Statistical Analysis and Prediction of Football Data

Chethan Tulsidas
chethan.tulsidas@gmail.com
Department of Computer Science
Ramaiah Institute of Technology
Bangalore

***Abstract*- It is widely proposed that past results and scores of football matches between the different teams in a league system can be used to predict future scores and results. The number of goals scored by each team can be approximated by a Poisson distribution, whose average can be used to quantify the attacking and defensive strength of either side. By using historical data and working backwards from the Poisson formula, it is possible to calculate the probability of a team scoring any number of goals. Using these probabilities and a simulation procedure, the outcome of a match can be predicted along with the odds for other popular statistics that are bet upon.**

Keywords- Probability, Statistics, Poisson Distribution, Goals, Football, Prediction

I. INTRODUCTION

Football, also known as association football or soccer is the biggest ball game in the world when considering the sheer number of participants and spectators. In Europe, every country has a club based football league, with some even comprising of multiple divisions or tiers. Each league has 20 teams who play each other twice - one at their home stadium and one at the opponent's home stadium for a total of 380 matches in a season. Each match can have one of three possible results - the home team wins, the away team wins, or the match ends in a draw. Due to the sport's immense popularity, several online fantasy leagues, betting agencies, and pundits try to successfully predict the outcome of matches.

It is widely proposed that past results and scores between the different teams in a league system can be used to predict future scores and results. The number of goals scored by each team can be approximated by a Poisson distribution, whose average can be used to quantify the attacking and defensive strength of either side. By using historical data and working backwards from the Poisson formula, it is possible to calculate the probability of a team scoring any number of goals. Using these probabilities and a simulation procedure, the outcome of a match can be predicted along with the odds for other popular statistics that are bet upon.

Despite having its limitations, the statistical analysis of football matches using a Poisson model is a solid starting point when trying to predict football results, or to find value bets by creating odds.

The Poisson distribution is a probability distribution that can be used to model data that can be counted numerically. If the frequency of something that is expected to happen is known, it is possible to find the probabilities that it happens any number of times. The number of goals a team scores in a football match are approximately Poisson distributed. This gives a method of assigning probabilities to the number of goals in a match and from this the probabilities for different match results can be calculated. The Poisson distribution does not always perfectly describe the number of goals in a match. It sometimes over or under estimates the number of goals, and some football leagues seems fit the Poisson distribution better than others.

II. RESEARCH BACKGROUND

There are several other authors that have worked in the field of predicting the outcomes of various sports. "Modelling Association Football Scores" written by **M.J. Maher** in 1982 was one of the earliest studies to investigate the correlation between a Poisson distribution model and the number of goals scored by teams in a league season. He proposed that teams can be represented by parameters such as "inherent attacking and defending strengths". Observed and expected frequencies of scores were compared and goodness of fit tests showed that although there were some small systematic differences, an independent Poisson model could be sued to give a reasonably accurate description of football scores. The inherent qualities of teams were inferred over the course of a

season based on the average number of goals scored and conceded by the home and away sides respectively. The mean of the Poisson distribution varies according to the quality of the team, thus if the distribution of goals scored by all teams were considered, a Poisson distribution with a variable mean would be obtained.

Andreas Heuer, and **Oliver Rubner** proposed "How Does The Past Of A Soccer Match Influence Its Future? Concepts And Statistical Analysis" They discovered that the number of draws is significantly 10% larger than expected from the assumptions of independent Poisson distributions. The efficiency of the home team to equalize was seen to improve if the away team held the lead in the middle of the match. Concurrently, if the away team still holds the lead during the final minutes of the game, dastic deviations from the Poisson expectation are observed with a greater chance of the home team conceding again. They concluded that the concept of score-insensitive goals rates as opposed to score-dependant match behavior is a very good approximation of a football match.

"Modelling Association Football Scores And Inefficiencies in the Football Betting Market" authored by **Mark J Dixon** and **Stuart G Coles** derives a method for estimating the probabilities of football results with the potential to achieve a positive expected return when used as the basis of a betting strategy against bookmakers odds. In the proposed model, a bet is placed on all outcomes for which the ratio of model to bookmaker's probabilities exceeds a specified level. For sufficiently high levels, it is seen that this strategy yields a positive expected return, even allowing for the built in bias in the bookmakers odds.

In "Modelling Scores In The Premier League : Is Manchester United Really The Best?", **Alan J Lee** discovered that although some of the matches are very evenly matched and some games are very close, the outcome essentially comes down to chance. A lucky goal or an unfortunate error may decide the game. He proposed that it is the long run advantage expressed as a probability that is important. The team that deserves to win the league could be thought of as the team that has the highest probability of winning. This is not necessarily the same as the team that actually won a particular fixture. The probability that a team wins the league could be calculated by considering the likely outcome when two teams compete. These probabilities depend on which teams are playing and also whether the game is at home or away. Teams can be rated by ranking their estimated probabilities of winning the competition.

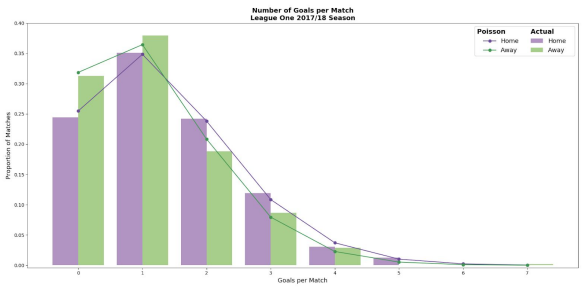
III. METHODOLOGY

The model is founded on the number of goals scored/conceded by each team. Teams that have been higher scorers in the past have a greater likelihood of scoring goals in the future. On average, the home team scores more goals than the away team. This is the so called 'home (field) advantage' and isn't specific to football. The Poisson distribution is a discrete probability distribution that describes the probability of the number of events within a specific time period (e.g 90 mins) with a known average rate of occurrence. A key assumption is that the number of events is independent of time. In this context, this means that goals don't become more/less likely by the number of goals already scored in the match. Instead, the number of goals is expressed purely as function an average rate of goals.

This can be represented by the mathematical formula

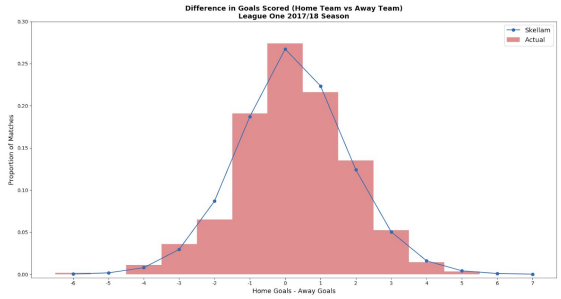
$$P(x) = \lambda^x e^{-\lambda}/x! \ , \lambda>0$$

The Poisson distribution is a mathematical concept for translating mean averages into a probability for variable outcomes across a distribution. The number of goals scored by the home and away team respectively can be represented as two independent Poisson distributions.



We can use this statistical model to estimate the probability of specific events.
The probability of a draw is simply the sum of the events where the two teams score the same number of goals. Similarly, the probability of the home or away team winning can be calculated, as well as outcomes pertaining to specific goal distributions. Note that the number of goals scored by each team is considered to be independent events. Each team plays at most 19 home and away games respectively. Due to a relatively variable sample size, the accuracy of this

approximation can vary significantly especially earlier in the season when teams have played fewer games.



To calculate the most likely score-line of a match, the average number of goals that each team is likely to score in that fixture must be calculated. This can be calculated by determining the "Attack Strength" and "Defence Strength" for each team and comparing them. Selecting a representative data range is vital when calculating Attack Strength and Defence Strength – too long and the data will not be relevant for the team's current strength, while too short may allow outliers to skew the data.

As a precursor to determining the Attack and Defense strength of both sides in a football match, a reference timeframe with the previous results of all the matches in the league must be determined. Then, the average number of goals scored per team, per team home game, and per away game are calculated. The next step is to determine the average number of goals conceded per game - for both home and away teams - which is the opposite of the average goals scored per game.

Average goals scored at home $(AGSH)$ =
Total home goals scored/Total no. of home games

Average goals scored away $(AGSA)$ =
Total away goals scored/Total no. of away games

Average goals conceded at home $(AGCH)$ =
Total home goals conceded/Total no. of home games

Average goals scored away $(AGCA)$ =
Total away goals conceded/Total no. of away games

Let the home team and away team be denoted by HT and AT respectively. The Attack and Defense strength of both teams are calculated using the following equations :

Home Team Attack Strength $(HTAS)$ =
Average home goals scored by HT/AGSH

Away Team Attack Strength $(ATAS)$ =
Average home goals scored by AT/AGSA

The ratio of a team's average number of goals scored and the league average number of goals scored is what constitutes "Attack Strength".

Away Team Defense Strength $(ATDS)$ =
Average goals conceded by AT/AGCA

Home Team Defense Strength $(HTDS)$ =
Average goals conceded by HT/AGCH

The ratio of a team's average number of goals conceded and the league average number of goals conceded is what constitutes "Defense Strength".

The number of goals to be scored in a particular match by the home and away side can be calculated using the reference table of attacking and defensive strengths of both sides. We call this the Goal Expectancy of the the respective side.

The average number of goals expected to be scored by the home side can be calculated as

Home Team Goal Expectancy $(HTGE)$ =
$HTAS$ x $ATDS$ x $AGSH$

The average number of goals expected to be scored by the away side can be calculated as

Away Team Goal Expectancy $(ATGE)$ =
$ATAS$ x $HTDS$ x $AGSA$

The goal expectancy tends to be a decimal, it is only an average and obviously, no football match can have such a scoreline. These averages must be converted into a probability. The Poisson distribution allows to spread 100% across multiple goal outcomes for each team.

In the mathematical formula,

$$P(x) = \lambda^x e^{-\lambda}/x! \ , \lambda>0$$

x represents the number of goals for which the probability is to be calculated and the λ parameter is the Goal Expectancy for the respective side.

The probability of the home/away team winning can be calculated by simply adding up the probabilities of all the outcomes where the home/away team scores more goals than the away/home team. Similarly, the probability of a draw is calculated by summing the probabilities of outcomes where both teams score the same number of goals. Other commonly placed bets such as "Both Teams To Score", "Over 2.5 Goals", and "Asian Handicap" can be easily computed from the probability distribution.

		AWAY TEAM										
HOME TEAM	Goals	0	1	2	3	4	5	6	7	8	9	10
	0	7.808	4.138	1.097	0.194	0.026	0.003	0.000	0.000	0.000	0.000	0.000
	1	15.772	8.359	2.215	0.391	0.052	0.005	0.000	0.000	0.000	0.000	0.000
	2	15.930	8.443	2.237	0.395	0.052	0.006	0.000	0.000	0.000	0.000	0.000
	3	10.726	5.685	1.507	0.266	0.035	0.004	0.000	0.000	0.000	0.000	0.000
	4	5.417	2.871	0.761	0.134	0.018	0.002	0.000	0.000	0.000	0.000	0.000
	5	2.188	1.160	0.307	0.054	0.007	0.001	0.000	0.000	0.000	0.000	0.000
	6	0.737	0.390	0.103	0.018	0.002	0.000	0.000	0.000	0.000	0.000	0.000
	7	0.213	0.113	0.030	0.005	0.001	0.000	0.000	0.000	0.000	0.000	0.000
	8	0.054	0.028	0.008	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	9	0.012	0.006	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	10	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

IV. LIMITATIONS

Unfortunately the model does have its limitations, some of which are listed below:

- Given the model uses past data to predict future results, it doesn't consider squad changes or manager movement.
- The only factor a Poisson model takes into consideration is the result. Results tell us the final score but are not always indicative of what happened in a match.
- A number of events both pre-game and during the game can affect a result. This model doesn't account for injuries, suspensions, fitness or weather - all of which can have an impact on a team's probability pre-game.
- Goal expectation can be affected once the game has started such as red cards, or an away goal - the team may then employ counter-attacking tactics etc.
- Correlations such as the condition of the pitch are also ignored, which highlights that specific matches can have a tendency for high or low-scoring outcomes.
- It is especially poor at predicting draws. Even when the two teams are expected to score the same number of goals the model rarely manages to assign the highest probability for a draw.

V. FUTURE SCOPE

Using the form of the players, strategic nuances in the formation, substitutions, player injures, and fatigue, a more accurate model can be created, along with which factors contribute heavily in a team's victory. Feature engineering other parameters such as the team form, and home advantage along with match information such as no. of attacking moves, through passes etc. into model would increase the accuracy of the predictions.

VI. REFERENCES

[1] "Towards The Perfect Prediction Of Soccer Matches", Andreas Heuer and Oliver Rubner

[2] "Football scores And The Poisson Distribution", Phil Scarf

[3] "Predicting Football Scores Via Poisson Regression Model", Erlandson F. Saraivaa, Adriano K. Suzuki, Ciro A. O. Filhob, Francisco Louzadab

[4] "Predicting And Retrospective Analysis of Soccer Matches in a League", Havard Rue, Oyvind Salvesen

[5] "A Ratings Based Poisson Model For World Cup Soccer Simulation", D. Dyte, S. R. Clarke

[6] "Modelling Scores In The Premier League : Is Manchester United Really The Best?", Alan J. Lee

[7] "Modelling Association Football Scores", M. J. Maher

[8] "Modelling Association Football Scores And Inefficiencies In The Football Betting Market", Mark J Dixon, Stuart G Coles

[9] "How Does the Past of a Soccer Match Influence Its Future? Concepts and Statistical Analysis", Andreas Heuer , Oliver Rubner