Universitat Politècnica de Catalunya - BarcelonaTech

Master Thesis

Forecasting Football Results

Author
Ricard Meyerhofer Parra

Supervisor
Alfredo Vellido Alcacena

A thesis submitted in fulfillment of the requirements for the Master in Innovation and Research in Informatics, Facultat d'Informàtica de Barcelona (FIB)

in the

IDEAI and SOCO Research Groups
Data Science Speciality

October 19, 2020

Abstract

The current study evaluates and compares different techniques to predict the outcomes of association football with the milestone to surpass the bookmakers forecast. There is a wide variety of approaches on this subject that have been covered on literature: Poisson distributions, Machine Learning techniques, ELO ratings, PGMs,... along with the study of many factors that influence the outcome of a match. Therefore, we perform an extensive study of the state-of-the-art techniques and variables together with the current available datasets. Given the lack of open-source extensive datasets, we opt for the one which seems to be the most complete one and up-to-date: the European Soccer Database which we extend with data from national cups and cross-european competitions.

Once the data source is chosen, we perform an exploratory data analysis to the dataset, extract event-log features regarding the shots, fouls, and crosses. Moreover, we generate variables seen on literature. Finally, we encapsulate everything in model which we approach thrice: as Poisson difference distribution in order to evaluate the goal difference as a machine learning model and as a pi-rating ELO system.

We evaluate the accuracy of our models with RPS were found that the best performing model we generated is a SVM with a RPS of 0.24318 against the 0.223418 from the bettingOdds. To conclude, we discuss the issue that data heterogeneity produces to the problem and showcase some soccer-logs which have some of their logs public are able to offer.

A la mare i al seu somriure incondicional,
el càncer no podrà mai apagar el caliu que irradies

Index

List of Figures

List of Tables

1 Introduction

1.1 Project Formulation

Sports are a transversal aspect of our society which comes from ancient times. There are many sports and all of them, gather a huge amount of fans across the globe. In the same way that sports have been with us for centuries, so does predicting their outcome. In this project, we aim to predict the final outcome of association football (soccer) matches. Among the many different sports possible, football is specially complex because of its nature. It is a low scoring sport, continuous, time varying and very strategic. It has already been proven that even top experts, cannot predict accurately results in a consistent fashion, because they tend to underestimate and overestimate their data[1][2] since they do not process public information properly and fail to make successful use of other unspecified information relevant to game outcomes[3][4]. As a result, there is an increasing tendency in the use of data science techniques not only to improve predictions but to improve sports performance in general.

This problem has been studied for decades and we can find plenty of literature trying to model this problem successfully: which variables make a team win, which model approach to follow, etc. Among these studies we can find that there has been an exhaustive study on how betting odds are or not accurate to the market and how representative of the actual outcomes are. In this project, we want to highlight that even we might use betting data, we are not focusing in profitability (e.g betting strategies) but predictability of a football match and how data can help us to achieve our goal. How far can we predict a football match with the data available? Which factors influence the most the final result? These are the kind of questions that we aim to resolve.

As aforementioned, data science is nowadays more present in sports and thanks to the big data revolution, we now can have access to information that previously could only be dreamed of. There are currently platforms that work in the generation of datasets which contain most inputs that occur in a match. These datasets, include the geolocalization of all the players when they have the ball, their passes, shots, in conclusion: most the events of a given game from every single player. Unfortunately, this information tends to not be publicly available and it is instead offered as a service and does not contain the information of what a player does when it does not have the ball.

Clubs also play a role on this by not facilitating their own tracking data[86].

As already mentioned, our aim is to accurately predict the chances that a team will win a match. This outcome is interesting to different stakeholders, for different reasons: The first one and most obvious, is the bettors. We do not want anyone to bet their money in betting. However, currently betting represents a fast growing market of hundreds of billions a year. Providing accurate results, will help these people to have a better understanding on their bets. The second stakeholder, is the bookmakers itself which need to work on their odds in order to be profitable and want to adjust the value of the odds to the real value. The third stakeholder, are the clubs and football. If you are able to understand the errors that make you win/lose, you might be able to work on them and address them, find your opponent weaknesses, understand what you could be doing better, etc. Note that, as an indirect consequence of this project, by being able to predict successfully the outcome of a match, we are contributing to the profitability problem since it is not anything but maximization problem with probabilities. Finally, it is very likely that some of the conclusions extracted from this project, can be extended to other sports as it happens that some novelties from football, actually come from models applied to other sports such as cricket[5], basketball[6] or even come from other studies such as animal behaviour[]

1.2 Project Main Goals

As mentioned, the main objective of the project is to predict the final outcome of a football match. This comes associated with several challenges which are going to be the guideline of our project.

- Perform an in depth study of the current state-of-the-art.
- Study the dataset market and end with a dataset(s) to work with.
- Creating a methodology which states how to measure the accuracy of our algorithms. Having a methodology that evaluates our progress in a consistent manner, will help us in iterating faster towards a goal.
- Try to over-perform the odds given by the bookmakers. Define which are the boundaries on forecasting accurately: which are the minimum and maximum boundaries, from a human/machine perspective. We also need to define which method are we going to use to evaluate the preci-

sion of our results.

- Apply different approaches and techniques to our problem, compare them and select one.

These are the main goals from our project, which will be detailed in the following sections once we have performed our state-of-the-art and dataset study since without them, we cannot state realistic goals.

2 State-of-the-art

Even that there is an heuristic named recognition heuristic[7] Goldstein & Gigerenzer (2002) that mainly states that in lay predictions one should go for the only team he recognizes, this is shown untrue[8]. Therefore, we need reliable models in order to be precise with our predictions.

Luckily for us, forecasting the results of sports events is not a novelty. We can find plenty of literature regarding football and other sports. Among the literature we can find up to the date, we can classify them in three big blocks[9] which depend on the kind of approach that is given to the problem.

- **Statistical models:** Which normally include ordered probit regression models and Poisson models.
- **Machine learning and probabilistic graphic models:** Here we have a bigger variety. We find Bayesian and Markov methods, genetic algorithms, neural networks, etc.
- **Rating systems:** This last approach is based on ELO rating systems and their variants which are pretty known and common across sports such as chess, tennis, online games, etc.

Aside from the previous three main blocks, which are focused in how we approach the problem to solve, we also can find studies that cover some factors that are relevant to our problem.

- **Study of factors and their relevance:** We are going to cover studies where a certain factor is studied e.g local factor, climate conditions, fatigue, rivalries, underdog effect, etc.
- **Other sports:** Clearly not everything that is applied to other sports, can be transferred from one sport to another (it might have no sense in our sport) but it is quite common that a given concept can be applied and that it makes sense. Therefore, it is important to look at other sports and have a grasp of what is being studied.
- **Football in-game analysis:** As we previously mentioned on the introduction, prediction is not the main aspect that is covered on association football. Nowadays, improving and providing an insight on games with data analysis is the trend. What can the players do to improve their

chance to score? How to pressure when defending? This are the kind of questions that most of teams focus in order to improve and in order to put science in their game play.

Since we think that is fundamental to have a clear idea of what has been previously done and that will help us to have a more clear idea, but before exhaustively analysing each of these categories and find out what can we use for our own work, we want to highlight some important information:

- A lot of papers are focused in profitability+betting strategies rather than predictability itself.
- Most of the papers use data from a single league or tournament.
- Regarding to the modelling part, there are two main ways of understanding predictions of a football match: Regression and classification.
  - We can treat the problem as a regression problem by modelling the goals scored and conceded by each team (which is the traditional approach).
  - We can treat the problem as a classification problem where we model the outcome of the match as Win/Draw/Lose.

  It is not difficult to see that both models are predicting the same but the way it is performed is slightly 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. As Goddard (2005) states[10], one model is not better than the other (there is not a significant difference), and apparently the best approach it is an hybrid approach.

2.1 Statistical Models

If we do a historical overview of what has been achieved in the area, we can find contributions starting from the 50's, Moroney (1956) demonstrated that the number of goals scored in a football game does not follow a Poisson distribution but a Negative Binomial[11]. This was confirmed by Reep, Pollard and Benjamin (1968) by testing it with English Football League First division for four seasons[12] and later on, extending this idea to other sports[13].

However, these publications affirm that "chance does dominate the game". This assumption was shown wrong by Hill (1974)[14] since everyone agrees that there is a luck factor and that not all chances nor the teams are equally relevant. Furthermore, it shows that experts were able to forecast the results of the league before the season started with some degree of success which indicates that skill rather than chance dominates the game. Aligned with this idea, we can find the idea that these inherent qualities of each team, can be identified by using for example, maximum likelihood[15] or by a linear models[16][17]. However, in 1982 Maher (1982) sets what will be the foundation for the following statistical models by using a Poisson. Therefore, we will structure it by Poissons distributions that have been used.

2.1.1 Double Poisson distribution

In Maher (1982) [18] sets the foundation for many of the following models. In this paper, Maher states "over a whole season, skill rather than chance dominates the game" (p.109). To showcase his theory, he evaluates teams in two aspects: offense and defense. In other words, the capacity of a team to score and to avoid conceding goals. In order to formulate this Maher (1982), assumes that each time a team has the ball, has the opportunity to attack and score with a probability p. If this p is constant and attacks are independent, the number of goals will be binomial and in these circumstances the Poisson will apply. Therefore, if a team i is playing against a team j and the observed score is (x_{ij}, y_{ij}), we can model the outcome of a football match as the following:

$$\begin{aligned} X_{ij} &\sim P(\alpha_i \beta_j) \\ Y_{ij} &\sim P(\gamma_i \delta_j) \end{aligned} \tag{1}$$

where:

- α_i is the attack strength of team i when playing home.
- δ_j is the attack strength of team j when playing away.
- β_j is the defense weakness of team j when the playing away.
- γ_i is the defense weakness of team i when playing home.

These variables are based on the number of goals scored and conceded in each team previous matches. In his initial model X_{ij} and Y_{ij} are independent as each team were playing a separate game. So we have a double Poisson distribution. Since the X and Y are independent, the estimation of α and β will come from x and y one, from δ and γ we have the following home team score equation:

$$log\ L(\alpha, \beta) = \sum_i \sum_{j \neq i} (-\alpha_i \beta_j + x_{ij}\ log(\alpha_i \beta_j) - log(x_{ij}!)) \tag{2}$$

Therefore,

$$\frac{\partial log(L)}{\partial \alpha_i} = \sum_{j \neq i} (-\beta_j + \frac{x_{ij}}{\alpha_i}) \tag{3}$$

and so, the maximum likelihood estimates, $\hat{\alpha}, \hat{\beta}$ satisfy:

$$\hat{\alpha_i} = \frac{\sum_{j \neq i} x_{ij}}{\sum_{j \neq i} \hat{\beta_j}} \text{ and } \hat{\beta_j} = \frac{\sum_{i \neq j} x_{ij}}{\sum_{i \neq j} \hat{\alpha_i}} \tag{4}$$

Finally, Newtwon-Raphson enable these MLE to be determined. These estimates show the advantage of playing at home, as each team attacking strenght is reduced when playing away. Later on in the same paper, Maher (1982) proves with MLE that there is no necessity of using the four parameters since we can describe with only one parameter per each: one for the quality of a team attack and another one for their defense. Because "althought home ground advantage is a highly significant factor, it applies with equal effect to all teams, and each team's inherent scoring power is diminished by a constant factor when playing away." (p. 113).

As we mentioned before, this model was a first major step towards the use of statistical models and because of the same reason, it has some major flaws. One of them is the oversimplification of the scoring rates which are assumed constant but are not. Dixon & Robinson (1998)[19], 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. So p fluctuates over time.

Other limitations of this model were that it was not able to predict ex-ante and that the attacking and defensive parameters for each team were estimated

ex-post (since it required a complete set for each season to work). We also find that the univariate distribution, there is a tendency of underestimating draws which is attributed to the interdependence between the goals scored by the home and away teams which is corrected by using a bivariate Poisson specification (which also is proposed on the same paper).

2.1.2 Bivariate Poisson distribution

The previous model, was treating a match as a two separated match. This is obviously a simplification as it obvious and Maher states in his paper[18] "A match does not consist of two independent games at opposite ends of the pitch; to the teams concerned, the result is all important, and so, for example, if a team is losing with ten minutes left to play, it must take more defensive risks in order to try to score". Therefore, he creates a new model, a bivariate Poisson. In this new model, the marginal distributions are still Poisson with the same means with means $\mu_{ij}(=\alpha_i\beta_j)$ and $\lambda_{ij}(=k^2\lambda_j\beta_i)$, but a correlation factor, ϱ, is added between the scores. The new model can be thought of as considering the difference in the number of goals scored, $Z_{ij} = X_{ij} - Y_{ij}$, resulting in a model with two dependent parts.

$$X_{ij} = U_{ij} + W_{ij} \quad \text{and} \quad Y_{ij} = V_{ij} + W_{ij}, \tag{5}$$

where U_{ij}, V_{ij} and W_{ij} are independent Poisson with means $(\mu_{ij} - \eta_{ij})$, $(\lambda_{ij} - \eta_{ij})$ and η_{ij}, respectively. η_{ij} being the co-variance between X_{ij} and Y_{ij}. Maher experimented with different values for ϱ, and found that the best fit would occur with $\varrho = 0.2$. The bivariate improves the results considerably, compared to the initial model and we can see that draws are better modelled.

Maher model is simple, since does not consider many factors that can affect a match. One of them is the attacking capability which we mentioned earlier, but there is another one that is the number of goals and if there is a correlation between match goals. There has been a lot of discussion regarding to the correlation between goals scored by each team. Karlis & Ntzoufras (2003)[20], discussed that as in many sports is more visible such basketball, since the two teams interact with each other during the match, the number of goals scored is influenced by both teams therefore, correlated. This is a concept that in a sport like basketball or handball, because are high scoring games and often there is a rhythm where both teams score each other constantly. However, football is low scoring so it is not so easy to proof. For instance, we

find that Mchale (2007)[22], which studies shots in football, finds no evidence of positive correlation. Since they found negative correlation between home team and away team shots.

Karlis & Ntzoufras (2003), worked also with a bivariate Poisson distribution. In this particular case, they chose to model slightly different and instead of adding the correlation factor separately from the distribution, they add it to the distribution itself. Which results in this model

$$\begin{aligned} (X_{ij}, Y_{ij}) &\sim BP(\lambda_i, \lambda_j, \varrho) \\ log(\lambda_i) = \mu + H + \alpha_i + \beta_j \quad \text{and} \quad log(\lambda_j) &= \mu + \alpha_j + \beta_i. \end{aligned} \tag{6}$$

This addition, improved the accuracy in prediction of draw games. Also they have improved more their model by inflating the diagonal probability (draws), which corrected results in case of overdispersion.

Koopman & Lit (2015)[23] also worked on a similar approach but introduced intensity coefficients for the number of goals scored by the two teams and a dependence coefficient for measuring the correlation between the two scores. The intensity coefficients depend on attack and defense strengths of the teams and they are allowed to evolve stochastically over time. The intensities are also subject to a fixed coefficient for home ground advantage. This model was put to test by a fixed betting strategy and was able to consistently gain profit.

2.1.3 Altered Poisson distribution

In 1997, Dixon & Coles (1997)[24], used Maher (1982) as basis but with small change, a home ground parameter H. This model is capable of generating probabilities for goals and match results and this time, ex-ante. It is focused in how many goals each team scores and these variables followed a univariate Poisson distributions to handle low-scoring matches an adhoc adjustment to the probabilities corrects for interdependence.

$$\begin{aligned} X_{ij} &\sim Poisson(\alpha_i \beta_j H) \\ Y_{ij} &\sim Poisson(\alpha_j \beta_i), \end{aligned} \tag{7}$$

where α_k and β_k are the attacking and defensive strengths of team k, and H the home ground advantage parameter. As mentioned, to handle low-scoring

matches, an adhoc adjustment to the probabilities is done

$$Pr(X_{ij} = x, Y_{ij} = y) = \tau_{\lambda,\mu}(x, y)\frac{\lambda^x exp(-\lambda)}{x!}\frac{\mu^y exp(-\mu)}{y!}, \tag{8}$$

where

$$\begin{aligned} \lambda &= \alpha_i \beta_j H \\ \mu &= \alpha_j \beta_i \end{aligned} \tag{9}$$

and

$$\tau_{\lambda,\mu}(x, y) = \begin{cases} 1 - \lambda\mu\varrho & \text{if } x = y = 0 \\ 1 + \lambda\varrho & \text{if } x = 0, y = 1 \\ 1 + \mu\varrho & \text{if } x = 1, y = 0 \\ 1 - \varrho & \text{if } x = y = 1 \\ 1, & \text{otherwise} \end{cases}. \tag{10}$$

ϱ is used as an dependence parameter. $\varrho = 0$ corresponds to independence. For $x \leq 1$ and $y \leq 1$, the independence distribution is altered.

Whereas the model of Maher had static strengths, Dixon & Coles (1997) incorporated to their model a weighting factor of recent matches.

Similar to this approach in Rue & Salvesen (2000) let the attack and defense parameters to variate through time randomly and the estimates update as new match outcomes are obtained. They added the use of Bayesian methods to update the estimates after each new match was played and in order to simulate, they use Monte Carlo and Markov Chain techniques to draw inference. Which Crowder, Dixon, Ledford, & Robinson (2002)[25] proposed a procedure for updating the team strength parameters which was less computer demanding.

More in detail, Rue & Salvesen (2000)[26] are also using Maher's foundation but as aforementioned, they represent the attacking and defensive strengths of team i as random variables α_i and β_i. They also represent $\mu_{\alpha,i}$ and $\sigma^2_{\alpha,i}$ as the prior mean and variance of α_i, and similar for defence. $e_i = (\alpha, \beta)_i$ represents the properties of team i. What is quite a novelty, is that they also added a psychological effect, modelling the underestimation when a superior team meets a team that is of supposed inferior quality. The psychological effect is given as

$$\Delta_{ij} = (\alpha_i + \beta_i - \alpha_j - \beta_j)/2,$$

which replaces the home ground advantage (however, the home ground advantage is part of an extended version of their model). An interesting insight of their investigation is that they assumed that some of the information from the final scoreline comes from the league itself, and not the actual result. To model this, they used a variable ϵ, which determines how much the league average contributes to the predicted number of goals. After some research, they found out that the best predictive results are achieved with $\epsilon = 0.2$.

2.1.4 Poisson difference distribution

Karlis & Ntzoufras (2009)[32], tried a Poisson difference distribution instead of a bivariate Poisson and instead of calculating the correlation between the performance and teams, they chose to model the goal difference. The goal difference Z_{ij} is calculated as a Skellam distribution (aside from the equation, model parameters are the same as defined in their previous paper).

$$Z_{ij} = X_{ij} - Y_{ij} \sim PD(\lambda_i, \lambda_j), \tag{11}$$

The reason why the Poisson difference is proposed, is that the marginal distributions of X and Y are in general defined as the convolution of a Poisson random variable with another discrete random variable, thus removing a large portion of the distributional assumptions concerning the number of goals scored by each team.

In order to incorporate prior distributions they use Bayesian methods. The information that they fed the model with was information regarding weather conditions, injuries on a team, how well they were performing. In case that the information was not available, they would use a normal prior distribution. The posterior predictive distributions were calculated with Monte Carlo.

The paper also mentions that there are multiple limitations in the previous models that used Poisson distribution and one way to address it is to use the goal difference instead of the number of goals scored and they also address the overestimation in draws.

Moreover, the author notes that since leagues are structured in a double round-robin distribution, the local effect results in distributions with one or both tails being too short or too long for the distribution. In order to overcome the limitations of the other Poisson-based models, a generalized

Poisson difference is proposed.

$$f(Z = X-Y = z|\lambda_1, \lambda_2, \theta_1, \theta_2) = e^{-\lambda_1-\lambda_2-\theta_1 z} \sum_{y=0}^{\infty} (\lambda_1, \theta_1)_{z+y} \, (\lambda_2, \theta_2)_y \, e^{-(\theta_1+\theta_2)y} \tag{12}$$

where

$$\forall z \in \mathbb{Z} \ , \ (\lambda, \theta)_x = \frac{\lambda(\lambda + x\theta)^{x-1}}{x!}. \tag{13}$$

To model the goal difference of a match, we can do it as follows:

$$\begin{aligned} E(Z_i) &= \mu_i = H + a_{h_i} - a_{v_i} \\ Var(Z_i) &= \sigma_i^2 = \gamma_1 + |a_{h_i} - a_{v_i}|, \end{aligned} \tag{14}$$

where a are the abilities of the home and visiting team in match, H is a fixed value, representing the home ground effect parameters. θ_1 and θ_2 are constant with respect to the team abilities. λ_1 and λ_2 are calculated as follows:

$$\lambda_{1,i} = \frac{[(1-\theta_2)^2 \ \sigma_i^2 + \mu_i](1-\theta_1)^3}{(1-\theta_1)^2 + (1-\theta_2)^2} \ ; \ \lambda_{2,i} = \frac{[(1-\theta_1)^2 \ \sigma_i^2 + \mu_i](1-\theta_2)^3}{(1-\theta_1)^2 + (1-\theta_2)^2} \tag{15}$$

2.2 Machine learning and PGM

2.2.1 Machine Learning

As it is known, machine learning gathers a lot of different techniques to the point that it is an umbrella term. In this section we are going to specify some techniques have been used and how successful they are. We will see in incoming sections that ML techniques are used for modelling other things such as expected goals, sentiment analysis, etc. But here in this section we are going to focus more in the predictability aspect and not in modelling a feature that will be included in our model.

If we look a bit at what has been done so far, we can find:

- Rotshtein, Posner, & Rakityanskaya (2005) created a fuzzy model for football prediction where the parameters were tuned using a combination of genetic algorithms and neural networks. They applied this model to the tournament data for the championship of Finland and showed that these tuning techniques for model parameter selection improved the results of their fuzzy logic model

- Aslan & Inceoglu (2007), trained two LVQ network models, in order to predict the results of the competitions for the Serie A 2001/02 league. The first half of the season was used as the training dataset, and the second half was used as the test data. The success obtained from the trained networks was 51.29% and 53.25%, respectively.
- Hucaljuk & Rakipovic (2011), worked on the multiclass classification problem for three possible results for the championship league competition. Presented six different artificial intelligence algorithms that were trained using 20 features. Some of those features were: the form of the team according to the last six games, the results of the previous competitions with the rivals, the status of the teams in ranking, the number of the injured players and the average number of goals scored per game for each team. His most successful classifier achieved a 65% accuracy.
- Esme & Servet (2018), applied bookmaker odds by using k-Nearest Neighbor Algorithm applied to the turkish league. Baboota & Kaur (2018), tried to beat the odds, and got very close to them. The best is gradient boosting with a 0.2156 RPS versus 0.2012 of the bookmakers (lower, better).

2.2.2 Probabilistic Graphic Models

Probabilistic Graphic Models (PGMs) are a rich framework for encoding probability distributions over complex domains. We can find the use of Markov and Bayesian Networks.

- Joseph, Fenton & Neil (2006), suggested the expert Bayesian network to predict the results of the competitions (home, draw, away) for the Tottenham Hotspur football team for the period 1995-1997. In this study, data such as whether prominent players would play in a game or not, their positions on the field, the attacking force of the team and average team quality were used as the variables. The average classification success of the model was determined to be 59.21%.
- Constantinou, Fenton, & Neil, (2012) evaluated for 6244 competitions in the English Premier League during 1993/94-2009/10. The Bayesian network was designed for 4 components, which were: team strength, team form, psychological impact, and fatigue. The ROI was between

2.87-9.48%.

- Another Bayesian network was proposed by Constantinou & Fenton (2013) based on their previous study Constantinou et al. (2012). They demonstrated that they were able to generate profitable returns by using a less complex model than their previous study.
- Another use of Bayesian networks was proposed by Owramipur, Eskandarian, & Mozneb (2013) predicting the football results of 92%, the Spanish football team, F.C. Barcelona with a very high accuracy . Their research utilized such as the weather conditions, psychological state of player. However, their model was for a single season, involving just 20 matches for the team under observation.

2.3 Rating systems

Rating systems are solutions that calculate the relative skills among competitors in zero-sum games based on a set of their preceding performances. As [33] says: "determining the relative ability between adversaries is probably the most important element prior to football match prediction, and the current league positions are widely assumed to be an accurate indication of this. However, league positions suffer from numerous drawbacks which makes them unreliable for prediction" (p. 37) and in general as he also mentions, fails to compare teams in different categories. Therefore, we cannot use leagues positions for this.

We are going to talk about 3 of the most relevant solutions that have been proposed. In general ratings is an unpopular topic in football whereas predicting has been more popular.

2.3.1 ELO

One of the most popular Rating systems is the ELO rating system created by Árpád Imre Élö, originally developed to assess chess players strength, which up to the date has been widely adopted in sports such as chess, tennis, football, videogames, etc. The ELO system is a system that was designed by chess. The central assumption of Elo's model is that the performance of each competitor, is a normally distributed random variable. Therefore, a competitor can one day perform better or worse but the mean value of

their performance, will only grow over time. This mean which represents the performance of a competitor, is the ELO rating which measures the current strength of a player. Since in chess you cannot really judge how good a player aside from the final result, the ELO system is based on the final result win/draw/lose and it is updated after every performance. Note that the ELO system needs a sufficient number of matches to be taken into consideration in order to be reliable, since it feeds from previous performances.

If we consider l_i^H and l_i^A the current ratings at the start of the match, of the home and away teams at a time i. The ELO rating defines a score system which sums 1, the expected home and away scores are defined by γ^H and γ^A which are calculated as:

$$\gamma^H = \frac{1}{1 + c^{(\ell_i^A - \ell_i^H)/d}} \quad \text{and} \quad \gamma^A = 1 - \gamma^H = \frac{1}{1 + c^{(\ell_i^H - \ell_i^A)/d}}. \tag{16}$$

c and d can be interpreted as setting the scaling of the rating. To calculate the ratings, the expected scores are compared to the observed scores, α^H and α^A respectively, given by

$$\alpha^H = \begin{cases} 1.0 & \text{if the home team won} \\ 0.5 & \text{if the match was drawn} \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

The actual score for the away team is $\alpha^A = 1 - \alpha^H$. The rating of both teams are updated in the same fashion once the match ends.The home team rating is:

$$\ell_i^H = \ell_{i-1}^H + k(\alpha^H - \gamma^H) \tag{18}$$

In order to use the ELO ratings for match prediction, they make use of an ordered logit regression model [34]. An initial set of matches is used to compute initial ratings for all the teams in the league. A second set of matches is used to estimate the parameters of the model. The rating difference, $x = \ell_i^H - \ell_i^A$ prior to the match serves as the covariate in the regression model. This system allows for updating both the ratings and the regression parameters, ensuring the most recent data is always utilized

This regression calculates the match predictions by assigning the corresponding probability for each outcome of the match, resulting in a probability dis-

tribution for the three outcomes. We can see an example in his paper where we can see that the model is able to capture the local effect. Where we can see that home win and away win probabilities are equal when the rating difference is about -80.

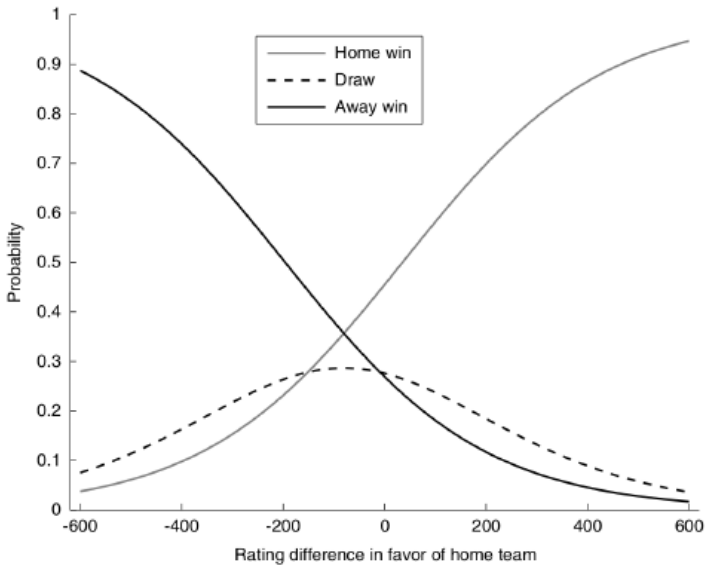


Figure 1: The home win and away win probabilities are equal when the rating difference is about -80

Hvattum and Arntzen (2010)[27] used rating systems in his studies in for match result prediction[55]. In his paper, he compares different rating systems and their variants:

- **Basic ELO (ELO_b) and its variant goal based ELO (ELO_g):** ELO_b is the same as the ELO rating we described so far. It is a normal ELO system. ELO_g is a variant of ELO_b where the rating update k, depends on the goal difference such that $k = k_0(1+\delta)^\lambda$, where k_0 and λ are fixed parameters bigger than 0, and δ the absolute goal difference.
- **FRQ:** Which takes the observed frequency of each outcome into account when making predictions.
- **UNI:** This method just assumes a uniform distribution on the outcomes.

- GOD_b, GOD_g Which uses Goddard's studies where he uses a ordered probit regression.
- **AVG:** Average odds from the bookmakers.
- **MAX:** Maximum odds from the bookmakers.

As presented, UNI assumes no information available, FRQ assumes previous results but without attention to details such as which teams are playing, number of goals, etc. AVG where for each outcome they compute the average odds and then take its inverse, prior to do so, they need to normalize the overround. MAX is a similar process but taking the maximum. This two methods are supposed to overpass the other ones, since if it were not the case, the bookmakers would be in an enormous trouble. However, it is known that the market may not be as efficient as we might think having bias.

To compare the different models, Hvattum & Arntzen (2010), used their predictions in combination with odds collected from various bookmakers. Match data from the English Premier League seasons 1993-2008 were used. The first two seasons were used for initial calculations of the ELO ratings. The five next seasons were used for estimating the parameters in the different prediction models. Finally, the eight remaining seasons were used for actual testing.

None of the methods presented was profitable and the paper concludes that none of the methods was able to surpass the ones based on the market odds but better than the other proposed methods.

2.3.2 Bradley-Terry modelling

The Bradley-Terry model[28] is a probability model representing the results of experiments in which responses are pairwise rankings of treatments. Given two individuals a and b, the defined probability that a is preferred over b $P(a > b)$ in a single comparison is the following:

$$P(a > b) = \frac{p_a}{p_a + p_b}, \tag{19}$$

where p_a and p_b are the scores assigned to a and b respectively.

This model was used by Cattelan, Varin, & Firth (2013) where they use descriptors of the competing team strenghts as the basis for the scores. This model was able to capture the evolution of team strengths.

The proposed model, calculates two team strengths: one for home, another when away. Initially, strengths of all teams start with the average number of points gained at home on their previous season. Once initialized, strengths are calculated according to the number of points scored recently on matches of the same type. For instance, home matches strength is estimated as following

$$\alpha_{h_i}(t_i) = \lambda_1 \mu_{h_i}(t_i) + (1 - \lambda_1)\alpha_{h_i}(t_{i-1}) \tag{20}$$

where $\alpha_{h_i}(t_i)$ is the home strength of team i at time t_i and t_{i-1} the time of the previous match played at home by team h_i. $\lambda_1 \in [0, 1]$ is used for determining how last result is weighted when estimating the team's home strength. The term $\mu_{h_i}(t_i)$ denotes the recent home strength of team i, based on the number of points earned in the last home match. $\mu_{h_i}(t_i)$ is defined as

$$\mu_{h_i}(t_i) = \beta_1 r_{h_i}(t_{i-1}),$$

where β_1 is a home-specific parameter, and $r_{h_i}(t_{i-1})$ the number of points earned in the last home match. The away strengths are estimated similarly. $\alpha_{h_i}(t_i)$ is then estimated using iterated back-substitution, thus incorporating the whole past of home matches. The same goes for the away strength. The values of λ_1 and λ_2 are estimated using maximum profile likelihood estimation.

To estimate the probabilities of each outcome, Cattelan, Varin, & Firth (2013)[29] use:

$$P(Y_i \leq y_i | Y_{i-1} = y_{i-1}, ..., Y_1 = y_1) = \frac{exp(\delta_{y_i} + \alpha_{h_i}(t_i) - \alpha_{v_i}(t_i))}{1 + exp(\delta_{y_i} + \alpha_{h_i}(t_i) - \alpha_{v_i}(t_i))}, \tag{21}$$

where $y_1 \in \{0, 1, 2\}$ denotes the outcome of the match (2 for home team victory, 1 for draw and 0 for away team victory). δ_{y_i} are cut-point parameters, where $\delta_0 < \delta_1 < \delta_2$. The cut-point parameters are needed for the Bradley-Terry model to support three outcomes. By setting $\delta_0 = -\delta$ and $\delta_1 = \delta$, with $\delta \geq 0$, one can ensure that two teams of the same strength playing at a neutral ground have the same probabilities of winning the match.

2.3.3 Pi-Rating

This paper proposes a solution which had a high impact an is a rating named "pi-rating" which unlike the ELO system we previously mentioned, obtained

profits. Pi-rating it is a rating very efficient with low complexity which can be used to treat our problem in both ways score-based and result-based match predictions. Furthermore, the pi-ratings can be incorporated into other more sophisticated models to increase their forecasting capability.

The paper wants to propose a rating which can accurately capture a team current ability considering:

- Home advantage factor
- Most recent results are more important than less recent results when estimating current ability.
- Winning is more important than increasing the goal gap.

and their proposal to tackle this musts are the following:

- Different ratings for when playing home and away. A learning rate γ which determines how much the newly acquired information based on home performance, influences a team rating and vice versa.
- Learning rate α which determines how much the newly acquired information of goal-based match results will override the old one
- High goal differences are exponentially diminished prior to update the pi-ratings.

With more detail, the overall rating of a team is the average rating between the home and away performance which we can define as:

$$R_\tau = \frac{R_{\tau H} + R_{\tau A}}{2} \tag{22}$$

where $R_{\tau H}$ is the local rating for team τ and the away one for $R_{\tau A}$. If we assume that we have a match between two teams, α and β, the home and away ratings are updated as follows:

Home team's home rating update $\rightarrow \hat{R_{\alpha H}} = R_{\alpha H} + \psi_H(e) \times \alpha$

Home team's away rating update $\rightarrow \hat{R_{\alpha A}} = R_{\alpha A} + (\hat{R_{\alpha H}} - R_{\alpha H}) \times \gamma$

Away team's home rating update $\rightarrow \hat{R_{\beta A}} = R_{\beta A} + \psi_A(e) \times \alpha$

Away team's away rating update $\rightarrow \hat{R_{\beta H}} = R_{\beta H} + (\hat{R_{\beta A}} - R_{\beta A}) \times \gamma$
(23)

where:

- $R_{\alpha H}$ and $R_{\alpha A}$ are the current home and away ratings for team α
- $R_{\beta H}$ and $R_{\beta A}$ are the current home and away ratings of team β
- $\hat{R_{\alpha H}}$, $\hat{R_{\alpha A}}$, $\hat{R_{\beta A}}$ and $\hat{R_{\beta H}}$ are the respective revised ratings
- ϵ is the error between predicted and observed goal difference $e = |g_D - \hat{g_D}|$ where $\hat{g_D} = \hat{g_{DH}} - \hat{g_{DA}}$ and $\hat{g_{DH}} = b^{\frac{|R_{\tau G}|}{c}} - 1$. This last formula, has no theoretical reasoning but gives them good results empirically.
- $\psi(e)$ is a function of ϵ also known as the weighting error, the main purpose of this function is to reduce importance of high score differences when updating the ratings. $\psi(e)$ is calculated as $\psi(e) = c \times \log_{10}(1+e)$. c in this particular case, it is 3.
- λ and γ are the learning rates. Which their reseach yielded $\lambda = 0.035$ and $\gamma = 0.7$

Some other important results that their investigation concludes is that in contrast with what was pretty much assumed which was that the home advantage factor remained costant as concluded Knorr-Held (1997, 2000); Koning (2000); Baio & Blangiardo (2010); Hvattum & Arntzen (2010); Leitner (2010). Which agrees with Clarke & Norman (1995) that in fact, reported that reported that teams can even develop a negative home advantage. In another area, they state that the development of the rating shows that two seasons of relevant historical outcomes, might be enough to converge into acceptable estimates on the basis of α and γ.

2.4 Study of factors and their relevance

Improvement and ideation of new models, have always been alongside to the study of some aspects of football and their relevance in the sport. How they

affect to the final result, how they can be incorporated to a model, etc. In this section, we are going to summarize some of the most relevant aspects that literature so far, has covered. There are many aspects to cover, I am going to prioritize their importance and how indicative I think they are of what has been done or what literature is capable of.

2.4.1 Local factor

One of the most studied and important factor is the local factor. It has been studied across the literature and there has been quite much contradiction on if this effect is constant, or not. As we previously mentioned, this effect, can even be negative Clarke & Norman (1995). Therefore, it really depends on the team. However, on average it tends to be a positive factor for teams where they tend to win more than lose Goddard (2006). Courneya & Carron (1992), presented a framework for game location research [**?**] which was later improved by Carron, Loughead, & Bray (2005). The proposed framework has been proven useful and considers the following aspects:

- **Game Location:** This point states that an artificial home ground i.e one designated by an organization, does not actually have the characteristics of a real home ground.
- **Game Location factors:**
  - **Crowd Factor:** Studies have shown that the public size is correlated with the home ground advantage. In the studies of Nevill, Newell, & Gale (1996) they saw that local teams had a bigger advantage when the crowd size was larger, whereas in cases where there was less public such as minor divisions, the effect was nearly absent. Also Pollard (2008), and Nevill, Balmer, & Williams (2002), state that crowd noise affect referees decisions which makes them bias towards them.
  - **Learning factor:** Basically this factor means that the visiting team has to get used to the field and factors such as the field size or the surface material, have been reported to affect the final result. In particular Carron, Loughead, & Bray (2005), stated that teams with a smaller field have an advantage over visiting teams with a bigger field. Also Barnett & Hilditch (1993)[54] stated that playing on an artificial pitch surface represents an advantage to

the local team if the visitor is not used to the surface. Another study Loughead et al., (2003) stated that relocating to a new venue, will diminish the local factors for teams that had a high winning percentage and will increase it for the teams that were not performing so well at home.

- **Distance factor:** Studies suggest that traveling contributes to the ground advantage. However, it is not very significant. Note that distance is not significant in derbies which could be one of the factors why derbies are always intense, together with that more supporters from the club will come from the visitor team.
- **Critical psychological and behavioural states:** Represent how the psychological and behavioral states of the teams are influenced by game location factors. There are very few studies on this area. Carron, Loughead, & Bray (2005) states in his paper that some studies showcased that players are in better physical and mental conditions when playing at home.
- **Performance outcomes:** Which is influenced by the previous mentioned aspects.

2.4.2 Expected Goals

Goals alongside with recent performances are normally used as a measure to predict. As we mentioned previously, football is low scoring and each goal has a lot of value. So one question arises, can we do something other than taking just into consideration the goals? Well, another simple solution which is already better would be to think: how many shots each team has performed? Were those shots were dangerous? Or just shots were the goalkeeper did not have to interfere. However, this measure is still naive. For instance, we can perform a very light centered shot with no danger and would count the same as a very well executed shot from the inner area.

Green (2012), introduces a new metric which tackles the problems that we mentioned previously. Expected goals (xG) is a metric which measures the probability that a shot will score on average at a given position. This is very useful because we can then have a deeper insight than just the goals scored, or the number of shots done that game. As a consequence of this model, we can remove some of the luck in the game and measure the shot chance

quality.

However, in order to measure the xG appropriately, we need some contextual information because after all, would not be fair to judge a shot from the position without considering, how is shot. For instance, it is known that shots which are assisted from crosses, are harder to convert that shots from other situation Knutson (2016).

There are some papers which have mentioned and improved contextual information that can be used:

Feature	Description
Distance	Distance to the mid-point between the goal posts
Visible Angle of the Goal	The angle formed between the shot location and the two goal posts
Passage of Play	Set formed by: open play, direct free kick, set play, corner kick, assisted, and throw-in
Assist Type	Set formed by: long ball, cross, through ball, danger-zone pass, and pull-back
Post take-on/ dribble	Whether the shot follows a previous attempt to beat a player
Rebound	Whether the shot follows a previous shot which has rebounded
Header	Whether the shot comes as a header
1vs1	A shot where there attacker has just one defensive player to pass
Big chance	A situation where a player should reasonably be expected to score. Normally very close to the goal posts, in 1vs1 with low pressure on the player

Table 1: Expected Goals metrics

Caley 2015 engineered new variables to the xG model. Moreover, he goes one step further and classifies shots in 6 categories:

- Regular shots.

- Shots from a direct free kick.
- Headed shots (excluding crosses).
- Headed shots from a cross.
- Shots from a cross
- Shots were the goalkeeper has been dribbled

In his studies also he notes that the game state (if a team is winning/drawing/losing) does not affect much the quality of a shot but for instance, the league does. This might seem surprising but the author attributes it to differences in pressure across leagues. In Soccermatics, David Sumpter[**?**] mentions that the counterattacks are the actions with the best chances of scoring.

Feature	Description
Fast Break	An attempt created after the defensive quickly turn defense into attack winning the ball in their own half[**?**]
Counterattack	Those actions that are not a fast break which begin with an open play turnover of possession, in which the attacking team moves steadily forward to the goal without recirculating the ball.
Established Possession	Attack that involves a minimum of five completed passes in the attacking half without passing the ball towards own the defensive zone.
Relative angle to the goal	The angle to the nearest post. If a player is in a central position, the angle is 1. If a player is at a 45º angle to the nearest post, the angle is 0.5.
Post take-on/ dribble	Whether the shot follows a previous attempt to beat a player
Interaction between the distance and angle	This is the distance to the goal multiplied by the relative angle to the goal.
Dribble distance	The distance a player has dribbled before taking the shot.
Error	Whether the shot follows an error by another player.
Body part	The body part used to take the shot
Game state	The game state is a feature that describes whether the team taking the shot is losing, drawing, or winning the match at the time of the shot.
League	A feature for the league, for example, the Bundesliga or the English Premier League
Log distance	The logarithm of the distance to the centre of the goal
Width of the goal mouth available to the shooter	The angle to the middle of the goal (American Soccer).

Table 2: Expected Goals metrics Caley, Kullowatz

So far we have only talked of features that can be used but we need to see how to model this features so that we have a model able to represent this variables. Therefore, if we analyze what has been done on the modelling aspect, we can see that Green, Caley, Kullowatz modelled this features with a logistic regression. This is a problem which contains some categorical variables which can be addressed by using one-hot encoding techniques. However, it is still

a problem because the non-linear relations such as pitch locations or angles, are not well handled. Because of this, decision trees are shown as an option.

Nevertheless, logistic regression is still a very solid option. The reason is that if we opt for trees, we will need to calibrate the model (which is provided by default in the logistic regression). Moreover, we lose also in interpretability, thing that the logistic regression provides in terms of odds ratios. Reason why Green, Caley, Kullowatz used logistic regression.

In terms of accuracy, Gelade (2007) suggests McFadden's pseudo-R2 or ROC curves, to validate the results k-fold cross-validation is recommended (with $k = 5$ or $k = 10$)

2.4.3 Expected Threat

As we previously mentioned, expected goals is a metric that measures the probability that a shot will score on average but is this enough? Can we do it better? Well the idea of expected threat (xT) Singh (2019)[**?**] is that yes, we can do it better. Instead of only measuring how dangerous is a shot from a given point $g(x, y)$, we are going to evaluate how dangerous is that point from a multiple stringing actions.

As Cervone mentions "Despite many recent innovations, most advanced metrics remain based on simple tallies relating to the terminal states of possessions like points, rebounds, and turnovers. While these have shed light on the game, they are akin to analyzing a chess match based only on the move that resulted in checkmate, leaving unexplored the possibility that the key move occurred several turns before. This leaves a major gap to be filled, as an understanding of how players contribute to the whole possession – not just the events that end it – can be critical in evaluating players, assessing the quality of their decision-making, and predicting the success of particular in-game tactics."

What Singh proposes informally is the following:

- **Reward individual player actions (passes, dribbles) in buildup play.** This will be done by assigning a player a score to each of his actions considering how much they contributed to the buildup play.
- **Operate on event-level data, due to availability constraints**. As we mentioned earlier we do not have access to player tracking data, we

only have event data. Therefore, we will have to work with event-data.

- **Reward actions independent of the end outcome of the possession**. A players failure should not imply that the previous players which did well, should be penalized. A dangerous pass it is a dangerous pass. In order to effectively providing a score of each action and since we only have event data (start and end location of an action), we are going to evaluate the difference of xG, so the score of an action that goes from point A to B, will be the differences of scores $xG_B - xG_A$.
- **Going further than xG and not only rewarding to move the ball in high-xG shooting positions, but also taking into consideration positions that can lead to high-xG shooting positions with high likelihood which are a threat.** We cannot take that from each point we are going to shoot, we have to think that in each point we can do two things: pass or shoot. We need to take into consideration this.

Therefore, when we talk of xT we can think that each position (x, y) has certain attributes:

- **Move probability** $m_{x,y}$ Probability that a player in (x, y) that it either passes or dribbles as their next action.
- **Shoot probability** $s_{x,y}$ Probability that a player in (x, y) shoots the ball.
- **Move transition matrix** $T_{x,y}$ Probability that the player once decides to move from (x, y) instead of shooting, goes to a nearby zone (x', y').
- **Goal probability** $g_{x,y}$ This basically represents what the xG metric is.

Lets assume that we have a field is a $M \times N$ grid, then we can formalize xT as follows:

Let $V_{x,y}$ be the value that xT assignes to zone (x, y). If we are in a position (x, y) we have the choices we mentioned earlier, shooting or moving. If we shoot, we know that the value is $g_{x,y}$. If we do not shoot, we know that we can either pass or move. However, we also have to decide something else: where do we move? There are $M \times N$ different zones where we can go. If we choose to go to another zone (x', y'), we know we will have a payoff of $V_{x',y'}$.

But we would need to compute all the payoff for the $M \times N$ positions. This is why the transition matrix $T_{x,y}$ was defined. So the pay off to move (x', y') is $T_{(x,y)\to(x',y')} \times V_{x',y'}$. Which for all the possible zones is calculated as follows:

$$\sum_{x'=1}^{M} \sum_{y'=1}^{N} T_{(x,y)\to(x',y')} \times V_{x',y'} \tag{24}$$

Now let's go back again to the beginning. We can shoot or move, we know now that shoot has a payoff of $g_{x,y}$ and moving has a payoff of $\sum_{z=1}^{M} \sum_{w=1}^{N} T_{(x,y)\to(x',y')} \times V_{x',y'}$. We know that there is a tendency to shoot $s_{x,y}$ and a tendency to move $m_{x,y}$. So the value of $V_{x,y}$ is the weight between these two probabilities. Which is the expected threat.

$$xT_{x,y} = (s_{x,y} \times g_{x,y}) + (m_{x,y} \times \sum_{x'=1}^{M} \sum_{y'=1}^{N} T_{(x,y)\to(x',y')} \times xT_{x',y'}) \tag{25}$$

However, the author emphasizes we cannot apply this formula directly, we need to first initialize the xT values. Otherwise we would be asking for a value which we do not know. This can be done by initializing $xT_{x,y} = 0$ and iterate until convergence. If we think about this iterative process, we can see that the first step is going to be by definition the xG model. Since we are scoring positions by how good they are if we shoot and not moving. From the second iteration, the model already understands that can move or pass, so we have a xT model which can consider "move then shoot". Which is one step further to the goal.

If we extend this, we will see that Specifically, $xT_{x,y}$ at iteration n represents the probability of scoring within the next n actions.

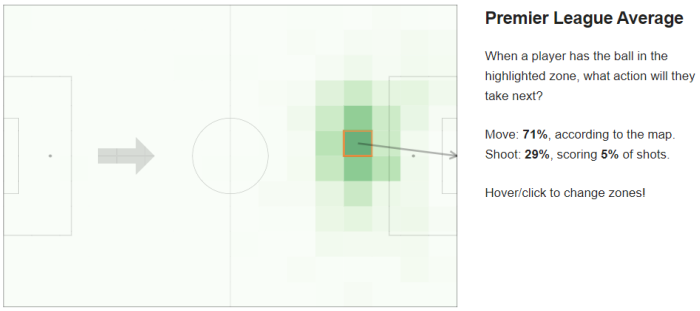


Figure 2: Expected threat, heat map[?]

2.4.4 Passes

Passes are very important in football and studies corroborate it[?]. Some facts about passes are:

- Teams that have a more centralized passing structure, score less goals than those which pass the ball more evenly (8%).
- The more a team passes the ball, the more likely to score will be. To showcase this, a team that passed on average 5 times a minute the ball, scored 20% more per season than those teams that only performed 3 passes per minute.
- Tenga, Ronglan, and Bahr (2010) provides an interesting framework to analyse passing effectiveness. The authors categorized passes into penetration, mixed, or non-penetrative only passes and analysed their effects of game performance.[47] Number of penetrative passes is correlated with scoring opportunities. A penetrative pass could be considered those passes that diminish the number of defenders between the ball carrier and the goal is positively correlated whereas a bigger number of defenders, it is negatively correlated with the idea.

There are many kind of networks which can we plotted. For instance, in the graphic below, we can see how passes are on teams with passing networks. Arrows show passes by players who completed 5 or more passes. Shapes connect average pass positions, show trends in movement. In the second image, we can see better the centrality and density of the passes from Portugal against Spain. On the third image, we can see the trend of pass per position.

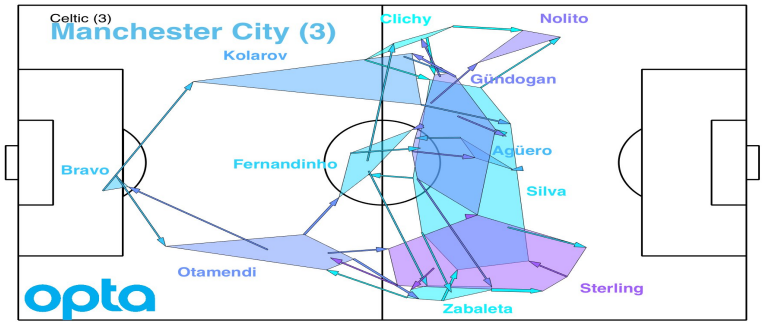


Figure 3: Man City passing networks vs Celtic[64]

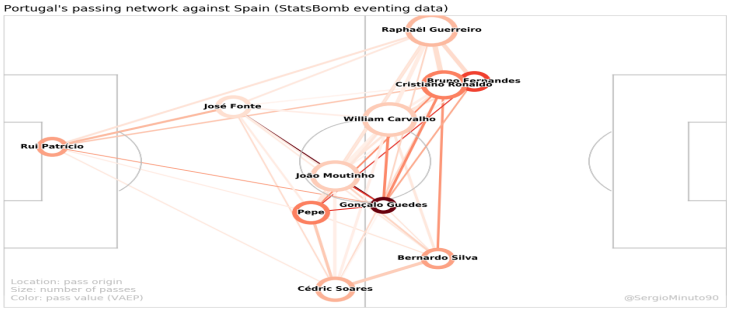


Figure 4: Passing Network[65]

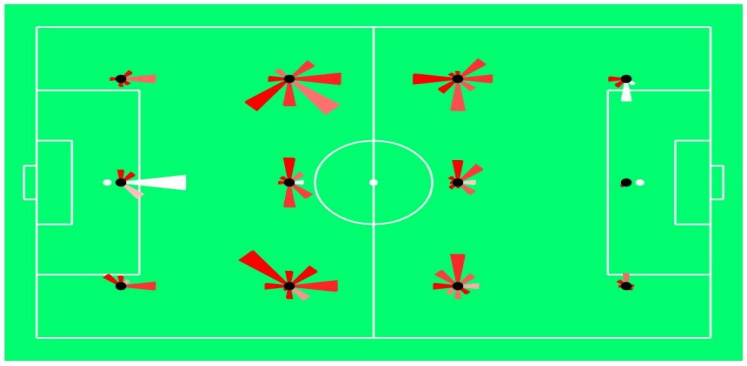
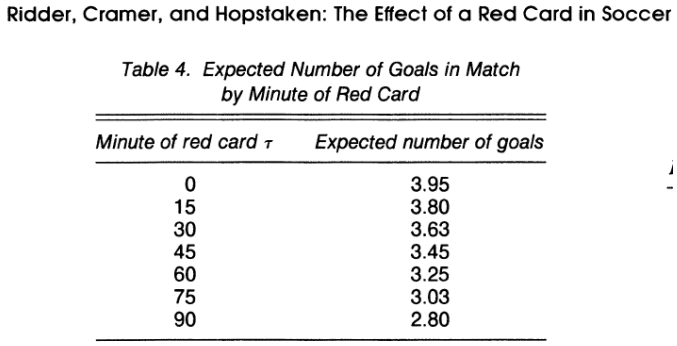


Figure 5: Polar Charts[66]

2.4.5 Red Cards

The effects of red cards on football were estimated by (Cramer and Hopstaken, 1994)[21]. Where basically concludes that there is evidence of the role of chance in the outcome. This implies that an early red card, increases the probability of victory substantially by the opposite team, the team that receives it decreases that chance of victory, and the chance of draw is relatively small.



Ridder, Cramer, and Hopstaken: The Effect of a Red Card in Soccer

Table 4. Expected Number of Goals in Match by Minute of Red Card

Minute of red card τ	*Expected number of goals*
0	3.95
15	3.80
30	3.63
45	3.45
60	3.25
75	3.03
90	2.80

Figure 6: Expected Number of Goals in a match by Minute of Red Card

Another interesting topic that the authors bring is that at some point of the match in a very dangerous attacking situation, the defender has the dilemma to do a foul -which could mean a red card- or to let it go, which would mean that could be a goal. As the paper states, the defender job is to minimize goals and there is a point of the match which the player should do the foul. However, this point depends on many factors such as their offensive capacities. The paper also concludes that weaker teams, are the ones that get more red cards for this same reason, since they have a stronger incentive.

2.4.6 Collective Knowledge

Evidence[56] suggests that tipsters do not process public information properly and that have little value (Forrest and Simmons, 2000; Pope and Peel, 1989).

Kampakis and Adamides, conducted a study over Sentiment Analysis of Twitter (2014) in order to predict the outcome of a match[60]. Schumaker, Jarmoszko and Labedz over wins and spreads with the same method (2015).[**?**].

They showed that crowdsourcing yielded better results than using domain experts, especially in wagering decisions. Godin et al. (2014) tackled the problem using a hybrid of collective knowledge and traditional statistical learning techniques.

2.4.7 Team studies

We can find different studies which regard to team, either in a psychological aspect or in a composition point of view. There are many of them (lots of results come always as part of some forecasting paper which aside from trying to improve the predictability, comes with a remark on some topic regarding to football)but I am going to mention those that I think are refreshing and that can give an idea of what has been studied.

- Ridder, Cramer, and Hopstaken (1994) showed that player dismissals have a negative effect on the match outcome for the teams affected.
- Audas, Dobson, and Goddard (2002) examined whether managerial change has any short-term impact on team results.
- Goddard and Asimakopoulos (2004) evaluated that important matches for a team (relegation, championship) make the players over-perform.
- Goddard and Asimakopoulos (2004) if a team loses a competition, we could think that will be able to be more focused on the remaining ones. However, the most normal scenario is that the team loses confidence and motivation, which affects their performances.
- Rue and Salvesen (2000) mentions that if a team is superior to another, the former tends to belittle the latter, which favors B. If the superiority is abysmal, the effect is the opposite. The latter, raises an inferiority complex and therefore, underperforms.
- We can see for instance, if a team is defending or not well by how correlated the arrow of the teammates are. Since studies found that teams that follow the same direction all together, defend better.
- Teams can be evaluated in terms of their weakest link. Which means that the overall quality of the team is the most important than extraordinary talented players on for instance, offensive roles.

2.5 Betting Odds

Shares are the quantities that the bookmakers pay for the different predictions available, which will be different between sports, matches and types of bets. This number goes from 1.01 to any number and represents how much you will win per 1 unit. There are many types of bets: simple, combined, system betting, live betting, etc. In our particular case, the betting odds provided are odds from different bookmakers which differ in the share that they provide. To exemplify how share work, we are going to put a practical example:

Figure 7: Bwin odds taken on 9$^{\text{th}}$ August 2020

In the following odds, we have that Atalanta has a share of 3.10. If we decide to gamble 5, we are going to win $5 \times 3.10 = 15.50$, which would give as a total profit of $15.50 - 5 = 10.50$. In case of win from Paris Sant-Germain we would win $5 \times 2.15 = 10.75$, which would end in 5.75 after subtracting the initial investment and similarly the same for the draw. In case that we bet for a result that is incorrect, we would earn anything and we would lose our investment of 5.

As we mentioned, we are not focusing in the profitability of the betting problem but rather, we are going to consider their forecasts as a reference. We are going to consider them because the bookmakers are the first ones interested in being accurate in their predictions. If we decompose what shares are, deep inside are the probabilities that the betting house things are giving to the event. So if we follow the example above, we can see that the probabilities by the previous matches according to Bwin are:

$$\text{Local Probability (LP)} = \frac{1}{3.10} \times 100 \approx 32.25\% \tag{26}$$

$$\text{Draw Probability (DP)} = \frac{1}{3.80} \times 100 \approx 26.31\% \tag{27}$$

$$\text{Away Probability (AP)} = \frac{1}{2.15} \times 100 \approx 46.51\% \tag{28}$$

$$\text{Total Probability } = LP + DP + AP \approx 104.97\% \tag{29}$$

As we can see, our total probability (booksum), is not over 100. This is because the bookmakers add what is called an overround which basically allows bookmakers to play safer with their money. But this is not a trivial problem because we cannot just assume that the overround is equally distributed across the different probabilities, it might follow a pattern. As specified in Clarke, Kovalchik & Ingram (2017)[] "Due to the overround, the implied probabilities form the bookmaker odds require an adjustment to obtain the actual expectations of bookmakers." (p. 45)

2.5.1 Bias

Aside from the overround which is the built-in margins that the bookmakers add, we can see another phenomenon which is biases. Biases imply that we cannot simply rely on what the bookmaker provides us since it may have a deviation from what should really be. There are three biases according to bib:constantinou-fenton-2013: Home-away bias, Most-likely bias and Favorite-longshot bias. However, most-likely and home-away bias, are favorite-longshot under different contexts.

Favorite-longshot bias This kind of bias is very common in horse racing but it is also present in association football as Cain, Law, & Peel (2000) showed. This bias basically occurs when an insufficient amount is bet upon the teams that are favoured to win an excessive amount is bet on the long shots, which distort the odds. This happens because there are scenarios where betting for a team that is very likely to win, only gives a small amount of money to the bettor which is not worthy the risk. For instance, if we have an odd of 1.08 and we invest €100, we are only going to obtain €8 whereas an unlikely win that has an odd of 35 will have a €100 investment €3500 in case that is correct. Which is very unlikely, but less likely is to play lottery and people still does.

2.5.2 Overround

We are going to put emphasis on four different methods for removing/incorporating the overround.

- **Additive Method:** In this model the overround is split evenly between the n outcomes. Thus, the true probability for the ith outcome, p_i, is

$$p_i = \pi_i - (\pi - 1)/n \text{ and } \pi_i = p_i + (\pi - 1)/n \tag{30}$$

  this method is rarely used in literature, this is due to the changes between the implied and adjusted probabilities for outsiders can be quite dramatic. Moreover, the additive method can produce negative probabilities for rank outsiders, which will happen whenever the ratio of the overround and implied probability is greater than the number of competitors, $(\pi - 1)/\pi_i > n$. The reverse process can also produce bookmaker probabilities greater than 1 for favourites.

- **Multiplicative Method:** Also known as the normalization method, this method splits the overround proportionally. Such that:

$$p_i = \pi_i \pi \quad \text{or} \quad \pi_i = \pi p_i \tag{31}$$

  It is the most commonly used method due to its simplicity. This method might be fine for many bets but it fails to take into consideration the favorite longshot bias (which are known to have a tendency to be overbet while favorites are underbet). Therefore, a greater proportion of overround needs to be removed/added to longshots than favorites. It also suffers from sometimes producing probabilities greater than 1 for favorites in the conversion from fair to bookmaker's probabilities.

- **Shin Method:** Shin proposed a correction method based on an assumed fraction z of knowledgeable punters. As given in, this results in

$$\pi_i = \sqrt{zp_i + (1-z)p_i^2} \quad \sum_{j=1}^{n} \sqrt{zp_j + (1-z)p_j^2} \quad \text{or} \quad p_i = \frac{\sqrt{\frac{z^2 + 4(1-z)\pi_j^2}{\pi - z}}}{2(1-z)}$$
$$\text{where } z = \frac{\sqrt{\frac{z^2 + 4(1-z)\pi_j^2}{n-2}}}{n-2} \tag{32}$$

This method is an iterative one. This method does not suffer from the favorite longshot bias, and has been shown to produce better predictive true probabilities than normalization method. However, it can be noted, that in the case of only two outcomes, the Shin method is equivalent to an additive method and it might suffer from the same problems than the aforementioned.

- **Power Method:** A natural extension of the additive method used in the additive approach (where probabilities are adjusted by a constant addition), and the multiplicative method used in normalization (where probabilities are adjusted by a constant multiplier), is to raise the probabilities to a constant power:

$$p_i = \pi_i^k \text{ or } \pi_i = \pi_i^{\frac{1}{k}} \tag{33}$$

This method assumes that bookmarker probabilities derived should satisfy the usual multiplicative law for independent events. In other words, the return to a punter from investing his winnings on subsequent events should be the same as a single investment on the joint event. When the n competitors are all equally likely, the value for k is calculated as, $k = \frac{\log(n)}{\log(\frac{n}{n})}$. However, in most cases iteration on k is necessary to ensure $\sum p_i = 1$, or the required booksum.

This method only produces probabilities on the [0, 1] range. This method also guarantees a great change to outsider probabilities than favorites. However, when compared to Shin it adjusts favorites and longshots more but middle-of-the-range priced teams less.

As [] explains, each method has its strong and weak points but in general trend, the power method is the one that performs the best.

2.6 Accuracy evaluation

Measuring the accuracy of a model is not a trivial question while it is a fundamental part to iterate and progress towards a better model. Very few papers talk about how to measure the accuracy of a football model since most of papers that tackle the forecasting issue, tend to use already established methods which are inadequate. However, Anthony C Constantinou, Norman E Fenton, et al. (2012)[31] introduced the Rank Probability Score (RPS) which is

an alternative to previous methods suggested which are shown not able to evaluate correctly the accuracy of a given model. What the RPS does different from other models is that it consider which scale we are working in. For instance, if we are playing the lottery and the numbers go from $\{1, 2, ..., 49\}$ the relevant scale is only *nominal* and if the correct number is 10 instead of 9, 9 is not closer than 25. They are both wrong and the scaling should be able to understand this. If we move to another domain for instance, the temperature of tomorrow in Celsius degree if we predict 34°C but actually tomorrow is 35°C, we must consider it closer than if our predict was 10°C since in this particular case the scale type is at least, ordinal (ranked). If we translate the RPS on a football domain, the crucial observation is that the set of observations { H, D, A } has to be considered an ordinal scale and not a nominal scale. Therefore, a draw (D) is close to a local win (H) than an away win (A) is to a local win (H). Which previous literature did not consider enough in their evaluations. Previous studies fall in 2 categories as Constantinou et al. (2012), mention in their paper:

- Those which consider only the prediction of the observed outcome (also known as local scoring rules). They are: Geometric Mean, Information Loss, and Maximum Log-Likelihood
- Those which consider the prediction of the observed as well as the unobserved outcomes. They are: Brier Score, Quadratic Loss function, and Binary decision.

Many of the state-of-the-art papers we covered, use these methods when evaluating their results. There are also other methods that are covered in literature which cannot provide a measure of accuracy for the prediction of a particular game. These methods are such as the error in cumulative points expected for a team after a number of matches, the RMS and Relative Rank Error of the final football league tables, and pair-wise comparisons between probabilities.

To informally showcase why RPS should be used as a validation method and not any other method, Constantinou et al. (2012), illustrates a set of benchmarks where:

Match	Model	p(H)	p(D)	p(A)	Result	Best model
1	α	1	0	0	H	α
	β	0.9	0.10	0		
2	α	0.8	0.10	0.10	H	α
	β	0.50	0.25	0.25		
3	α	0.35	0.30	0.35	D	α
	β	0.60	0.30	0.10		
4	α	0.60	0.25	0.15	H	α
	β	0.60	0.15	0.25		
5	α	0.57	0.33	0.10	H	α
	β	0.60	0.20	0.20		

Figure 8: Hypothetical forecasts for α and β, and results for matches 1 to 5.

- **Match 1:** (Taking account of perfect accuracy) Model α predicts the actual outcome with total certainty and hence must score better than any other, less perfect, predicted outcome.
- **Match 2:** (Taking account of predicted value of the observed outcome) Both models α and β assign the highest probability to the winning outcome H, with the remaining two outcomes evenly distributed. Since the observed value of α is higher than that of β, it must score higher.
- **Match 3:** (Taking account of distribution of the unobserved outcomes) Given that the observed outcome here is D, both of the unobserved outcomes are equally distanced from the observed one. Hence, the ordering concern here is eliminated. Still, a scoring rule must identify that model α is more accurate since its overall distribution of probabilities is more indicative of a draw than that of β (which strongly predicts a home win).
- **Match 4:** (Taking account of ordering when the set of unobserved outcomes are equal) Both models α and β assign the same probability to the winning outcome H. This time, however, they also assign the same probability values (but in a different order) to the unobserved outcomes (0.25 and 0.15). But, a scoring rule must identify that model α is more accurate since its overall distribution of probabilities is more

indicative of a home win.

- **Match 5:** (Taking account of overall distribution) Although α predicts the actual outcome H with a lower probability than β the distribution of α is more indicative of a home win than β. This match is the most controversial, but it is easily explained by considering a gambler who is confident that the home team will not lose, and so seeks a lay bet (meaning explicitly that the bet wins if the outcome is H or D). Assuming that α and β are forecasts presented by two different bookmakers, bookmaker α will pay less for the winning bet (this bookmaker considers that there is only 10% probability the home team will lose, as opposed to bookmaker β who considers it a 20% probability).

If we apply the benchmark to the previous methods described, we can see that actually none of the is able to predict correctly for all the 5 scenarios.

Match (Model)	Binary Decision Score	Brier Score	Geometric Mean Score	Information Loss Score	MLLE Score
1	✓	✓	✓	✓	✓
(α)	1	0	1	0	0
(β)	0	0.0200	0.9000	0.1520	-0.1054
2	×	✓	✓	✓	✓
(α)	1	0.0600	0.80	0.3219	-0.2231
(β)	1	0.3750	0.50	1	-0.6931
3	×	✓	×	×	×
(α)	0	0.7350	0.30	1.7369	-1.2039
(β)	0	0.8600	0.30	1.7369	-1.2039
4	×	×	×	×	×
(α)	1	0.2450	0.60	0.7369	-0.5108
(β)	1	0.2450	0.60	0.7369	-0.5108
5	×	× ×	× ×	× ×	× ×
(α)	1	0.3038	0.57	0.8110	-0.5621
(β)	1	0.0240	0.60	0.7369	-0.5108

Figure 9: Applying the specified scoring rules to each benchmark presented

Formally, the RPS which was introduced by Epstein (1969)[30], represents the difference between the cumulative distributions of forecasts and observations,

and the score is subject to a negative bias. Since the scoring is sensitive to the distance, the score penalty increases the more the cumulative distribution forecasted differs from the actual outcome.

$$RPS = \frac{1}{r-1}\sum_{i=1}^{r-1}(\sum_{j=1}^{i}(p_j - e_j))^2 \tag{34}$$

where r is the number of potential outcomes, p_j and e_j are the forecasts and observed outcomes at position j. The previous formula is based on the squared distance which for model 3 has some issues. The issues are that when that when H and A are equally distanced from D (order penalty eliminated), the squared measurement of probabilities results in scores that are higher (worse forecasts). The author states that would be wrong to assume that in a draw, victory of both teams would be at the same distance since normally one team is dominating and due to luck, ends up or not scoring, or getting a goal and the match ends as a draw. Reason why it suggests a corrected formula with the absolute distance instead of the squared distance.

$$RPS = \frac{1}{r-1}\sum_{i=1}^{r-1}(\sum_{j=1}^{i}(|p_j - e_j|))^2 \tag{35}$$

3 Methodology

We understand by methodology the general research strategy that outlines the way in which research is to be undertaken and, among other things, identifies the methods to be used in it[]. In our particular case, we are going to apply the Cross Industry Standard Process for Data Mining (CRISP) methodology since it is a robust and well-proven methodology. However, we are going to complement the CRISP evaluation phase, with some evaluation methods introduced by Andrew Ng in his course Structuring Machine Learning Projects from his Deep Learning Specialization[87] which we think that are particularly useful.

Figure 10: CRISP phases

We are going to drive our document by the CRISP phases (with few emphasis on the business understanding and Deployment part). However, we are going to compress the iterations and show the results of these. As we can see, there are 6 steps which are iterated until the project ends which are Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. Each phase has several sub-phases which can be found in the corresponding Appendix C. We are going to compare and evaluate our two models by the aforementioned RPS.

4 Project

4.1 Data Understanding

The main objective of this step is to search for different sources to work with and overall to chose a dataset to work with. We think that integrating several sources can help to the overall performance and might provide a more transversal understanding of the problem. For each of the sources we have found, we have considered the following factors:

- **Completeness of the dataset**.
- **Accessibility:** since one of the biggest problems when trying to approach is the lack of public data. There are some datasets available but the most important ones are offered as services.

On the Appendix A, we can find the results of the research where we list and explain briefly each source. But summarizing, data can be collected in different ways: soccer logs, video-tracking data and GPS. From the 3 sources, GPS data is not available publicly because it is data that clubs do not share. What we can find the most publicly are soccer logs which to showcase better it is basically listing actions that happen in a match: passes, shots, fouls, etc. However, we do not have access to the position of the players when they do not intervene on the game. For instance, we can have a player which is pressing exceptionally well, but this will not be registered. It will be listed who cuts the other team possession, who commits a foul, etc.

Having this into consideration, there are few sources available for free of soccer logs. We can find that many websites which provide information to the public, but are all using the same data provider which in most cases, it is Opta which is a membership platform and we should scrap them off. However, there are two datasets available on the internet with highly detailed information of each match which are from StatsBomb and Wyscout. The problem is that in total, there are not this many matches nor seasons to train with. Luckily, we found on Kaggle a dataset that is basically an open-source project which scraps different sources the crawler itself, it is currently outdated but the dataset is available. The dataset contains:

- Football matches, end of game statistics and in-game events. Extracted from football-data.mx[83].

- Player attributes from EA Sports FIFA video game extracted from SOFIFA[67].
- Betting odds from Football-data.co.uk[69].

As a result of this crawling process, the following information from European leagues (missing tournaments such as Champions/Europa Leagues), is available:

- +25000 matches.
- 10000 players.
- 11 countries from Europe.
- Seasons 2008-2016.
- Players and Team updates from FIFA, including weekly updates that the game performs on the players according to each week results.
- Betting Odds from 10 providers.
- Detailed match events for +10000 matches.

Which honestly, I think that is better than crawling a website. Considering that might be a problem when not used for personal purposes. Furthermore, as we were mentioning before, heterogeneous data sources can be an enrichment and in this case we have FIFA and teams information, we have also betting odds which are going to be our performance to beat. We will also use the advanced soccer-logs in some of the approaches in order to see if they suppose an improvement or not.

The data is organized in different tables:

Variable	**Description**
id	Integer that represents a league
country_id	Integer that represents a country
name	Name of the league

Table 3: League Table

Variable	Description
id	Integer that identifies a team
team_api_id	Integer id that corresponds to a team
team_fifa_api_id	Integer id that corresponds to a fifa team
team_long_name	Long name of the team
team_short_name	Short name of a team

Table 4: Team Table

Variable	Description
id	Integer that identifies a player
player_api_id	Player identifier
player_name	Name of the player
player_fifa_api_id	FIFA player ID
birthday	Day were the player was born YYYY-MM-DD
height	Height of the player in cms
weight	Weight of the player in lbs

Table 5: Player table

Variable	Description
id	Id of the given team
team_fifa_api_id	Team FIFA ID
team_api_id	Team api ID
date	Year where the attribute is added
buildUpPlaySpeed	The speed in which attacks are put together
buildUpPlaySpeedClass	Classification according to buildUpPlaySpeed values
buildUpPlayDribbling	Ability to dribble the opponents
buildUpPlayDribblingClass	Classification according to buildUpPlayDribbling values
buildUpPlayPassing	Affects passing distance and support from teammates
buildUpPlayPassingClass	Classification according to buildUpPlayPassing values
buildUpPlayPositioningClass	Classification according positioning
chanceCreationPassing	Amount of risk in pass decision and run support
chanceCreationPassingClass	Classification according to chanceCreationPassing values
chanceCreationCrossing	The tendency / frequency of crosses into the box
chanceCreationCrossingClass	Classification according to chanceCreationCrossing values
chanceCreationShooting	The tendency / frequency of shots taken
chanceCreationShootingClass	Classification according to chanceCreationShooting values
chanceCreationPositioningClass	A team's freedom of movement in the final third of the pitch
defencePressure	Affects how high up the pitch the team will start pressuring
defencePressureClass	Classification according to defencePressure values
defenceAggression	Affect the team's approach to tackling the ball possessor
defenceAggressionClass	Classification according to defenceAggression values
defenceTeamWidth	Affects how much the team will shift to the ball side
defenceTeamWidthClass	Classification according to defenceTeamWidth values

Table 6: Team attributes table, information from FIFA[]

Variable	Description
id	Table id
player_fifa_api_id	FIFA id
player_api_id	API id
date	Data of insertion
overall_rating	Overall rating of a player
potential	Defines potential of a player
preferred_foot	Defines preferred foot of a player
attacking_work_rate	Offensive rating of a player
defensive_work_rate	Defensive rating of a player
crossing	Evaluates crosses
finishing	Evaluates goal ability
heading_accuracy	Defines how good a player is at heading
short_passing	Evaluates short passing player abilities
volleys	Evaluates how good a player volleys are
dribbling	Evaluates a player dribbling ability
curve	Evaluates curve passes
free_kick_accuracy	Evaluates accuracy in free kicks
long_passing	Evaluates long passing
ball_control	Evaluates a player ball control
acceleration	Evaluates how fast a player accelerates
sprint_speed	Evaluates how fast a player sprints
agility	Evaluates how agile a player is
reactions	Evaluates how fast a player reacts
balance	Evaluates balance from a player
shot_power	Evaluates how strong a player kicks
jumping	Evaluates how good a player is at jumping
stamina	Evaluates how resistant a player is
strength	Evaluates how strong a player is
long_shots	Evaluates how good long shots are from a player
aggression	Evaluates how aggressive a player is
interceptions	Evaluates how well a player intercepts the ball
positioning	Evaluates how good a player positions on the field
vision	Evaluates vision of a player
penalties	Evaluates how good a player is at penalties
marking	Evaluates how good a player marks another
standing_tackle	Evaluates quality of standing tackles
sliding_tackle	Evaluates quality of sliding tackles
gk_diving	Diving abilities
gk_handling	Handling of the goalkeeper
gk_kicking	Kicking of the goalkeeper
gk_positioning	How good goalkeeper positioning is
gk_reflexes	Reflexes of the goalkeeper

Table 7: Player attributes table

Variable	Description
id	Match identifier
country_id	Country identifier
league_id	League identifier
season	Year in which the matches takes place
stage	Stage of the season the match occurred
date	Date in ISO 8601 format (YYYY-MM-DD)
match_api_id	Match api identifier
home_team_api_id	Home team id
away_team_api_id	away team id
home_team_goal	Number of goals scored by the local team
away_team_goal	Number of goals scored by the visiting team
home_player_X1-11	X coordinates of the local players
away_player_X1-11	X coordinates of the visitor
home_player_Y1-11	Y coordinates of the local players
away_player_Y1-11	y coordinates of the visitor
home_player_1-11	Id's of the players
away_player_1-11	Id's of the players
goal	XML of goals
shoton	XML of shots on
shotoff	XML of shots off
foulcommit	XML of fouls commited
card	XML of cards in the match
cross	XML of the crosses in the match
corner	XML containing information of the corners
possession	XML containing information of the possession
Betting odds (W/D/A)	For B365, BW, IW, LB, PS, WH, SJ, VC, GB, BS

Table 8: Team table

4.2 Model to beat

As we mentioned previously, the model that we are going to try to surpass is a model which comes from the odds data. This data is generated by the bookmakers and it is known that is a model that they generate and rely on.

4.2.1 Data used

For this particular problem, we have used a subset of the data we previously described which is:

Variable	Description
league	Name of the competition
country	Name of the country
season	Year where the season develops
stage	Number corresponding to the
date	Date where the match develops
match_api_id	Id of the match
home_team_api_id	Id of the local team
away_team_api_id	Id of the visitor team
local_goal	Goals from the local team
away_goal	Goals from the visitor team
bookmakers odds (W/D/A)	B365, BW, IW, LB, PS, WH, SJ, VC, GB, BS

Table 9: Dataset variables

There are some variables (e.g season/date/stage) which we are actually just using them for the joins since we are going to dismiss them. This model is

4.2.2 EDA

Regarding to the betting odds, for each of the bookmakers, we can see that we have the same number of missings in Home/Draw/Away which is the following:

Variable	Missings	Percentatge (%)
PS	14811	57.0
BS	11818	45.5
GB	11817	45.5
SJ	8882	34.2
IW	3459	13.3
LB	3423	13.2
VC	3411	13.1
WH	3408	13.1
BW	3404	13.1
B365	3387	13.0

Table 10: Betting odds B365 summaries

As we can see, we have a significative number of missings in PS, BS, GB and SJ. This is because this dataset comes from a scrapper and there is data that from the website from which extracts this information, is not available or directly the betting house did not provide a bet for that match. This affects our data in some ways as per example, that PS/BS/GB miss a lot of data and for instance, if we create a variable to see how accurate each betting house is assuming that the smallest odd is the one that we should take, we have the following results which seem directly related to the number of missing values. Regarding to the missings, we cannot impute the odds to those who have missings, as we will explain later. Moreover, if we pay more attention to the data, we can actually see that there are missings on our data are happening across the data. We could consider imputing those that have other bookmakers but imputing those that are missing across all our data does not make sense.

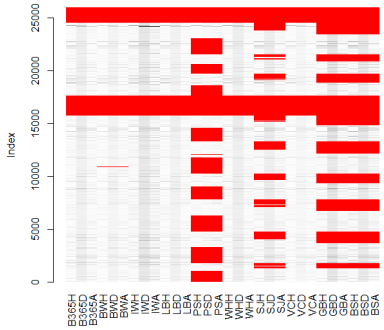


Figure 11: Missing values from Match table

Missing values aside, we can also take a look at how odds are distributed across bookmakers. We can see in the boxplot that there is a difference in the interquartile ranges across results Home/Draw/Away where in case of Draw, it is smaller than Home and significantly smaller than Away. However, from a same result point of view, they tend to be similar even that we can see that there are some bookmakers that have more variability than others (specially those with more missings have the most variability) and we can see that this trends, hold across results (figures 6-7)

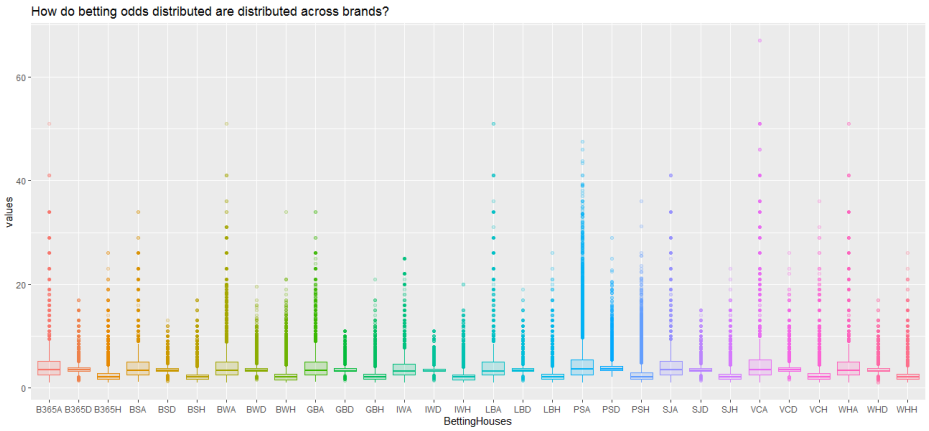


Figure 12: Odds across bookmakers and across results

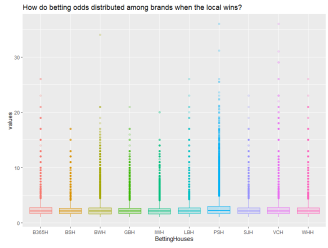


Figure 13: Wins

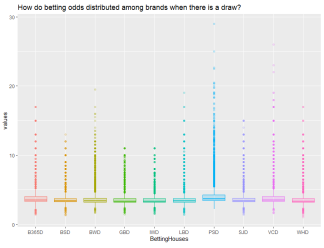


Figure 14: Draw

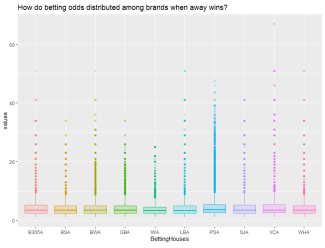


Figure 15: Away

If we pay attention to how this odds are when are correct and incorrect, we can see for instance case of B365, we can see that in general as we can see by the IQR, those bets that tend to miss more, are those that were already in general higher in revenue (betting house guesses right).

Type.	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Win + Correct	1.040	1.440	1.850	2.075	2.300	17.000
Draw + Correct	1.400	3.250	3.400	3.615	3.750	13.000
Away + Correct	1.080	1.910	2.700	3.139	3.600	34.000
Win + Incorrect	1.060	1.910	2.380	3.099	3.400	26.000
Draw + Incorrect	1.620	3.300	3.500	3.916	4.000	17.000
Away + Incorrect	1.140	2.880	3.800	5.278	6.000	51.000

Table 11: Betting odds B365 summaries

We can be surprised by how big some max values actually are, specially the ones that were guessed correctly. For instance, if we look at the data, we can find that for instance one of this matches is a F.C Barcelona vs Hércules C.F where F.C Barcelona actually lost 0-2 in the 2^{nd} stage of the 2010-2011 Spanish League. For this specifical result for instance, we can find that the other bookmakers, estimated as 20-ish the result (instead of 34).

However, it is normal that bookmakers are more prone to have higher bets on the loses of local and draws. It is very simple, there is an effect that is named local effect and we can see clearly that this happens in football. If we check our dataset, we can see that in our +25000 matches, the local wins in a 45.9% of the matches. We can also see that draws are the least frequent outcome of a football match with a 25.4% followed by a 28.7% of the local losing. So we can say that definitely this is something to take in consideration and so the bookmakers do and if we divide across all the different leagues that we have in our dataset, we can see that in all leagues we have similar proportions.

Figure 16: Win/Draw/Lose proportion across the leagues

4.2.3 Imputations

The process that we thought when considering imputing is to first ignore those matches that have missings in all betting columns and to consider imputing the remaining ones. Since all variables with missing values are continuous variables. There are different methods for imputing their values:

- **Deterministic methods** such as KNN and PCA imputation.
- **Stochastic methods** such as MICE and Random Forest.

It is known that the methods presented, have parameters. In this dataset, it was computationally very expensive to try them but we are going to do it anyways. For each method we can configure the following:

- In the case of KNN, we can configure the number of neighbors, the method that will be used to calculate the distance and if we scale or not the data.
- For the case of PCA imputation, we can configure similarly to KNN

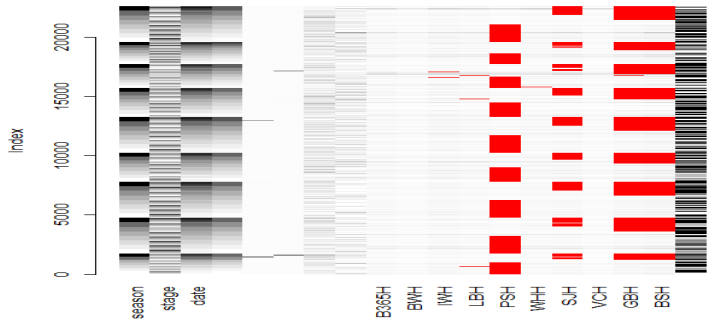


Figure 17: Data once without eliminated rows

the number of components to predict (ncp), if scaling or not and the method which can be Regularized or EM, there are other parameters which can be checked in imputePCA documentation[].

- In the case of MICE imputation, we can configure the number of multiple imputations (m), the method, number maximum of iterations, etc. This method is more complete than the previous one and therefore has more parameters.
- We are not going to use random forest, since it takes too much to compute, since the R library does not provide parallelism.

What we are going to do, is to once fixed some parameters such as scaling or not or which method to use, we are going to iterate thought the number of neighbors in KNN, the ncp for PCA and the m for MICE. What we did is to generate imputations for k, ncp and m for the range of $[2, 10]$.

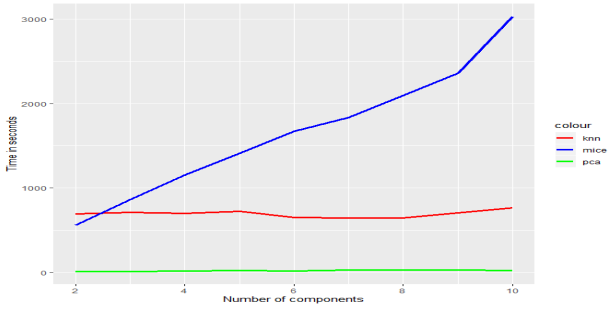


Figure 18: Imputation time

The problem that we ignored and found once generated the imputations, is that we cannot really impute the data. The bookmakers generate odds which add 1 or more. But imputations do not guarantee this. The implications that this has is that when taking the overround by any of the methods, this can give error which is quite an inconvenient. Moreover, considering that literature advices to use the average of odds, we chose to do the average off the odds once we removed the overround. We can do the average not taking into consideration missings with the parameter $na.rm = TRUE$

4.2.4 Outlier detection

When can we consider that these odds are outliers? Can we consider that for instance, a match between F.C Barcelona and Hércules C.F with an away odd of 34 is an outlier? Well in this particular case, we are not going to consider outliers. We thought of removing those odds that are out of the IQR ($Q1 - 1.5IQR$, $Q3 + 1.5IQR$) where $IQR = Q3 - Q1$. Also we thought of applying techniques such LOF. However, we decided not to do so. We are not going to do so, for the following reasons:

- The data we use, it is data that bookmakers offered in their platforms, is not incorrect. It has a reason behind, which is how unlikely the bookmakers thing it is which in this case, their prediction that Hércules wins, is of $\frac{1}{34} \times 100 = 2.94\%$. Furthermore, we are going to evaluate their model against ours so we should not interfere in their analysis.
- As we saw in our state-of-the-art, it is normally advised to bet not with maximum odds but with the average odds. Therefore, we are going to use the average odd for our model, this will blur this cases where a bookmaker proposes a "crazy odd".

4.2.5 Feature selection and engineering

As we mentioned earlier, we are only going to use the bookmakers information. However, this information is still quite exhaustive. From the state-of-the-art, we learned that normally it is advised to use the Average of the odds. We are first going to to perform a principal component analysis on the odds to see if they are as correlated as we expect they will.

What we can see and is somehow predictable is that odds of the same group, are highly related which we can also see when we perform a correlation anal-

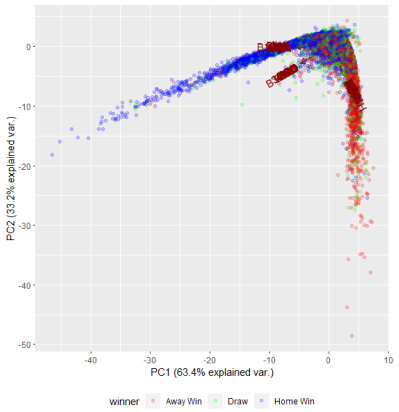


Figure 19: Missing values from Match table

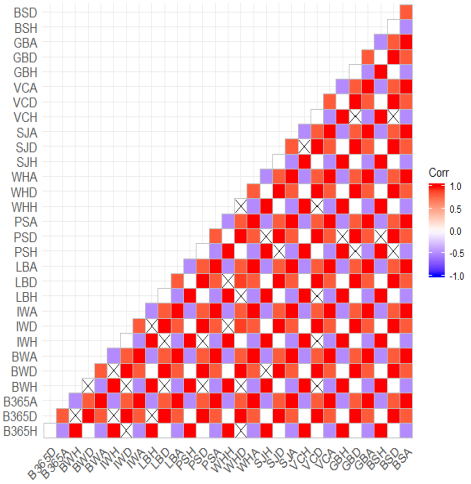


Figure 20: Betting odds "accuracy"

ysis. If pay attention to the PCA, we have mainly 3 groups which are the Home/Draw/Away grouped across the bookmakers. This simplifies a lot our problem since we are drastically reducing its dimensions. Since the PCA suggests so, what we are going to do, is to create 3 new variables: AvgH, AvgD, AvgA as literature suggests. This three variables are going to be the average of the Home/Draw/Away odds from all the bookmakers. However, as we introduced earlier there are several ways to calculate the overround. Therefore, we are going to calculate these averages per each of the 4 methods (additive, multiplicative, shin and power methods) and compare their performance to see if there is a significant difference when comparing those.

One could think that in order to improve the accuracy, we could add some additional variables such the goal difference. But as we mentioned before, if we are using already the odds from the bookmarkers, do we really need anything aside from this? Are not those variables already included in their models? In summarization, we are going to stick to the minimum information we need which are the odds and the final result of the match. In conclusion, the dataset we are going to work with is going to looks like the following:

Variable	**Description**
Avg Betting Odds (W/D/A)	Average odds across bookmakers

Table 12: Dataset variables

4.2.6 Evaluation

Since the RPS will help us to compare the bookmakers model with the model we will develop, we are now only going to compare across the different overround removal methods we mentioned earlier. As we can see by the table results, the results are very similar which is normal since these methods are simply working in how this overround is distributed.

Overround method	**Sum RPS**	**Mean RPS**	**[Min, Max]**
Additive	4437.892	0.1964017	(0.003618406, 0.8433418)
Cumulative	4431.697	0.1961275	(0.00157856, 0.9014916)
Shin	4432.541	0.1961649	(0.002032044, 0.8844678)
Power	4431.454	0.1961168	(0.000973493, 0.9083094)

Table 13: RPS evaluation (lower better)

We can see that the additive method performs way worse than the other methods. Between the 3 remaining methods, we can see that the Cumulative, Power and Shin are very similar. Considering that Shin method has a bigger computation cost, I do not think it is so worthy.

4.3 Our Model

4.3.1 Introduction

In this model what we aim to is to with summary data from matches, predict the final outcome of the given match. Previously, we worked only with the betting odds, but in this case we are going to work with event data such as the possession, shots, fouls committed, yellow and red cards, corners, penalties, etc. We also dispose of FIFA ratings for the players and teams information. Summarizing, we are going to use the data introduced in 4.1.

4.3.2 Data used

As in the previous case, we are going to work with the European Soccer Database, but in this particular case we have some problems. As we mentioned, this dataset works by scrapping from some websites and in this particular case we have around 50% of data of the 25000 matches regarding to values such as possession, fouls, corners... which we will refer as event data. We decided that we are still going to use it because basically the data that we have is data from some leagues with almost no missings and those leagues are the most important ones (Italy/Spain/Germany/England). Even this affects the size of our dataset, we still have a significant number of matches (11784).

Country	**% Available**
England	100 %
France	66.6 %
Germany	100 %
Italy	99.9 %
Netherlands	21.7 %
Poland	0.365 %
Scotland	0.713 %
Spain	100.0 %
Switzerland	7.03 %
Belgium	0 %
Portugal	0%

Table 14: Event data availability

The incident data that we can find in the dataset (goal/shoton/shotoff/foul-

commit/card/cross/corner/possession) is far from being definitive. It needs a lot of processing since the format that appears per row has an XML format. In the following XML, we can see an example of one goal from a given match.

```
<goal>
...
  <value>
     <comment>n</comment>
     <stats>
       <goals>1</goals>
       <shoton>1</shoton>
     </stats>
     <elapsed_plus>3</elapsed_plus>
     <event_incident_typefk>393</event_incident_typefk>
     <elapsed>90</elapsed>
     <player2>46353</player2>
     <subtype>shot</subtype>
     <player1>40755</player1>
     <sortorder>3</sortorder>
     <team>8586</team>
     <id>1566459</id>
     <n>275</n>
     <type>goal</type>
     <goal_type>n</goal_type>
  </value>
...
</goal>
```

As we can see there is a lot of information that we dispose of but at the same time, there is no documentation which makes it a tedious task. In any case what we had to do in order to be able to study the data that we dispose, has been converting these XML's in a new table.

Variable	Description
id	Id of event
type	Type of play, factor
subtype1	Subtype of the type
subtype2	Subtype of the type
player1	Player that performs the action
player2	Player that contributes
team	ID corresponding to the team
lon	longitude on the field
lat	latitude on the field
elapsed	Time from 0-90 of a match
elapsed_plus	Extra time
half	First or second half
half_elapsed	Result from elapsed + elapsed plus

Table 15: Incident data transformation

4.3.3 EDA

In this section we are going to perform the exploration data analysis of our dataset. Since the dataset it is very extensive, we are possibly not going to be able to document all we can notice but, what we have seen will be reflected on the selection and engineering phases. We are going to split the analysis process in 4 different parts:

- **Leagues:** Are some leagues different than others?
- **Teams and team attributes analysis** Can we extract insightful features from this tables?
- **Players and player attributes data** Can FIFA data provide us an insight about how good really teams and players are?
- **Event data** What can we learn from it? Is the data detailed enough? What can we built with it

4.3.3.1 Leagues

One of the first variables we have considered is league_id and the question is: are all leagues equal? We have seen for instance in the state-of-the-art

studies where league was needed. Furthermore, according to the Dolores system results[9], we should see that there is a difference between leagues which will make difficult to be able to generalise a single model.

First and foremost, we can see that not all the countries have the same number of teams which results in an imbalance on the number of matches which from a yearly perspective, can represent having up to 50% less of matches.

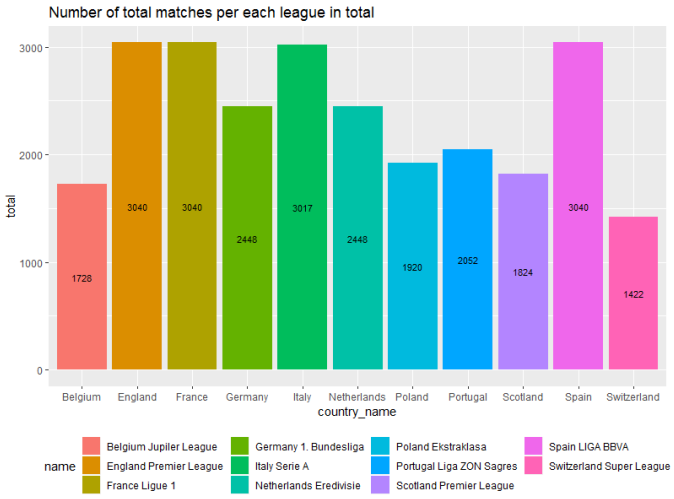


Figure 21: Total number of matches in our dataset across leagues

If we evaluate the different leagues one thing that can be an indicator that there are different play styles among different leagues could be the goals. We have calculated the average number of goals per season that each league has scored. First we have divided them in two groups in order to be able to visualize them better:

- **"Major" leagues:** where we include English, French, German, Italian and Spanish leagues.
- **"Minor" leagues:** where we include Belgium, Poland, Scotland, Netherlands, Portugal, Switzerland leagues.

If we pay attention to the average goals across leagues, we can see that there are some countries that look alike. Moreover, we have also analysed

the number of points that a team does at home per away point were we can observe that some leagues are more similar than others.

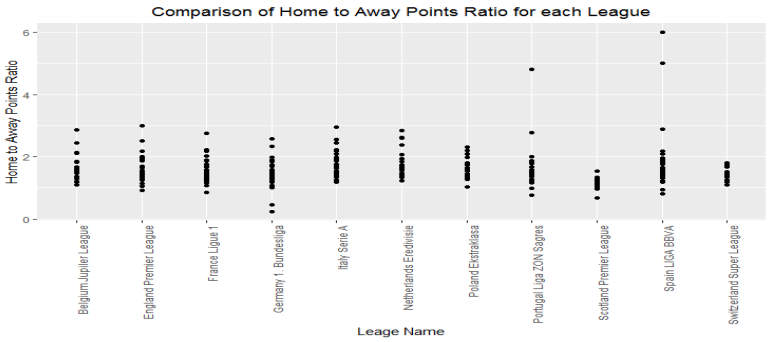


Figure 22: Ratio of Home/Away points across leagues

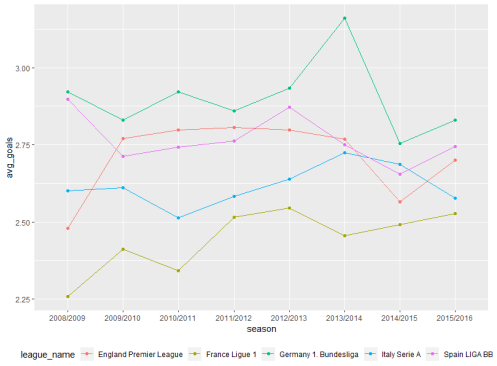


Figure 23: Average Goals across the major leagues

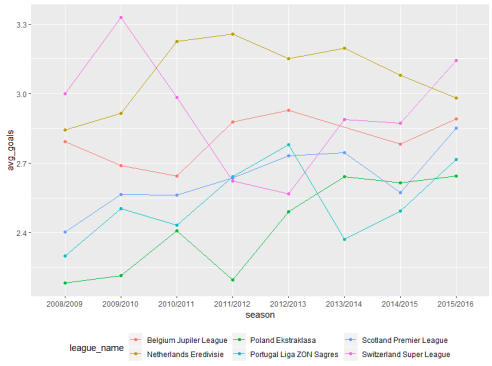


Figure 24: Average Goals across the minor leagues

4.3.3.2 Matches

Matches is the biggest table from the database that we are using. From the original table, we have extracted the event data as we mentioned. Therefore, in the table we have the following attributes left:

Variable	Description
id	Match identifier
country_id	Country identifier
league_id	League identifier
season	Year in which the matches takes place
stage	Stage of the season the match occurred
date	Date in ISO 8601 format (YYYY-MM-DD)
match_api_id	Match api identifier
home_team_api_id	Home team id
away_team_api_id	away team id
home_team_goal	Number of goals scored by the local team
away_team_goal	Number of goals scored by the visiting team
home_player_X1-11	X coordinates of the local players
away_player_X1-11	X coordinates of the visitor
home_player_Y1-11	Y coordinates of the local players
away_player_Y1-11	y coordinates of the visitor
home_player_1-11	Id's of the players
away_player_1-11	Id's of the players

Table 16: Team table

As we can see, we are not including the XML's mentioned before nor the betting data which leaves us with a significantly minor number of features to study on this phase.

However, there is still a lot to see on this table. We can see that the coordinates of the players is also information that as it is, is not very insightful but which could be used to bring insight in our analysis. What we have done is analysed the 44 columns regarding to home and away players location (home_player_X1-X11, home_player_Y1-Y11, away_player_X1-X11, away_player_Y1-Y11). If we see the data and we research on data from the corresponding matches, we can see that the coordinates X(1–9) and Y (1, 3, 5–11) position the players in the pitch according to their role. The role

of a player is mostly defined by the Y coordinate. We have assigned each player a role according to this rules: assigned to each player according to the following:

- If $Y = 1$ we are going to assign the role of goalkeeper.
- If $Y = 3$ we are going to assing the defender role.
- If $5 \leq Y$ and $Y \leq 9$ is a midfielder.
- If $10 \leq Y$ and $Y \leq 11$ forward.

For instance, we can see which are the most common formations across the years and across leagues for home and away scenarios. We can see that there is variability across years. Also across leagues it changes, there are trends in formations. Moreover, if we recall a xG models, there was a component league, because there were some features that depended of the league which the model could not observe.

Figure 25: Formations 2011

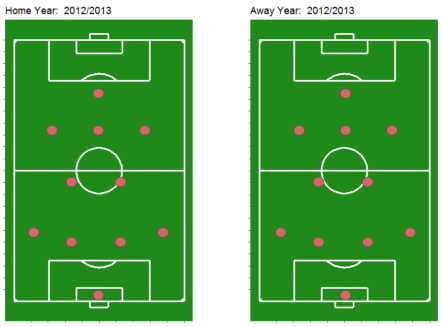


Figure 26: Formation 2012

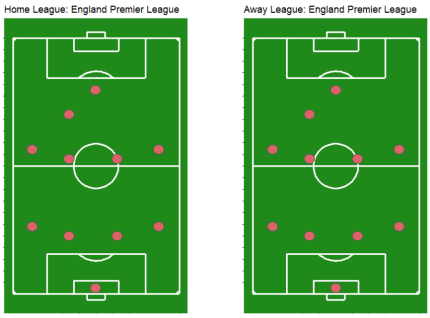


Figure 27: Formations England

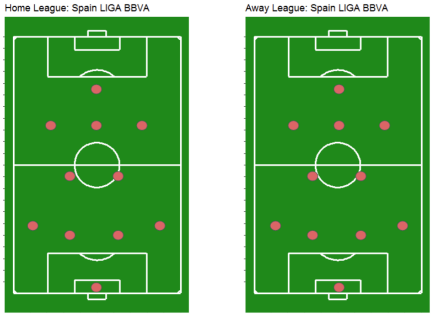


Figure 28: Formations Spain

Since this variables are very consuming in terms of space, we are going to create a variable named formation which will be the local formation and away formation e.g 4-3-1-2. Moreover, aside form their formation, we can extract which players are playing the game on a given match since we have the $home\_player\_1-11$ and $away\_player\_1-11$ information which contains their ID's.

4.3.3.3 Players

If we take a look at the player attributes we find more than 40 qualities that have a rating per player such as: agility, stamina, vision, strength, etc. We took a look at all of them and we can see that there are many that are highly correlated.

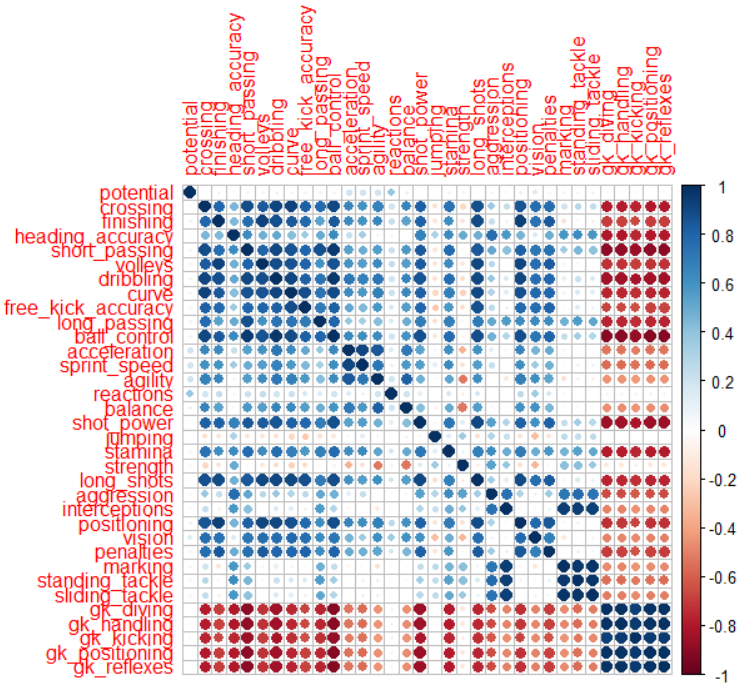


Figure 29: Correlation matrix

We can see that there are some attributes that correspond to goalkeepers and some others that correspond to other roles and that in general, there seems

to be a high correlation between most of values. In order to see more clearly how correlated each variable is, we have analysed the correlation per role. From this individual analysis we have seen that most goalkeeper attributes and most defense attributes are very correlated whereas midfield and attacker attributes, have a bigger number of attributes and there is less correlation among them.

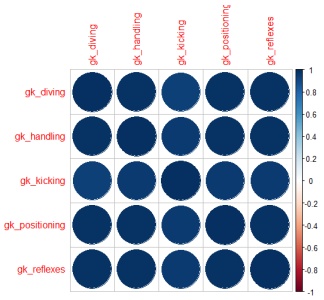


Figure 30: Correlation Goalkeeper

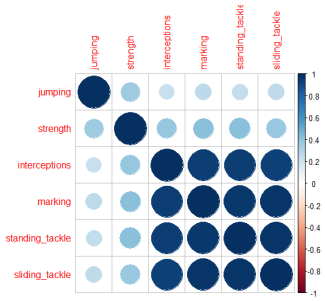


Figure 31: Correlation defense

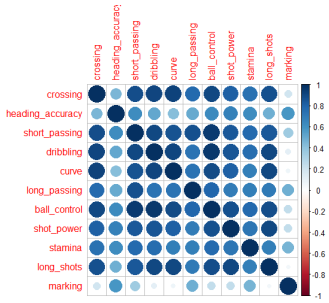


Figure 32: Correlation midfielder

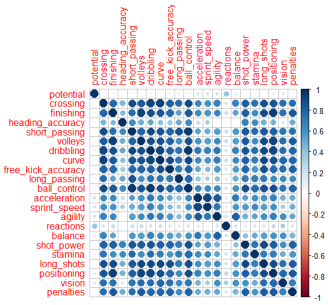


Figure 33: Correlation offense

In order to see which variables are correlated in a general way, we have also performed a PCA which shows that

- Goalkeepers attributes are completely opposite to player attributes.
- Among the players variables, we can see two directions. The ones that go downward correspond to attributes from defensive players whereas the ones that aim upwards, go from strikers to a more middle field position the closer to the center.

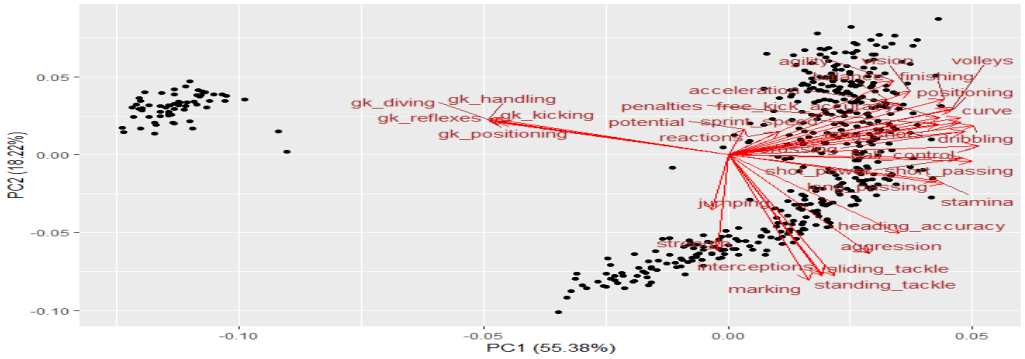


Figure 34: Principal component analysis

4.3.3.4 Teams

Team and its attributes mostly describe the play style of a given team. If we pay attention to our variables, we can see that half of our variables are actually a category which classifies our numerical variable. Therefore, what we did is to remove those categorical variables which are highly correlated to our numerical variables and just left the numerical ones. Since we have the years in which each attribute was collected, we can see if a team for instance, improves over time or does not. Moreover, in the same fashion than with the teams and its attributes, we can see the correlation between their variables where we see a minor correlation among variables.

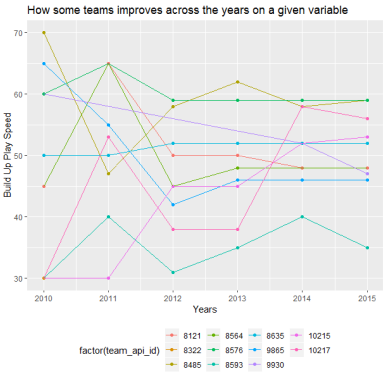


Figure 35: Build up Play speed improvement over the years

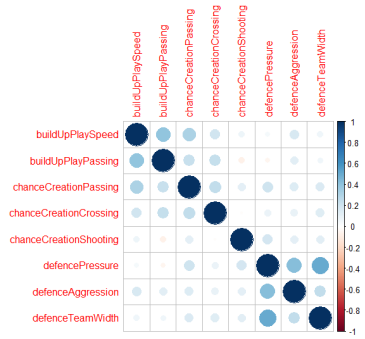


Figure 36: Correlation among team variables

4.3.3.5 Event Data

Regarding at the event data, as we mentioned before, we have the following events:

- **Goals:** Are those events that result in goal. We have many types of goals as we will introduce but the common factor is that they are goals.
- **Shots On:** Shots On are those shots that the goalkeeper has to do something about given their danger.
- **Shots Off:** Shots Off are shots that do not require of the goalkeeper intervention. Since they would not represent a goal.
- **Corners:** Corners realized during a match.
- **Crosses:** A cross is a medium- to-long-range pass from a wide area of the field towards the centre of the field near the opponent's goal.[**?**]
- **Fouls:** Actions that a player did on the field that are considered a foul and can be or not given a yellow or red card.
- **Throw-in:** Is a method of restarting play in a game of football (or soccer) when the ball has exited the side of the field of play.[**?**]

In general, most of this data has locations. However, we can find a quite high number of them not having longitude and latitude which is a shame.

Goals

If we study how the incidence of goals, is distributed and how affects win or lose, we can see first how goals are distributed where we can see that actually there are more goals on the second part 56% of goals on the second half.

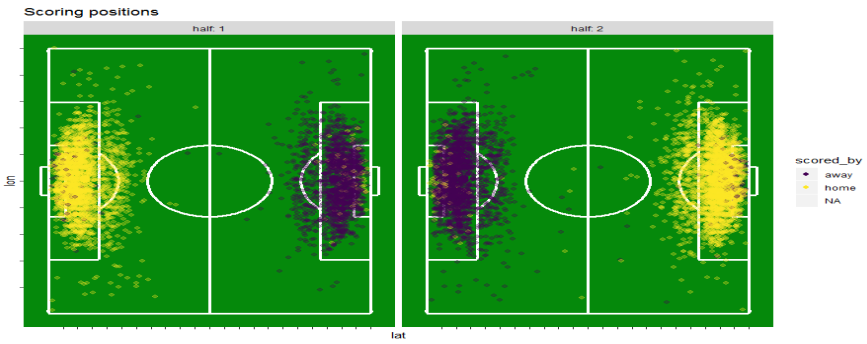


Figure 37: Goals plot

As we can see most of goals are close to the area and quite central and the farther from this area, the least goals we have in our plot.

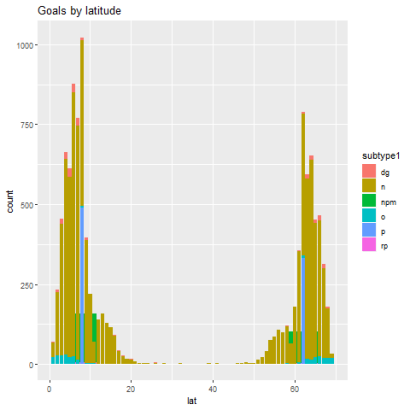


Figure 38: Goals by Latitude

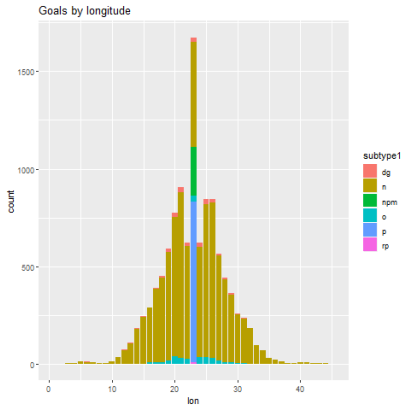


Figure 39: Goals by Longitude

If we look at the timeline with detail, we can see that the beginnings of parts, tend to be very calm, specially on the extra times, where teams tend to be exhausted and are looking not to get scored. We can also see that the local factor is a real factor to consider as we mentioned previously and here again we can see that at every single moment of the match, is more likely that the local will score than the visitor.

We can even see in the data of a match that if we compare the results from a match on half time vs end time, the most common output is that the team that is winning half time, wins the match with a 63% of probability. We

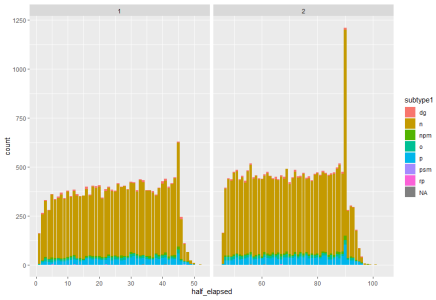


Figure 40: Goals over time

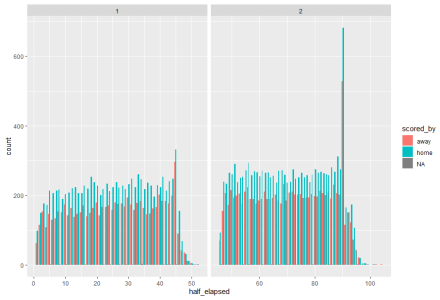


Figure 41: H vs A goals over time

see that when there is a draw, almost every scenario is possible. However, scenarios such as comebacks were the local is losing by the first half and ends winning or vice-versa, are very unlikely being only 6.53% of the matches. This is due to the low scoring nature of football without a doubt. On the density plot located right hand side, we can see that on average the local team scores 1.72 goals versus 1.29 goals of the visitor.

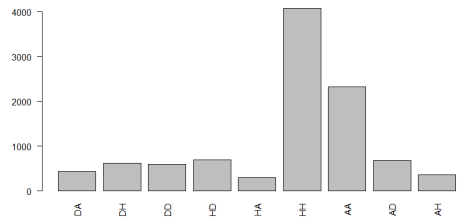


Figure 42: Goals over time

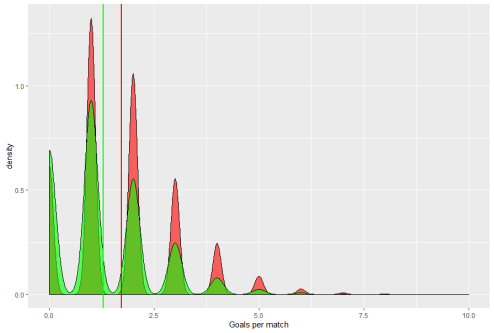


Figure 43: H vs A goals over time

Finally but not less important, if we apply a Poisson distribution to the number of goals, we can see that Maher was correct in assuming a Poisson distribution over goals.

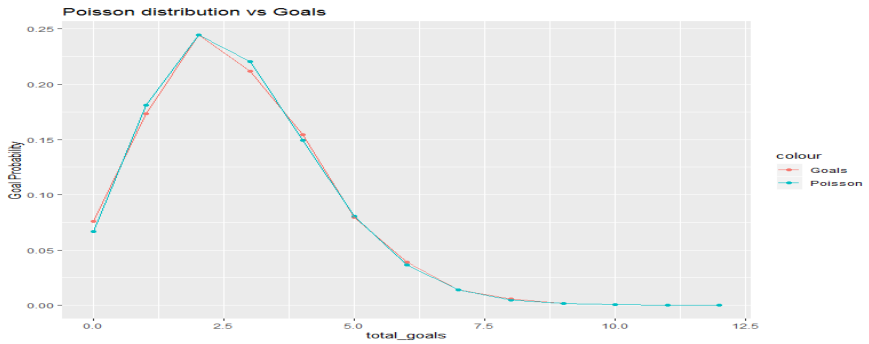


Figure 44: Goals following a poisson distribution

We can also see that there are multiple types of goal: "dg", "n", "npm", "o", "p", "psm" and "rp". What do this stand for? Well it is a difficult question actually but it seems that not all the goals are actually goals. We can find that the only types of goals are n (normal), p (penalty) and o (own goals). On the contrary, the other goals are not goals at all, are missed or saved goals penalties (npm psm), disallowed goals (dg), etc. However, as we can see below, the number of shots is very small and we decided that we will not remove them because they might affect the data in cases like xG modelling.

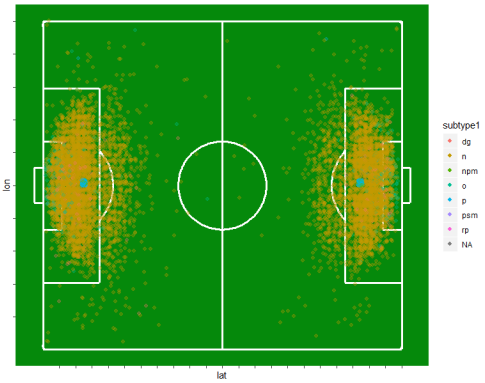


Figure 45: All goals

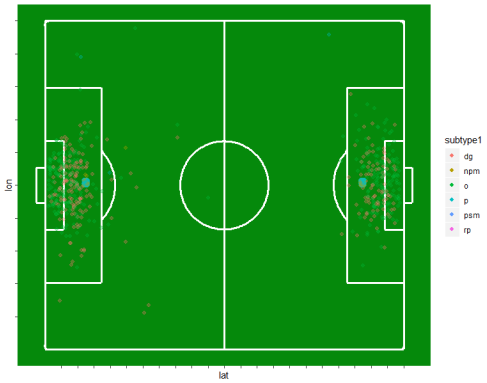


Figure 46: All goals minus "n" type

Shots On

As we previously mentioned, shots on are by definition those shots that go between the posts. If we take a look at the number of shots on that we have

in our dataset, we can see that the number is way more significant than goals and also we can see that the spread is bigger, which explains why is not a goal in part.

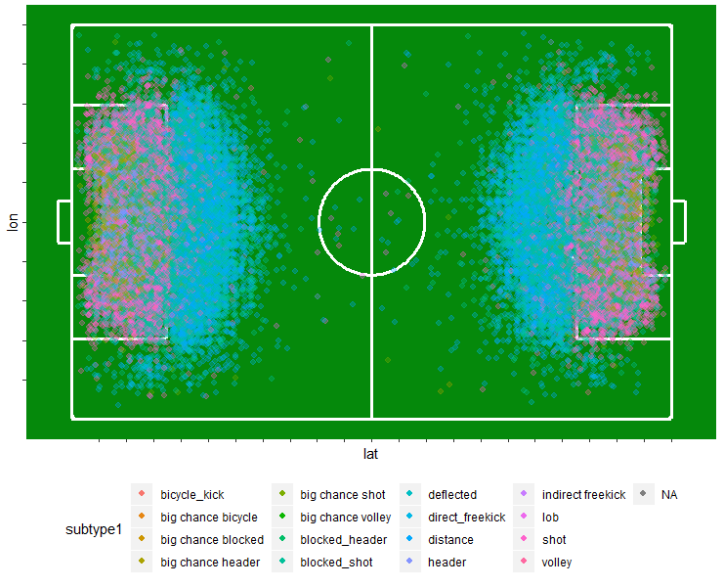


Figure 47: Shots On location

If we analyse the shots by longitude and latitude, and compare it with goals, we see latitudes are more in the center with a double spike and regarding to longitude we can see a thicker spread with actually less shots on the center than from the sides.

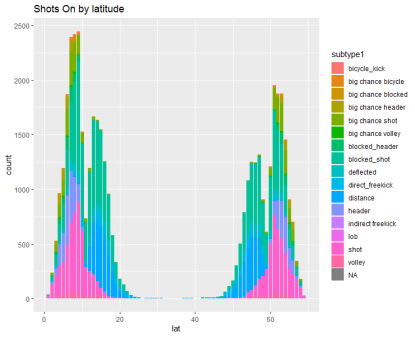


Figure 48: Shots on by Latitude

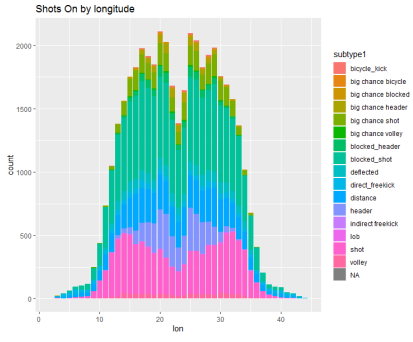


Figure 49: Shots on Longitude

Shots Off

Similarly than as shots On, Shots Off are the shots that in this case, did not got between he posts. If we see again the location of the Shots off, we can see that the image is quite similar to shots on but again we see a bigger spread, particularly near the middle of the field.

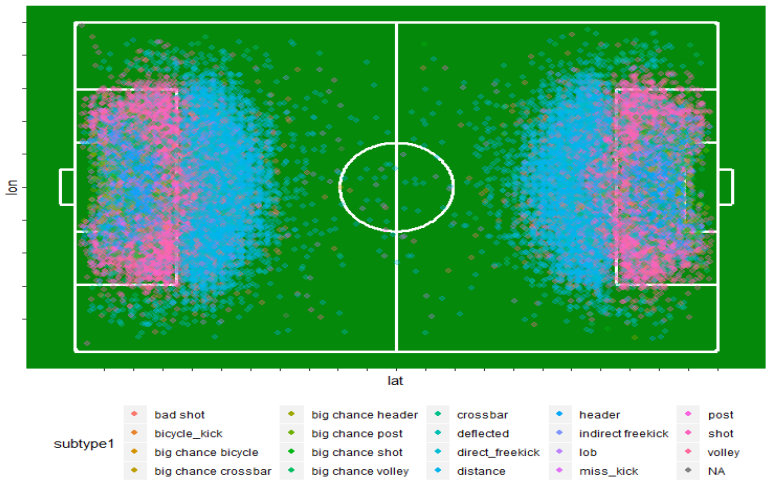


Figure 50: Shots Off location

This can be more clearly seen if we plot the shots on and goals by latitude and longitude, we can see that the spread is surprisingly similar to shots on.

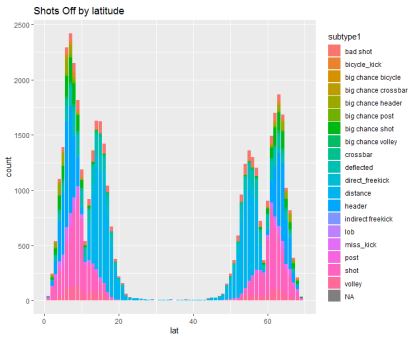


Figure 51: Goals by Latitude

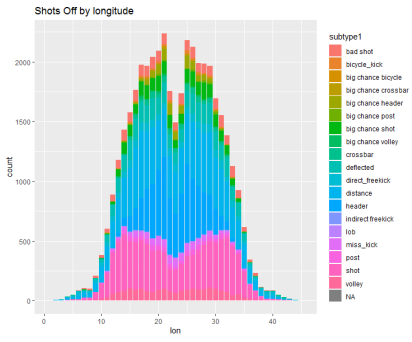


Figure 52: Goals by Longitude

Goals, Shots On and Shots Off

As we can see, there is a difference in the sharpness of locations when scoring or not. Shots that are more central and closer, to the goal. However, we can for instance ask ourselves. Why some shots are off and others on? The spreads are honestly, very similar. Well, as we have seen in our state-of-the-art, the contextual information of a shot is very important. With the data that we have, we cannot actually judge if the player for instance did a bad shot due to a very intense pressure. Indeed, if we build an expected goals model and we see the accuracy of a given shot, shots by rule, have a very low score rate.

Another insight and limitation on the industry is that by default, GPS data is not found. And this particular dataset, has this limitation. However, from 2018 StatsBomb collects information (named pressure events) on the players at 5 yards of the shooters position at the time that a shot is taken so that expected goals have less bias and more insight on the actual plays. One example of this shot freezes is the following image

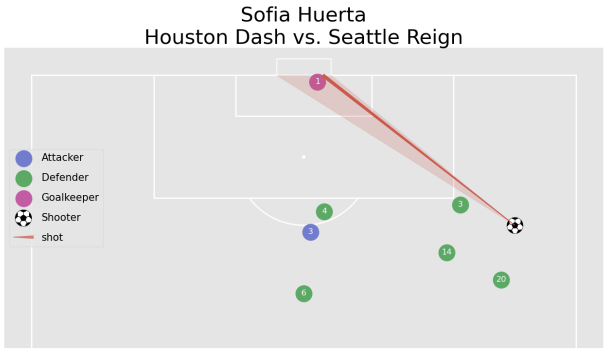


Figure 53: Freeze Frame, extracted from Houston Dash vs. Seattle Reign

Corners

"A corner kick is the method of restarting play in a game of association football when the ball goes out of play over the goal line, without a goal being scored and having last been touched by a member of the defending

team. The kick is taken from the corner of the field of play nearest to where it went out".[?]

In order to see if corners do really make a difference in games, we are going to study their effect. First we can see that the corners data has some outliers which we are going to remove:

Figure 54: Corners from the dataset

First we can see that there are short corners, this corners we cannot track actually what happens with them. Do they convert in successful crosses? We can not know with our data. We know that they are a threat but not more than this. Then if we focus on the other crosses, the only think that we can actually think of is from shots off for instance we have big chance headers, big chance volley, header, volley and from the shots on we can add big chance blocked. From goals we do not really have the information. We can guess that those shots are from corners but that would be an inference. Since we do not have a clear idea, we will let the modelling part to decide if it is an important feature to keep or it is not.

Crosses

As we mentioned, crosses are medium- to-long-range pass from a wide area towards the centre of the field near the opponent's goal. In this particular case, our crosses do not have the player that receives the cross, only the one that does it which does not provide much of an insight, since we cannot really follow the play. If we pay attention to the field, we can see that crosses are

pretty much done from anywhere, specially on the sides, since they are a part of wingers job.

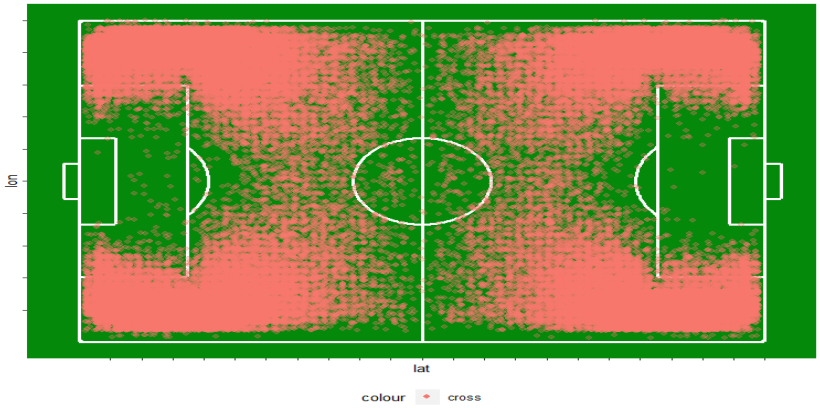


Figure 55: Corners from the dataset

If we evaluate if the team that wins performs more crosses, we can see that teams with more crosses, tend to be superior and therefore, they have more tendency to win. So we are going to incorporate it to our model to evaluate.

Throw-In

Throw-In are completely un-insightful. It happens with some data across the events data which lacks of lon and lat coordinates but in this particular case, we do not have any single one. There are around 8000 of them (versus 235000 crosses), and all of them are labeled as crosses because by definition they quite are. However, we are going to analyse if they can represent a measure for instance of how much a team is pressing and if it is associated with more goals.

If we analyse the throw-in's, we do not actually see that they are significant on the game. If we think in football mentality, normally if you lose a lot the ball you get more throw-in's, is not necessarily good, it also does not imply you are inferior since you could be pressing all time and therefore you could get many throw-in.

Fouls Committed and Cards

We are going to first going to focus in the Cards that players get from fouls and their effects. As its known, not all fouls receive a yellow or red card. Therefore, a foul can be punished with a yellow or red card, but not all yellow and red cards are necessarily caused by a foul, they can be given due to a misconduct.

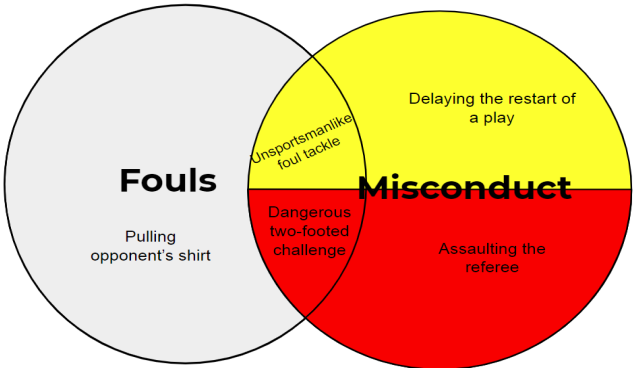


Figure 56: Simplified Venn diagram of fouls

If we pay attention to the cards, they can be given for many reasons but can only be of these three kinds:

- **Yellow card:** Which . On average we have 4,21 yellow cards per game which constitutes a 94.36% of the total cards given.
- **Second yellow card:** On average we see 0.12 second yellow cards per game.
- **Red card:** On average we see 0.12 red cards per game.

So as we can see, the biggest class among the 3 are yellow cards which are common on a football match. If we pay attention to when these cards are given, we can see that per half, normally there are more cards on the end of that half and that there are significantly more cards on the second half with 63.17% of the yellow cards, 85.83% of the second yellow cards and a 71% of the red cards. So we can see that the second half has an important weight on this factor.

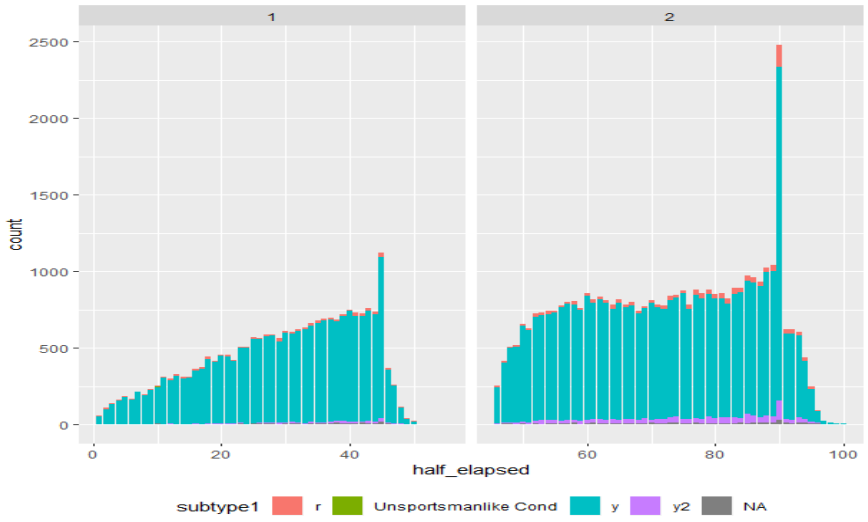


Figure 57: Principal component analysis

If we evaluate each of this cards why it is given we can find the following information at the subtype of each card, where we can see that some actions prevail in all the types of cards but specially we can see that second yellow card is normally given to a player that commited a serious foul and that a red card is to someone that was aggressive either verbally or in the field (in-game or not). The mean time to get a card is minute 66.

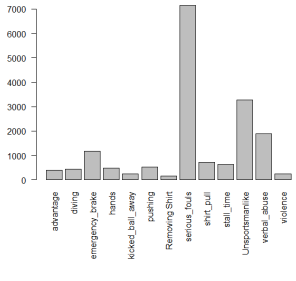


Figure 58: Yellow cards

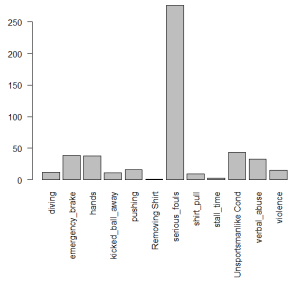


Figure 59: 2nd Yellow

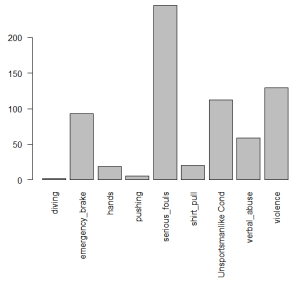


Figure 60: Red cards

If we look at which teams are the ones that play dirtier, we can see that the top 15 teams with the most cards, are all Spanish or Italian which makes

sense because they are the leagues with most matches and we did not have also all our leagues for the event data as we stated previously.

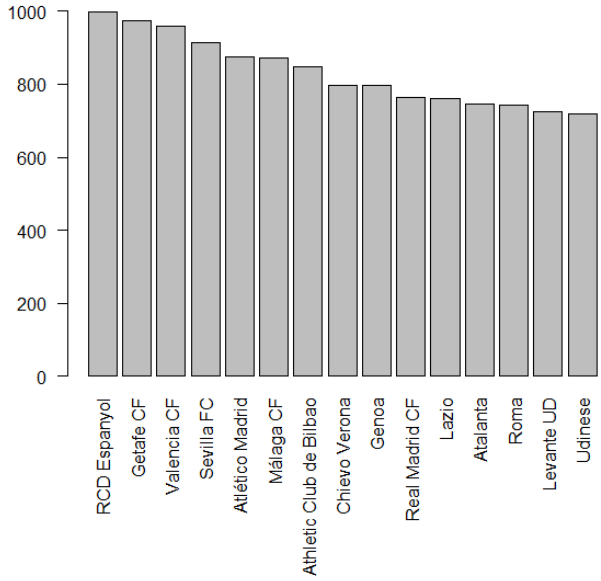


Figure 61: Top 15 teams with the most cards received

If we pay attention to faults, we have the latitude and longitude of most of them so that we can see how each event distributes across the field which we could filter per each subtype. However, what we are really interested in this particular case is, how this faults relate with the yellow cards and if it does affect the outcome of a match. As we can see in the below image, we have many faults but of course, not all of those are cards, on the right hand side, we can see which of those, are actually yellow cards, and from the below plots we can see how uncommon second yellow cards and red cards are but for instance, we see that second yellow cards are all across the map while red cards, are way closer to dangerous points.

We saw in our state-of-the-art the effects of red cards but, do this effects hold for our data? The answer is yes, we still see that it has a considerable effect on the match outcome and the earlier it is given, the more impact it has.

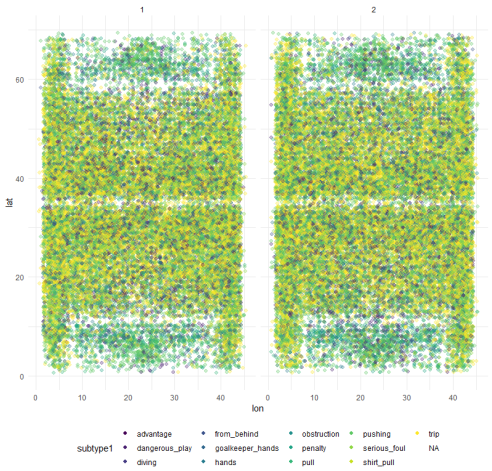


Figure 62: Faults by halves

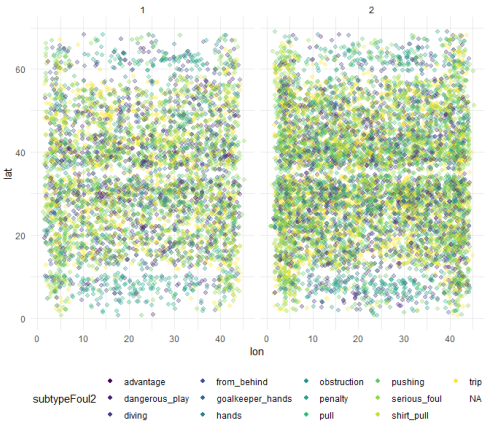


Figure 63: Faults with yellow cards

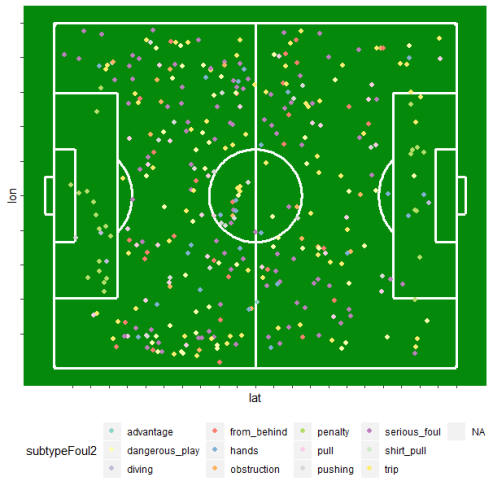


Figure 64: Second Yellow Cards

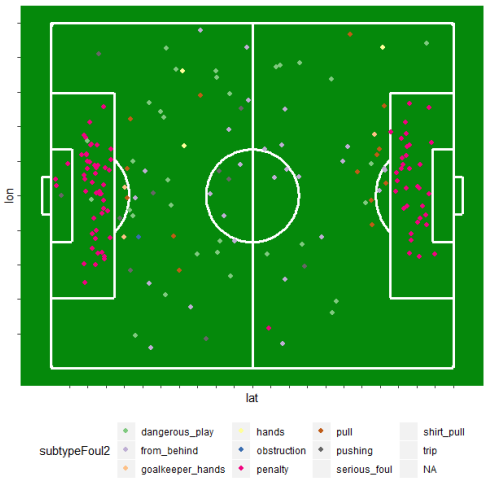


Figure 65: Red cards

4.3.4 Feature Engineering

We are going to generate our features over the match table that is the table that we are going to use and predict from. Matches is the biggest table and is the one that contains each match. Therefore, makes sense that is in this table where we execute and modify the features that we can extract from the other tables that we have seen previously.

Goals related

From the match table, as we have seen, the goal difference is normally a variable that is taken into account when modelling, therefore, we are going to create a feature named goalDiff which will be equal to the difference of goals between the local and away team, $goalDiff = home\_team\_goal - away\_team\_goal$. We are also going to engineer another variable related with goals named the adjusted goals. $Adjusted\_goals$. This variable what does is to consider that goals should be adjusted to the conditions in which was scored. In particular I am going to follow the same directions as FivethirtyEight [**?**] does for their SPI index:

- If there are more players from the team that scores, that goal has a value of 0.8.
- If the team leading, scores in the late moments of a match. Those goals, do not have the same importance as a goal to draw or to win. These goals are a bit "filling" and the losing team probably gave up. Therefore we are going to evaluate that after the 80' until the 89', a goal is worth 0.75. A goal in the 90' or later is worth 0.5.

Also from the same table, we are going to extract the score from the first part. This might be able to explain for instance how superior a team is from the other. As we mentioned before, it is more important a win than improving the score between two teams. Therefore, we are going to consider that factor because maybe the winner of the match, was already happy with their advantage on the first half and chose not to risk.

Player ratings related

From the player table what we are going to extract is for a given match, the quality of each player. We are going to try different approaches:

- Create a variable named $weakest\_link$, which will basically be the minimum quality of the player on the field for each team.
- Do the average of the players for each team.
- Do a weighted average where the defense is more important than the offense. Since offensive players can miss, it is in the nature of trying to

score but being a defender, can not let space to errors.

Team qualities

From the team table we are going to do the average of all the qualities of the team.

Event data

From the events table we are going to consider the following variables:

- We are going to model a naive expected goals model, we lack of context but it might more explanatory than goals. We are also going to apply to xG.
- Red and Yellow Cards and the earliest minute.
- Number of corners, fouls, crosses and throw-in's

In conclusion, the variables that we are going to consider in our modelling phase, are the following ones:

Variable	Description
league_id	League identifier
season	Year in which the matches takes place
stage	Stage of the season the match occurred
home_team_api_id	Home team id
away_team_api_id	Away team id
goal_difference	Goal difference
adjusted_goalsH	Local adjusted goals
adjusted_goalsA	Away adjusted goals
xGH	Expected goals local team
xGA	Expected goals away team
formationH	Local formation
formationA	Away formation
goalsLocalHalfTime	Goals scored by the local by half time
goalsAwayHalfTime	Goals scored by the visitor by half time
localFouls	Number of fouls committed by the local team
awayFouls	Number of fouls committed by the away team
homeYellowCards	Number of cards given to the local team
awayYellowCards	Number of cards given to the away team
homeRedCards	Number of red or 2nd yellow cards given to the local
awayRedCards	Number of red cards or 2nd yellow given to the away team
firstYellowMinH	Earliest minute yellow card given to the local team
firstYellowMinA	Earliest minute yellow card given to the away team
homeExpulsion	Earliest minute of expulsion from the local player
awayExpulsion	Earliest minute of expulsion from the away player
localCorners	Number of corners of the local team
awayCorners	Number of corners of the away team
localCrosses	Number of crosses of the local team
awayCrosses	Number of crosses of the away team
homeTI	Local team throw-ins
awayTI	Away team throw-ins
homeTeamQuality	Average team quality of the local team
awayTeamQuality	Average team quality of the away team
homeWeakestLink	home player with the worst rating
awayWeakestLink	visiting player with worst rating
averageHomePlayers	Average quality of the players playing home
averageAwayPlayers	Average quality of the players playing away
waverageHomePlayers	W. Average quality of the players playing home
waverageAwayPlayers	W. Average quality of the players playing away

Table 17: League Table

Expected Goals Modelling

However, prior to proceed to the modelling phase, we are going to define how xG is done in this section. In our current dataset, we do not have as much contextual data as we would like to. Nevertheless, we do have some information. We have the **types of shots** and which ones missed, we have the **longitude and latitude** from quite some of them, therefore we know the **relative angle to the goal** that we have, we know the **log distance** and the **interaction between the distance and angle**

Our first approach was to just consider the shots given a position, those that missed and those that went in, which shows us a bit what we could have seen before from the goals, shot on and shot off. Where we only have a high xG on the area where goals truly happen. Aside from that, the accuracy outside from the area, diminishes. To perform this graphic, we have folded the field to showcase in only one side, we have removed the penalty shots which have a high success rate and we have eliminated goals from some positions where there are very few shots and they were goals, we also found some outlayers which we removed.

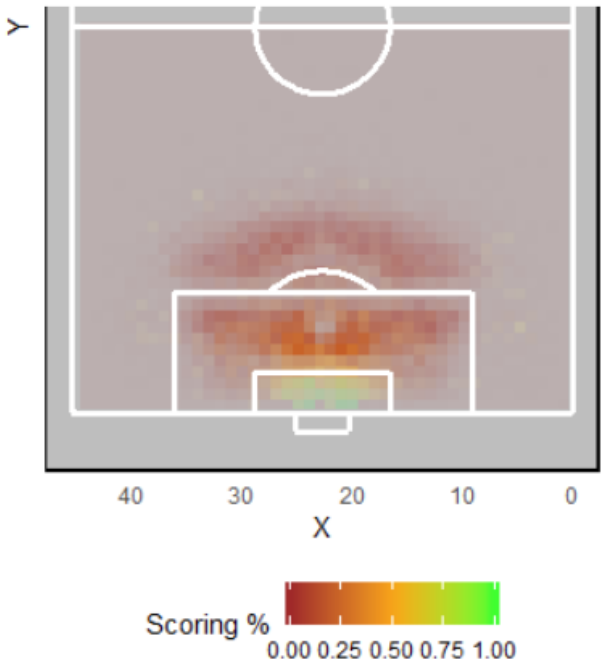


Figure 66: Expected Goals visualization

In the more advanced approach where we actually modelled the problem with a random forest. What we have done also is to fake data points in order to enrich our model[?]. This provides us of more information to train with. We used random trees instead of the logistic regression because we want to encode the information raw without angles and distances. This cannot be done in a logistic regression since as we mentioned on the state-of-the-art,

does not handle well the non-linearity. As a drawback we had to calibrate our model which we will do with isotonic regression, which comes for guaranteed in the logistic model. When training the model I chose to split it in 80-20. Finally validated with a 5-fold CV tuned for mtry and used AUC-ROC to evaluate my model.

In this approach mostly what we do that we did not on the previous one, is to take into consideration the type of shot that we are performing and to fake data points so that the model has more to train with. The results are not much completely different (they are smoother and more accurate) if we do print them as we did previously but the differences is that now, what for instance could be a header from a corner which has very few chance to score, will actually have that score instead of having the score of a very close shot next to the goalkeeper in the inner area.

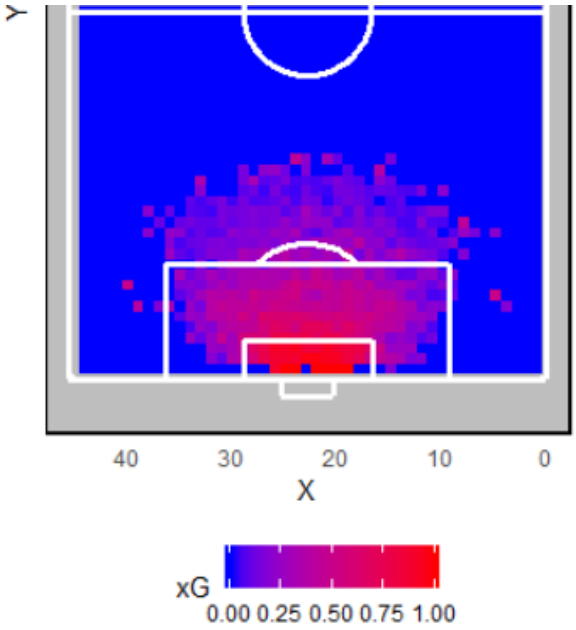


Figure 67: Expected Goals visualization

4.3.5 Modelling

The problem we are trying to solve is a supervised problem. We can model it in two different ways: classification or regression. We are going to treat it

as a classification problem were we are going to classify the final output of the match (H/D/A). There is a huge variety of algorithms which we need to train and all of them have its particularities as we will mention later.

In order to model our data, we are going to split the data in 3 sets: Train, Development (dev) and Test. The train set is where we are going to try ideas and train different models. The dev set is where we are going to evaluate the different ideas an pick one and keep iterating to improve the dev-set performance until we are satisfied (our results are close to the metric we aim) and finally we are going to evaluate it on the test set. Note that is important that actually the test and the dev set have the same distribution because otherwise we might not be working on a set that is representative enough and not being efficient, since our effort spend on optimizing some algorithms, could easily not be good for the test set whereas in the dev set does it well. Regarding to the size of each of the sets, we are going to use a 60% on the training set, and a 20% for the test and the dev set. Since we think that seasonality can be a component of our model, we are not going to perform cross-validation and we are going to sort our dataset by date.

4.3.5.1 Train Set

In this set is where we have tried our ideas and different models. Regarding to the algorithms that we can work with, we have to consider that we previously mentioned we would work with RPS. RPS works with an output probability so when choosing the 3 approach for our models, we are going to need to take this into consideration. While training our set, we have seen that the number of features that we were having was way too big. Therefore we applied VIF tests to check multicolinearity in our variables, realised PCA's and tested their predictive capacity. In order to test our model, we have used Naive Bayes.

After iterating with our train set, we have found that the model which best represents our data is the following:

Variable	Description
league_id	League identifier
season	Year in which the matches takes place
stage	Stage of the season the match occurred
home_team_api_id	Home team id
away_team_api_id	Away team id
xGH	Expected goals local team
xGA	Expected goals away team
goalsLocalHalfTime	Goals scored by the local by half time
goalsAwayHalfTime	Goals scored by the visitor by half time
homeExpulsion	Earliest minute of expulsion from the local player
awayExpulsion	Earliest minute of expulsion from the away player
localCorners	Number of corners of the local team
awayCorners	Number of corners of the away team
waverageHomePlayers	W. Average quality of the players playing home
waverageAwayPlayers	W. Average quality of the players playing away

Table 18: League Table

4.3.5.2 Dev Set

In order to optimize our algorithms we have basically searched which was the tuning which was most adequate witht the hyperopt and gridsearchCV for the SVM's in case of machine learning. In case of pi-rating we have a r library named *piratings* which also tunes which alpha and beta are the bests (optimize_pi_ratings). Finally, for Karlis Ntzoufras approach, we used PyStan a package for Bayesian inference using the No-U-Turn sampler, a variant of Hamiltonian Monte Carlo.

4.3.6 Statistical Approach: Karlis Ntzoufras

Among the different models that we had looked on our state-of-the-art, the Karlis Ntzoufras was one of the bests and I wanted to test how it works. Other models, were very simple and I did not consider implementing them because they were already outdated.

4.3.7 Machine Learning Approach

The fact that we use RPS implies that the algorithm that we might want to use, has to actually be able to provide the inside we are mentioning, more

concretely a probabilistic classifier.

Some classification models, such as naive Bayes, logistic regression and multilayer perceptrons (when trained under an appropriate loss function) are naturally probabilistic. Other methods such as SVM's are not but can be turned into probabilistic classifiers. For instance, we have decided not to approach decision trees and boosting methods, since they produce distorted class probability distribution. Regardless that for instance XgBoost is one of the best algorithms one could possibly apply to their dataset.

Therefore, we are going to work with the following algorithms:

4.3.7.1 Naive Bayes

Naive Bayes is a classification technique based on the independence assumption between predictors (Bayes theorem). In other words, this classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature regardless if those features are interdependent, which simplifies the computation which makes it a fast learner that requires few data which has worked quite well in many real-world situations.

4.3.7.2 SVM

Informally, a SVM is a method that is one of the best known and most widely methods used for classification and regression analysis. More concretely, a SVM model is a representation of data in space, mapped in a way that separate categories are divided by a clear gap that is as wide as possible. Since it is not always possible, we might need to leave a soft-margin.

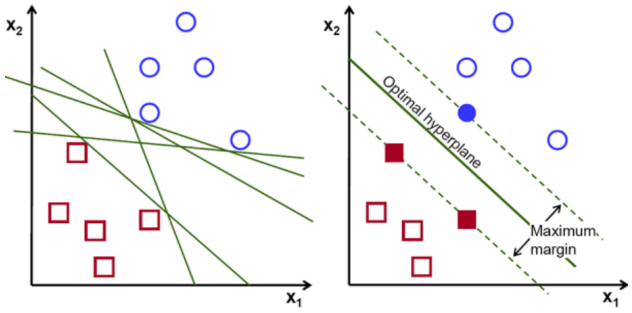


Figure 68: Hyperplanes that are a solution VS optimal

4.3.8 ELO Approach: Rating System

We are going to use a Rating System. As we mentioned in our state-of-the-art, there different rating systems. One of them was the Pi-Rating.

As we previously mentioned, Pi-Rating is capable of:

- Considering the Home advantage factor
- Give more relevance to recent matches.
- Understand that winning is more important than increasing the goal gap.

In this particular case, we are not going to implement the Pi-Rating by ourselves because there is a library in R named **piratings** which implements the work mentioned from Constantinou and Fenton (2013).

4.3.9 Improvement: Competitions addition

In order to improve our accuracy and to take into consideration certain factors such as $fatigue$, or mental state when getting $knocked\_out$ after a competition (variables we are introducing in this improvement). We are going to need data of competitions aside from the ones we have. Currently, we only have data from leagues. Having data from only leagues, implies that we do have a steady flow of matches of around 7 days. When a team is fighting for other tournaments such as Europa Champions League, UEFA, or their national Cup, is when they might start to feel strain in their bodies, save some players for the important matches, underperform because they got knocked

out of one of those competitions where they had high expectations, etc. So we decided to add data from those competitions and see if it helps our model.

Our dataset is going to be a very simplistic one and is not going to contain lots of information since what our models will be dealing with is mostly league matches. As we said, we want the data so that we can evaluate the consequences it might bring insight to those league matches. We have taken the information from Scoreboard.

For each competition, UEFA, Champions and national Cups (except for the Polish, Scottish, Swiss, Belgian and Portuguese leagues, which have no event data) we have collected the following information:

Variable	**Description**
season	Year in which the matches takes place
stage	Stage of the season the match occurred
date	day where the competition takes place
home_team_api_id	Home team id
away_team_api_id	Away team id
competition	name of the competition
home_team_goal	Number of goals scored by the local team
away_team_goal	Number of goals scored by the visiting team

Table 19: Champions, UEFA and inner Cups Table

Also we have added some variables to the team table which are the coordinates of their city and the name of the city. At this way, we are going to be able to understand if the away team has a long trip to dispute their following match and if this contributes to the away effect. Moreover, we have also added another table named *rivalries* where we added only the biggest rivalries. This is because from one part, we have the *distance* that we added on teams and another thing is that if we look for rivalries on the internet, we will find that there are way too many considered rivalries and I do not agree with the classification that for instance, Wikipedia offered. The rivalries table looks as follows:

Variable	**Description**
home_team_goal	Number of goals scored by the local team
away_team_goal	Number of goals scored by the visiting team

Table 20: Rivalries table

We are also going to create a variable for the local and visitor which is the match *importance*. We are building this ratio according to the following variables:

- **Crucial match:** Does the team win the league? Does avoid relegation? Does enter Champions/UEFA positions? Is it a direct rival?
- **Distance between the cities:** We can know if we have kind of a derby if the distance is close in case of a local derby.
- **Is it a rivalry?**

We are also going to add a variable name *streak* which is represents how many points the local/visitor team scored during the last 5 matches. Another variable will be the $fatigue$ which is going to be calculated according to how many days a team had their last match. Another one that is $knocked\_down$ which if they lost a competition such as Champions, national Cup, etc.

We studied their affect together with the other variables of our model, and found $fatigue$ and $knocked\_down$ not to be relevant. However, *importance* and *streak* did.

Therefore our final model which we trained for is the following:

Variable	**Description**
league_id	League identifier
season	Year in which the matches takes place
stage	Stage of the season the match occurred
home_team_api_id	Home team id
away_team_api_id	Away team id
xGH	Expected goals local team
xGA	Expected goals away team
goalsLocalHalfTime	Goals scored by the local by half time
goalsAwayHalfTime	Goals scored by the visitor by half time
homeExpulsion	Earliest minute of expulsion from the local player
awayExpulsion	Earliest minute of expulsion from the away player
localCorners	Number of corners of the local team
awayCorners	Number of corners of the away team
waverageHomePlayers	W. Average quality of the players playing home
waverageAwayPlayers	W. Average quality of the players playing away
importanceH	importance of the match for the local
importanceA	importance of the match for the away
streakH	points in the last 5 matches
streakA	points in the last 5 matches

Table 21: League Table

4.3.10 The Statsbomb and Wyscout plus

In this particular case, we are going to work with soccer-logs. As we have mentioned in our appendix, there are two sources which are soccer-logs and are open-source. We are not going to add this in our work directly but I thought that was very insightful to showcase what not having appropiate datasets available does to the researcher.

- **Wyscout**: Which includes all the spatio-temporal events (passes, shots, fouls, etc.) that occur during all matches of an entire season of seven competitions (La Liga, Serie A, Bundesliga, Premier League, Ligue 1, FIFA World Cup 2018, UEFA Euro Cup 2016).[62]
- **Statsbomb**: Which includes all the spatio-temporal events that occur during matches as Wyscout but for 1999-2019 Champions League, 2018-2019 FA Women's Super League, 2018 FIFA World Cup, 2019 Women's World Cup, La Liga BBVA from 2004-2018, NWSL 2018 and 2003/2004 Premier League.[**?**]

With more detail Statsbomb, does not have all the data from the competitions we mentioned. The data that has is the following:

Competitions	**Description**
2018 Men's FIFA World Cup	64 games
2019 Women's FIFA World Cup	52 games
UEFA Champions League finals	14 finals
FA Women Super League	195 games
National Women's Soccer League	46 games (some games)
Spanish La Liga 2004/05-2017/18	452 games, FC Barcelona focused
English Premier League 2003/04	32 games, Arsenal focused

Table 22: League Table

As we can see, there is not much variability in our dataset. Without going further, FC Barcelona has around 30% of the shots of the entire dataset. With almost a 10% belonging to Messi. This is supposed not to affect our xG and xT models but it has a bias that I thought that should be mentioned.

Wyscout on the other hand, has a less biased information proportioned but only for one season of the already mentioned competitions:

Competitions	**Total Games**
2016 Men's UEFA Euro	51 games
2018 Men's FIFA World Cup	64 games
English Premier League 2017/18	380 games
French Ligue 1 2017/18	380 games
German Bundesliga 2017/18	380 games
Italian Serie A 2017/18	380 games
Spanish La Liga 2017/18	380 games

Table 23: League Table

If we pay attention to the data, we can see that there are some overlappings since they both covered for instance, the 2018 Men's FIFA World Cup. In total the overlapping occurs in 98 matches. The 64 from the Men's Cup and 34 from 2017/18 La Liga BBVA. As Rowlinson (2020) concludes, the difference between both datasets is that Wyscout records fewer non-penalty shot attempts than StatsBomb and concludes that StatsBomb has a better coverage. This is because the minor number of attempted shots, actually what does is to change the xG scores since we are actually scoring with less performed shots. Aside from un-recorded shots, we also have some missplacements. This last fact I think that is natural that happens because event data up to the date, is taken manually by people watching the match and recording the events that take place. In case of Opta which relates to StatsBomb, we can see that their accuracy is proven to be excellent as this paper showcases [48].

Rowlinson (2020), states in his work and as he later concludes, that merging both datasets (which he does) in order to have more data which is suboptimal since the data comes from different sources. In my particular case, I am then just going to work with StatsBomb data. My purpose is to showcase how better the xG modelling can get if we change from the European Soccer Database to a event-data that actually has everything that happens. Therefore what I am going to do is to calculate xT, since the first iteration it is xG.

4.3.10.1 Expected Threat Showcase: Quantifying pressure

As we mentioned, expected threat, quantifies not only the threat that a shot from a given position represents, but the threat that this position can create if they chain a given number of passes. In our particular case, we defined xT in a generic way where the dimensions where a grid of $N \times M$. In this particular case, we are going to use xT in addition of pressure events and turnovers so that we can understand how players perform under pressure.

If we print the pressure and turnover event locations, we can see that as we expect, each of the events has its maximum pressure around the opponent area.

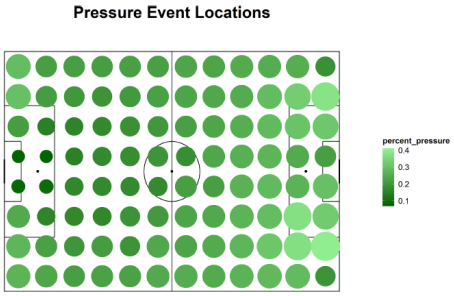


Figure 69: Pressure Event Locations

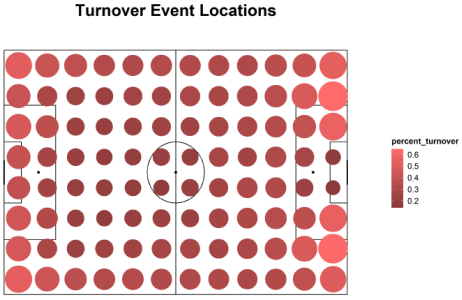


Figure 70: Turnover Event Locations

We can see for instance from the data below, how players react when under pressure or without. Even thought, this metrics provide some insight, defensive pressure can actually be changing the behaviour of players into something more conservative and creating non-threating situations. If we want for instance to know which players generate threat regardless of the pressure, we can see this with xT.

Therfore, we are going to implement xT as we introduced in our state-of-the-art but to save ourselves a bit of time, we are going to use the xG model from StatsBomb. Also we are going to apply to the model the pressure that we are talking about and consider the turnovers.

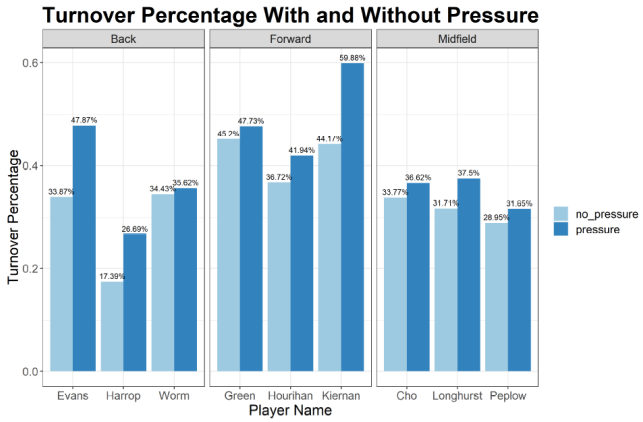


Figure 71: Turnover with vs without pressure

$$xT_{x,y} = (s_{x,y} \times g_{x,y}) + (m_{x,y} \times \sum_{x'=1}^{12} \sum_{y'=1}^{8} T_{(x,y)\to(x',y')} \times xT_{x',y'}) \tag{36}$$

where now $g_{x,y}$ is substituted by $\frac{\sum i=1^{n} xG_{x,y\,i}}{n}$:

Player	**Team**	**xT**	**xTpressure**
Vivianne	Arsenal	4.78	-1.97
Nikita Parris	Manchester City	4.25	-0.874
Allie Long	Seattle Reign	0.554	-0.121
Rachel Corsie	Utah Royals	0.578	0.457

Table 24: League Table

This makes that our model, now incorporates the pressure in the decision making and we can actually quantify which players are generating more risk regardless of the pressure.

For instance, if we apply to our data the xT with pressure, we can see that players behave differently under pressure in a whole match. The following table shows the cumulative xT with and without pressure:

In the following table we can see that for instance Corsie, did not have a high expected threat without pressure, seems to do same well with pressure. Whereas Miedema chokes and underperforms very hard.

4.4 Did we beat them?

In this last section we are going to answer the question we have been wondering this whole project. Have we over-performed the betting odds accuracy? In order to see if we over-performed the betting odds, we are going to compare their RPS performance for each of the improvements versus the model to beat.

Model	**RPS**
NB	0.27255
SVM	0.27485
Pi-rating	0.28752
K. Ntzoufras	0.293447
NB + Pi-rating	0.2556
SVM + Pi-rating	0.24318
bettingOdds (Power)	0.223418

Table 25: League Table

As we can see, we do not win the betting odds but we get quite close to it. The best performing method is the SVM with the pi-rating with an RPS of 0.24318.

5 Conclusions & Future Work

5.1 Regarding to the project

I think that there is not much left to do regarding to this project, if we think of it as with this dataset. We could have tried more approaches even I think we covered the state-of-the-art, we could have tried more machine learning algorithms. I personally think that with the current data, and the fact that a better data from many seasons which currently is not publicly available, we could have done slightly better but not significantly better. We could have scrapped private data as mentioned, which would have helped our results, since we could have tried and engineered many other metrics which someone thought before or that we could think ourselves.

I think that the biggest problem on this project and in football analytics, is data heterogeneity and availability of those. As we mentioned several times across the project, in association football, it is very important and the difference between a good and a bad prediction, relies on those. Lacking of data such as pressure events, passes, GPS data, limits us a lot and hurts our model.

Regarding to our objectives, we said that we wanted to perform an in depth study of the state-of-the-art which we did, we studied the data available on the market as we can see on the appendix, we used a methodology in order to iterate towards our goals and finally, we almost reached the odds in performance, we applied different techniques from different approaches and we were able to do it, our objectives were fulfilled.

Unfortunately, COVID-19 has been very harsh on everyone and there is no doubt of this. However, in terms of football I think that it has created a be very interesting chance to investigate how this pandemic has affected our preconceived ideas. For instance, at the end of this season, there have been many matches that were disputed in neutral ground or in a smaller stadium from the second team. Moreover, public events are restricted so we did not have public. As we mentioned in our state-of-the-art, this effects contribute to the home ground advantage which is highly probable that has been diminished. This for instance, arises an interesting question on the fairness of the final league result since all teams endure the same restrictions, but not really the same. Because as mentioned, a league should never be taken as

criteria until is finished due to the imbalances that has (e.g you could have been playing very strong teams at home and lost, but now that you need wins at home, and you are against similar level or lower, you lose the home ground factor). This can also be extended to leagues that terminated, once COVID began.

5.2 Personal Conclusions

One of the things I learned is that **problems can be solved in many different ways**. I have to admit that I am shocked by the quantity of researched conducted on association football, specially during the last years, the number of papers that have been published have incremented widely.

Moreover, I am surprised that some areas of study, end up having some relationship with a football aspect. When I started, I was very naive and I thought that football would be less dimensional and have only some approaches but I can see now, how extremely complex can be to give an explanation to something as simple as saying which team will win. I use the word simple because it is something that from the ignorance, we belittle a lot, we maybe see some statistics and think that we are already able to guess what will happen. But the truth is that reality is far more complex than we think, which I think that adds beauty to the game.

I already knew and assumed that was a difficult topic, I was interested in the problem and I personally liked that the domain was generally known. I feel that sometimes when you work on a topic that is very specific, it blurs a bit what you are really trying to achieve. Reason why I am happy that always all I did, I could have a sense if it made or not sense. Moreover, if I had to read as much literature in another topic as on this one, I think I would have had a harsher time getting to know well the domain and probably I would have required of help from an expert of it.

However, I would like to emphasize how hard has been to me to in a way summarize and get to know as much as I could football, it has been very consuming in terms of time and mentally, specially because **I committed a mistake in naively not scoping more my thesis.** While working on the state-of-the-art it felt like working in a puzzle but with pieces from 20 mixed puzzles. But I think that precisely of not being a too scoped project, it taught me how hard research can be.

I think that there is a lot to do, there are a lot of aspects in football that can improve. However, while there is not an open-source extensive dataset, I think that progress will be somehow on-hold. It is true that Statsbomb and Wyscout offer data and that Opta and other platforms provide data to researchers.

I would like to mention that towards this direction, David Sumpter is performing an excellent job in promoting football analytics into football. He is the author of a book Soccermatics where he brings some insight to ground level people who might just like football, he recently co-started a YouTube channel along with Github repositories regarding to football statistics (unluckily for me, this started half way of my thesis) where him and other football experts that work for Clubs, share their knowledge which otherwise can cost way more hours to acquire or even to be aware of its existence. Moreover, some days prior to the delivery of this thesis, he has also announced a collaboration with a data provider which I think that is very important to democratize the access of data to everyone, not just to associate researchers. Aside from the academics, there is quite a community around football analytics in Medium and Twitter.

Figure 72: David Sumpter recent tweet[63]

My personal opinion on football analytics, is that clubs will keep integrating analysis to strengthen their abilities and that football will not be stripped of its magic, but we will see that maybe some things like a goal from a place with low percentage scoring, a very centralized passing network, how teams press... will change and become more uniform (and efficient) over time. It may have not been emphasized enough on this thesis because it is not the subject of

study, but many clubs are actually relying every time more on data in order to do things such as understand their inefficiencies and mistakes, scouting players or even trainers and I think that football will become even a more tactical sport where talent will still matter but teams with a great analytical work behind will be able to achieve great things too.

Regarding to the prediction problem, I think that there is a ceiling, we have seen a paper from 2011 with a 65% which is a nice result, and I am sure that someone has recently achieved somewhere near 70% but I do not see someone crossing the 80% barrier soon. I think that it will not happen soon because football, it is a complex problem and that while in football analytics things can be useful just by visualizing some insights from the game so that the coach teaches their players, it is not so clear that some variables will actually provide a better insight on predicting an outcome.

Without going further, we could for instance analyse with Voronoid diagrams if a defense is perfect or not at a given moment, are there openings at that moment? Was the pass realised the best one?

Figure 73: BirdsPyView library example[63]

In this play for instance, we can see a big opening on the upper-side for F.C Barcelona but there are so many factors to consider yet. Is the defender near

him, faster than Barcelona's striker? Will he be able to receive well? What I refer is that, even we could probably agree in that if a player has a lot of space on the field, will most likely be a dangerous action, it will still be a matter of chance+skill and stripping chance from the game would be killing its magic.

To conclude, we mentioned this before but, in such a low-score game, goals have too much value. If this were basketball, american football, cricket, handball, e-sports, etc. It would probably be much easier to obtain a high accuracy. This is not something I say to belittle other sports, but to give perspective on how much a goal is worth and how a shot with few chance, can actually change an outcome of a match, it is incredible. It might sound ridiculous after writing so much about football providing insights but I sort of think that the magic behind football is this, that there will always be a big chance on it.

References

[1] Jochems, D. B. (1962). Forecasting the outcomes of soccer matches: A statistical appraisal of the dutch experience. Metrika, 5(1), 194–207. doi:10.1007/bf02616199

[2] 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

[3] Forrest, D., Simmons, R. (2000). Forecasting sport: the behaviour and performance of football tipsters. International Journal of Forecasting, 16(3), 317–331. doi:10.1016/s0169-2070(00)00050-9

[4] Pope, P. F., Peel, D. A. (1989). Information, Prices and Efficiency in a Fixed-Odds Betting Market. Economica, 56(223), 323. doi:10.2307/2554281

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

[6] Alamar, B. (2006). Basketball on Paper: Rules and Tools for Performance Analysis. Journal of Sport Management, 20(1), 120–123. doi:10.1123/jsm.20.1.120

[7] Goldstein, D. G., Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. Psychological Review, 109(1), 75–90. https://doi.org/10.1037/0033-295X.109.1.75

[8] 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

[9] Constantinou, A.C. Dolores: a model that predicts football match outcomes from all over the world. Mach Learn 108, 49–75 (2019). https://doi.org/10.1007/s10994-018-5703-7

[10] John Goddard, Regression models for forecasting goals and match results in association football, International Journal of Forecast-

ing, Volume 21, Issue 2, 2005, Pages 331-340, ISSN 0169-2070, https://doi.org/10.1016/j.ijforecast.2004.08.002.

[11] Moroney, M. J. (1956) Facts from Figures, 3rd edn. London: Penguin

[12] Reep, C., Benjamin, B. (1968). Skill and Chance in Association Football. Journal of the Royal Statistical Society. Series A (General), 131(4), 581. doi:10.2307/2343726

[13] Reep, C., Pollard, R., Benjamin, B. (1971). Skill and Chance in Ball Games. Journal of the Royal Statistical Society. Series A (General), 134(4), 623. doi:10.2307/2343657

[14] Hill, I. D. (1974). Association Football and Statistical Inference. Applied Statistics, 23(2), 203. doi:10.2307/2347001

[15] Thompson, M. (1975). On Any Given Sunday: Fair Competitor Orderings with Maximum Likelihood Methods. Journal of the American Statistical Association, 70(351a), 536–541. doi:10.1080/01621459.1975.10482468

[16] Harville, D. (1977). The Use of Linear-Model Methodology to Rate High School or College Football Teams. Journal of the American Statistical Association, 72(358), 278. doi:10.2307/2286789

[17] Leeflang, P. S. H., Praag, B. M. S. van. (1971). A procedure to estimate relative powers in binary contacts and an application to Dutch Football League results. Statistica Neerlandica, 25(1), 63–84. doi:10.1111/j.1467-9574.1971.tb00134.x

[18] Maher, M. J. (1982). Modelling association football scores. Statistica Neerlandica, 36(3), 109–118. doi:10.1111/j.1467-9574.1982.tb00782.x

[19] M. Dixon and M. Robinson. "A birth process model for association football matches". In: Journal of the Royal Statistical Society: Series D (The Statistician) 47 (3 1998), pp. 523–538.

[20] D. Karlis and I. Ntzoufras. "Analysis of sports data by using bivariate Poisson models". In: Journal of the Royal Statistical Society: Series D (The Statistician) 52 (3 2003), pp. 381–393.

[21] Ridder, G., Cramer, J. S., Hopstaken, P. (1994). Down to Ten: Estimating the Effect of a Red Card in Soccer. Journal of the American Statistical Association, 89(427), 1124. doi:10.2307/2290942

[22] I. McHale and P. Scarf. "Modelling soccer matches using bivariate discrete distributions with general dependence structure". In: Statistica Neerlandica 61 (4 2007), pp. 432–445.

[23] S. J. Koopman and R. Lit. "A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League". In: Journal of the Royal Statistical Society: Series A (Statistics in Society) 178 (1 2015), pp. 167–186.

[24] M. J. Dixon and S. G. Coles. "Modelling Association Football Scores and Inefficiencies in the Football Betting Market". In: Journal of the Royal Statistical Society: Series C (Applied Statistics) 46 (2 1997), pp. 265–280.

[25] Crowder, Martin, et al. "Dynamic Modelling and Prediction of English Football League Matches for Betting." Journal of the Royal Statistical Society. Series D (The Statistician), vol. 51, no. 2, 2002, pp. 157–168. JSTOR, www.jstor.org/stable/3650316. Accessed 12 Oct. 2020.

[26] H. Rue and Ø. Salvesen. "Prediction and Retrospective Analysis of Soccer Matches in a League". In: Journal of the Royal Statistical Society: Series D (The Statistician) 49 (3 2000), pp. 399–418.

[27] L. M. Hvattum and H. Arntzen. "Using ELO ratings for match result prediction in association football". In: International Journal of Forecasting 26 (2010), pp. 460– 470.

[28] Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of pairedcomparisons. Biometrika,39, 324-345

[29] M. Cattelan, C. Varin, and D. Firth. "Dynamic Bradley–Terry modelling of sports tournaments". In: Journal of the Royal Statistical Society: Series C (Applied Statistics) 62 (1 2013), pp. 135–150.

[30] Edward S Epstein. "A scoring system for probability forecasts of ranked categories". In: Journal of Applied Meteorology 8.6 (1969), pp. 985–987.

[31] Anthony C Constantinou, Norman E Fenton, et al. "Solving the problem of inadequate scoring rules for assessing probabilistic football forecast

models". In: Journal of Quantitative Analysis in Sports 8.1 (2012), pp. 1559–0410.

[32] Dimitris Karlis and Ioannis Ntzoufras. "Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference". In: IMA Journal of Management Mathematics 20.2 (2009), pp. 133–145.

[33] Constantinou, A., Fenton, N. (2013). Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries, Journal of Quantitative Analysis in Sports, 9(1), 37-50. doi: https://doi.org/10.1515/jqas-2012-0036

[34] RichardWilliams et al. "Generalized ordered logit/partial proportional odds models for ordinal dependent variables". In: Stata Journal 6.1 (2006), p. 58.

[35] 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

[36] 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

[37] 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

[38] 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

[39] 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

[40] 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

[41] Esme, E., Kiran, M.S. (2018). Prediction of Football Match Outcomes Based On Bookmaker Odds by Using k-Nearest Neighbor Algorithm. International Journal of Machine Learning and Computing, 8, 26-32.

[42] 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

[43] 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

[44] 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

[45] 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

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

[47] Rein, Robert Raabe, Dominik Memmert, Daniel. (2017). "Which pass is better?" Novel approaches to assess passing effectiveness in elite soccer. Human movement science. 55. 172-181. 10.1016/j.humov.2017.07.010.

[48] Liu, Hongyou Hopkins, Will Ruano, Miguel Molinuevo, Javier. (2013). Inter-operator reliability of live football match statistics from OPTA Sportsdata. International Journal of Performance Analysis in Sport. 13. 803-821. 10.1080/24748668.2013.11868690.

[49] Kuypers, Tim. (2000). Information and Efficiency: An Empirical Study of a Fixed Odds Betting Market. Applied Economics. 32. 1353-63. 10.1080/00036840050151449.

[50] Asimakopoulos, Ioannis Goddard, John. (2004). Forecasting Football Results and the Efficiency of Fixed-odds Betting. Journal of Forecasting. 23. 51-66. 10.1002/for.877.

[51] Lee, A. J. (1997). Modeling scores in the Premier League: Is Manchester United really the best? Chance,10(1), 15–19.

[52] Karlis, D., Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. Journal of the Royal Statistical Society: Series D (The Statistician), 52(3), 381–393.

[53] Tsakonas, A., Dounias, G., Shtovba, S. Vivdyuk, V. (2002). Soft computing-based result prediction of football games. In The first international conference on inductive modelling (ICIM2002), Lviv, Ukraine.

[54] Barnett, V., Hilditch, S. (1993). The Effect of an Artificial Pitch Surface on Home Team Performance in Football (Soccer). Journal of the Royal Statistical Society. Series A (Statistics in Society), 156(1), 39. doi:10.2307/2982859

[55] L. M. Hvattum and H. Arntzen. "Using ELO ratings for match result prediction in association football". In: International Journal of Forecasting 26 (2010), pp. 460–470.

[56] 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

[57] 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

[58] 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

[59] 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

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

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

[62] Pappalardo, L., Cintia, P., Rossi, A. et al. A public data set of spatio-temporal match events in soccer competitions. Sci Data 6, 236 (2019). https://doi.org/10.1038/s41597-019-0247-7

[63] David Sumpter tweet
https://twitter.com/Soccermatics/status/1314188608774057984

[64] David Sumpter twitter
https://twitter.com/Soccermatics/

[65] Passing Networks in Python
https://github.com/Friends-of-Tracking-Data-FoTD/passing-networks-in-python

[66] Soccer analytics package
https://github.com/CleKraus/soccer_analytics

[67] Sofifa's website
http://sofifa.com.

[68] Football-data's website
http://football-data.mx-api.enetscores.com/

[69] Football-data.co.uk's website
http://www.football-data.co.uk

[70] Goal.com's interview regarding to fifa player ratings
https://www.goal.com/en-ae/news/fifa-player-ratings-explained-how-are-the-card
1hszd2fgr7wgf1n2b2yjdpgynu

[71] FIFA18 complete dataset
https://www.kaggle.com/thec03u5/fifa-18-demo-player-dataset

[72] FIFA19 complete dataset
https://www.kaggle.com/karangadiya/fifa19

[73] Football DB repository
https://github.com/openfootball/quick-starter

[74] FB-Ref website
https://fbref.com/en/

[75] WhoScored ebsite
ttps://www.whoscored.com/

[76] Understat website
https://understat.com/

[77] Soccerstats website
https://www.soccerstats.com/

[78] Engsoccerdata website
https://github.com/jalapic/engsoccerdata

[79] Eloratings website
http://www.eloratings.net/

[80] 2017 Soccer Prediction Challenge
https://osf.io/ftuva/

[81] FiveThirtyEight repository
https://github.com/fivethirtyeight/data/tree/master/soccer-spi

[82] How fivethirtyeight predictions work
https://fivethirtyeight.com/methodology/how-our-club-soccer-predictions-work/

[83] Football-data
https://www.football-data.org/

[84] Transfermarkt
https://www.transfermarkt.com/

[85] Statsbomb free data
https://github.com/statsbomb/open-data/blob/master/doc/Open%20Data%20Competitions%20v2.0.0.pdf

[86] Sumpter, D (2017b) Soccermatics. UK:Bloomsbury Sigma.

[87] Structuring Machine Learning Projects
https://www.coursera.org/learn/machine-learning-projects

A Datasets

Since we did not have a clear intuition in which data we wanted to exactly use and our results will be directly affected by the data we train with, we have performed a deep study in most of sources that are available and that we were able to find. As we will see, it seems that there is a big variety of sources but in reality there are not so many different sources since, because most sources are using the same data providers. Furthermore, as we will see there are very few soccer datasets available for scientific research. The purpose of this appendix is to document this process and to shed some light about which data could be useful for a given purpose and which source we can use for scientific purpose.

Football data can be divided into these 3 main types[62]:

- **Soccer logs** Which describes the events that occur during a match and are collected through proprietary tagging software.
- **Video-tracking data** Which describes the trajectories of players during a match and are collected through video recordings.
- **GPS data** Which describes the trajectories of players during training sessions and are collected through GPS devices worn by the players.

A.1 FIFA's Dataset

We can find information regarding to the player characteristics online that has been scrapped from websites like SOFIFA. We can find for several editions of FIFA, the complete datasets of their players and the updates of these players over the season that FIFA performs[71][72].

We might think: why do we need a game information for our dataset? Well we must consider that as the Head of Data Collection licensing Michael Mueller-Moehring declared to Goal[70]: "We have many leagues in the game; no stats provider could offer us data for all these leagues, teams and players," told ESPN. "This is also the reason why we use this online database, because it is not possible to buy this data some way - it just does not exist."

For reference, FIFA20 has data from more than 30 official leagues, 700+ teams, and 17,000+ authentic players. This data is constituted by EA Sports

which employs a team of 25 EA Producers and 400 outside data contributors, who are led by Michael Mueller-Moehring. This team is responsible for ensuring all player data is up to date, while a community of over 6000 FIFA Data Reviewers or Talent Scouts from all over the world are constantly providing suggestions and alterations to the database.

All the data is overseen the data team of EA Producers and has to be backed up and verified before even a minor change is made. Each player in the game has over 300 fields as well as over 35 specific attributes which ultimately determine the rating seen in the game.

Therefore, we think that it can be considered a very updated and reliable and diverse source to consider a player performance from literally almost any league. However, this online datasets do not contain this internal 300 fields. They do contain the ratings that are seen in the game (around 50 attributes where many we can already assume that will be colinear).

A.2 Wyscout

Wyscout is a private platform that offers all kind of analytics in order for clubs to be able to scout players. Wyscout some years ago, released what is the biggest collection of soccer-logs.S occer-logs describe match events, each containing information about its type (pass, shot, foul, tackle, etc.), a timestamp,the player(s), the position on the field and additional information (e.g., pass, accuracy). More concretely, contains all matches of an entire season of seven competitions (La Liga, Serie A, Bundesliga, Premier League, Ligue 1, FIFA World Cup 2018, UEFA Euro Cup 2016).[62]

A.3 European Soccer Database

This dataset is the result of the execution of the football-data-collection open source project. Which crawls from two sources:

- Football matches, end of game statistics and in-game events. Extracted from football-data.mx[83].
- Player attributes from EA Sports FIFA video game extracted from SOFIFA[67].
- Betting odds from Football-data.co.uk[69].

As a result of this crawling process, the following information from European leagues (missing tournaments such as Champions/Europa Leagues), is available:

- +25000 matches.
- 10000 players.
- 11 countries from Europe.
- Seasons 2008-2016.
- Players and Team updates from FIFA, including weekly updates that the game performs on the players according to each week results.
- Betting Odds from 10 providers.
- Detailed match events for +10000 matches.

If we do a brief analysis from the data, we can see that the data does not contain a lot of missing values but particularly has NA's in some player attributes also in the match table, we have some of bettors data missing and there are columns that are completely empty which is because they are related with other tables.

A.4 Football DB

Football DB is a free and open public domain football database and schema[73] for use in any (programming) language. It contains information from pretty much all the years possible of the following competitions:

- National Teams
  - World Cup
  - Copa América
  - Copa Oro / Gold Cup
  - European Football Championship
- Football Clubs
  - Copa Libertadores
  - Copa Sudamericana

- Champions League
- Europa League

The data that the API offers can be summarized in the schema that they provide in their guide. We can see that regardless that there are missing competitions, we are actually missing game information.

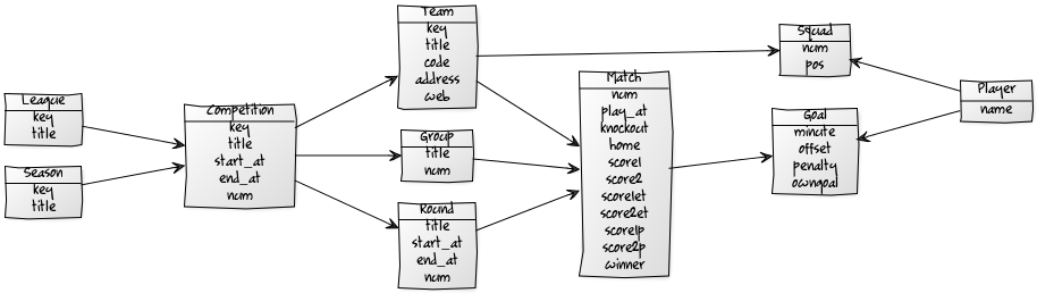


Figure 74: Football DB data model

A.5 Opta

Opta is a private data source that collects data from almost every league and that in general, includes almost every data of what happens in a football match. They collect and distribute full, time-stamped, contextual data live, featuring complete x/y co-ordinates (even z co-ordinates where applicable), and a granularity of event type unique amongst data providers, all this is achieved by their extensive team of analyst which collect the information of each match on the field which is a manual process.

There is a platform from Opta named StatsBomb which provides of free data for some competitions. This data is very detailed and contains ingame stats and it is actually what we are looking for. Without going very in depth on what this data exactly contains (note that you can find a documentation on what this data have in their repository[85]), we can highlight aside from the competitions and seasons tables, the events and lineups. Where for instance in events we have all the events that happen in a match with their corresponding timestamp, which kind of action is (Ball Receipt/Ball Recovery/ Dispossessed/ Duel/ Block/...) which is the team in possession of the ball,

the play pattern (from corner/from free kick/ from throw in/...), location, duration of the play, position, related event to that event, tactic formation, etc.

A.6 FB Ref

FB Ref[74] is a database where we can find information regarding clubs, matches, players, competitions. The problem is that is not information offered in an API but in a website and we should scrap this information from the website which uses StatsBomb API.

A.7 WhoScored

WhoScored[75] is a website specializing in the in-depth analysis of detailed football data. Provided with unique stats, they compile and create comprehensive analysis on the major European divisions, as well as providing data on over 500 leagues and 15000 teams and 250000 players. They offer aside from live results, statistics facilitated by Opta for the top 5 leagues in Europe, ratings based on events recorded in the game calculated by their own algorithms and characters where a player receives some strengths and weaknesses.

A.8 SoccerStats

SoccerStats[77] features football statistics, results, form tables and team stats on national leagues and international soccer competitions worldwide. It provides results from many leagues and it gives information regarding to the match like what each team in average has being doing, historical context on previous matches, statistics of each team very focused in goals, but it does not give more insight aside from the final result of the matches and statistics. For instance we have information in of "% matches over 2.5 goals" but no information regarding to the possession. So we think that this source might be interesting to extract some KPI's but not for the data itself.

A.9 Engsoccerdata

This R package[78] is mainly a repository for complete soccer datasets, along with some built-in functions for analyzing parts of the data. Currently I in-

clude three English ones (League data, FA Cup data, Playoff data - described below), several European leagues (Spain, Germany, Italy, Holland, France, Belgium, Portugal, Turkey, Scotland, Greece) as well as South Africa and MLS.

A.10 Understat

Understat[76] it is a website that calculates the expected goals (xG) for a given team for the major leagues. Expected goals (xG), is a metric which measures the quality of a shot based on a number of different variables. This metric gives an insight to how many goals a player or team should have scored on average based on the chances they have had. Therefore you might have an idea of how many goals you can expect from a team. This metrics appears because in football because in a low-scoring game such as football, the final match score does not provide a clear picture of performance. Reason why measuring the quality of the chances created and conceded is important.

The problem of this source is that it does not mention the source from its results. I personally reached this author to know which sources does use, but I have never received an answer.

A.11 2017 Soccer Prediction Challenge

The data[80] consist of a training dataset which incorporates 216743 match instances from different football leagues throughout the world, and a test dataset of 206 match instances that occurred between March 31 and April 9 in 2017. For each sample, the dataset provides information about the name of the home and away teams, the football league, the date of the match, and the final score in terms of goals scored. The following table, illustrates the leagues captured by the training and test datasets, which incorporate missing data as part of the challenge. Specifically, cells in background colour:

- Yellow represent leagues captured by data.
- Grey represent leagues not captured by data.
- Red represent missing data; i.e., missing match results for a whole season. A total of seven seasons of match results are omitted for model training as part of the challenge in the competition, which is expected to negatively influence the predictive accuracy of the model.

- Blue represent ongoing leagues captured by the test dataset.

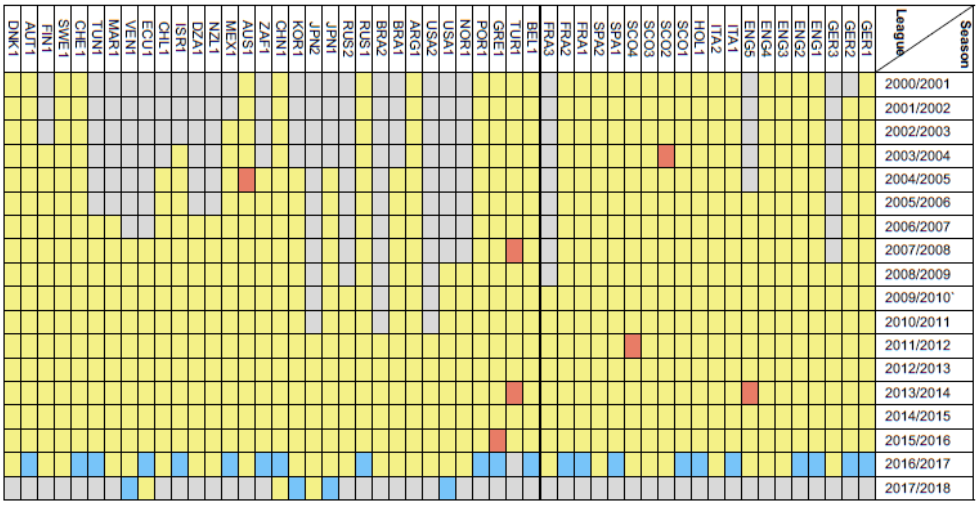


Figure 75: Football leagues captured by the training and test datasets.

A.12 Transfermarkt

Transfermarkt[84] is a website that focuses in the price of football market and their estimations by considering different characteristics and statistics of the market and the players. This kind of information can be useful for instance, as FiveThirtyEight states when calculating their SPI: "We've found that a team's market value — relative to their league's average value — is strongly correlated with its end-of-season SPI rating."[81]

To assess the relative strength of domestic leagues, we use recent matches played between teams from different leagues, supplemented with league market values from Transfermarkt, to assign a strength rating to every league for which we have data.

A.13 FiveThirtyEight

FiverThirtyEight is a website that focuses on opinion poll analysis, politics, economics, and sports blogging. However, we can find that they have been working in association football prediction from 2017 and that by using En-

gsoccerdata, Opta and Transfermarkt they calculate their indexes[82] which are publicly available[81] in their repository.

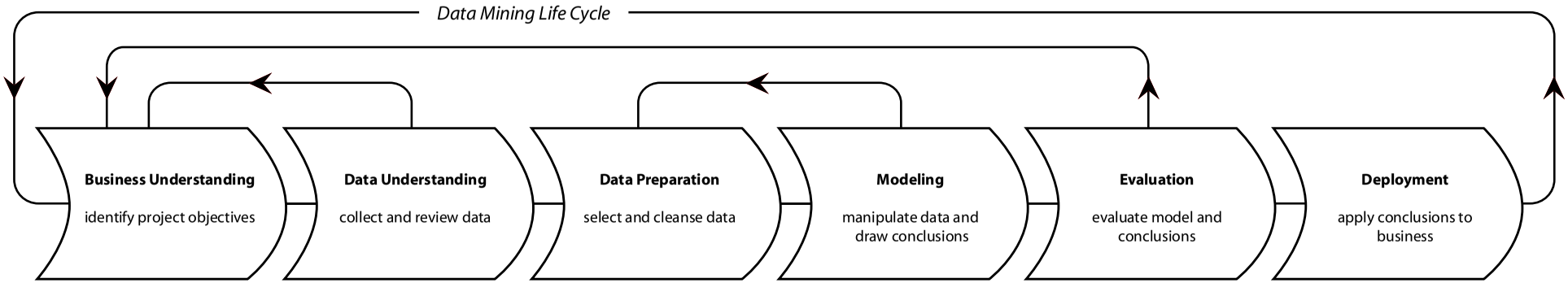
A.14 Football-data

Football-data[83] is a private API that offers the main leagues information for free. Offers information from competitions, teams and matches and some basic information regarding to players. It has 4 tiers:

- **Tier 1 - Free:** 12 competitions with results/fixtures/tables.
- **Tier 2 - Non-free:** 25 competitions with livescores, results/fixtures/tables + squads and match events (player stats (assists, scorers, cards, lineups, substitutions)
- **Tier 3 - Non-free:** 50 competitions with livescores, results/fixtures/tables + squads and match events (player stats (assists, scorers, cards, lineups, substitutions)
- **Tier 4 - Non-free:** 125 competitions with livescores, results/fixtures/tables + squads and match events (player stats (assists, scorers, cards, lineups, substitutions)

As we can see only the first tier is available and the information it provides is non-existent in regards to in-game stats.

Phases



Determine Business Objectives
Background
Business Objectives
Business Success Criteria
(Log and Report Process)

Assess Situation
Inventory of Resources, Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits
(Log and Report Process)

Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria
(Log and Report Process)

Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques
(Log and Report Process)

Collect Initial Data
Initial Data Collection Report
(Log and Report Process)

Describe Data
Data Description Report
(Log and Report Process)

Explore Data
Data Exploration Report
(Log and Report Process)

Verify Data Quality
Data Quality Report
(Log and Report Process)

Data Set
Data Set Description
(Log and Report Process)

Select Data
Rationale for Inclusion/ Exclusion
(Log and Report Process)

Clean Data
Data Cleaning Report
(Log and Report Process)

Construct Data
Derived Attributes
Generated Records
(Log and Report Process)

Integrate Data
Merged Data
(Log and Report Process)

Format Data
Reformatted Data
(Log and Report Process)

Select Modeling Technique
Modeling Technique
Modeling Assumptions
(Log and Report Process)

Generate Test Design
Test Design
(Log and Report Process)

Build Model Parameter Settings
Models
Model Description
(Log and Report Process)

Assess Model
Model Assessment
Revised Parameter
(Log and Report Process)

Evaluate Results
Align Assessment of Data Mining Results with Business Success Criteria
(Log and Report Process)

Approved Models
Review Process
Review of Process
(Log and Report Process)

Determine Next Steps
List of Possible Actions
Decision
(Log and Report Process)

Plan Deployment
Deployment Plan
(Log and Report Process)

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan
(Log and Report Process)

Produce Final Report
Final Report
Final Presentation
(Log and Report Process)

Review Project
Experience Documentation
(Log and Report Process)

Generic Tasks
Specialized Tasks
(Process Instances)

a visual guide to CRISP-DM methodology

SOURCE CRISP-DM 1.0
http://www.crisp-dm.org/download.htm
DESIGN Nicole Leaper
http://www.nicoleleaper.com