Who Will Win It?
An In-game Win Probability Model for Football

Pieter Robberechts[1], Jan Van Haaren[2], and Jesse Davis[1]

[1] Dept of Computer Science, KU Leuven, Belgium
{first.lastname}@cs.kuleuven.be
[2] SciSports, Amersfoort, The Netherlands
j.vanhaaren@scisports.com

Abstract. In-game win probability is a statistical metric that provides a sports team's likelihood of winning at any given point in a game, based on the performance of historical teams in the same situation. In-game win-probability models have been extensively studied in baseball, basketball and American football. These models serve as a tool to enhance the fan experience, evaluate in game-decision making and measure the risk-reward balance for coaching decisions. In contrast, they have received less attention in association football, because its low-scoring nature makes it far more challenging to analyze. In this paper, we build an in-game win probability model for football. Specifically, we first show that porting existing approaches, both in terms of the predictive models employed and the features considered, does not yield good in-game win-probability estimates for football. Second, we introduce our own Bayesian statistical model that utilizes a set of eight variables to predict the running win, tie and loss probabilities for the home team. We train our model using event data from the last four seasons of the major European football competitions. Our results indicate that our model provides well-calibrated probabilities. Finally, we elaborate on two use cases for our win probability metric: enhancing the fan experience and evaluating performance in crucial situations.

Keywords: Association football · Win probability · Sports analytics.

1 Introduction

"A 2-0 lead is the worst lead" is a cliché commonly used in association football. This saying implies that the chances a team will lose (or at best draw the game) are maximized compared to other leads.[3] The underlying idea is that a team leading 2-0 will have a false sense of security and therefore become complacent. In contrast, a team leading 1-0 will tend to concentrate and play with intensity to protect or extend their narrow lead, whilst teams leading by three or more goals have a sufficiently large buffer that comebacks are unlikely.

---

[3] https://en.wikipedia.org/wiki/2-0_lead_is_the_worst_lead

This is an interesting theory and one might wonder whether there is any statistical proof to it. An adequate answer obviously depends on several parameters, such as the time remaining in the match and the relative strengths of both teams. For example, the win probability of a team that leads by two goals with less than a minute remaining in the game should be approaching 100%. How would that probability differ if the team leads by two goals at half time, or with ten minutes remaining? Or with 15 minutes and 31 seconds remaining and having had one player red carded? Such questions can be answered by an in-game win-probability model, which provides the likelihood that a particular team will win a game based upon a specific game state (i.e., score, time remaining, . . .).

In-game win probability models have become increasingly popular in a variety of sports over the last decade. Nowadays, in-game win probability is widely used in baseball, basketball and American football. It has a number of relevant use-cases within these sports' ecosystems. First, the win probability added (WPA) metric computes the change in win probability between two consecutive game states. It allows one to rate a player's contribution to his team's performance [17,11], measure the risk-reward balance of coaching decisions [13,15], or evaluate in-game decision making [14]. Second, win probability models can improve the fan experience by telling the story of a game.[4] For example, they can help identify exciting or influential moments in the game [22], which may be useful for broadcasters looking for game highlights. Third, they are relevant to in-game betting scenarios. Here, gamblers have the option to continue to bet once an event has started, and adapt their bets depending on how the event is progressing. This became a popular betting service in many countries, and is estimated to account for over one-third of online betting gross gambling yield in Britain [5].

While well established in these American sports, in-game win probability is a relatively new concept in association football. It first emerged during the 2018 World Cup when both FiveThirtyEight and Google published such predictions. The lack of attention in win probability in association football can probably be attributed to its low-scoring nature and high probability of ties, which makes the construction of a good in-game win probability model significantly harder in comparison to the aforementioned sports. Unfortunately, FiveThirtyEight and Google do not provide any details about how they tackled those challenges.

We present a machine learning approach for making minute-by-minute win probability estimates for association football. By comparing with state-of-the-art win probability estimation techniques in other sports, we introduce the unique challenges that come with modelling these probabilities for football. In particular, it involves challenges such as capturing the current game state, dealing with stoppage time, the frequent occurrence of ties and changes in momentum. To address these challenges we introduce a Bayesian model that models the future number of goals that each team will score as a temporal stochastic process. We

---

[4] ESPN includes win probability graphs in its match reports for basketball (e.g., http://espn.com/nba/game?gameId=401071795) and American football (e.g., http://espn.com/nfl/game?gameId=401030972)

evaluate our model on event stream data from the four most recent seasons of the major European football leagues. Finally, we introduce two relevant use cases for our win probability model: a "story stat" to enhance the fan experience and a tool to quantify player performance in the crucial moments of a game.

2   Related Work

Win probability models emerged in Major League Baseball as early as the 1960s [12]. Baseball can be easily analyzed as a sequence of discrete, distinct events rather than one continuous event. Each point of the game can be characterized by a game-state, which typically captures at least the following information: the score differential, the current inning, the number of outs and which bases are occupied by runners. Given the relatively low number of distinct game-states and the large dataset of historical games,[5] it is possible to accurately predict the probability of winning based on the historical final results for that game-state.

However, defining the game-state in other sports is not as straightforward as it is in baseball. Most sports, like football, basketball, and ice hockey, run on continuous time, and both score differentials and game states can be extremely variable. Typically, win probability models in these sports use some regression approach to generate predictions for previously unseen game states.

Stern [23] proposed one of the first win probability models in American football. Using data from the 1981, 1983 and 1984 NFL seasons, he found that the observed point differential (i.e., the winning margin) is normally distributed with a mean equal to the pregame point spread and a standard deviation of around 14 points. The probability that a team favoured by p points wins the game can then be estimated from this distribution. PFR [18] adjusted this model to incorporate the notion of expected points (EP). The EP captures the average number of points a team would expect to score on its current drive based on the game state (down, distance, quarter, time left, etc.). It adjusts the current score difference by adding the EP to it, which yields a de facto "current expected margin" based on the game conditions. The more recent publicly available models either use a linear logistic regression [2,16] or non-linear random forest model [13] using various features such as the score difference, time remaining, field position, down, the Las Vegas spread and the number of time-outs remaining.

Basketball win probability models are quite similar to the ones employed for American football. The "standard" model of NBA win probability considers game time, point differential, possession, and the Vegas point spread and predicts the match outcome using a logistic regression model.[6] However, several authors have noted that while a single logistic regression model works reasonably well for most of the game, such an approach performs poorly near the end of a game. Seemingly, this occurs because the model misses the fact that non-zero score

---

[5] Baseline records go back more than 100 years.

[6] http://www.inpredictable.com/2015/02/updated-nba-win-probability-calculator.html

differentials at the end of games are deterministic. The crux of this issue lies in the fact that there is a non-linear relationship between time remaining and win-probability. Both Bart Torvik[7] and Brian Burke[8] have solved this issue by partitioning the game into fixed-time intervals and learning a separate logistic regression model for each interval.

Recently, Ganguly and Frank [7] have identified two other shortcomings of the existing basic win probability models. They claim that they (1) do not incorporate sufficient context information (e.g., team strength, injuries), and (2) lack a measure of uncertainty. To address these issues, they introduce team lineup encodings and an explicit prediction of the score difference distribution.

In ice hockey – which has the most similar score progression compared to association football, win probability models are less well established. A basic one was proposed by Pettigrew [17]. The bulk of his model estimates the win probability from an historical average for a given score differential and time remaining, but the model also takes into account power plays by using conditional probabilities.

A third approach to win probability prediction estimates the final score line by simulating the remainder of the game. For example, Rosenheck [21] developed a model that considers the strength of offence and defence, the current score differentials and the field position to forecasts the result of any possession in the NFL. This model can be used to simulate the rest of the game several times to obtain the current win probability. Štrumbelj and Vračar [25] did something similar for basketball. These approaches can be simplified by estimating the scoring rate during each phase of the game instead of looking at individual possessions. For example, Buttrey et al. [3] estimate the rates at which NHL teams score using the offensive and defensive strength of the teams playing, the home-ice advantage, and the manpower situation. The probabilities of the different possible outcomes of any game are then given by a Poisson process. Similarly, Stern [24] assumes that the progress of scores is guided by a Brownian motion process instead, and applied this idea to basketball and baseball. These last two approaches are most similar to the model that we propose for football. However, in contrast to the aforementioned methods, we consider the in-game state while estimating the scoring rates and assume that these rates change throughout the game and throughout the season.

3   Task and Challenges

This paper aims to construct an in-game win probability model for football games. The task of a win probability model is to predict the probability distribution over the possible match outcomes given the current game state. In

---

football, this corresponds to predicting the probability of a win, a draw and a loss.

A naïve model may simply report the cumulative historical average for a given score differential and time remaining, i.e., the fraction of teams that went on to win under identical conditions. However, such a model assumes that all outcomes are equally likely across all games. In reality, the win probabilities will depend on several features of the game state, such as the time remaining and the score difference as well as the relative strengths of the opponents.

While the same task has been solved for the major American sports, football has some unique distinguishing properties that impact developing a win probability model. We identify four such issues.

1. Describing the game state. For each sport, win probability models are influenced by the time remaining, the score differential and sometimes the estimated difference in strength between both teams. The remaining features that describe the in-game situation differ widely between sports. American football models typically use features such as the current down, distance to the goal line and number of remaining timeouts, while basketball models incorporate possession and lineup encodings. Since there are no publicly available models for association football, it is unclear which features should be used to describe the game state.

2. Dealing with stoppage time. In most sports, one always knows exactly how much time is left in the game, but this is not the case for football. Football games rarely last precisely 90 minutes. Each half is 45 minutes long, but the referee can supplement those allotted periods to compensate for stoppages during the game. There are general recommendations and best practices that allow fans to project broadly the amount of time added at the end of a half, but no one can ever be quite certain.

3. The frequent occurrence of ties. Another unique property of football is the frequent occurrence of ties. Due to the low-scoring nature, football games are often very close, with a margin less than or equal to a single goal. In this setting the win-draw-loss outcome provides essentially zero information. At each moment in time, a win or loss could be converted to a tie, and a tie could be converted to a win or a loss for one of both teams. Therefore, a setting that directly predicts the win-draw-loss outcomes breaks down in late game situations.

4. Changes in momentum. Additionally, the fact that goals are scarce in football (typically less than three goals per game) means that when they do arrive their impact is often game-changing in terms of the ebb and flow of the game thereafter, how space then opens up and who dominates the ball – and where they do it. The existing win probability models are very unresponsive to such shifts in the tone of a game.

4   A Win Probability Model for Football

In this section, we outline our approach to construct a win probability model for association football. First, we discuss how to describe the game state. Second, we introduce our four win probability models. The first three models are inspired by the existing models in other sports. With a fourth Bayesian model, we address the challenges described in the previous section.

4.1   Describing the game state

To deal with the variable duration of games due to stoppage time, we split each game into $T = 100$ time frames, each corresponding to a percentage of the game. Each frame can capture a reasonable approximation of the game state, since the events that have the largest impact on the game state (goals and red cards) almost never occur multiple times within the same time frame. In our dataset of 6,712 games, only 6 of the 22,601 goals were scored within the same percentage of the game and no two players of the same team were red carded. Next, we describe the game state in each of these frames using the following variables:

1. **Base features**
    - Game Time: Percentage of the total game time completed.
    - Score Differential: The current score differential.
2. **Team strength features**
    - Rating Differential: The difference in Elo ratings [9] between both teams, which represents the prior estimated difference in strength with the opponent.
3. **Contextual features**
    - Team Goals: The number of goals scored so far.
    - Yellows: Number of yellow cards received.
    - Reds: The difference with the opposing team in number of red cards received.
    - Attacking Passes: A rolling average of the number of successfully completed attacking passes (a forward pass ending in the final third of the field) during the previous 10 time frames.
    - Duel Strength: A rolling average of the percentage of duels won in the previous 10 time frames.

The challenge here is to design a good set of contextual features. The addition of each variable increases the size of the state space exponentially and makes learning a well-calibrated model significantly harder. On the other hand, they should accurately capture the likelihood of each team to win the game. The five contextual features that we propose are capable of doing this: the number of goals scored so far gives an indication of whether a team was able to score in the past (and is therefore probably capable of doing it again); a difference in red cards represents a goal-scoring advantage [4]; a weaker team that is forced to defend can be expected to commit more fouls and incur more yellow cards; the percentage

of successful attacking passes captures a team's success in creating goal scoring opportunities; and the percentage of duels won captures how effective teams are at regaining possession. Besides these five contextual features, we experimented with a large set of additional features. These features are listed in the appendix.

4.2   Applying existing win probability models to football

At first, association football seems not very different from basketball and American football. In all these sports, two teams try to score as much as possible while preventing the other team from scoring. After a fixed amount of time, the team that scored most wins. Therefore, it seems straightforward that a model similar to the ones used in basketball and American football could be applied to association football too. We consider three such models:

1. **Logistic regression model [16,2] (LR).** This is a basic multi-class logistic regression model that calculates the probability of the win, tie and loss outcomes given the current state of the game:

$$P(Y = o|x_t) = \frac{e^{\boldsymbol{w}^T \boldsymbol{x}_t}}{1 + e^{\boldsymbol{w}^T \boldsymbol{x}_t}}, \tag{1}$$

   where Y is the dependent random variable of our model representing whether the game ends in a win, tie or loss for the home team, $\boldsymbol{x}_t$ is the vector with the game state features, while the coefficient vector $\boldsymbol{w}$ includes the weights for each independent variable and is estimated using historic match data.
2. **Multiple logistic regression classifiers [1] (mLR).** This model removes the remaining time from the game state vector and trains a separate logistic classifier per time frame. As such, this model can deal with non-linear effects of the time remaining on the win probability.
3. **Random forest model [13] (RF).** Third, a random forest model can deal with non-linear interactions between all game state variables.

4.3   Our model

Most existing win probability models use a machine learning model that directly estimates the probability of the home team winning. Instead, we model the number of future goals that a team will score and then map that back to the win-draw-loss probability. Specifically, given the game state at time t, we model the probability distribution over the number of goals each team will score between time $t + 1$ and the end of the match. This task can be formalized as:

Given: A game state $(x_{t,home},\ x_{t,away})$ at time t.
Do: Estimate probabilities

- $P(y_{>t,home} = g \mid x_{t,home})$ that the home team will score $g \in \mathbb{N}$ more goals before the end of the game
- $P(y_{>t,away} = g \mid x_{t,away})$ that the away team will score $g \in \mathbb{N}$ more goals before the end of the game

such that we can predict the most likely final scoreline $(y_{home},\ y_{away})$ for each time frame in the game as $(y_{<t,home} + y_{>t,home},\ y_{<t,away} + y_{>t,away})$.

This formulation has two important advantages. First, the goal difference contains a lot of information and the distribution over possible goal differences provides a natural measure of prediction uncertainty [7]. By estimating the likelihood of each possible path to a win-draw-loss outcome, our model can capture the uncertainty of the win-draw-loss outcome in close games. Second, by modelling the number of future goals instead of the total score at the end of the game, our model can better cope with these changes in momentum that often happen after scoring a goal.

We model the expected number of goals that the home $(y_{>t,home})$ and away $(y_{>t,away})$ team will score after time t, as independent Binomial distributions:

$$
\begin{aligned}
y_{>t,\text{home}} &\sim B(T - t, \theta_{t,\text{home}}), \\
y_{>t,\text{away}} &\sim B(T - t, \theta_{t,\text{away}}),
\end{aligned}
\tag{2}
$$

where the θ parameters represent each team's estimated scoring intensity in the tth time frame. These scoring intensities are estimated from the current game state $x_{t,i}$. However, the importance of these game state features varies over time. At the start of the game, the prior estimated strengths of each team are most informative, while near the end the features that reflect the in-game performance become more important. Moreover, this variation is not linear, for example, because of a game's final sprint. Therefore, we model these scoring intensity parameters as a temporal stochastic process. In contrast to a multiple regression approach (i.e., a separate model for each time frame), the stochastic process view allows sharing information and performing coherent inference between time frames. As such, our model can make accurate predictions for events that occur rarely (e.g., a red card in the first minute of the game). More formally, we model the scoring intensities as:

$$
\begin{aligned}
\theta_{t,home} &= \text{invlogit}(\boldsymbol{\alpha}_t * x_{t,home} + \beta + \text{Ha}) \\
\theta_{t,away} &= \text{invlogit}(\boldsymbol{\alpha}_t * x_{t,away} + \beta)
\end{aligned}
\qquad
\begin{aligned}
\alpha_t &\sim N(\alpha_{t-1}, 2) \\
\beta &\sim N(0, 10) \\
\text{Ha} &\sim N(0, 10)
\end{aligned}
\tag{3}
$$

where $\boldsymbol{\alpha}_t$ are the regression coefficients and Ha models the home advantage.

Our model was trained using PYMC3's Auto-Differentiation Variational Inference (ADVI) algorithm [10]. To deal with the large amounts of data, we also take advantage of PYMC3's mini-batch feature for ADVI.

5   Experiments

The goal of our experimental evaluation is to: (1) explore the prediction accuracy and compare with the various models we introduced in the previous section and (2) evaluate the importance of each feature.

5.1   Dataset

The time remaining and score differential could be obtained from match reports, but the contextual features that describe the in-game situation require more detailed data. Therefore, our analysis relies on event stream data. This kind of data is collected manually from watching video feeds of the matches. For each event on the pitch, a human annotator records the event with a timestamp, the location (i.e., a (x, y) position), the type of the event (e.g., pass, shot, foul, . . .) and the players that are involved. From these data streams, we can extract both the game changing event (i.e., goals and cards), and assess how each team is performing at each moment in the game.

We use data provided by Wyscout from the English Premier League, Spanish LaLiga, German Bundesliga, Italian Serie A, French Ligue 1, Dutch Eredivisie, and Belgian First Division A. For each league, we used the 2014/2015, 2015/2016 and 2016/2017 seasons to train and validate our models. This training set consists of 5967 games (some games in the 2014/2015 and 2015/2016 season were ignored due to missing events). The 2017/2018 season was set aside as a test set containing 2227 games. Due to the home advantage, the distribution between wins, ties and losses is unbalanced. In the full dataset, 45.23% of the games end in a win for the home team, 29.75% end in a tie and 25.01% end in a win for the away team.

To asses the pre-game strength of each team, we scraped Elo ratings from http://clubelo.com. In the case of association football, the single rating difference between two teams is a highly significant predictor of match outcomes [9].

5.2   Model evaluation

We trained all four models using 3-fold cross validation on the train set to optimize model parameters with respect to the Ranked Probability Score (RPS) [6]:

$$RPS_t = \frac{1}{2} \sum_{i=1}^{2} (\sum_{j=1}^{i} p_{t,j} - \sum_{j=1}^{i} e_j)^2, \tag{4}$$

where $\boldsymbol{p}_t = [P(Y = \text{win} \mid x_t),\ P(Y = \text{tie} \mid x_t),\ P(Y = \text{loss} \mid x_t)]$ are the estimated probabilities at a time frame t and $\boldsymbol{e}$ encodes the final outcome of the

game as a win ($e = [1, 1, 1]$), a tie ($e = [0, 1, 1]$) or a loss ($e = [0, 0, 1]$). This metric reflects that an away win is in a sense closer to a draw than a home win. That means that a higher probability predicted for a draw is considered better than a higher probability for home win if the actual result is an away win.

We then evaluate the quality of the estimated win probabilities on the external test set, which is a challenging task. For example, when a team is given an 8% probability of winning at a given state of the game, this essentially means that if the game was played from that state onwards a hundred times, the team is expected to win approximately eight of them. This cannot be assessed for a single game, since each game is played only once. Therefore, we calculate for all games in the test set where our model predicts a win, draw or loss probability of x% the fraction of games that actually ended up in that outcome. Ideally, this fraction should be x% as well when averaged over many games. This is reflected
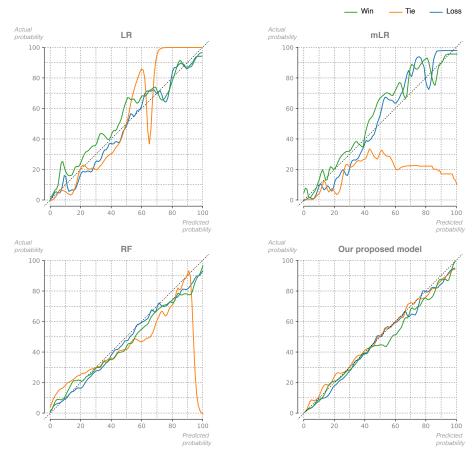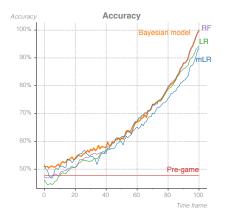

Fig. 1: Only the Bayesian classifier has well calibrated win, draw and loss probabilities.

in the probability calibration curves in Figure 1. Only our proposed Bayesian classifier has a good probability calibration curve. Among the three other models, the RF classifier performs well, but the predictions for the probability of ties break down in late game situations. Similarly, the LR and mLR models struggle to accurately predict the probability of ties. Additionally, their win and loss probabilities are also not well calibrated.

Besides the probability calibration, we also look at how the accuracy and RPS of our predictions on the test set evolve as the game progresses (Figure 2). To measure accuracy, we take the most likely outcome at each time frame

$$\underset{o\in\{\text{win,tie,loss}\}}{\operatorname{argmax}} \quad P(Y = o \mid x_t) \tag{5}$$

and compare this with the actual outcome at the end of the game. Both the RPS and accuracy of all in-game win probability models improve when the game progresses, as they gain more information about the final outcome. Yet, only the Bayesian model is able to make consistently correct predictions at the end of each game. For the first few time frames of each game, the models' performance is similar to a pre-game logistic regression model that uses the Elo rating difference as a single feature. Furthermore, the Bayesian model clearly outperforms the LR, mLR and RF models.
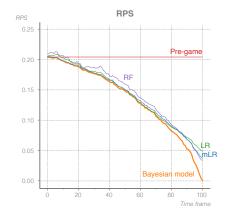

Fig. 2: All models' performance improves as the game progresses, but only our Bayesian model makes consistently correct predictions at the end of each game. Early in the game, the performance of all models is similar to an Elo-based pre-game win probability model.

Finally, we apply our Bayesian statistical model on the 48 games of the 2018 World Cup group stage and compare our predictions against the ones by FiveThirtyEight. Due to FiveThirtyEight's better pre-game team ratings, their model performs better in the early stages of the game. Our model has a similar

performance in the later stages. More details about this evaluation can be found in the appendix.

5.3   Feature importance

In addition to the predictive accuracy of our win probability estimates, it is interesting to observe how these estimates are affected by the different features. To this end, we inspect the simulated traces of the weight vector $\boldsymbol{\alpha}$ for each feature. In the probabilistic framework, these traces form a marginal distribution on the feature weights for each time frame. Figure 3 shows the mean and variance of these distributions. Primarily of note is that winning more duels has a negative effect on the win probability, which is not what one would intuitively expect. Yet, this is not a novel insight.[9] Furthermore, we notice that a higher Elo rating than the opponent, previously scored goals, yellow cards for the opponent and more successful attacking passes all have a positive impact on the scoring rate. On the other hand, receiving red cards decreases a team's scoring rate. Finally, the effect of goals, yellows and attacking passes increases as the game progresses. Red cards and duel strength have a bigger impact on the scoring rate in the first half. For the difference in Elo rating, mainly the uncertainty about the effect on the scoring rate increases during the game.

6   Use Cases

In-game win-probability models have a number of interesting use cases. In this section, we first show how win probability can be used as a story stat to enhance fan engagement. Second, we discuss how win probability models can be used as a tool to quantify the performance of players in the crucial moments of a game. We illustrate this with an Added Goal Value (AGV) metric, which improves upon standard goal scoring statistics by accounting for the value each goal adds to the team's probability of winning the game.

6.1   Fan Engagement

Win probability is a great "story stat" – meaning that it provides historical context to specific in-game situations and illustrates how a game unfolded. Figure 4 illustrates how the metric works for Belgium's illustrious comeback against Japan at the 2018 World Cup. As can be seen, the story of the game can be told right from this win probability graph. It shows how Japan managed to take the lead, shortly after a scoreless first half in which neither team could really threaten the opponent. The opening goal did not faze the Belgians. Their win probability increased as Eden Hazard hit a post. Nevertheless, Japan scored a second on a counter-attack. In the next 15 minutes, Belgium hit a lull, which

---

[9] https://fivethirtyeight.com/features/what-analytics-can-teach-us-about-the-beautiful-game/
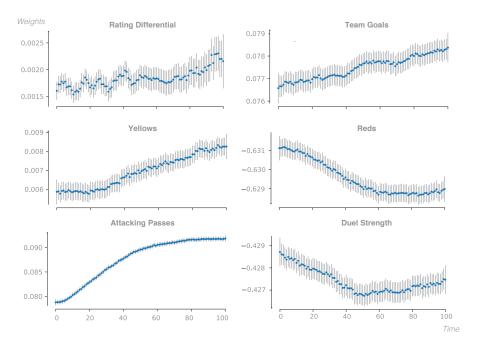
Fig. 3: Estimated mean weight and variance for each feature per time frame.

further increased Japan's win probability. Right when a Belgium win seemed improbable, Belgium got the bit of luck they probably deserved when Vertonghen's header looped over the Japanese keeper. This shifted the momentum of the game in Belgium's favour and five minutes later Belgium were level, before snatching the win with a stunning counter attack in the last second of the game. With this comeback, Belgium created a little bit of history by becoming the first team to come from two goals down to win a World Cup knockout match since 1970.
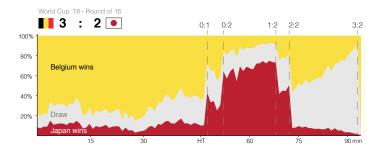

Fig. 4: Win probability graph for the 2018 World Cup game between Belgium and Japan.

Undoubtedly fans implicitly considered these win probabilities too as the game unfolded. Where football fans and commentators have to rely on their intuition and limited experience, win probability stats can deliver a more objective view on these probabilities. Therefore, win probability could be of interest to fans as they watch a game in progress or afterwards, to put the (un)likeliness of certain game situations into a historical context. For example, one could wonder whether Belgium's comeback was truly that exceptional. After all, only 145 World Cup knockout games were played since that one game in 1970 of which very few (if any) had a similar scenario, making this statistic not very valuable. According to our model, Belgium had a win probability of about only 8% right before their first goal. This indicates that it was indeed an exceptional performance, although perhaps not as exceptional as the "once in 50 years" statistic suggests.

Similarly, win probability can be used to debunk some of the most persistent football myths, such as the earlier introduced "2-0 is the worst lead" myth. Perhaps not unsurprisingly, a 2-0 lead turns out to be a very safe lead, much safer than a 1-0 or 2-1 lead. The appendix includes the details of this analysis.

6.2   Quantifying Performance under Mental Pressure

"Clutch" performance, or performance in crucial situations is a recurring concept in many sports – including football. Discussions about which players are the most clutch are popular among fans[10] and teams define the ability to perform under pressure as a crucial asset.[11] However, such judgements are often the product of short-term memory among fans and analysts. Perhaps the most interesting application of win probability is its ability to identify these crucial situations. By calculating the difference in win probability between the current situation and the win probability that would result after a goal, one can identify these specific situations where the impact of scoring or conceding a goal would be much greater than in a typical situation [19]. It is a reasonable assumption that these situations correspond to the crucial moments of the game.

To illustrate this idea, we show how win probability can be used to identify clutch goal scorers. The number of goals scored is the most important statistic for offensive players in football. Yet, not all goals have the same value. A winning goal in stoppage time is clearly more valuable than another goal when the lead is already unbridgeable. By using the change in win probability[12] when a goal is scored, we can evaluate how much a player's goal contributions impact their team's chance of winning the game. This leads to the Added Goal Value metric

---

[10] https://www.thetimes.co.uk/article/weight-of-argentinas-collapse-is-too-much-for-even-messi-to-shoulder-wl7xbzd83

[11] https://api.sporza.be/permalink/web/articles/1538741367341

[12] We remove the pre-game strength from our win probability model for this analysis. Otherwise, games of teams such as PSG that dominate their league would all start with an already high win probability, reducing a goal's impact on the win probability.

below, similar to Pettigrew's added goal value for ice hockey [17].

$$\text{AGVp90}_i = \frac{\sum_{k=1}^{K_i} 3 * \Delta P(\text{win}|x_{t_k}) + \Delta P(\text{tie}|x_{t_k})}{M_i} * 90 \qquad (6)$$

where K_i is the number of goals scored by player i, M_i is the number of minutes played by that same player and t_k is the time at which a goal k is scored.

This formula calculates the total added value that occurred from each of player i's goals, averaged over the number of games played. Since both a win and a draw can be an advantageous outcome in football, we compute the added value as the sum of the change in win probability multiplied by three and the change in draw probability. The result can be interpreted as the average boost in expected league points that a team receives each game from a player's goals.

Figure 5 displays the relationship between AGVp90 and goals per game for the most productive Bundesliga, Ligue 1, Premier League, LaLiga and Serie A players who have played at least the equivalent of 20 games and scored at least 10 goals in the 2016/2017 and 2017/2018 seasons. The diagonal line denotes the average AGVp90 for a player with a similar offensive productivity. The players with the highest AGVp90 are Lionel Messi, Cavani, Balotelli, Kane and Giroud. Also, players such as Neymar, Lewandowski, Lukaku, Mbappé and Mertens have a relatively low added value per goal; while players such as Austin, Balottelli, Dybala, Gameiro and Giroud add more value per goal than the average player.

Fig. 5: The relation between goals scored per 90 minutes and AGVp90 for the most productive Bundesliga, Ligue 1, Premier League, LaLiga and Serie A players in the 2016/2017 and 2017/2018 seasons.

7   Conclusions

This paper introduced a Bayesian in-game win probability model for football. Our model uses eight features for each team and models the future number of goals that a team will score as a temporal stochastic process. Our evaluations indicate that the predictions made by this model are well calibrated and out-perform the typical modelling approaches that are used in other sports. The model has relevant applications in sports story telling and can form a central component in the analysis of player performance in the crucial moments of a game.

References

1. Burke, B.: Modeling win probability for a college basketball game. http://wagesofwins.com/2009/03/05/modeling-win-probability-for-a-college-basketball-game-a-guest-post-from-brian-burke/ (2009)
2. Burke, B.: (WPA)explained. http://archive.advancedfootballanalytics.com/2010/01/win-probability-added-wpa-explained.html (2010)
3. Buttrey, S.E., Washburn, A.R., Price, W.L.: Estimating NHL scoring rates. JQAS **7**(3) (2011)
4. Červenỳ, J., van Ours, J.C., van Tuijl, M.A.: Effects of a red card on goal-scoring in world cup football matches. Empirical Economics **55**(2), 883–903 (2018)
5. Commission, G.: In-play (in-running) betting: Position paper. https://gamblingcommission.gov.uk/pdf/In-running-betting-position-paper.pdf (2016)
6. Constantinou, A.C., Fenton, N.E.: Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models. JQAS **8**(1) (2012)
7. Ganguly, S., Frank, N.: The problem with win probability. In: Proc. of the 12th MIT Sloan Sports Analytics Conf. (2018)
8. Govan, A.Y., Langville, A.N., Meyer, C.D.: Offense-defense approach to ranking team sports. JQAS **5**(1) (2009)
9. Hvattum, L.M., Arntzen, H.: Using Elo ratings for match result prediction in association football. Int. J. Forecast **26**(3), 460–470 (2010)
10. Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., Blei, D.M.: Automatic differentiation variational inference. JMLR **18**(1), 430–474 (2017)
11. Lindholm, S.: Using WPA to measure pitcher effectiveness. https://www.beyondtheboxscore.com/2014/5/19/5723968/chicago-white-sox-chris-sale-wpa-major-league-baseball-pitcher-effectiveness (2014)
12. Lindsey, G.R.: The progress of the score during a baseball game. J. Am. Stat. Assoc **56**(295), 703–728 (1961)
13. Lock, D., Nettleton, D.: Using random forests to estimate win probability before each play of an NFL game. JQAS **10**(2) (2014)
14. McFarlane, P.: Evaluating NBA end-of-game decision-making. J. Am. Stat. Assoc **5**(1), 17–22 (2019)
15. Morris, B.: When To Go For 2, For Real. https://fivethirtyeight.com/features/when-to-go-for-2-for-real/ (2017)
16. Pelechrinis, K.: iWinRNFL: A Simple, Interpretable & Well-Calibrated In-Game Win Probability Model for NFL. arXiv:1704.00197 [stat] (2017)
17. Pettigrew, S.: Assessing the offensive productivity of NHL players using in-game win probabilities. In: Proc. of the 9th MIT Sloan Sports Analytics Conf. (2015)

18. Pro-Football-Reference: The P-F-R win probability model. https://www.pro-football-reference.com/about/win_prob.htm (2012)
19. Robberechts, P., Bransen, L., Van Haaren, J., Davis, J.: Choke or Shine? Quantifying Soccer Players' Abilities to Perform Under Mental Pressure. In: Proc. of the 13th MIT Sloan Sports Analytics Conf. (2019)
20. Robberechts, P., Davis, J.: Forecasting the FIFA World Cup–combining result-and goal-based team ability parameters. In: Machine Learning and Data Mining for Sports Analytics ECML/PKDD 2018 workshop. vol. 2284, pp. 52–66 (2018)
21. Rosenheck, D.: Tom Brady, 4.8-time Super Bowl champion. The Economist (2017)
22. Schneider, T.: What Real-Time Gambling Data Reveals About Sports: Introducing Gambletron 2000. http://www.gambletron2000.com/about (2014)
23. Stern, H.: On the probability of winning a football game. Am. Stat. **45**(3), 179–183 (1991)
24. Stern, H.S.: A brownian motion model for the progress of sports scores. J. Am. Stat. Assoc **89**(427), 1128–1134 (1994)
25. Štrumbelj, E., Vračar, P.: Simulating a basketball match with a homogeneous markov model and forecasting the outcome. Int. J. Forecast **28**(2), 532–542 (2012)

A   Game State Features

Besides the final set of features listed in Section 4.1, we considered a large list of additional game state features. These are listed below.

Offensive strength

- Offensive strength: A team's estimated offensive strength, according to the ODM model [8] (relative to the competition)
- Number of shots: The total number of shots a team took during the game.
- Number of shots on target: The total number of shots on target a team took during the game.
- Number of well positioned shots: The number of shots attempted from the middle third of the pitch, in the attacking third of the pitch.
- Opportunities: The total number of goal scoring opportunities (as annotated in the event data).
- Number of attacking passes: The total number of passes attempted by a team in the attacking third of the field.
- Attacking pass success rate: The percentage of passes attempted in the attacking third that were successful.
- Number of crosses: The total number of made crosses by a team.
- Balls inside the penalty box: Number of actions that end in the opponents penalty box.

Defensive strength

- Defensive strength: A team's estimated defensive strength, according to the ODM model (relative to the competition).
- Tackle success rate: The percentage of all attempted tackles that were successful.

Playing style

- Tempo: The number of actions per interval.
- Average team position: Average x coordinate of the actions performed.
- Possessions: The percentage of actions in which a team had possession of the ball during the game.
- Pass length: The average length of all the attempted passes by a team.
- Attacking pass length: The average length of all attacking passes attempted by a team.
- Percentage of backward passes: The percentage of passes with a backward direction.
- Number of attacking tackles: The total number of tackles attempted in the attacking third.
- Attacking tackle success rate: The percentage of tackles attempted in the attacking third that were successful.

Game situation

- Time since last goal: The number of time frames since the last scored goal.

B   World Cup Predictions

FiveThirthyEight's win probability estimates for the 2018 World Cup are the only publicly available[13] in-game win probability estimates for football. To compare our model against these predictions, we apply our Bayesian statistical model on the 48 games of the 2018 World Cup group stage using the event data provided by StatsBomb.[14]. We only look at the group stage, because the knockout stage does not allow for ties and has extra time. This would require a different modelling approach.
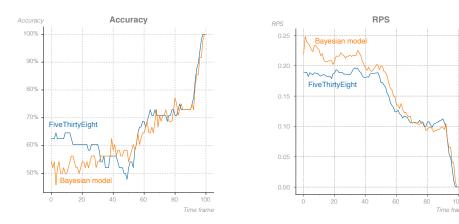

Fig. 6: A comparison between FiveThirtyEight's in-game predictions for the group stage of the 2018 World Cup and our Bayesian model.

Figure 6 compares both models in terms of accuracy and RPS. The differences can mainly be attributed to FiveThirtyEight's better pre-game prediction model. Their SPI ratings, which are the underlying basis for the pre-game projections, are based on a combination of national team results and an estimate of each team's individual player abilities using club-level performance. At least for the 2018 World Cup, these outperformed a simple Elo-based system [20]. Near the end of each game, when the importance of these pre-game ratings decreases, our model performs similarly to FiveThirtyEight's. Another interesting observation is that both RPS and accuracy have a sudden increase in the final 10% of the game, indicating that most games in the World Cup could go either way until the final stages of the game. For domestic league games, this increase in accuracy over time has a much more gradual increase.

---

[13] FiveThirtyEight's predictions can be found at https://projects.fivethirtyeight.com/2018-world-cup-predictions/matches/

[14] Data repository at https://github.com/statsbomb/open-data

C   "A 2-0 lead is the worst lead"

Win probability can be used to debunk some of the most persistent football myths, such as "2-0 is the worst lead", "10 do it better" and "No better moment to score a goal than just before half time". Perhaps not unsurprisingly, none of these appear to be true. In this section we focus on the "2-0 is the worst lead" myth, which we introduced in the introduction of the main article. To debunk this myth, we adapt the game state of all the games in our dataset such that the scoreline is 2-0 at each moment in the game. This allows us to capture a large range of possible game states (i.e., strong team leading, weak team leading, multiple players red carded, . . .). In Figure 7, we show the win probabilities for all resulting match scenarios. A 2-0 lead appears to be a very safe lead, much safer than a 1-0 lead or 2-1 lead. A team leading 2-0 has a minimal win probability of 70%, but in most scenarios (and especially in the second half) the probability is 90% of higher.
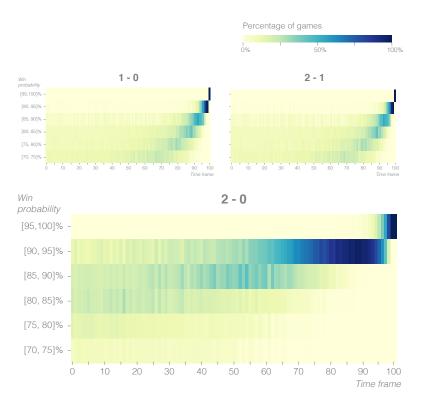

Fig. 7: "2-0 is the worst lead", at least according to the cliché . In reality, a 2-0 lead is much safer than a 1-0 or 2-1 lead, which are the true "worst leads". A 2-0 lead in the second half guarantees a win probability close to 90%.