The use of forecasting in the world of football

Adrian Astăluș

Football is one of the most popular sports in the world, if not even the most popular. For ninety minutes, twenty-two players run across the pitch, giving their best for their team to win and to be one step closer to winning trophies. From its earliest days, football has managed to get people talking about which team is better, who is the best player, but most importantly, who will win. Whether we're talking just about a casual discussion regarding a football match, or about placing ambitious bets, predicting the result of a match has always been a topic of interest in this sport. Despite this, even the best pundits, who possess vast footballing knowledge often struggle to correctly predict the victor of a game. Simply put, the winner of a match is the team that scores more goals than the opposing team, but there are numerous factors that determine whether a goal is scored, ranging from the strength of a team's attack or defence, home ground advantage, and specific events that take place during the match[1].

This is where forecasting comes into play. In the past, forecasting was a neglected area of sports sciences[2], but in recent years, it has proved essential in predicting sporting results and events. Although the idea of predicting a match result seems impossible, recent studies have discovered that there are factors which have an influence in forecasting outcomes, such as prior performance, home advantage, and win-loss record among others[2]. One of the factors towards which most of the attention has been directed is the home advantage that a team has during a match. At first glance, it can be considered that the home advantage of a team means that the players are more familiar with the environment they are playing in, or avoid the fatigue created by travelling. This is most likely still true, but a greater impact is offered by the supporting fans, especially in domed arenas where their effect is more pronounced. In some cases, a vocal home crowd can even have an influence on referee decisions in favour of the home team, thus enhancing the home advantage[2].

The papers chosen for this report aim to offer solutions to predicting the results of football matches by analyzing what methods perform best and what parameters have the greatest influence when deciding the outcome of a game. In the end, both papers provide important insights into this topic. The findings in these papers represent a solid baseline for further development of new methods for predicting match results.

In their paper, Koopman and Lit develop a new model for the analysis of football match results. They analyze which variable provides the best forecasting results: the pairwise count of the number of goals, the difference between the number of goals, or the category of the match result (a win, a loss or a draw). Each of these variables require a different distributional assumption: Bivariate Poisson distribution, Skellam distribution and Ordered Probit models respectively. The authors also take into consideration the fact that the goals scored by a team should not be treated as independent events, as a team's strength is likely to be determined to some degree by its performance in recent matches. A team's composition can vary from match to match, and some teams grow in strength over the course of several years. Therefore, a fast approach to the dynamic modelling of teams' strengths in attack and defence is developed using score-driven models. The time-varying parameter of the model is adapted to the three variables proposed initially, thus creating an adaptable framework suited for time series analysis and forecasting of match results. To demonstrate the performance of this model, the authors provide an estimation of how two of the biggest teams in football perform in terms of attack and defence over the seasons. The resulting estimations showcase a yearly persistence of 0.9, which is considered to be realistic. This means that both teams have been competitive over the seasons, with their strengths increasing steadily. Moreover, the estimations illustrate that Barcelona has kept a slight edge over Real Madrid, fact also showcased by the fact that the Catalan team has a win probability of 0.5, while their adversaries only have a 0.35 probability of winning the duel.

The model is paired with a vast dataset. It contains all the match results from the top six European football competitions: the national leagues of England, Germany, Spain, Italy, France and the Netherlands over the course of seventeen seasons, from the 1999-2000 season up to the 2015-2016 season. Moreover, the dataset slightly increases as every season there are teams that relegate, and teams that promote into the first division of each league. The teams that move either up or down a division are still kept in the dataset as they can re-appear in future seasons. From all the seasons, the first ten seasons (1999-2009) are set as the in-sample seasons, which are used to estimate the parameters. The other seasons are used to study the forecasting precision.

Forecasting starts from the first matches of the season 2009-2010, and are based on parameter estimates from the previous ten seasons. Since the authors already have the match results, they can evaluate a loss function which is used to measure the precision of the forecast. When forecasting the next round of matches, the parameters are re-estimated after including the last match results, ensuring that as much data as possible is used. The loss function used is the rank probability score (RPS), which takes into consideration the other probabilities as well. The RPS is computed for all football matches of a round and the average RPS is taken as the loss function. One alternative of the RPS is verifying the prediction results against the bookmakers’ odds, i.e., if the model is able to generate a profit over the odds (a betting simulation).

The results of the forecasting show that the first two distributions are almost always preferred over the third one. A conclusion coming from this observation is that information tends to be lost when data is recorded in a more condensed manner. In more direct terms, this means that the goals scored by both teams in a match or the difference between the goals scored by each team are more useful and provide more information. When verifying if the model is able to beat the bookmakers’ odds, the failure of it showcased that there are still improvements to be made in the field, but not in the direction of creating better prediction models, but in finding variables that provide more information. The betting simulation also shows that the Skellam model is slightly better than the Bivariate Poisson model and is more suited to bigger leagues, such as England, Germany and Spain.

The second article, by Yiannakis et al, makes use of ARIMA as a base for a multivariate time series analysis procedure in order to determine the most influential factors that determine a win, a loss or a draw. Initially, 26 independent input variables were chosen. Despite most of them being binary, some variables also consisted of either ordinal or interval properties. Nevertheless, the appropriate ARIMA models were determined with the help of autocorrelations. Moreover, the authors also compare different methods of forecasting the result of matches, in search for the most precise.

In contrast to the dataset used in the previous paper, this one uses a much smaller one: the matches from the 1997-98 season of the English Premier League. From these matches, the authors picked three teams from across the table, one from each third (Chelsea, Derby and Tottenham). One possible reason for this is that the picked teams are impacted from external factors differently, whether we are talking about rest days, squad performance level, home crowd and so on. Using three different types of teams increases the accuracy with which it can be said that a factor influences a team’s possibility to win.

The first method used on the procedure was a single distribution, where each match result was associated with a numerical value (Win = 1, Draw = 0, Loss = -1). This method was applied for a team from the middle of the table. The first 28 games of the season were considered as the Test Model and the procedure was tasked with predicting the outcome of the last ten games. Out of these games, only six were predicted correctly, which, although better than predicting the outcomes by chance, yields poor results nevertheless.

An alternate method implies the creation of different distributions, one for each match result type. Therefore, wins, draws or losses are coded as ones or zeroes, where a 0 indicates the absence of that result and 1 is the presence of that result. To be consistent with the definition of probability, the prediction values were transformed into values ranging from 0 to 1. In this case, the authors considered values over 0.5 as ‘hits’, values below 0.5 as ‘misses’ and values of exactly 0.5 as ‘indeterminate’.

Using this method, the authors found that winning, drawing and losing a game are influenced by different factors, even if some of them overlap. Moreover, the forecasts generated using this model had better success rates when predicting the last ten games of the season for all three teams that were chosen.

When applying the model to predict winning, the success rate was 90 percent for the first two teams and 60 percent for the team in the lower third of the table. As to the major variables that contribute to winning for the teams, some of them overlap, such as playing at home for Derby and Chelsea or previous wins for Tottenham and Derby. Some variables were particular to a single team, like a loss in the previous game for Tottenham, or bigger crowds for Derby.

Similar models were applied to predict losing and drawing. Predicting losses turned out to be more precise for all three teams, with the success rate being either 80 or 90 percent. For losses, some variables were determined to have an influence on all teams, such as fewer rest days or playing away, while other variables applied only to one or two teams: playing a superior team for Chelsea, or a lower resolve to win for Chelsea and Tottenham. Predicting draws provided good success rates, ranging from 70 percent for Tottenham to 90 percent for Derby and 100 percent for Chelsea. As in the previous cases, predicting draws appears to be a combination of both shared and unique variables. For instance, facing a stronger side that played at home the game before, combined with a previous draw or loss helps predict a draw for both Tottenham and Derby. Playing after a draw or loss, combined with playing an opponent who enters the game with a prior draw seem to be major contributors to predicting a draw for Chelsea.

Analyzing these two papers, it is clear that both provide valuable insights regarding forecasting match results, offering two different methodologies and perspectives which yield notable results. Both papers highlight the importance of choosing appropriate models and variables, as well as a good dataset,in order to accurately predict the outcome of a football game. Nevertheless, they also highlight areas for improvement, particularly in refining variable selection, as illustrated in the first paper. This suggests that while current approaches are effective, further advancements could enhance the accuracy and utility of predictions, especially in the context of football analytics.

Bibliography

[1] Yiannakis, Andrew & Selby, Michael J.P. & Douvis, John & Han, Joon Young. (2006). Forecasting in Sport The Power of Social Context — A Time Series Analysis with English Premier League Soccer. International Review for the Sociology of Sport. 41. 89-115. 10.1177/1012690206063508.

[2] Koopman, Siem Jan & Lit, Rutger. (2019). Forecasting football match results in national league competitions using score-driven time series models. International Journal of Forecasting. 35. 10.1016/j.ijforecast.2018.10.011.