Universitat Politècnica de Catalunya

Barcelona School of Informatics

Master in Innovation and Research in Informatics

Data Analysis and Knowledge Discovery

Football Forecasting: State-of-the-art and approach

Author
Ricard Meyerhofer Parra

Lecturer
Alfredo Vellido Alcacena

March 31, 2019

Index

Abstract

The aim of this project is to study the methods and techniques that have been used during the last years in forecasting sports and more concretely in football. We start by conducting a study regarding football forecasting and later and by considering the relevant information that we gathered with this study, create our own dataset that is a modification of the Football-Data[20] with the addition of several factors that can be relevant to the problem. Finally we a Naive Bayes models and evaluate them.

1 Introduction

Forecasting is an activity that comes from the ancient Greece and the Romans, grows during the XVIII - XIX century and finally explodes with the irruption of internet. Who will win the fight? Which horse is the fastest? Is Miami Heat going to win in the last quarter when is losing 74-88? From long time ago, sports betting is a thing and any sport is good to bet and hopefully win some money regardless of the level of expertise of the bettor. There are many sports where one can try to forecast the result but this project is going to focus in a sports team, concretely in the most popular one: Football(soccer).

In football there are many different leagues that depend on the quality of the teams, the gender of the players, age of the players and the country. There are also national teams that compete against other countries in events like FIFA World Cup, Olympic games, etc. Among all these possibilities, the project is going to focus in the Spanish first division league named LaLiga which is formed by males. LaLiga is composed by 20 teams that compete to each other two times: one as a visitant (away) and another at home. This makes a total of 38 rounds with 10 matches per round which makes a total of 380 games per season. At the end of each season, the three bottom teams from the first division are descended to the second division and are replaced with the three top teams of the second division in a similar fashion the same happens to every league.

There are a lot of rules and technicalities in football that are not relevant to our scope which is **predicting the outcome of a football match**. This outcome is decided by which team scores more goals. If both teams score the same number of goals, the final result is a draw. Note that we are not aiming to predict the exact result e.g 4-2. We are only going to consider Win/Draw/Lose as possible outcomes and we are not going to predict a result while the match elapses, we are going to predict the final result of a match with the previous results before the former starts.

2 State-of-the-art

Even that there is an heuristic named recognition heuristic (Goldstein, D. G., Gigerenzer, G. (2002)) that mainly states that in lay predictions one should go for the only team he recognizes, this is shown untrue[16]. Therefore, we need reliable models in order to be precise with our predictions.

Forecasting is an activity transversal to all sports that is developed in sports such as Tennis, American Football, Basketball, and a large etcetera. What differentiates models between sports is that each of them, has their own particularities. Without digging further, football has a very special characteristic that is the low number of goals scored and the implications that each has. Consequently, we are going to focus in football forecasting literature. There is plenty of literature regarding to this topic which John Goddard summarizes in his paper[9]. We can find literature either in modelling and the impact of a given variable.

Regarding to the modelling progress, we could say that there are two main ways of understanding predictions of a football match:

- Modelling the goals scored and conceded by each team (which is the traditional approach).
- Modelling Win/Draw/Lose outcome of the match.

It is obvious that one model and the other are predicting the same but the way it is done is different. One is focusing in how many goals one team will receive and make. Once you have the predicted value, you can decide which is the result. The other model is just focused in the points that each team will score regardless of the goal count. In John Goddard's paper, he states that one model is not better than the other (there is not a significant difference), and that the best approach it is a hybrid approach.

If we do a historical overview of what has been achieved in the area, we can find contributions from the early 80's where Maher (1982) uses univariate and bivariate Poisson distributions, with means reflecting the attacking and defensive capabilities of the two teams. The limitations of this model were that was not able to predict before the match. By 1997 Dixon and Coles developed a forecasting model capable of generating probabilities for goals and match results and this time, before the match starts. It is focused in how many goals each time scores and these variables followed a univariate Poisson distributions to handle lowscoring matches an adhoc adjustment to the probabilities corrects for interdependence.

In 2004 Dixon and Pope compared probabilistic forecasts obtained from the Dixon–Coles model with probabilities inferred from UK bookmaker's prices for fixed-odds betting. Similar to this approach in 2000 Rue and Salvesen let the attack and defense parameters to variate through time randomly and the estimates update as new match outcomes are obtained. In order to simulate, they use Monte Carlo and Markov Chain techniques. In 2002 Crowder, Dixon, Ledford, and Robinson proposed a procedure for updating the team strength parameters which is less computer demanding.

More recent studies, rather than modelling goals, have been more focused in discrete choice regression models to model Win/Draw/Lose outcome.

- Forrest and Simmons (2000) investigated the quality of tipster's results forecasts, and the performance for postponed matches of the pools panel in providing hypothetical results.

- Audas, Dobson, and Goddard (2002) examined whether managerial change has any short-term impact on team results.

- Goddard and Asimakopoulos (2004) and Kuypers (2000) investigated the efficiency of prices quoted by high-street bookmakers for fixed-odds betting on results. Cain, Law, and Peel (2000) and Dixon and Pope (2004) in the case of goals. These findings suggest that the forecasts embedded in the bookmakers odds are inefficient. This is because bookmakers set the odds at the beginning of the week and do not change during the whole period.

- Kampakis and Adamides, conducted a study over Sentiment Analysis of Twitter (2014) in order to predict the outcome of a match[12]. Schumaker, Jarmoszko and Labedz over wins and spreads with the same method (2015).[17]

- Several machine learning based solutions such as k-Nearest Neighbor Algorithm[6] and in general all kinds of algorithms(2018)[1]

- Anthony Constantinou and Norman Fenton[3] (2017) with an smart-data approach that predicts at the beginning of the season the winner. This approach tries to difference of ML approaches by only predicting those necessary values and substituting some inaccuracy that can be lost with ML and substitute it with knowledge engineering approach.

- Bing predicts correctly all outcomes of 2018 FIFA WC group stage matches.

Regarding to the impact of some specific factors in the final result of a match, there have been multiple studies about them:

- Barnett and Hilditch (1993) investigated whether artificial playing surfaces gave advantage to the local team.
- Ridder, Cramer, and Hopstaken (1994) showed that player dismissals have a negative effect on the match outcome for the teams affected.
- Clarke and Norman (1995) quantified the local effect on match outcomes.
- Dixon and Robinson (1998) investigated how scoring rates of the home and away teams fluctuate during a match. And states that at any time are dependent on the time elapsed, and on which team is leading.
- Dyte and Clarke (2000) used 1998 international World Cup's data to study the relationship between a previous performance and the World Cup tournament performance.
- Carmichael et al (2000) studied the relation of specific types of plays with the number of goals.
- Numerous studies of rating systems such as the ELO rating system[11] have been studied and appear to be less reliable than the bookmakers odds. [18].
- As mentioned in this paper[18] Forrest, Goddard and Simmons studied how payrolls affect the results (2005).
- As stated in the paper Issues in sports forecasting[18]. Evidence suggests that tipsters do not process public information properly and that have little value Forrest and Simmons, 2000; Pope and Peel, 1989. In contrast, Forrest et al. (2005) demonstrates that there is no difference between forecasts of the odd makers and forecasts of a complex model which is consistent with other authors (Kuypers 2000). This topic is unclear.

3 Project's work

In order to be successful with the project. We have followed the CRISP methodology therefore, we are going to organize the project with each of the CRISP phases but compressing the cyclic constant iteration of the whole process when writing about it. The deployment phase makes no sense in this project because is not for a real organization and it is just an academical project therefore, I will not include it.

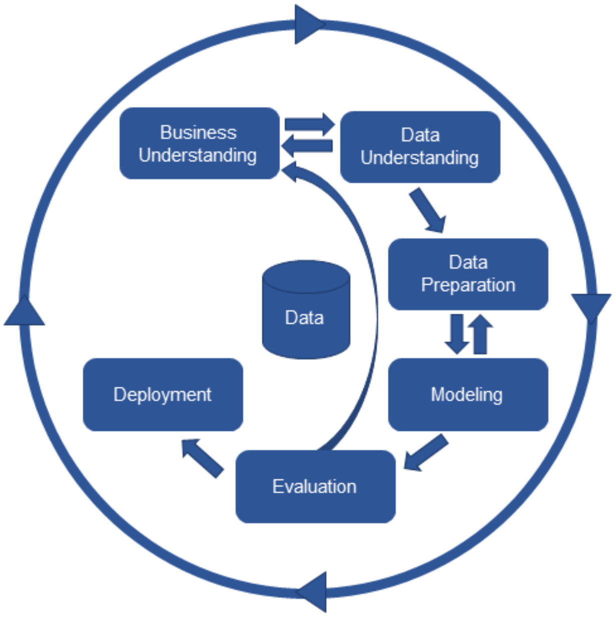

Figure 1: CRISP phases

3.1 Problem Understanding

As aforementioned, we are going to focus in forecasting the result of Spanish LaLiga football league. I am going to treat it a classification problem so where we are going to decide if a team wins, loses or draws against another.

In order to see which is the best algorithm possible we are going to try some different models to see which one performs the best. Once we have a model, the success criteria will be to see if we predict well matches or not by comparing to the real results. We are going to consider ourselves successful if we have an accuracy over 50% which I think it is a difficult goal to achieve because of the sport we are trying to forecast and the dimensions of my project.

More concretely what is going to be evaluated is the 4 last journeys the 2017-2018 league which consists in the following matches:

Home Team	**Away Team**	**Outcome(H/D/A)**
Levante	Sevilla	H
Espanyol	Las Palmas	D
Real Sociedad	Athletic	H
Real Madrid	Leganés	H
Villarreal	Celta	H
Getafe	Girona	D
Alavés	Atlético	A
Valencia	Eibar	D
Deportivo	Barcelona	A
Betis	Málaga	H

Table 1: Journey 35 LaLiga Santander 2017-2018

Home Team	**Away Team**	**Outcome(H/D/A)**
Sevilla	Real Sociedad	H
Girona	Eibar	A
Athletic	Betis	H
Celta	Deportivo	D
Villarreal	Valencia	H
Málaga	Alavés	A
Atlético	Espanyol	A
Las Palmas	Getafe	A
Barcelona	Real Madrid	D
Leganés	Levante	A

Table 2: Journey 36 LaLiga Santander 2017-2018

Home Team	**Away Team**	**Outcome(H/D/A)**
Real Sociedad	Real Sociedad	H
Alavés	Leganés	H
Deportivo	Villareal	H
Eibar	Las Palmas	A
Betis	Sevilla	D
Getafe	Atlético	A
Girona	Valencia	A
Real Madrid	Celta	H
Espanyol	Málaga	H
Levante	Barcelona	H

Table 3: Journey 37 LaLiga Santander 2017-2018

Home Team	**Away Team**	**Outcome(H/D/A)**
Celta	Levante	H
Leganés	Betis	H
Las Palmas	Girona	A
Málaga	Getafe	A
Sevilla	Alavés	H
Villareal	Real Madrid	D
Valencia	Deportivo	H
Athletic	Espanyol	A
Atlético	Eibar	D
Barcelona	Real Sociedad	H

Table 4: Journey 38 LaLiga Santander 2017-2018

In order to carry out the project I will use R. Mainly because I feel more comfortable with it than in Python for instance and also because R is a better data exploration tool than Python, especially if we consider plots and in general data exploration.

3.2 Data Understanding

Probably the most important step in the whole project is deciding which is the data that we are going to use. In this context, retrieving data from football is not an easy task. There are not this much complete datasets and what I have decided is to create my own one since I think that all the ones I found, were missing something.

Since the data I was looking for is not available at least in an open-source unified dataset, I had to create my own. I took as base the Spanish LaLiga datasets from Football-Data[20].In addition to the dataset variables, (which can be found in the Appendix A) I aggregated the following variables that I think that might be relevant. These variables have been selected during several iterations and even that I am not showing plots from them, I am taking into consideration what I read about football forecasting.

- **Budget of the clubs:** It is highly related with the quality of a team. Normally a team that has a bigger budget, can afford better players and in case of bad results will always be able to get better players. Probably will also have better physical trainers and a better planning to avoid strain.
- **Number of foreigner players:** Normally a foreigner player is better than a local player because otherwise, there is no reason to hire a player from another country.
- **SPI:** It is a dataset which gives the score that the team from FiveThirtyEight website calculate for each international team. It provides and offensive and defensive score and a total score.[24]
- **Derby**: Which indicates if the match between two teams is a derby.

I want to make note that I am not using all the variables from the Football-Data dataset. I am using some of them which can be found in an appendix with their respective explanation (the one provided by football-data). Summarizing it is a dataset that aggregates data from the match and also it aggregates data from different 1x2 betting odds (1x2 odds are those with betting option for Win/Lose/Draw). In this case we are just going to keep one of them which is Bet365 one.

Some variables that were initially considered but have been turned down during the iterations:

- **Sentimental analysis from Twitter:** Since we chose Spanish LaLiga league and the main language and source of information is Spanish, I decided not to deal with sentiment analysis in a language different than English.

- **Average age:** I thought it was an interesting variable since the younger a team is, the more motivated can be but at the same time, lacks of experience and other factors that one only gets as the time passes. But then I saw that all teams have a very similar average so its impossible that I get something significant from there.

- **Distance between Home and Away team:** This variable was considered at first and I read in some literature where was studied but honestly, with nowadays transportation, you can be in any place of Spain within two hours if you take a plane so this is not an issue anymore.

- **Average quality of players considering FIFA datasets:** This information has been added with the premise that a team with better players will win and that a team more less has a whole team with similar quality (few variance between players). Initially I aggregated this data in order to have an index that could be updated and I thought that taking this indexes from the game or any API that provides them is an affordable option. Since we are doing just Spanish first division league I think that SPI is better because at least I have a justification of how the calculus are done and it has free access to the public. Furthermore I have a historical report where I can see for each match how it evolves.

- **Different tipster's predictions:** I thought would be interesting to have an expert prediction in my dataset so that my model could feed from it but I did not find any dataset nor easy way to collect the data.

- **Weather conditions at the day of the match:**It is know that weather affects the playing style of a match since it makes the ball go faster, it makes some strategies better than other, etc. Therefore, I wanted to see if this variable really affects the results or it just really affects both teams for equal and does not affect the outcome but it actually supposes a lot of work and it very hard to find if in the match the climate conditions were good or not without looking one by one by hand.

3.3 Data Preparation

Given the dimensions and the scope of the project, I will only use seasons 2016-2017 and 2017-2018 (league in which I will use some of the journeys to test my model) and I will remove some variables that are not useful to us since are variables of odds in between the match.

I am only doing two seasons because obtaining the data for a season is an effort and for illustrative purposes it does not variate doing it with 15 seasons or 2. In the result based point of view it does affect the result. In order to be able to model the data, I had to normalize it so that I could merge the different attributes from the different datasets (SPI and the budget and foreigner players of the clubs). I did this part with excel because honestly there were too many columns and it was easier to filter and merge data. What I did is to match the data with both datasets and then merge columns. I merged the budget of each club by first normalizing the names and then add them. From the SPI matches dataset I only ended up with the SP Index of each team.

Also data had to be split in two: training set and test set. The training set will be the one where we will create a model from and the test set is where we are going to test the accuracy of the created model. As aforementioned, I am going to test the last 4 journeys of 2017-2018 league which means that will be removed from the training set (otherwise it would be a massive error)

At the end among the many columns that we could have chosen, we ended up arround 30 columns that could be summarized in 4 big blocks:

- Betting data.
- Index data.
- Match data.
- Economic data.

3.4 Modelling

The problem we have to solve it is a classification problem. We are going to compare their performances in order to see which one apparently behaves better and with which percentage of accuracy. There are many algorithms that we can try but we are going only going to do Naive Bayes.

3.5 Evaluation

We compared the real results with the results that the Naive Bayes gave us and we obtained:

- Journey 35: 40% accuracy
- Journey 36: 50% precision
- Journey 37: 60% precision
- Journey 38: 50% precision

4 Conclusions and future work

As could be seen in the state-of-the-art, there is a huge number of studies performed. Not all of them go in the same direction and some of them contradict each other. Moreover, the quantity of studies is honestly overwhelming but I think that the state-of-the-art provided is sufficient and clear enough

Regarding to the results, in order to achieve better results it would probably help to have a better dataset it was a too out of scope work with some effects such as the weather of the match day. More complete and more precise. I think that having the dataset is the most challenging step of the whole project and it is more difficult than I would expect because even it is somehow public information, it is not all gathered and ready to go for the user that wants just to create a model of it. More specifically I think that having different leagues would be beneficial to improve our performance if in every league the forecasting behaviour is the same (thing that should be proved). Also further detail and study in the state-of-the-art would let to a deeper knowledge of the area which even that I think it has been quite extensive, can be improved.

In relation with the implementation, I would like to note that the code that has been done is not a scalable code. This is because is just for analysis purposes and it is not a code that expects to be deployed nor to be applied for other leagues/sports. I would have also liked to apply more than one method.

Overall I would like to conclude that I am happy with the results given the scope of the project and that it was not an ambitious project. It is a humble approach to complement a project that could have been just a state-of-the-art of football forecasting but I thought that putting some of the concepts in practise and to do my own version would be more educational than just doing a summary of it and I really enjoyed doing it. When I started I felt curious about the topic and this subject, brought me the opportunity of exploring the area. I do not discard doing my final master thesis of the same topic.

A Football-Data.co.uk dataset

League Statistics Date = Match Date (dd/mm/yy)
HomeTeam = Home Team AwayTeam = Away Team FTHG and HG = Full Time Home Team Goals
FTAG and AG = Full Time Away Team Goals
FTR and Res = Full Time Result (H=Home Win, D=Draw, A=Away Win)
HTHG = Half Time Home Team Goals
HTAG = Half Time Away Team Goals
HTR = Half Time Result (H=Home Win, D=Draw, A=Away Win)

Match Statistics (where available)
Attendance = Crowd Attendance
Referee = Match Referee HS = Home Team Shots AS = Away Team Shots
HST = Home Team Shots on Target
AST = Away Team Shots on Target
HHW = Home Team Hit Woodwork
AHW = Away Team Hit Woodwork
HC = Home Team Corners AC = Away Team Corners HF = Home Team Fouls Committed
AF = Away Team Fouls Committed
HFKC = Home Team Free Kicks Conceded
AFKC = Away Team Free Kicks Conceded
HO = Home Team Offsides AO = Away Team Offsides HY = Home Team Yellow Cards
AY = Away Team Yellow Cards
HR = Home Team Red Cards AR = Away Team Red Cards

Note that Free Kicks Conceded includes fouls, offsides and any other offense committed and will always be equal to or higher than the number of fouls. Fouls make up the vast majority of Free Kicks Conceded. Free Kicks Conceded are shown when specific data on Fouls are

Key to 1X2 (match) betting odds data:
B365H = Bet365 home win odds
B365D = Bet365 draw odds B365A = Bet365 away win odds

References

[1] Baboota, R., Kaur, H. (2018). Predictive analysis and modelling football results using machine learning approach for English Premier League. International Journal Of Forecasting. doi: 10.1016/j.ijforecast.2018.01.003

[2] Baker, R., McHale, I. (2013). Forecasting exact scores in National Football League games. International Journal Of Forecasting, 29(1), 122-130. doi: 10.1016/j.ijforecast.2012.07.002

[3] Constantinou, A., Fenton, N. (2017). Towards smart-data: Improving predictive accuracy in long-term football team performance. Knowledge-Based Systems, 124, 93-104. doi: 10.1016/j.knosys.2017.03.005

[4] Constantinou, A., Fenton, N., Neil, M. (2013). Profiting from an inefficient association football gambling market: Prediction, risk and uncertainty using Bayesian networks. Knowledge-Based Systems, 50, 60-86. doi: 10.1016/j.knosys.2013.05.008

[5] Dixon, M., Pope, P. (2004). The value of statistical forecasts in the UK association football betting market. International Journal Of Forecasting, 20(4), 697-711. doi: 10.1016/j.ijforecast.2003.12.00

[6] Esme, E., Kiran, M. (2018). Prediction of Football Match Outcomes Based On Bookmaker Odds by Using k-Nearest Neighbor Algorithm. International Journal Of Machine Learning And Computing, 8(1), 26-32. doi: 10.18178/ijmlc.2018.8.1.658

[7] F.I., A., J.C, O. (2015). English Premier League (EPL) Soccer Matches Prediction using An Adaptive Neuro-Fuzzy Inference System (ANFIS). Transactions On Machine Learning And Artificial Intelligence. doi: 10.14738/tmlai.32.1027

[8] Forrest, D., Goddard, J., Simmons, R. (2005). Odds-setters as forecasters: The case of English football. International Journal Of Forecasting, 21(3), 551-564. doi: 10.1016/j.ijforecast.2005.03.003

[9] Goddard, J. (2005). Regression models for forecasting goals and match results in association football. International Journal Of Forecasting, 21(2), 331-340. doi: 10.1016/j.ijforecast.2004.08.002

[10] Grant, A., Johnstone, D. (2010). Finding profitable forecast combinations using probability scoring rules. International Journal Of Forecasting, 26(3), 498-510. doi: 10.1016/j.ijforecast.2010.01.002

[11] Hvattum, L., Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. International Journal Of Forecasting, 26(3), 460-470. doi: 10.1016/j.ijforecast.2009.10.002

[12] Kampakis, S., Adamides, A. (2014). Using Twitter to predict football outcomes. CoRR, abs/1411.1243.

[13] Kampakis, S., Thomas, W. (2015). Using Machine Learning to Predict the Outcome of English County twenty over Cricket Matches. arXiv preprint arXiv:1511.05837.

[14] Kovalchik, S., Reid, M. (2018). A calibration method with dynamic updates for within-match forecasting of wins in tennis. International Journal Of Forecasting. doi: 10.1016/j.ijforecast.2017.11.008

[15] Leitner, C., Zeileis, A., Hornik, K. (2010). Forecasting sports tournaments by ratings of (prob)abilities: A comparison for the EURO 2008. International Journal Of Forecasting, 26(3), 471-481. doi: 10.1016/j.ijforecast.2009.10.001

[16] Pachur, T., Biele, G. (2007). Forecasting from ignorance: The use and usefulness of recognition in lay predictions of sports events. Acta

Psychologica, 125(1), 99-116. doi: 10.1016/j.actpsy.2006.07.002

[17] Schumaker, R., Jarmoszko, A., Labedz, C. (2016). Predicting wins and spread in the Premier League using a sentiment analysis of twitter. Decision Support Systems, 88, 76-84. doi: 10.1016/j.dss.2016.05.010

[18] Stekler, H., Sendor, D., Verlander, R. (2010). Issues in sports forecasting. International Journal Of Forecasting, 26(3), 606-621. doi: 10.1016/j.ijforecast.2010.01.003

[19] Štrumbelj, E. (2014). On determining probability forecasts from betting odds. International Journal Of Forecasting, 30(4), 934-943. doi: 10.1016/j.ijforecast.2014.02.008

[20] Football-Data.co.uk website. Base from the dataset used.
http://www.football-data.co.uk/spainm.php

[21] Website used to get the budgets from clubs.
https://www.transfermarkt.es

[22] Fifa dataset from which we calculate the team average quality, age and number of foreigners.
https://www.kaggle.com/thec03u5/fifa-18-demo-player-dataset

[23] Kaggle dataset from fifa players.
https://www.kaggle.com/thec03u5/fifa-18-demo-player-dataset

[24] This website has some sports blogging and has a forecasting system for the LaLiga League.
https://projects.fivethirtyeight.com/predicciones-de-futbol/la-liga