiding it to a conclusive end where the winner is decided by a certain event, that is, scoring goals. We have tried to come up with two prospective models to determine the prediction of the final Premier League tables based on team records and their past history to guide them to the championship title. The first approach talks about the simple Poisson distribution to figure out the number of goals a team can score based in on its past record and characteristics. For the second approach we combine a Bayesian statistics with a Hidden Markov Model to predict the outcome of a match. With these two methods, we have explained the simplicity with which statistics can govern even the most unpredictable events and their ubiquity in our daily lives.

1 Workload distribution

Abhijeet and Dave: Poisson Distribution method
Arjun: Analysis of dataset and calculation of probabilities, report editing
Felipe: Dirichlet Distribution method

2 Introduction

With a potential global audience of 4.7 billion people across over 200 countries and territories, the Premier League is by far the most watched sports league in the world. The league represents England's highest division of soccer with 20 teams competing in a season spanning from August of one calendar year to May of the following year.

The competition follows a "double round-robin" format, in which each team plays the other 19 teams twice (once at home and once away) for a total of 38 games played by each team per season. Teams are awarded three points for each win, one point for each draw and zero points for each loss. The league table is sorted in descending order of points tally. Should two or more teams have an equal number of points, their standings are decided by highest to lowest goal difference (total number of goals scored minus total number of goals conceded). At the end of the season, the team at the top of the table is declared to be the league champions.

Like most other soccer leagues in Europe, the Premier League also follows a "promotion and relegation" system in which at the end of each season, the bottom three teams in the league are demoted (or relegated) to a lower division. Similarly, the top three teams in the lower division are then promoted to the Premier League to compete in the following season. As such, strong competition among teams exists at both the top of the league, where teams vie for the championship as well as at the bottom of the league where teams aim to avoid being in the bottom three.

As such, each season begins with much speculation and discussion over which team will be the champions and which three teams will be relegated. Predicting a final league table before a season starts (and indeed, even during a season itself) can be a challenging task, which this project will attempt to do.

3 Problem Statement

Predicting the final league table essentially involves predicting results of games between individual teams. While sporting events are inherently unpredictable with many possible sources of variation, there is information that can be used

to suggest which team has a higher chance of winning. This report will attempt to use historical performances of teams to predict their performance in an upcoming season, based on the final league standings.

4 Data source

The match results and final league standings for every season of the Premier League since its inception in 1992 can be found at most sporting websites. For this project, we used the past results and standings from the 2013-14 through to the 2017-18 seasons as training data and used the 2018-19 season to test the models (both mid-season and at the end of the season). The Premier League's Kaggle dataset was used as a source for the data.

5 Exploratory Data Analysis

In predicting the result of any game between two teams, the following criteria are considered:

- The location of the game: which team is at home and which team is away. It is expected for a team to perform better at home, where it will be more familiar with the climate and playing field conditions, will avoid the inconvenience of travelling to the venue and will have more fans supporting them in the stadium. Equally, a team playing away is usually perceived to have a disadvantage compared to when it is playing at home.

- Historical results between the two teams: a team that consistently beats another team when they play each other can be considered to be "better" and would probably be expected to win when they play each other next. Of course, there can be other factors too. Teams could now have new players in their roster or have existing players missing through injury among other things which can make historical results unreliable.

- Recent results of the teams playing: a team that is on a winning streak will be expected to approach a new game with confidence and may even consider themselves capable of beating otherwise much superior opposition. On the other hand, a team that is not winning games is likely to be more nervous, have lower morale and may be more likely to lose or draw games that they would have normally been expected to win.

As such, we can consider two different methods to predict a result of a match: predicting the number of goals that will be scored by either team to determine a result and attempting to directly predict the probability of a win, loss or draw for a team based on its historical performance against a particular opponent at a particular venue.

Based on our findings from the recently concluded season, we see that there is a significant correlation of around **0.9852** between goal difference, that is the difference of goals scored and goals conceded, shows a direct correlation to our final standings. A team with basically higher points will have a higher goal difference.

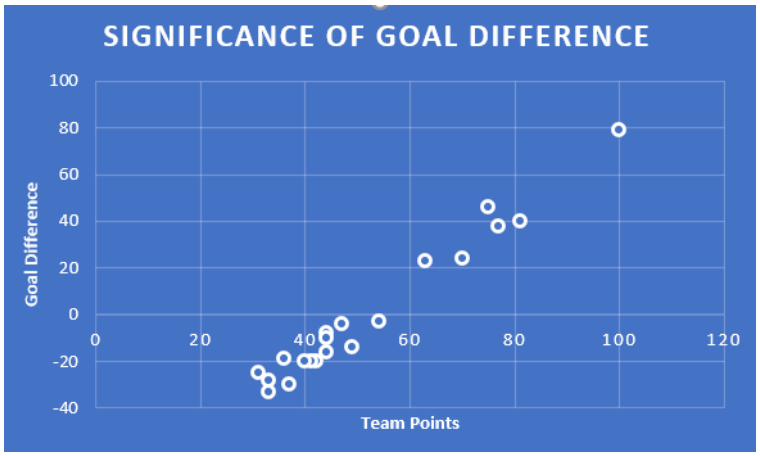

Figure 1: Points vs Goal Difference

Moreover shots taken by a team did not show any significant correlation, just around **0.703** which shows that just taking shots wildly is not how once can win games.

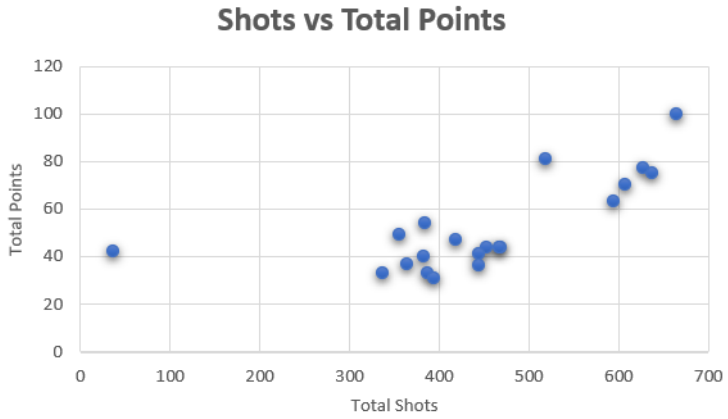

Figure 2: Points vs Shots

An interesting statistic emerges from when we have **shot creating actions** such as defence-piercing through balls, long crosses, key passes and lob passes.

These actions have a high correlation **(0.8738)** to a team's success on points board as shown below.

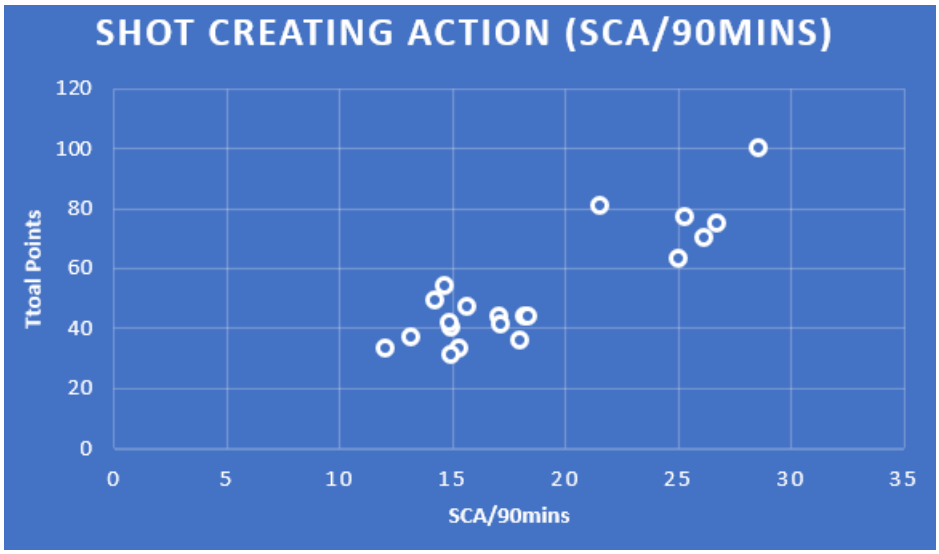

Figure 3: SCA(per 90 mins) vs Shots

Hence, we will proceed with analysing the total amounts of goals difference in our approach for poisson distribution approach as that is the parameter which has high correlation with total points. Obviously, there are other parameters also that influence the game but we will show how simplistic the approach is with goal difference as parameter.

6 Bayesian Analysis of Matches using Dirichlet Distribution

The explanatory data analysis showed certain trends on the outcomes when two teams played. This gave us the idea of simplifying the problem into one where instead of having to know the final score we could simply try to predict the outcome (Win, Draw Loss).

6.1 Bayesian Model

Given that there is still plenty of uncertainty on an outcome of a match we chose a Bayesian approach to model the probabilities of each of the outcomes:

$$\theta_{ij} \sim Dir(0.5, 0.5, 0.5)$$
$$y_{ij}|\theta_{ij} \sim Categorical(3, \theta_{ij})$$

The categorical variable y_{ij} represents the outcome of the match when team i plays team j (Win, Draw, Loss). The Dirichlet variable θ_{ij} is the prior for the probabilities of each team winning. As the priori, we saw all outcomes equally likely. As such, we chose an uninformative prior for the Dirichlet distribution. We want to know the posterior of θ_{ij} given the outcomes of all the matches these teams have played. Using Bayes theorem:

$$P(\theta_{ij}|y_{ij}) = \frac{P(y_{ij}|\theta_{ij})P(\theta_{ij})}{P(y_{ij})}$$

Because the Dirichlet distribution is a conjugate prior of the categorical variable we know the posterior without having to compute it:

$$\theta_{ij}|y_{ij} \sim Dir(0.5 + i\ wins, 0.5 + draws, 0.5 + j\ wins) \tag{1}$$

Where $i\ wins$ is the number of times team i has won this fixture, $draw$ is the number of draws between the two team and $j\ wins$ is number of times team j won this fixture.

6.2 Hidden States

As we still felt that the current state of the team is important when predicting an outcome we decided to further improve our model. To do this we created a dataset for each team with the following data for each match the team has played: *number of wins on the last 5 matches*, *number of losses on the last 5 matches*, *average number of goals in favor on the last 5 matches*, *average number of goals against on the last 5 matches*, *average number of attempts on the last 5 matches*. Using K-Means we created clusters for the matches. We then repeated this procedure separately for each team given that a good state for one team at a given time might not be the same as for another team. An example of such a clustering is shown in the figure below for the team "Arsenal".

Figure 4 shows an example of the clusters derived for Arsenal. Each point represents a match played by Arsenal. We can see that the cluster divide nicely between winning averages and goals against. We use each of the clusters as the hidden state of the team. To calculate transition probabilities we use the following formula:

$$TP_{ij}^{(k)} = \frac{\sum_{t=1}^{n} I(cluster_t = i, cluster_{t+1} = j)}{\sum_{t=1}^{n} I(cluster_t = i)} \tag{2}$$

Where $TP_{ij}^{(k)}$ is the transition probability of going from state i to state j for team k. $I()$ is the indicator function and $cluster_t$ is the cluster in match t. n is the number of matches played by team k.

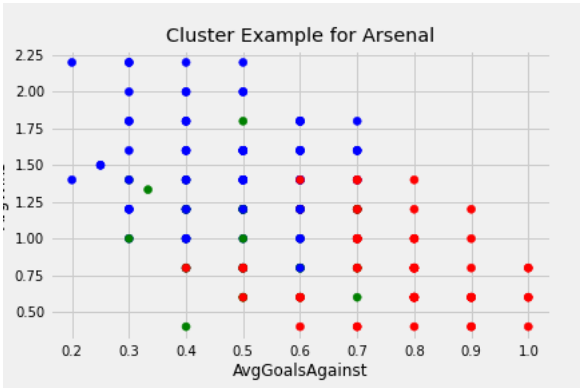

Figure 4: Matches clusters for Arsenal

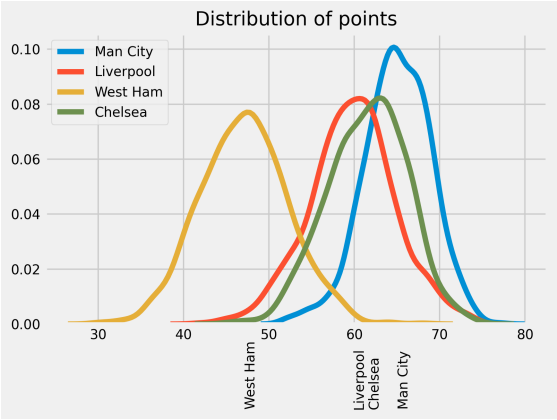

Figure 5: Distribution of points for Manchester City, Liverpool, West Ham, and Chelsea

6.3 Combining Bayesian Model and Hidden States

To add the effect of the current state on the outcome probability we simply add the average number of wins on the hyperparemeters of the Dirichlet distribution we described before. Therefore, the final outcome probability distribution is:

$$\theta_{ij}|y_{ij}, s_i, s_j \sim Dir(0.5{+}i\ wins{+}s_i\ avgwins, 0.5{+}draws, 0.5{+}j\ wins{+}s_j\ avgwins)$$

Where s_i, s_j are the state of teams i and j respectively and $s_k\ avgwins$ is the average number of wins for team k in its current state.
Finally to estimate the final position of the team in the table we simulated the season several times using the winning and transition probabilities. We get a distribution the number of points at the end of the season.

6.4 Results for the Premier League 19/20

To check how the model is working we are going to compare the predicted position on the table versus their true position. For example, on Figure 6 we can see the predicted distribution versus the true value (dashed line).

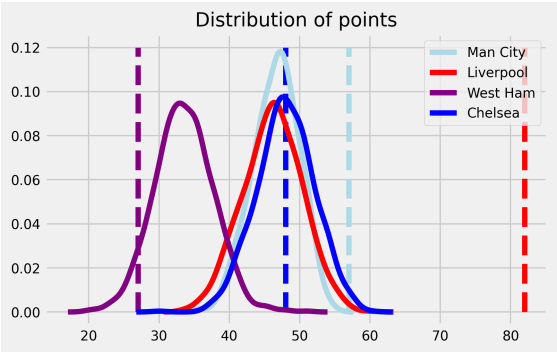

Figure 6: Distribution of points for Manchester City, Liverpool, West Ham, and Chelsea at the middle of the season

In general the true value is within the calculated distribution with the exception of Liverpool that has had an exceptional season for any standard. To evaluate the result for all the team we are going to use the following loss function:

$$Loss = \frac{1}{20}\sum_{i=1}^{20} |Standing(predicted)_i - Standing(true)_i| \tag{3}$$

Where the $Standing(predicted)_i$ is the position based on the expected value of the points. We get a value of 4.3. This means that the predicted position is within four positions of the real one.
There are still plenty of options to improve the model. For example, we could use a different prior based on how good the team has been on the last decades. Also the effect of the state on the outcome probability could be refined using cross-validation.

7 Simple, but Effective: Poisson Distribution

The exploratory data analyses done above suggests that **goal difference** in each match plays a major role in determining a team's position above anything else. It makes sense intuitively also that this factor is far more important than anything else because that's how a game is decided in the end: the more you score and the fewer you concede, higher will be your chance of winning.

So, the basic idea now shifts to determine the goals a team can score in each match at the end of 90 minutes. This is the basic definition of a Poisson distribution: **the number of times an event happens in a specific timeframe.** Specifically, the Poisson distribution works with a constant rate λ and is independent of previous events which makes sense since each match comes with its own unique challenges even though things like confidence and form that are carried over from the previous match but we have tried to simulate based on data that we are given.

The Poisson distribution is defined as follows:

$$P(X = x) = \frac{\lambda^x e^{-x}}{x!} \tag{4}$$

We set constant rate (λ) reflects the performance of a team. This rate needs to depend on attacking strength of the team and defensive strength of the opponent. Also, we also take into consideration that a team generally plays better at home ground. Therefore, the better team has a higher rate of scoring goals on average. We followed the methodology defined by Apaar Gupta in [1]. The model is highly intuitive and tractable to determine the final standings in 2 scenarios.

7.1 Methodology

We first need to define the following parameters for each of the teams to calculate their average scoring rate λ:

- Home Scored = Average number of goals scored in home matches
- Home Conceded = Average number of goals conceded in home matches
- Away Scored = Average number of goals scored in Away matches
- Away Conceded = Average number of goals conceded in Away matches

We also compute League Home Scored , League Home Conceded , League Away Scored , League Away Conceded to normalized parameters to each team and named them as Home Attack Strength, Home Defense Strength, Away Attack Strength, Away Defense Strength.

Then, the multipliers would be:

- Overall Goals Scored of Home = (Home Scored / Away Conceded) /2
- Overall Goals Scored Away = (Home Conceded / Away Scored)/2

The above defined statistics clearly talk about the significance of goals per team. So, for each team i: we calculate the scoring rates as follows:

λ (Home)=Home Attack Strength $*$ Away Defense Strength $*$ Overall Goals Scored of Home

λ (Away) = Away Attack Strength $*$ Home Defense Strength $*$ Overall Goals Scored of Away

We have also **assumed that all scorelines are independent on each other**, and hence the probabilities are summed together to simulate a match between Home (H) and Away (A):

1) P(H wins) = $\sum$ P (Home team wins if (Home goals $>$ Away goals))

2) Prob (Away wins) = $\sum$ P (Away team wins if (Home goals $<$ Away goals))

3) Prob (Tie) = $\sum$ P (Tie(Home goals = Away goals))

Clearly if we simulate all the matches, then we will get the final table. after assigning the right points for win, draw and loss per team. We simulated our results using the code that can be found in the Appendix

7.2 Evaluation

There are 2 ways to evaluate the working efficiency of this model: **full season prediction based on past 5 years data** or, **final season prediction after the mid-table standings are decided for a season.**

We have collated data for the past 5 season so that the teams do not show much variability in terms of their results and are more robust to changes. We predicted the results using our data for the final table at the end of May 2019 for both the situations stated above. The final table looked as shown below, this is basically the truth:

In order to check for the error, we have used the L_1 norm distance for each team from their final position.

$$Loss = \frac{1}{20}\sum_{i=1}^{20} |Standing(predicted)_i - Standing(true)_i| \tag{5}$$

Clearly, Manchester City and Liverpool title race will be remembered for years as they put in spectacular show towards the end and did not lose a single point in their last 11 games. While this phenomena was highly surprising, their standings weren't. Let's analyse the standings evaluated by our model from the start of the season.

Team	GP	W	D	L	GF	GA	GD	PTS
Manchester City	38	32	2	4	95	23	72	98
Liverpool	38	30	7	1	89	22	67	97
Chelsea	38	21	9	8	63	39	24	72
Tottenham Hotspur	38	23	2	13	67	39	28	71
Arsenal	38	21	7	10	73	51	22	70
Manchester United	38	19	9	10	65	54	11	66
Wolves	38	16	9	13	47	46	1	57
Everton	38	15	9	14	54	46	8	54
Leicester City	38	15	7	16	51	48	3	52
West Ham United	38	15	7	16	52	55	-3	52
Watford	38	14	8	16	52	59	-7	50
Crystal Palace	38	14	7	17	51	53	-2	49
Newcastle United	38	12	9	17	42	48	-6	45
Bournemouth	38	13	6	19	56	70	-14	45
Burnley	38	11	7	20	45	68	-23	40
Southampton	38	9	12	17	45	65	-20	39
Brighton	38	9	9	20	35	60	-25	36
Cardiff City	38	10	4	24	34	69	-35	34
Fulham	38	7	5	26	34	81	-47	26
Huddersfield Town	38	3	7	28	22	76	-54	16

Figure 7: Final Premier League Table for 2018/2019

From the above table the loss comes out to be: **42/20**, which equals **2.1**. Basically, our predicted model was off by 2 standings for each team. This is a pretty good estimate of a team's potential ranking at the start of the season based on a simplistic model. Let's see if we can better our results if we know what the mid-season table looked like for 2018/2019. The mid-season table looked like this.

Clearly, Liverpool was on a road to victory in mid-season. Let's see what the prediction look like taking mid-season data into account.

The final table loss here is **26/20**, which equals **1.3**. Clearly, we are off by almost just **one standing** for this season. The model, simplistic yet powerful, predicted the standings using the basic principles of probability.

Team	Points
Man City	86.09
Liverpool	76.08
Chelsea	74.33
Tottenham	72.17
Arsenal	71.76
Man United	69.16
Everton	57.24
Leicester	54.91
Wolves	53.67
Southampton	53.19
West Ham	49.26
Crystal Palace	46.99
Bournemouth	44.13
Newcastle	43.43
Watford	43.01
Burnley	41.58
Brighton	37.30
Cardiff	28.84
Fulham	27.50
Huddersfield	25.42

Table 1: **Prediction**: Final 2018/19 Table from the **start** of the season

7.3 Practical Aspects and Model Drawbacks

The data is impeccably strong in determining the forces of uncertainty around an event. The Premier League works on the principles of signing new players and making dynamic transfers that can boost the morale of our team. This model looks at previous data and is pretty smart in predicting the outcomes of matches using simple Poisson distribution. As expected, the model cannot account for all sources of uncertainty. The following are the major drawbacks of the model:

1. **Morale** of a team is something that is always a hidden variable to account for their collective performance. An even more sophisticated model can be built around taking the morale of a team into consideration.

2. **Latest results** have a more lasting impression on a team's outlook when they set out to face an opponent. This could be a combination of their previous few meetings, the venue or the standings of the other team which can escalate or dent their confidence.

3. Taking **time** into account is essential and probably the use of time dummies to determine the results based on previous few results could help us in finding even more accurate results.

Team	Points
Liverpool	54
Man City	47
Tottenham	45
Chelsea	43
Arsenal	38
Man United	35
Wolves	29
Leicester	28
Watford	28
Everton	27
West Ham	27
Bournemouth	26
Brighton	25
Crystal Palace	19
Newcastle	18
Cardiff	18
Southampton	15
Burnley	15
Fulham	14
Huddersfield	10

Table 2: Mid Season Table for 2018/19

8 Conclusion

Using a Dirichlet distribution to model the outcome probabilities has the advantage that takes into consideration the uncertainty present on a professional match. There are many scenarios when this can be helpful. Unfortunately, our approach of using the Dirichlet distribution has worst performance for this case than the simpler method of using a Poisson Distribution. There are definitely avenues of improvements for the Dirichlet model such as more informative priors.

Poisson Distribution is so simplistic in its working but at the same time has been able to provide really intuitive results that are not high on deviation from the final standings of the season. We see from our results that the final standings are generally dependent on the mid-season performance also. A team that can perform well in the first half of the season is likely to perform well in the second half also carrying the momentum forward of a good start.

There is always room for improvement but a model that is technically sound and is built on the theoretical foundations of probability theory can only improve with addition of pertinent information to it.

Team	Points
Liverpool	90.47
Man City	88.34
Tottenham	79.62
Chelsea	77.55
Arsenal	72.28
Man United	67.26
Wolves	54.89
Everton	54.2
Leicester	53.54
West Ham	50.82
Watford	48.38
Bournemouth	47.34
Brighton	43.07
Crystal Palace	41.64
Southampton	40.76
Newcastle	37.96
Burnley	34.37
Cardiff	31.17
Fulham	26.19
Huddersfield	21.65

Table 3: **Prediction:** Final Predictions post the **midseason** table for 2018/19

9 References

[1] Predicting Premier League Standings - Apaar Gupta

10 Appendix - software codes

[1] Shared drive folder for all codes