Assessing the Probability of the Soccer Match Outcome

Zhao Xinyuan
xizh5315@colorado.edu
University of Colorado Boulder
BOULDER, CO

Yu Shanqing
shyu1770@colorado.edu
University of Colorado Boulder
BOULDER, CO

Arman Aydemir
aray7750@colorado.edu
University of Colorado Boulder
BOULDER, CO

ABSTRACT

The article is about a project to access the probability of the soccer match outcome. It includes the problems to be solve, previous work has been done on this subject, the work we have done, and the result we have found. The Dataset and tools used are also included in this article. This project is based on the Dixon Coles model. We apply this mode to our work and build up a model to predict the result of a soccer game.[1].

KEYWORDS

Datasets, Statistical model, Data mining, Poisson distribution, Soccer matches, Betting strategy

1 INTRODUCTION

Since the soccer betting industry went online in the 1990s, it has experienced tremendous development in all areas. The number of bookmakers is continually increasing, with several hundred already in the industry. Needless to say, the customers seek to win money from bookmakers while bookmakers want to create a steady return in long run. From the customer's point of view, a computer program for finding good bets could be a very valuable tool to make informed decisions. It is also very interesting statistical problem to formalize the methods to evaluate the odds offered by bookmakers and decide if the bet should be made. From the bookmaker's point of view, a tool which could determine more accurate probability distribution for setting odds on sports event would also be a useful instrument. Bookmakers are on constant alert for the latest news in the sport world, and this could help them stay a step ahead when offering betting odds.

2 PROBLEM STATEMENT

The goal of this project is to investigate the correlation between a soccer teams history and the probability distribution for each possible match outcome in a given soccer match. The examination of the soccer result and odds data is to lead to the establishment of a model assessing the probability of each possible match outcome in a soccer match. The model will only use historical data match result data, and has no other prior information on the match.

The questions we try to answer include:

① Are there significant betting opportunities for soccer betting?

② Is it possible to build a successful model for predictions of the results for soccer matches?

3 DATASETS

This project aims to to build up a model based on the last 10 years' record of teams that plays for premier league, and predict the outcome of a given game.

We have verified the authenticity of this dataset and determined that this dataset is basically reliable and relatively complete.

The data set includes various information about a soccer match such as scores, attempt shots, home and away goals and the odds from different bookmakers, which, we considered, is sufficient to use for different kinds of analysis. Although our dataset does not contain information such as weather, temperature and players of each game. However, we think the effect of weather and temperature will be presented in the other attribute such as total running distance. We can draw similar inference for player absences. For instance, the lack of the best striker will affect the ratio of attempt shots and shots on target. The lack of midfielder will decrease the probability of his team to have a shot, which will decrease the number of attempt shots . The structure of the attributes we will be working with are :

- Date - Interval (String)
- Home Team - Nominal (String)
- Away Team - Nominal (String)
- Final Result - Nominal (Char)
- Away Goals, Home Goals, and other goal statistics - Nominal (Int)
- Other game statistics - Nominal (Double)
- All Betting Odds - Nominal (Double)

This project will use the record of the English Premier League in the last 10 years. There is a few reasons for our choice. First, compared with other league like Spanish La Liga and French Ligue 1, the ability of teams in the Premier League is much closer. This means building a successful model based on the Premier League is more challenging than the other leagues and it will make our result more convincing. Second, over the last 10 years, at least 8 teams have played only in the Premier League, having not been relegated in any of those years. This means we will have more data to train our model. Our dataset is available here:

https://datahub.io/sports-data/english-premier-league.

4 RELATED WORK

In recent years, the challenge of modelling soccer outcome has gained attention. This task may be achieved by adopting two different modelling strategies: the 'direct' models, for the number of goals scored by two competing teams. the 'indirect' models for estimating the probability of the categorical outcome of a win, draw or a loss.

The basic assumption of the direct models is that the numbers of goals scored by the two teams follow two Poisson distributions. Their dependence structure and the specification of their parameters are the most relevant issues. Maher (1982) used two conditionally independent Poisson distributions, one for the goals scored by the home team, and another for the away team. Dixon and Coles (1997) expanded upon Maher's work and extended his model, introducing a parametric dependence between the scores.[2][3]

The second common assumption is the inclusion in the models of some teams' effects to describe the attack and the defence strengths of the competing teams. Generally, they are used for modelling the scoring rate of a given team, and in much of the aforementioned literature they do not vary overtime. Of course, this is a major limitation. Dixon and Coles (1997) tried to overcome this problem by downweighting the likelihood exponentially overtime in order to reduce the impact of matches far from the current time.[1]

We are interested in both the estimation of the models parameters, and in the prediction of a new set of matches. Intuitively, the latter task is much more difficult than the former, since football is intrinsically noisy and hardly predictable. However, we believe that combining the betting odds with an historical set of data on match results may give predictions that are more accurate than those obtained from a single source of information.

4.1 Dixon-Coles Model

Our target is to build up a model to predict the outcome of a given soccer game accurately. And several features will be required in our model.

1. We need to find a way evaluate the ability of different teams based on their recent performance, and get more attribute involved into this evaluation.

2. We need to add time weighting parameter into the process of ability evaluation. The most recent performance of a team will have the greatest impact on this process.

3. Both teams' ability will be considered to give a prediction result

The work of Dixon and Coles gave a model based on the recent result of a team to predict the result of a particular game. In their work, they divided the ability of a specific team into two parts: attack ability and defend ability and considered the attack ability and defend ability are independent Poisson variables. For a given team A and a given team B, we can set $X_{A,B}$ and $Y_{A,B}$ to be the number of goals scored by the home and away sides respectively. Combined with "home effect", the fact that home team always enjoy some advantage over away team, the goal can be given as:

$$X_{A,B} \sim Poisson\left(\alpha_A \beta_B \gamma\right) \tag{1}$$

$$Y_{A,B} \sim Poisson\left(\alpha_B \beta_A\right) \tag{2}$$

α_A and β_A measure the attack and defend parameter of team A, and α_B and β_B measure the attack and defend parameter of team B. The γ is the home effect rate. And the probability of different result is given by:

$$Pr\left(X_{A,B} = x, Y_{A,B} = y\right) = \tau\left(x, y, \gamma\right) \frac{\lambda^x exp\left(-\lambda\right)}{x!} \frac{\mu^x exp\left(-\mu\right)}{y!} \tag{3}$$

where τ is a step function about whether consider the 'home effect'

$$\tau\left(x, y, \gamma\right) = \begin{cases} 1 - \lambda\mu\rho & ifx = y = 0 \\ 1 + \lambda\rho & ifx = 0, y = 1 \\ 1 + \mu\rho & ifx = 1, y = 0 \\ 1 - \rho & ifx = y = 1 \\ 1 & otherwise \end{cases}$$

here,in Dixon Coles model, ρ is in range:

$$max(-1/\lambda, 1/\mu) <= \rho <= min(-1/\lambda\mu, 1) \tag{4}$$

ρ is a dependence parameter, and when $\rho = 0$, that means there exists an independence between the team's goals as a home team and an away team , and :

$$\lambda = \alpha_A \beta_B \gamma \tag{5}$$

$$\mu = \alpha_B \beta_A \tag{6}$$

The functions we list above are based on two assumptions:

①: Home goals and away goals are independent Poisson variables.

②: The frequency of goals scored satisfy Poisson Distribution.

And whether these assumptions are valid will be discussed in section8

5 MAIN TECHNIQUES APPLIED

5.1 Data Cleaning

Only the information about the numbers of the goal scored, when and where the score happens is relevant, other information such as the referee's name, the number of yellow or red cards will be discarded.

We will also eliminated the match result which a large number of players has been shifted compared to previous games and those games happened during extreme bad weathers.

5.2 Data Prepossessing

The ratio of frequencies of home wins, draws and away wins will be used to determine the home advantage. The dependency between home and away score will be checked.

A Dixon Coles model and a negative binomial will be developed based on the available data set. Numerical and graphical checks for our model will be provided.

5.3 Data Integration

A large amount of data results from different seasons and from different leagues will be integrated together to build the model. The model will be both descriptive and predictive. The results and predictive accuracy of the model on different leagues will be checked.

6 EVALUATION METHODS

This project aims at predicting the outcome of a given soccer game and try to find a betting strategy. Once the model for determining the probabilities of each soccer match outcome is developed, the result of our predicting model will be compared with the bookmaker's assessment with regards to setting odds. The performance of the proposed model and betting strategies performed on actual odds will be examined.

7 TOOLS

We will use Python in this project and the tools we will use can be listed as follows:

①pandas

②numpy
③scrapy (A web scraper to scrap online data)
④sklearn
⑤matplotlib

8 KEY RESULTS

Just as we clarified in section4.1, The work of Dixon and Coles is based on two assumptions: the first is the goals of away team and home team are independent, and the second is the frequency of different score satisfy Poisson Distribution.

In this section, we will first verify the two assumptions made by Dixon and Coles. Then we will determine which contributes will be contained in our model.

8.1 Poisson Distribution

Given the data of the last 20 seasons in premier league, for goals from 0 to 7, we calculate the relative frequency,and present it as percentage in figure1 .

Dixon and Coles' assumption that the marginal distribution of random match scores is Poisson can be proved to be valid through this figure. At least it is plausible for our dataset. The relative frequency shown in figure 1 nearly perfect fit the Poisson Distribution. And the 3d model of Empirical estimates of each score probability can be shown in figure 2

8.2 Independency of Away and Home Goals

To prove the validity of the assumption that home goals and away goal are independent. For a given home scores(i=0,1,2...7) and away scores(j=0,1,2...7), we need to calculate:

$$\frac{\widetilde{f}(i,j)}{\widetilde{f_H}(i)\,\widetilde{f_A}(j)} \tag{7}$$

where $\widetilde{f}(i,j)$ is the joint probability functions for given i and j, and $\widetilde{f_H}(i)$,$\widetilde{f_A}(j)$ is the marginal empirical probability functions for home goals and away goals. And the result of the ratio is calculated in figure 3.

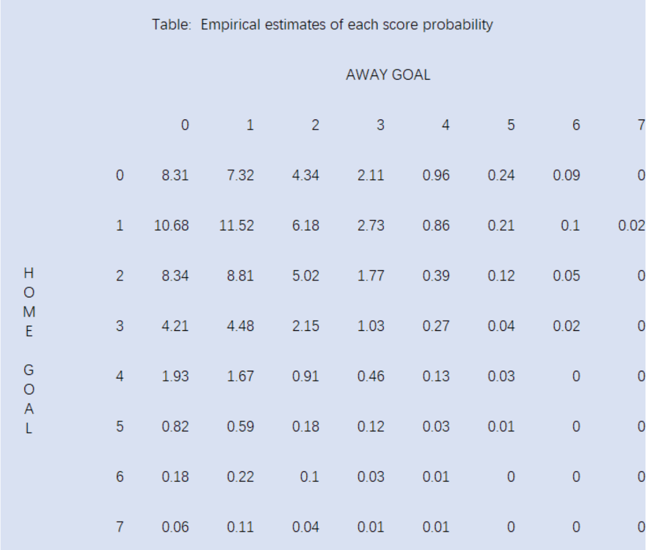

Table: Empirical estimates of each score probability

		AWAY GOAL							
		0	1	2	3	4	5	6	7
HOME GOAL	0	8.31	7.32	4.34	2.11	0.96	0.24	0.09	0
	1	10.68	11.52	6.18	2.73	0.86	0.21	0.1	0.02
	2	8.34	8.81	5.02	1.77	0.39	0.12	0.05	0
	3	4.21	4.48	2.15	1.03	0.27	0.04	0.02	0
	4	1.93	1.67	0.91	0.46	0.13	0.03	0	0
	5	0.82	0.59	0.18	0.12	0.03	0.01	0	0
	6	0.18	0.22	0.1	0.03	0.01	0	0	0
	7	0.06	0.11	0.04	0.01	0.01	0	0	0

Figure 1: Empirical estimates of each score probability

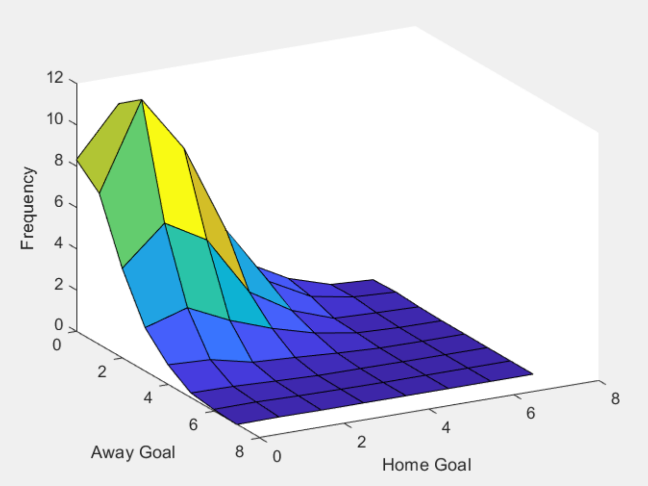

Figure 2: 3D distribution of score probability

Through this figure, we can find most ratios are close to one. That means the home goal and away goal are independent.

8.3 Attributes

The primary attributes we are using are

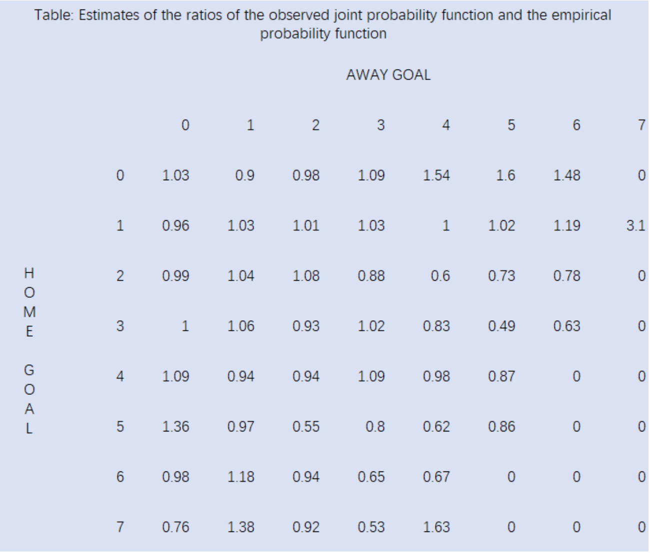

Table: Estimates of the ratios of the observed joint probability function and the empirical probability function

AWAY GOAL

		0	1	2	3	4	5	6	7
	0	1.03	0.9	0.98	1.09	1.54	1.6	1.48	0
	1	0.96	1.03	1.01	1.03	1	1.02	1.19	3.1
HOME GOAL	2	0.99	1.04	1.08	0.88	0.6	0.73	0.78	0
	3	1	1.06	0.93	1.02	0.83	0.49	0.63	0
	4	1.09	0.94	0.94	1.09	0.98	0.87	0	0
	5	1.36	0.97	0.55	0.8	0.62	0.86	0	0
	6	0.98	1.18	0.94	0.65	0.67	0	0	0
	7	0.76	1.38	0.92	0.53	1.63	0	0	0

Figure 3: Ratios of the observed joint probability function and the empirical probability function

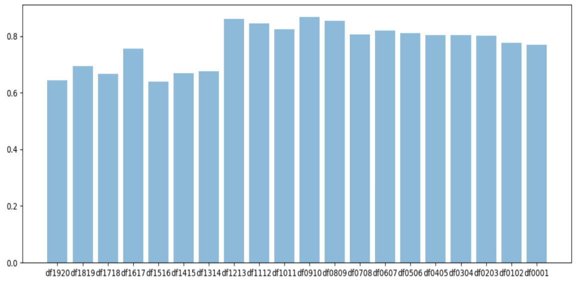

Figure 4: Correlation coefficient between full time total goals and full time total shots on target

- Date - Interval (String)
- Home Team, Away Team - Nominal (String)
- Final Result - Nominal (Char)
- Goals, Shots On Target, Total Shots, Fouls, Yellow Cards, Red Cards (for both Home and Away) - Nominal (Int)
- All Betting Odds - Nominal (Double)

We are also working on adding more stats from a secondary source. However, for these statistics we will not have as many seasons of data to work with.

- Possession - Nominal (Double)

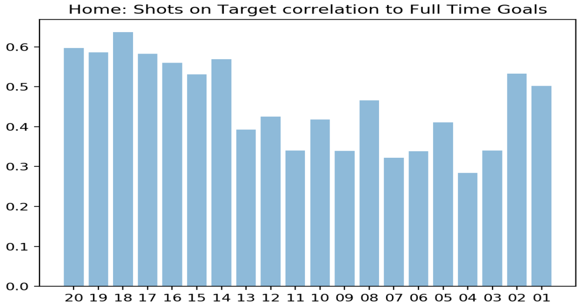

Figure 5: Home: Correlation coefficient between full time total goals and full time total shots on target

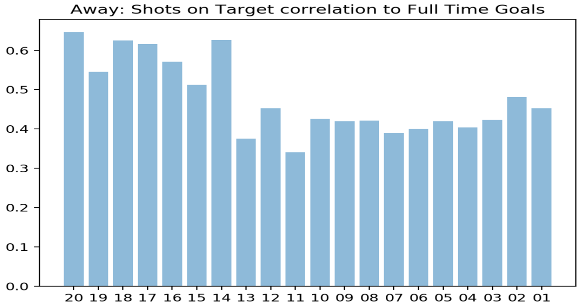

Figure 6: Away: Correlation coefficient between full time total goals and full time total shots on target

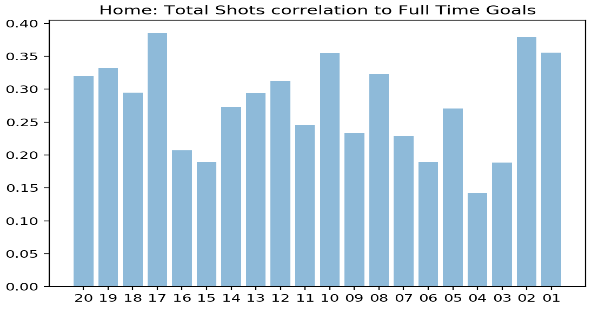

Figure 7: Home: Correlation coefficient between full time total goals and full time total shots

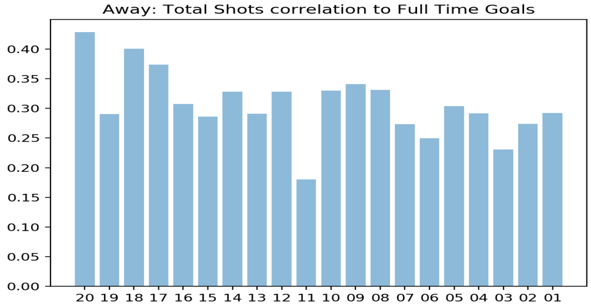

Figure 8: Away: Correlation coefficient between full time total goals and full time total shots

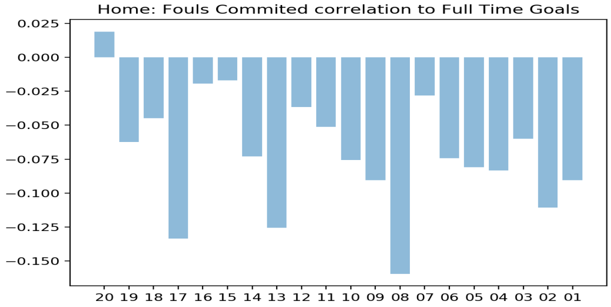

Figure 9: Home: Correlation coefficient between full time total goals and fouls committed

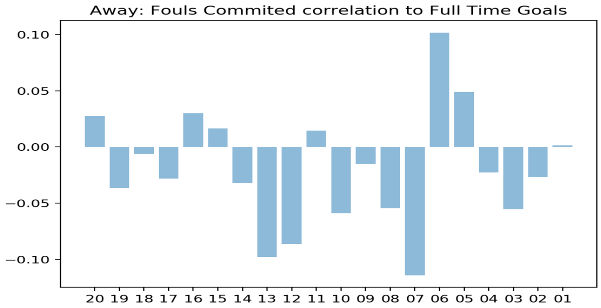

Figure 10: Away: Correlation coefficient between full time total goals and fouls committed

- Total Passes - Nominal (Int)

- Tackles - Nominal (Int)

- Attacks, Dangerous Attacks - Nominal (Double)

These are the main attributes we are focused on adding. However, as we further evaluate the effectiveness of our methods we could add or exclude some variables from this secondary source.

For shots on target. we calculate the Correlation coefficient between full total goals and full time total shots on target in the last 20 seasons. And paint a histogram of it in figure 10.And then we calculate the the correlation coefficient between home team and away team's final goals and other attribute separately, like shots on target, time of shots and number of fouls.

From these graphs above we can find that first, there is no obvious relation between the number of fouls and the finals goal. And second point is that, as for the relation between shots,shots on target and final goals, whether for all the teams, or for the home team and away team separately, the correlation coefficient is approximate 0.6, which means we can only say the final goals and shots on target is positive related. And after our discussion, there is no need to include that in our assessment process.

8.4 Basic Poisson Model Improvement

During the previous segment, we checked the dependency between the number of goals scored and the number of goals conceded. It reaches to a conclusion that Basic Poisson model can make a good approximation of the goal distribution. In the their paper, Mark Dixon and Sturat Coles proposed two specific improvements to the Basic Poisson Model.

① Introduce an interaction term to correct underestimated frequency of low scoring matches.

② Apply the time decay component so that recent matches weighs stronger the matches played before

The paper says that low score results (0-0, 1-0, 0-1 and 1-1) are misreported by the BP model. The matrix below shows the average difference between actual and model predicted scorelines for the 2005/2006 seasons to the 2019/2020 season. Green cells imply the Basic

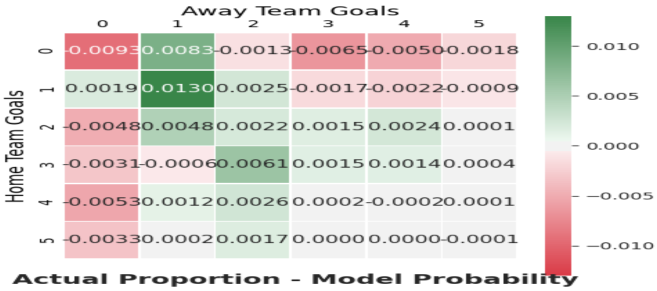

Figure 11: Departure from Basic Poisson

Poisson model underestimated those scorelines, while red cells suggest overestimation. The color strength suggest the level of disagreement.

There seems to be an departure from the Basic Poisson model with low scoring draws and it is less apparent with 0-1 and 1-0. The correction can be done through following:

$$\tau\left(x,y,\gamma\right)=\begin{cases}1-\lambda\mu\rho & if x=y=0\\ 1+\lambda\rho & if x=0,y=1\\ 1+\mu\rho & if x=1,y=0\\ 1-\rho & if x=y=1\\ 1 & otherwise\end{cases}$$

$$max(-1/\lambda, 1/\mu) <= \rho <= min(-1/\lambda\mu, 1) \tag{8}$$

The main difference over the Basic Poisson Model is the τ function. It is highly dependent on ρ, which controls the strength of the correction.

8.5 Model Inference

To calculate the model coefficients that exists in that model. We need to construct the likelihood function and find the coefficients that maximise it, a technique called Maximum Likelihood Estimation.

$$L\left(\alpha_i,\beta_i,\rho,i=1,2,3,4..\right)=\prod_{k=1}^{N}\left\{\tau_{\lambda_k}\left(x_k,y_k\right)\right. \\ exp\left(-\lambda_k\right)\lambda_k^{x_k}exp\left(-\lambda_k\right)\mu_k^{x_k} \tag{9}$$

in this equation, i(k) and j(k) represents the indices of the home and away teams in a given match k. For numerical precision, it'll be maximised using log-likelihood function. As the logarithm is a strictly increasing function. Both to evaluate a team's strength by attack and defend parameters. For n teams and their attack parameter α_1 α_2 α_3...α_n, defend parameter β_1 β_2 β_3 ... β_n. To standardize attack and defend parameter, we impose a constraint on them.

$$n^{-1}\sum_{i=1}^{n}\alpha_i=1 \tag{10}$$

And for a given home team i, α_i and β_i will be its attack and defend parameter. Consider recent k games, and index of matches can be set as k=1,2,3,...N. The scores of a corresponding match k can be shown as (x_k, y_k). The γ is the home effect rate. ρ is a dependence parameter.

$$\lambda_k=\alpha_{i(k)}\beta_{j(k)}\gamma \tag{11}$$

$$\mu_k=\alpha_{j(k)}\beta_{i(k)}\gamma \tag{12}$$

$\alpha_{i(k)}$ and $\beta_{i(k)}$ is the attack parameter of home team in kth match. And $\alpha_{j(k)}$ and $\beta_{j(k)}$ is the defend parameter of away team in kth match.

For 20 teams in Premier League, and 380 games in 19-20 seasons, we can generate the 20 attack parameters and 20 defend parameters as shown in table 1. We can find that the mean of attack parameter is equal to 1.

8.6 Time weighting parameter

There are several problems with the equation9. First, consider the fact that there is always a fluctuation in teams' performance and their performance is always dynamic. The attack and defend parameter generated from equation 9 remain static. To find a more accurate way to describe the strength of different team, Dixon and Cole added time weighting parameter. Based on the equation 9,we make two assumptions in this project:

① The attack and defend strength of any team will keep relatively constant in three days.

② All the historical performance of a team will be counted for their strength assessment. And we assume

Table 1: Attack parameter α and Defend parameter β in 19-20 season

Team	$log\,(\alpha)$	$log\,(\beta)$
Arsenal	1.134194461	-0.9388646131
Aston Villa	0.8428578212	-0.6189790245
Bournemouth	0.8120741353	-0.6519458441
Brighton	0.7809206483	-0.8319848409
Burnley	0.8736217223	-0.9169827088
Chelsea	1.3406671569	-0.8095874718
Crystal Palace	0.5402903267	-0.9276208686
Everton	0.904040090	-0.7978082567
Leicester	1.3057729257	-1.0774228343
Liverpool	1.53881068376	-1.2883023793
Man City	1.7197223805	-1.2127458303
Man United	1.2833265314	-1.2167372006
Newcastle	0.7560667438	-0.7732833065
Norwich	0.38629771197	-0.5236456429
Sheffield United	0.7571901683	-1.1712131162
Southampton	1.046302694	-0.7286538126
Tottenham	1.216391527	-0.9507146131
Watford	0.7122211110	-0.6644120721
West Ham	1.0118122088	-0.692856079
Wolves	1.0374189490	-1.1199782123
Mean	1.0000	-0.8957

that the historical performance possessed less weight in the process of assessment.

for a set time t, the equation considering time weighting parameter can be written as

$$L\left(\alpha_i, \beta_i, \rho, i=1,2,3,4..\right) = \prod_{k \epsilon A_t} \{\tau_{\lambda_k}\left(x_k, y_k\right) \\ exp\left(-\lambda_k\right) \lambda_k^{x_k} exp\left(-\lambda_k\right) \mu_k^{x_k\,\{\phi(t-t_k)\}} \tag{13}$$

In this function, $t - t_k$ is the time that match k was played,$A_t = \{k : t_k < t\}$. We need to clarify that α_i, α_j, β_i, β_j, τ and ρ are all function of t.

Several time weighting functions can be considered, and the most simplest one can be shown as follow:

$$\phi\left(t\right) = \begin{cases} 1 & ift <= t_0 \\ 0 & ift > t_0 \end{cases}$$

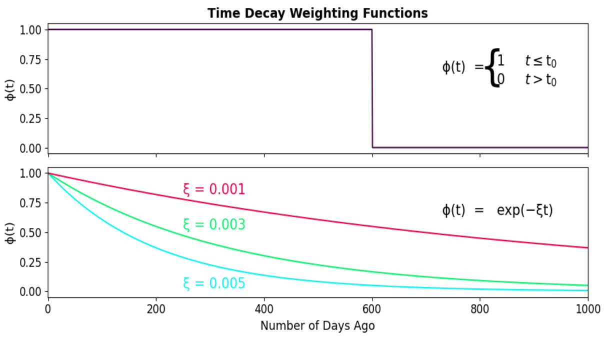

Figure 12: Time Weight Funtion

In this case, for any previous results(result before t_0) will be considered to have the same weight in the assessment process. If we apply our assumption② to it, we can write the time weighting function as:

$$\phi(t) = -exp\left(-\xi t\right) \tag{14}$$

ξ here is the downweighted exponential factor. When $\xi = 0$ is the static case we mentioned above and when $\xi > 0$, this equation describe the process where the previous result' weight drop exponentially. And the value of ξ can be chosen by make our prediction of outcome accurate enough. By using different time weight parameter ξ from [0,0.1] to make predictions and compare predictions with actual match results, the most optimal ξ is calculated as 0.0065. And the time weight function with different ξ can be shown as 12

8.7 Parameter Fluctuation

We can generate our attack and defence parameter by maximizing the function 13 at a given time t. the attack and defence parameter represent the strength of attack and defence of a specific team at given time t. Moreover, if we calculate these parameter across the history, the attack and defence strength of a specific team across the history can be observed.

The fluctuation of the attack and defence ability in 100 days can be shown in figure13 figure14 and figure15.From that we can tell a team's attack and defence ability remain pretty much constant in half season.

For the figure13 and figure14, Green line is for Newcastle, yellow line is for Brighton and blue line is for

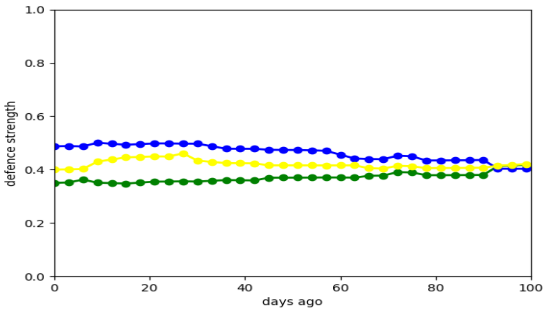

Figure 13: Fluctuation of defence parameter over 100 days. Newcastle(Green), Brighton(Yellow) Bournemouth(Blue).

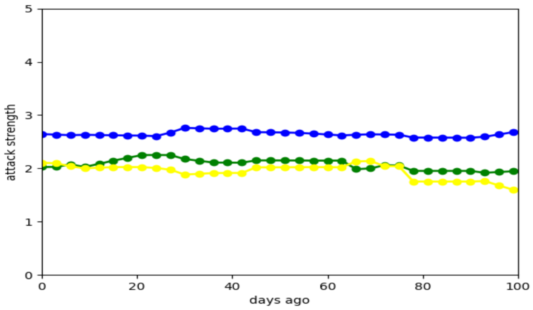

Figure 14: Fluctuation of attack parameter over 100 days. Newcastle(Green), Brighton(Yellow) Bournemouth(Blue).

Bournemouth. In figure15, the green line is about the change of home advantage factor γ over time and the blue line is the dependence factor ρ .

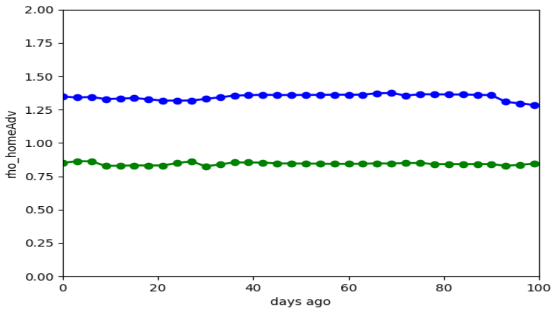

Figure 15: Fluctuation of dependence factor ρ(Blue) and home advantage factor γ(Red) over 100 days

9 APPLICATION

Given the model, we can simulate future matches using previous match results as training set. The model can give the probability of each possible score so the probability of home team win, away team win and draw can be derived from that. Using the model, the prediction on final day matches are given as follows: The model gives reasonable prediction as the team performed better before is more possible to win the match. Because the whole model is based on previous matches result. As to how accurate the prediction is, it should be compared with bookmaker's odd data to determine whether or not it is comparable to bookmaker's prediction.

During the processing of building this model, we've gained much insight about football:the validation of in-dependency between the goal scored and the goal conceded gives us the basic Poisson model. By introducing the correction parameter rho and the time weight function, A lot of Improvements has been made on it. Finally, A model which can make prediction and quantify the probability result is produced.

From the observation of attack and defence parameter of different teams, we find the strength of a team whether its attack or defence relatively stay constant during a season. It tells that the soccer games result are not total random, ultimately team with better skills will have more chance to win the game.

Table 2: The prediction of given teams

Home Team	Away Team	Probability of Home Win	Probability of Away Win	Probability of Drawn
Arsenal	Watford	0.68435	0.10981	0.20591
Burnley	Brighton	0.51685	0.18444	0.29869
Chelsea	Wolves	0.39325	0.3345	0.26822
Crystal Palace	Tottenham	0.18543	0.53088	0.28367
Everton	Bournemouth	0.56457	0.18075	0.25467
Leicester	Man United	0.34908	0.37817	0.27274
Man City	Norwich	0.95967	0.00547	0.02986
Newcastle	Liverpool	0.14106	0.64365	0.21222
Southampton	Sheffield United	0.39826	0.28137	0.32036
West Ham	Aston Villa	0.58609	0.18475	0.22913

REFERENCES

[1] Mark J.Dixon and Stuart G.Coles. Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 1997.

[2] M. J. Maher. Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118, 2008.

[3] Richard Pollard. 69.9 goal-scoring and the negative binomial distribution. *The Mathematical Gazette*, 69(447):45–47, 1985.