Profiting from an inefficient association football gambling market: Prediction, risk and uncertainty using Bayesian networks.
Constantinou, AC; Fenton, NE; Neil, M

doi:10.1016/j.knosys.2013.05.008

For additional information about this publication click this link.
http://qmro.qmul.ac.uk/xmlui/handle/123456789/10778

Information about this research object was correct at the time of download; we occasionally make corrections to records, please therefore check the published record when citing. For more information contact scholarlycommunications@qmul.ac.uk
Profiting from an inefficient association football gambling market: Prediction, risk and uncertainty using Bayesian networks

Anthony Costa Constantinou *, Norman Elliott Fenton, Martin Neil

Risk & Information Management (RIM) Research Group, Department of Electronic Engineering and Computer Science, Queen Mary, University of London, London E1 4NS, United Kingdom

A R T I C L E   I N F O

Article history:
Received 4 December 2012
Received in revised form 20 May 2013
Accepted 21 May 2013
Available online 4 June 2013

Keywords:
Bayesian networks
Expert systems
Football betting
Football forecasts
Subjective information

A B S T R A C T

We present a Bayesian network (BN) model for forecasting Association Football match outcomes. Both objective and subjective information are considered for prediction, and we demonstrate how probabilities transform at each level of model component, whereby predictive distributions follow hierarchical levels of Bayesian inference. The model was used to generate forecasts for each match of the 2011/2012 English Premier League (EPL) season, and forecasts were published online prior to the start of each match. Profitability, risk and uncertainty are evaluated by considering various unit-based betting procedures against published market odds. Compared to a previously published successful BN model, the model presented in this paper is less complex and is able to generate even more profitable returns.

© 2013 The Authors. Published by Elsevier B.V. Open access under CC BY license.

1. Introduction

Association Football (hereafter referred to as simply football) is the most popular sport internationally [10,27,11], and attracts an increasing share of the multi-billion dollar gambling industry; particularly after its introduction online [6]. This is one of the primary reasons why we currently observe extensive attention paid to football odds by both academic research groups and industrial organisations who look to profit from potential market inefficiencies. While numerous academic papers exist which focus on football match forecasts, only a few of them appear to consider profitability as an assessment tool for determining a model’s forecasting capability.

Pope and Peel [30] evaluated a simulation of bets against published market odds in accordance with the recommendations of a panel of newspapers experts. They showed that even though there was no evidence of abnormal returns, there was some indication that the expert opinions were more valuable towards the end of the football season. Dixon and Coles [8] were the first to evaluate the strength of football teams for the purpose of generating profit against published market odds with the use of a time-dependent Poisson regression model based on Maher’s [26] model. They formed a simple betting strategy for which the model was profitable at sufficiently high levels of discrepancy between the model and the bookmakers’ probabilities. However, these high discrepancy levels returns were based on as low as 10 sample values; at lower discrepancy levels and with a larger sample size the model was unprofitable. The authors suggested that for a football forecast model to generate profit against bookmakers’ odds without eliminating the in-built profit margin, “it requires a determination of probabilities that is sufficiently more accurate from those obtained by published odds”. A similar paper by Dixon and Pope [9] was also published on the basis of 1993–1996 data and reported similar results. Rue and Salvesen [32] suggested a Bayesian dynamic generalised linear model to estimate the time-dependent skills of all the teams in the English Premier League (EPL) and English Division 1. They assessed the model against the odds provided by Intertops, a firm which is located in Antigua in the West Indies, and demonstrated profits of 39.6% after winning 15 bets out of a total of 48 for EPL matches, and 54% after winning 27 bets out of a total of 64 for Division 1 matches.

In an attempt to exploit the favourite-longshot bias for profitable opportunities, Poisson and Negative Binomial models have been used to estimate the number of goals scored by a team [3]. The conclusion was that even though the fixed odds offered against particular score outcomes did seem to offer profitable betting opportunities in some cases, these were few in number. Goddard and Asimakopoulos [17] proposed an ordered probit regression model to forecast EPL match results in an attempt to test the
weak-form efficiency of prices in the fixed-odds betting market. To evaluate the model they considered seasons 1999 and 2000. Even though they reported a loss of $-10.5\%$ for overall performance, the model appeared to be profitable (on a pre-tax gross basis) at the start and at the end of every season.\footnote{Gross pre-taxed returns of $+3.1\%$ and $+1.5\%$ for respective seasons beginning 1999 and 2000, and gross returns of $+8\%$ for respective seasons ending 1999 and 2000.} Using a benchmark statistical model with a large number of quantifiable variables relevant to match outcomes Forrest et al. \cite{15} examined the effectiveness of forecasts based on published odds and forecasts generated. They considered five different bookmaking firms for five consecutive seasons (1998–2003) and demonstrated that the model generated negative returns ranging from $-10\%$ to $-12\%$ depending on the bookmaking firm, but the loss was reduced to $-6.6\%$ when using the best available odds by exploiting arbitrage between bookmaking firms.

\cite{19} attempted to investigate the rationality of bookmakers’ odds using an ordered probit model to generate predictions for EPL matches. By considering William Hill odds, they followed the betting strategy introduced in \cite{8,9} and reported negative returns ranging from $-2.5\%$ to $-15\%$ for all discrepancy levels during seasons 2004–2006. In the absence of any consistently successful model against market odds, the authors claimed that “if it was successful, it would not have been published”.\cite{21} considered the ELO rating system for football match prediction, although it was initially developed by \cite{12} for assessing the strength of international chess players. Even though the ratings appeared to be useful in encoding the information of past results for measuring the strength of a team, resulting forecasts reported negative expected returns against numerous seasons of published odds using various betting strategies. However, Constantinou and Fenton \cite{5} later developed a novel rating technique (called pi-rating) that outperformed considerably the two ELO rating variants of \cite{21}, in terms of profitability, over a period of five EPL seasons.

\cite{7} recently presented a Bayesian network model that was used to generate forecasts about the EPL matches during season 2010/2011, by considering both objective and subjective information for prediction. Forecasts were published online \cite{29} prior to the start of each match, and this was the first academic study to demonstrate profitability that was consistent against published market odds over a sufficiently high number of betting trials without eliminating the bookmakers’ profit margin.

In this paper we present a Bayesian network model for forecasting football outcomes that is based on the approach in \cite{7}, but with reduced complexity and higher forecasting capability (which we explain in detail in Sections 2–4). The paper is organised as follows: Section 2 describes the model; Section 3 presents the various betting procedures along with a Bayesian network component for assessing the risks involved under each of the procedures; Section 4 discusses the results; Section 5 provides our concluding remarks.

### 2. The model

In this section we first provide a brief overview of the model summarising the main differences to the approach in \cite{7}. We then describe the technical components of the model in subsections.

As in \cite{7} we have used the AgenaRisk Bayesian network tool to build the model. The most important differentiator between AgenaRisk and other Bayesian network tools is its ability to properly incorporate continuous variables, without any constraint, and without the need for static discretisation. It does this through

<table>
<thead>
<tr>
<th>$S_r$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>...</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
</table>

**Table 1** How $S \rightarrow S_r$ is defined in 14 predetermined ranks, based on \cite{7}.

![Fig. 1. Simplified model topology of the overall Bayesian network.](image)
Hence, the underlying formulations on the basis of intervals can select the number of iterations and convergence criteria, and discretisations. In the implementation of the algorithm the user can select the number of iterations and convergence criteria, and its revolutionary dynamic discretisation algorithm that produces results with far greater accuracy than is possible otherwise. The dynamic discretisation algorithm uses entropy error as the basis for approximation. In outline, the algorithm follows these steps:

1. Convert the BN to a Junction Tree (JT) and choose an initial discretisation for all continuous variables.
2. Calculate the Node Probability Table (NPT) of each node given the current discretisation.
3. Enter evidence and perform global propagation on the JT, using standard JT algorithms.
4. Query the BN to get posterior marginals for each node, compute the approximate relative entropy error, and check if it satisfies the convergence criteria.
5. If not, create a new discretisation for the node by splitting those intervals with highest entropy error.
6. Repeat the process by recalculating the NPTs and propagating the BN, and then querying to get the marginals and then split intervals with highest entropy error.
7. Continue to iterate until the model converges to an acceptable level of accuracy.

This dynamic discretisation approach allows more accuracy in the regions that matter and incurs less storage space over static discretisations. In the implementation of the algorithm the user can select the number of iterations and convergence criteria, and hence can go for arbitrarily high precision (at the expense of increased computation times). Details about the role of qualitative judgments and how inference is done are provided in [13,28].

The model is constructed on the basis of three generic factors: team strength, form, and fatigue with motivation. There are model components corresponding to each of the three factors. The components are inferred hierarchically and at each level of hierarchy a match forecast is generated. This helps us understand how the probabilities transform at each level and determine the effectiveness of each model component by assessing the probability distributions generated at each level. Specifically:

1. At level 1, match forecasts of type \( \{ p(H), p(D), p(A) \} \) are generated based on each team's strength \((S)\), where an \( S \) prior is formulated according to observed and expected results \((P)\) of relevant match instances of the current season, and team inconsistencies \((I)\) given relevant final league point totals from the five most recent seasons;
2. At level 2, posterior predictive distributions of \( S \) (from level 1) are formulated based on team form \((F)\);
3. At level 3, posterior predictive distributions of \( S \) (from level 2) are formulated based on team fatigue and motivation \((M)\).

Thus, the model follows hierarchical levels of Bayesian inference such that \( S_1 \rightarrow S_2 \rightarrow S_3 \), where \( S_1 = p(S|P, I) \), \( S_2 = p(S|S_1, F) \), and \( S_3 = p(S|S_2, M) \).

The variable \( S \) is a \( \text{Normal}(\mu, \sigma, 0, 114)^3 \) probability density function and at each level of hierarchy represents a prediction of the team's strength which is measured in total league points. The distribution of \( S \) is summarised in 14 predetermined ranks \((S_n)\) as presented in Table 1, whereby the granularity of the 14 ranks ensures that, for any match combination of parameters \( S_r \), sufficient data points exist for a reasonably well informed match forecast prior. In particular, match forecasts given \( S_r \) are formulated on the basis of relevant historical match outcomes.\(^4\) Hence, the underlying

---

\(^1\) Corresponding to home win, draw, and away win.
\(^2\) Truncated Normal where the endpoints are the respective minimum and maximum number of points a team can accumulate in an EPL season (38 games with three points for a win).
\(^3\) The database consists of the home, draw and away results of all the EPL matches from season 1993/1994 to 2010/2011 inclusive (a total of 6624 occurrences). This information is available online at [14].
approach generates forecasts that are ‘anonymous’ in the sense that historical outcomes are not restricted by the name of the team. For example, given a match between Manchester United and Newcastle United, and assuming their respective $S$ values are $85(S_E = 2)$ and $62(S_E = 7)$, the resulting forecasts will represent: “a team with a probability density function $S(S_E)$ and a maximum likelihood estimation of $85(2)$ plays against a team with a probability density function $S(S_E)$ and a maximum likelihood estimation of $62(7)$” instead of: “Man United plays against Newcastle”. Accordingly, a team’s $S$ distribution varies throughout the season, and it is possible for teams to share similar such distributions at certain periods throughout the season.

Fig. 4 illustrates a simplified model topology of the overall Bayesian network model. Fig. G.1 presents the actual outcomes of the Arsenal vs. Liverpool match as forecasted on August 20th 2011 and demonstrates how match forecasts transform on the basis of hierarchical posterior predictive distributions of $S$ beliefs. The observed outcome was $A$ (score was 0–2). Table G.1 provides a brief description of all the model parameters.

The primary differences with the BN model proposed in [7] are:

1. $P$, which formulates the prior predictive distribution of $S$, is now measured using a straightforward $Beta – Binomial$ approach (which we describe in detail in Section 2.1 below), rather than the complex non-symmetric Bayesian parameter learning approach;

2. Model components which correspond to each of the generic factors have been both decreased in number and simplified in an attempt to reduce model complexity. Specifically:

(a) The number of variables in component $F$ has decreased from 21 (10 for each team plus one representing discrepancy) to 10 (5 for each team). In particular, instead of providing distinct subjective beliefs about the availability of the Primary Key Player, the Secondary Key Player, the Tertiary Key Player, and the Remaining First Team Players, we now introduce a single subjective variable called $Availability$ of players who resulted in current form. Further, there is no $Home$ Form and $Away$ Form, but rather a single variable called $Form$ which represents the most recent (and overall) form of a team. This variable is taken into consideration when playing both home and away. As we demonstrate in Section 4.2 below, this modification not only simplified the model, but also resulted in notably increased profitability.

(b) The two previously proposed model components of $Fatigue$ and $Psychological$ Impact have been merged into the single component $M$, and the number of variables (which correspond to each of the two competing teams) has been decreased from a total of 28 to 18 (9 for each team). In particular, instead of requiring indications about the number of first team players rested during the last match in an attempt to measure Restness, which is later used to diminish tiredness, we now directly provide this information in a single subjective variable called $Toughness$ of previous match (i.e. the subjective indication of toughness will be lower if we already know that some first team players were rested). Further, the beliefs regarding managerial impact, team spirit, motivation, and the expert’s degree of certainty with regards to their subjective indications are now replaced by a single subjective variable called $Motivation$ (this follows the same rational as with the toughness of previous match).

3. In the previous model, the subjective components were used to directly revise the probability distribution of match forecasts in an ordinal manner (i.e. a function was used to skew the predictive distribution of a match forecast towards a home win or an away win based on subjective proximity). Instead of modifying the match forecast directly, we now let the subjective components modify each team’s distribution $S$, which serve as parents for formulating match forecasts. Thus, match forecasts are now always formulated based on relevant historical data (i.e. no skewness), but given modified distributions of $S$ as a consequence of one or more positive or negative subjective indications. This change does not necessarily reduce model complexity, but rather improves model sophistication and thus, forecasting capability.

4. In the previous model, the values of each of the subjective components (i.e. form, fatigue and psychological impact) were compared between the two teams, and a revision in prediction was only made on the basis of discrepancies between the components (e.g. the team with better form received an
increased probability to win). This implies that match forecasts are not revised if both teams have high (or low) levels of form (i.e. no discrepancy). The problem with this is that, if we assume that high levels of form increase a team’s distribution $S$ by 20 points, there is still a difference between the match $S_{\text{HOME}} = 60$ versus $S_{\text{AWAY}} = 30$, and another $S_{\text{HOME}} = 80$ versus $S_{\text{AWAY}} = 50$. Consequently, in the new model we do not perform comparisons when it comes to subjective components, but we instead allow all of the components to have a direct impact on each team’s distribution $S$. Again, this change does not necessarily reduce model complexity but rather improves model sophistication.

5. The impact of each model component is now inferred hierarchically; implying that model components now follow a nonlinear weighting when revising distribution $S$ (i.e. model components computed first have less impact); in contrast to the previous model, where the three subjective components had identical impact on match forecasts. The hierarchical computation also decreases the time required to calculate posterior probabilities.

2.1. Level 1 component: team performance ($P$) and inconsistency ($I$)

At level 1, $P$ is modelled using the Beta – Binomial approach. The Beta distributions serve as conjugate distributions of the Binomial distributions, formulating a compound distribution such that the $p$ parameter of the Binomial distribution is being randomly drawn from the Beta distribution. In our model, the Beta distributions $p(Win), p(Draw)$ and $p(Lose)$ (with hyperparameters $\alpha$ and $\beta$ priors based on relevant historical data), serve as the $p$ parameters of the Binomial distributions $Number\ of\ Wins, Number\ of\ Draws$, and $Number\ of\ Loses$ respectively.\(^{7}\)

The posterior Beta distributions are then used to estimate team expectations for the residual match instances of the current season. These team expectations allow expert modifications based on subjective beliefs regarding the difficulty of residual opponents (this indication allows the expert to diminish the bias in cases where the results are formulated on the basis of mostly poor/high quality opponents). Observed and expected cumulative match points are then considered for formulating the prior distributions of $S$. This is the first part (out of two) of level 1. The Bayesian network component $P$ is illustrated in Fig. 2, where:

(a) the variables $p(Win), p(Draw)$ and $p(Lose)$ are the Beta distributions. For example, in the case of $p(Win)$ the hyperparameters are $\sim Beta(\omega + 1, d + 1 + 2)\(^6\)$ where $\omega$ is the number of previous season’s wins, $d$ is the number of draws, $I$ is the number of losses, and values 1 and 2 are introduced for minimal Laplacian smoothing so that we avoid overfitting by ensuring that posterior parameters $\alpha$ and $\beta$ are positive for all teams;

(b) the variables $Number\ of\ Wins, Number\ of\ Draws$ and $Number\ of\ Loses$ are $\sim Binomial(n, p)$. For example in the case of $Number\ of\ Wins$, $n$ is the number of matches played during the current season and $p$ is the probability of success for each trial (for this example $p$ is the Beta distribution $p(win)$);

(c) the variable $Expected\ Residual\ Points(\pi_p)$ represents the points a team expects to accumulate over the current season’s residual match instances and hence, $\pi_p$ is dependent on the $Number\ of\ residual\ matches$ and the posterior Beta beliefs of $p(Win), p(Draw)$ and $p(Lose)\(^7\)$;

(d) the variable ERP given opponent difficulty ($\pi_p$) is a $\pi_p$ posterior given the Difficulty of residual opponents($\psi$), whereby $\pi_p$ may receive adjustments for up to $\pm 10\%$ based on a 7-level subjective belief, and it is defined as the case function of:

---

6 Hyperparameters are provided as node-inputs and are not shown in Fig. 2.

7 We do not perform convolution but we instead perform aggregation of averages (which means that the variance might be overestimated) in order to keep the complexity of the model at significantly lower levels.
Table 3
Unit based profit and profitability rates based on BP1 and BP2.

<table>
<thead>
<tr>
<th>Discrep. levels (%)</th>
<th>Betting Procedure 1 (BP1)</th>
<th>Betting Procedure 2 (BP2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bets/trials</td>
<td>Win rate (%)</td>
</tr>
<tr>
<td>0</td>
<td>379</td>
<td>34.30</td>
</tr>
<tr>
<td>1</td>
<td>359</td>
<td>34.54</td>
</tr>
<tr>
<td>2</td>
<td>316</td>
<td>34.49</td>
</tr>
<tr>
<td>3</td>
<td>272</td>
<td>34.19</td>
</tr>
<tr>
<td>4</td>
<td>227</td>
<td>35.24</td>
</tr>
<tr>
<td>5</td>
<td>193</td>
<td>35.23</td>
</tr>
<tr>
<td>6</td>
<td>168</td>
<td>35.71</td>
</tr>
<tr>
<td>7</td>
<td>144</td>
<td>37.50</td>
</tr>
<tr>
<td>8</td>
<td>129</td>
<td>38.76</td>
</tr>
<tr>
<td>9</td>
<td>107</td>
<td>37.38</td>
</tr>
<tr>
<td>10</td>
<td>97</td>
<td>39.18</td>
</tr>
<tr>
<td>11</td>
<td>84</td>
<td>35.71</td>
</tr>
<tr>
<td>13</td>
<td>53</td>
<td>30.19</td>
</tr>
<tr>
<td>14</td>
<td>38</td>
<td>34.21</td>
</tr>
<tr>
<td>15</td>
<td>33</td>
<td>36.36</td>
</tr>
</tbody>
</table>

Fig. 7. Unit-based returns based on BP1 and BP2, and according to the specified level of discrepancy.

The component inconsistency (I) approximates a team’s inconsistency based on respective league point totals over the five most recent seasons, and the resulting variance is added to the prior predictive distribution of S and together formulate S1, Fig. 3 presents a naive parameter learning procedure for approximating a team’s inconsistency, where:

(a) the variables Season Y1 to Y5 are $\sim T\text{Normal}(\mu, \sigma^2, 0, 114)$;
(b) the variable Inconsistency (Variance) (V) is a $\sim$Uniform (0, 150)$^8$ and serves as the input $\sigma^2$ for the $T\text{Normal}$ distributions of (a) above;
(c) the variable Overall Performance is a $\sim$Uniform (0, 114) and serves as the input $\mu$ for the $T\text{Normal}$ distributions of (a) above.

Moreover, the variable Confidence in Historical Inconsistency (C), presented in Figs. 1 and G.1, gives the option to the expert to diminish the additional variance V if the expert feels that the team is not currently as inconsistent as it used to be over the period of the last five seasons. The case function below shows how V, which serves as an input for $S_{11}$, can diminish in value based on the subjective indication of C:

$$ S_{11} = \begin{cases} T\text{Normal}(S, \frac{V}{T}, 0, 114), & S, C = \text{Low} \\ T\text{Normal}(S, \frac{V}{T}, 0, 114), & S, C = \text{Medium} \\ T\text{Normal}(S, V, 0, 114), & S, C = \text{High} \end{cases} $$

2.2. Level 2 component: team form (F)

At level 2 posterior predictive distributions of $S_{12}$ are formulated given $S_{11}$ and a posterior team-form ($\Phi$), as presented in Figs. 4 and G.4, where $\Phi$ is a continuous variable on a scale that goes from 0 to 1. A value close to 0.5 suggests that the team is performing as expected, whereas a higher value indicates that the team is performing better than expected (and vice versa). The expectations are determined by the forecasts generated by this model, and $\Phi$ is measured on the basis of the five most recent gameweeks.$^9$

The $\Phi$ posterior is formulated hierarchically based on the Availability of players who resulted in current form ($L_A$) and the Important players return ($L_k$), where both variables follow ordinal scale distributions with subjective indications as illustrated in Figs. 4 and G.4 and the case functions below. The variable Expected Form given player availability ($\Phi_{L_A}$) is the case function:

$$ \Phi_{L_A} = \begin{cases} T\text{Normal}(\Phi \times 0.0001, 0.1), & \Phi, L_A = \text{VeryHigh} \\ T\text{Normal}(\Phi \times 0.001, 0.1), & \Phi, L_A = \text{High} \\ T\text{Normal}(\Phi \times 0.01, 0.1), & \Phi, L_A = \text{Medium} \\ T\text{Normal}(\Phi \times 0.01, 0.1), & \Phi, L_A = \text{Low} \\ T\text{Normal}(\Phi \times 0.02, 0.05, 1), & \Phi, L_A = \text{Very Low} \end{cases} $$

and the variable Expected form given further important players ($\Phi_{L_k}$) is the case function:

$^8$ Upper bound is 150 rather than 114 to account for the limited number of parameters learned.

$^9$ A complete EPL season consists of 38 gameweeks.
2.3. Level 3 component: fatigue and motivation (M)

At level 3 posterior predictive distributions of $S_{L3}$ are formulated given $S_{L2}$ and $U_{LR}$ as presented in Fig. 5. A Prior Fatigue ($G_p$) is first measured given $EU$ match Involvement ($E$) (which represents team involvement in European tournaments) and Toughness of previous match ($T$), where $E$ and $T$ follow ordinal scale distributions with subjective indications as illustrated in Figs. 5 and G.5, and the case function of $G_p$ below:

$$G_p = \begin{cases} 
\text{TNormal}(T, 0.001, 0.1), & T, E = \text{None} \\
\text{TNormal}(T + (1 - T) \times \frac{1}{2}), 0.001, 0.1), & T, E = \text{Very Low} \\
\text{TNormal}(T + (1 - T) \times \frac{2}{3}), 0.001, 0.1), & T, E = \text{Low} \\
\text{TNormal}(T + (1 - T) \times \frac{2}{3}), 0.001, 0.1), & T, E = \text{Medium} \\
\text{TNormal}(T + (1 - T) \times \frac{2}{3}), 0.001, 0.1), & T, E = \text{High} \\
\text{TNormal}(T + (1 - T) \times \frac{2}{3}), 0.001, 0.1), & T, E = \text{Very High} 
\end{cases}$$

The Expected Fatigue ($G_e$) is a posterior $G_p$ value which diminishes on the basis of Days Gap since previous match ($\delta$), and increases with National Team Involvement ($\lambda$), where $\delta$ and $\lambda$ are ordinal scale...
distributions with subjective indications as illustrated in Figs. 5 and G.5, and the case function of $G_p$ below:

$$G_p = \begin{cases} 
\text{TNormal}((G_p - G_p \times \delta), 0.001, 0.1), & G_p, \delta, \lambda = \text{None} \\
\text{TNormal}((G_p - G_p \times \delta) + (1 - (G_p - G_p \times \delta)) \times 0.1), 0.001, 0.1), & G_p, \delta, \lambda = \text{Low} \\
\text{TNormal}((G_p - G_p \times \delta) + (1 - (G_p - G_p \times \delta)) \times 0.2), 0.001, 0.1), & G_p, \delta, \lambda = \text{Medium} \\
\text{TNormal}((G_p - G_p \times \delta) + (1 - (G_p - G_p \times \delta)) \times 0.3), 0.001, 0.1), & G_p, \delta, \lambda = \text{High} \\
\text{TNormal}((G_p - G_p \times \delta) + (1 - (G_p - G_p \times \delta)) \times 0.4), 0.001, 0.1), & G_p, \delta, \lambda = \text{Very High} 
\end{cases}$$

Fig. 11. Cumulative unit-based returns based on $BP_3$, for component levels 1, 2 and 3.

Fig. 12. Cumulative unit-based returns based on $BP_4$, for component levels 1, 2 and 3.

Table 5
Team-based returns relative to overall returns for the specified betting procedure.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team</th>
<th>Betting procedure</th>
<th>1 (%)</th>
<th>2 (%)</th>
<th>3 (%)</th>
<th>4 (%)</th>
<th>5.1 (%)</th>
<th>5.2 (%)</th>
<th>5.3 (%)</th>
<th>5.4 (%)</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Man Utd</td>
<td>-37.57</td>
<td>-14.46</td>
<td>21.83</td>
<td>24.01</td>
<td>6.35</td>
<td>6.34</td>
<td>8.05</td>
<td>6.47</td>
<td>-3.37</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Arsenal</td>
<td>111.74</td>
<td>49.49</td>
<td>68.91</td>
<td>59.98</td>
<td>4.82</td>
<td>4.93</td>
<td>7.18</td>
<td>16.71</td>
<td>40.47</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Tottenham</td>
<td>-23.84</td>
<td>15.97</td>
<td>32.22</td>
<td>8.78</td>
<td>12.14</td>
<td>12.07</td>
<td>12.39</td>
<td>9.77</td>
<td>9.69</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Newcastle</td>
<td>76.20</td>
<td>19.77</td>
<td>83.19</td>
<td>39.44</td>
<td>10.43</td>
<td>10.33</td>
<td>13.22</td>
<td>19.25</td>
<td>33.98</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Chelsea</td>
<td>-97.38</td>
<td>-9.16</td>
<td>-108.64</td>
<td>-112.74</td>
<td>11.51</td>
<td>11.80</td>
<td>9.60</td>
<td>3.11</td>
<td>-36.49</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Everton</td>
<td>-32.39</td>
<td>-12.66</td>
<td>-27.98</td>
<td>-30.82</td>
<td>13.82</td>
<td>13.82</td>
<td>12.75</td>
<td>9.55</td>
<td>-6.74</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Arsenal</td>
<td>175.87</td>
<td>76.32</td>
<td>192.25</td>
<td>237.84</td>
<td>27.83</td>
<td>27.69</td>
<td>29.59</td>
<td>36.40</td>
<td>106.47</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Fulham</td>
<td>25.18</td>
<td>17.84</td>
<td>-10.08</td>
<td>7.17</td>
<td>5.66</td>
<td>5.94</td>
<td>5.30</td>
<td>7.18</td>
<td>1.73</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>West Brom</td>
<td>62.23</td>
<td>8.55</td>
<td>23.44</td>
<td>31.22</td>
<td>14.67</td>
<td>14.67</td>
<td>15.96</td>
<td>21.00</td>
<td>11.15</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Swansea</td>
<td>59.54</td>
<td>2.68</td>
<td>-7.67</td>
<td>7.64</td>
<td>7.29</td>
<td>7.09</td>
<td>6.31</td>
<td>6.29</td>
<td>6.92</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Norwich</td>
<td>-55.93</td>
<td>2.45</td>
<td>-47.79</td>
<td>-32.66</td>
<td>7.72</td>
<td>7.85</td>
<td>6.92</td>
<td>4.51</td>
<td>-13.39</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Stoke</td>
<td>16.79</td>
<td>36.62</td>
<td>15.39</td>
<td>-12.31</td>
<td>6.88</td>
<td>7.24</td>
<td>7.41</td>
<td>5.75</td>
<td>10.47</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Wigan</td>
<td>-211.84</td>
<td>4.38</td>
<td>3.66</td>
<td>95.50</td>
<td>9.06</td>
<td>9.22</td>
<td>7.78</td>
<td>8.09</td>
<td>1.98</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Aston Villa</td>
<td>-70.95</td>
<td>20.29</td>
<td>-25.35</td>
<td>-20.34</td>
<td>7.23</td>
<td>7.73</td>
<td>6.39</td>
<td>4.33</td>
<td>-8.83</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>QPR</td>
<td>128.59</td>
<td>17.80</td>
<td>59.61</td>
<td>91.06</td>
<td>4.88</td>
<td>4.69</td>
<td>6.01</td>
<td>19.33</td>
<td>41.50</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Bolton</td>
<td>29.70</td>
<td>2.62</td>
<td>5.47</td>
<td>-2.27</td>
<td>7.36</td>
<td>7.25</td>
<td>7.65</td>
<td>9.16</td>
<td>8.37</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Blackburn</td>
<td>-5.84</td>
<td>-24.90</td>
<td>-33.66</td>
<td>-52.58</td>
<td>11.20</td>
<td>10.95</td>
<td>9.99</td>
<td>3.39</td>
<td>-10.68</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Wolves</td>
<td>59.87</td>
<td>15.64</td>
<td>-2.34</td>
<td>-29.65</td>
<td>16.73</td>
<td>16.62</td>
<td>15.45</td>
<td>8.03</td>
<td>12.55</td>
<td></td>
</tr>
</tbody>
</table>
Finally, $G_e$ is revised into Fatigue and Motivation ($G$) given Motivation $(\kappa)$ and Head-to-Head Bias($\omega$), where $\kappa$ and $\omega$ follow ordinal scale distributions that go from 0 to 1 with subjective indications as illustrated in Figs. 5 and G.5, and the case function of $G$ below:

$$G = \begin{cases} 
T \text{Normal}(\frac{\omega}{\kappa}, 0.01, 0.1), & \kappa, \omega, G_e = \text{Very Rested} \\
T \text{Normal}(\frac{\omega}{\kappa} \times 0.9), 0.01, 0.1), & \kappa, \omega, G_e = \text{Rested} \\
T \text{Normal}(\frac{\omega}{\kappa} \times 0.8), 0.01, 0.1), & \kappa, \omega, G_e = \text{Normal} \\
T \text{Normal}(\frac{\omega}{\kappa} \times 0.7), 0.01, 0.1), & \kappa, \omega, G_e = \text{Tired} \\
T \text{Normal}(\frac{\omega}{\kappa} \times 0.6), 0.01, 0.1), & \kappa, \omega, G_e = \text{Very Tired} 
\end{cases}$$

### 3. Forecast performance based on profitability and risk

In this section we describe how the forecasting capability of the model was assessed on the basis of profitability and relevant risks involved. Profitability is measured on the basis of a set of predetermined betting procedures. For market odds we have considered the odds with the highest payoff as recorded by [14] for the matches of the EPL season 2011/2012. The number of bookmaking firms considered for recording maximums ranged from 26 to 49 per match instance.\(^{10}\)

 Naturally, the performance of a football forecast model is determined by its ability to generate profit against market odds. However, many researchers also consider (or solely focus) on various scoring rules for this purpose in an attempt to determine the accuracy of the forecasts against the observed results [8,32,20,16,24,16,15,22,19,21]. Forecast assessments based on scoring rules have been heavily criticised because different rules may provide different conclusions about the forecasting capability of football forecast models [4]. Furthermore, in financial domains researchers have already demonstrated a weak relationship between various accuracy and profit measures [25], whereas [40] suggested that it might be best to combine accuracy and profit measures for a more informative picture.

In this paper we are interested in the profitability of the model relative to market odds. For this to happen, market odds have to be sufficiently less accurate (or inefficient) relative to those generated by our model so that the bookmakers’ profit margin, where present, can be overcome. The bookmakers’ profit margin, sometimes also called ‘over-round’, refers to the margin by which the sum of published market probabilities of the total outcomes exceeds 1. For example, if the true (i.e. the initially measured) probabilities for a match instance are $p(\text{Win}) = 0.50$, $p(\text{Draw}) = 0.25$ and $p(\text{Lose}) = 0.25$, a bookmaker’s published probabilities might be $p(\text{Win}) = 0.52$, $p(\text{Draw}) = 0.26$ and $p(\text{Lose}) = 0.26$ (which result in lower odds for payoff) and hence, the sum of published probabilities exceeds 1. The bookmaker’s profit margin here is simply $(p(\text{Win}) + p(\text{Draw}) + p(\text{Lose})) − 1$; which in this case would be 4%.

Since profitability is not only dependent on the forecasting capability of a model relative to market odds but also on the specified betting methodology, we have introduced an array of such betting procedures. For each procedure, we introduce sensible modifications relative to the standard betting strategy that was proposed and considered by the vast majority of the previous relevant published papers, whereby a bet is placed when expectations exceed a predetermined level [30,8,32,9,17,15,19,21,7].

#### 3.1. Defining profitability

We measure the profitability on the basis of the quantity of profit (or net profit which is stated as unit-based returns), rather than on the basis of percentage returns relative to respective stakes. The example below illustrates the rationale behind our preference.

**Example:** Suppose we have two football forecast models $\alpha$ and $\beta$. We want to compare their performance on the basis of profitability given the set of five match instances $\{M_1, M_2, M_3, M_4, M_5\}$. Table 2 presents a hypothetical betting performance between the two models over those match instances.

<table>
<thead>
<tr>
<th>Model</th>
<th>Bets</th>
<th>Win (%)</th>
<th>P/L (Units)</th>
<th>Profit Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\beta$</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

After considering the five match instances we observe the following results\(^{11}\):

- Model $\alpha$ suggested two bets and both were successful (100% winning rate), returning a net profit of £200 which represents a profit rate of 100% relative to total stakes.
- Model $\beta$ suggested five bets and four of them were successful (80% winning rate), returning a net profit of £300 which represents a profit rate of 60% relative to total stakes.

\(^{10}\) Betfair odds are not considered within the dataset since Betfair is a betting exchange company whereby published odds constantly fluctuate. These odds are normally the best possible odds (i.e. with the highest payoff) a bettor can find online. However, unlike traditional bookmakers Betfair will deduct a fixed % from your winnings which ranges from 2% to 6% depending on membership status [2].

\(^{11}\) For simplification we assume identical stakes (£100) and odds for payoff (even; or 2.00 in decimal form).
An evaluation based on the percentage profit rates would have erroneously considered model $\beta$ as being inferior at picking winners than model $\alpha$. But, such an evaluation fails to consider the possibility that model $\alpha$ might have failed to discover potential advantages against the market for all of the match instances. The reality is that model $\beta$ managed to simulate riskier bets that reduced the percentage rates of winning and profit, but increased net profit due to the larger number of successful bets.

We have to choose which model is best to follow; model $\alpha$ with a higher winning rate on bets and a higher profit rate between stakes and returns, or model $\beta$ with a higher (33.33%) net profit? If the ultimate aim is to make money, then every bettor would have preferred model $\beta$ over model $\alpha$ for betting against the market. Therefore, we suggest that a bettor should be increasing net profit rather than establishing good winning percentage rates, and for this to happen a bettor is expected to consider all of his advantages presented at every match instance rather than choosing the ‘best’ of his advantages that occasionally arise.

Consequently, in this paper we measure profitability on unit-based returns (net profit) over $n$ match instances (in our case $n = 380$, the total number of matches played in the EPL season of 2011/2012). The betting procedures are defined in the following section.

3.2. Defining the betting procedures

We define the following set of betting procedures for evaluating the profitability of the model against the market:

1. ($BP_1$): For each match instance, place a fixed bet equal to a single unit on the outcome with the highest absolute percentage discrepancy, where the model predicts the higher probability, if and only if the discrepancy is $\geq n\%$ (where $n$ is an integer $0 \leq n \leq 15$);

2. ($BP_2$): For each match instance, place a fixed bet equal to a single unit on every outcome the model predicts with higher probability, if and only if the absolute discrepancy is $\geq n\%$;

3. ($BP_3$): For each match instance, place a bet equal to $U$ units for each outcome the model predicts with higher probability, where the stake of the bet is a real number equal to the absolute discrepancy percentage between outcomes multiplied by $U$ (e.g. if an absolute discrepancy of 4.45% and 1.17% is observed for outcomes $H$ and $D$ respectively while $U = 1$, then bets of £4.45 and £1.17 are simulated for a home win and a draw respectively);

4. ($BP_4$): For each match instance, place a bet equal to $U$ units for each outcome the model predicts with higher probability, where the stake of the bet is a real number equal to the relative discrepancy percentage between outcomes multiplied by $U$ (e.g. if a relative discrepancy of 4.45% and 1.17% is observed for outcomes $H$ and $D$ respectively while $U = 1$, then bets of £4.45 and £1.17 are simulated for a home win and a draw respectively);

5. ($BP_{5,1}$, $BP_{5,2}$, $BP_{5,3}$, $BP_{5,4}$): These apply only to match instances where arbitrage opportunities are discovered. Repeat 1, 2, 3 and 4 but substitute the betting procedure with arbitrage bets whereby the total amount of the three bets is equal to the bankroll available at that time (a bankroll specification is required prior to initialising the betting simulation, and tests are performed for different bankroll values).

If a betting procedure $A$ indicates higher profitability than another $B$ over a fixed number of match instances, it does not necessarily suggest that we should always choose $A$ over $B$. This is true if we are also interested in the risks involved and the level of uncertainty over the posterior predicted distribution of unit-based returns (i.e. the magnitude of potential losses and winnings as well as the probability associated with such events). Accordingly, we have constructed a simple Bayesian network component (Fig. 6) that measures the risk of ending with less than, or equal to, a specified number of units over a specified number of match instances. Fig. 6 illustrates, as an example, the risk of ending with $U \leq 0$ after bets are simulated (given $BP_1$ at discrepancy levels of 0%) on the 380 match instances. This assumes relevant model performances as demonstrated in Section 4 below. In particular,

(a) the variable Match Instances represents the number of match instances over which the risk is measured;
(b) the variables $p$ (profitable) and $p$ (unprofitable) are Beta distributions with $alpha$ and $beta$ hyperparameters representing the probability to profit (and not to profit) for each match instance simulated;
(c) the variables Estimated Unprofitable Instances and Estimated Profitable Instances are Binomial distributions with $n$ number of trials equal to (a) above, where input $p$ is the respective Beta distribution of (b) above;
(d) the variables Profit Rate and Loss Rate are averaged values associated with observed profit and loss for respective match instances;
(e) the variables Expected Loss and Expected Profit are posterior predictive density functions which represent the overall loss/profit given (c) and (d) above;
(f) the variable Estimated Profit & Loss is the summary probability density function given (e);
(g) the variable Less than, or Equal to 0 Units is the probability of ending at, or below the specified value of $U$ given (f) above.

4. Results and discussion

In this section we demonstrate and discuss the resulting performance of the model. In Section 4.1 we demonstrate the profitability of the model along with the relevant risks involved with each of the betting procedures; in Section 4.2 we evaluate the effectiveness of the model components based on the transitions of profitability at each hierarchical component level; in Section 4.3 we provide evidence of market inefficiency based on specific football teams; finally, in Section 4.4 we compare the performance of the model against the model presented in [7].

4.1. Model performance

Table 3 presents the amount of bets simulated and unit-based returns (along with the frequency rates of successful bets and profit rate relative to stakes for procedures $BP_1$ and $BP_2$) at the specified discrepancy levels. Fig. 7 illustrates a summary comparison between the two betting procedures. In general, under both procedures the model appears to be profitable at discrepancy levels up to 10%, but unprofitable thereafter. In particular, for $BP_1$, the profitability appears to be consistent up to that point, with the highest returns of £17.45 and £17.34 observed at discrepancy levels of 6% and 1% respectively. In contrast, $BP_2$ generated maximum returns that are substantially higher relative to $BP_1$; returns of
U47.71 and U47.13 at discrepancy levels of 0% and 1% respectively. Figs. A.1 and A.2 compare the cumulative returns over the season between the two betting procedures; the results show that $BP_2$ consistently generates higher returns than $BP_1$ throughout the period and at almost every discrepancy level.

At discrepancy levels of $\geq 11\%$ $BP_2$ essentially mimics the betting simulation of $BP_1$ since it becomes unlikely for probabilities of paired match instances (model and market) to encompass more than one outcome at such high discrepancy levels. At discrepancy levels of $\geq 10\%$ the model appears to be unprofitable, with betting trials in the range of 33 and 84. However, it would not be safe to formulate conclusions on the basis of model performances at such high discrepancy levels. We explain why next.

For $BP_1$ and $BP_2$, it is important to note that we are much more confident about results generated at lower discrepancy levels, since at those levels the number of bets simulated is sufficiently high for us to formulate safe conclusions. As the discrepancy levels increase, the number of betting trials inevitably decreases. Yet, at higher discrepancy levels we actually require more betting trials to formulate conclusions that are as safe as those at the lower levels. To understand why, assume that we have simulated 50 bets at uncertainty generated on the basis of numerous bets simulated. In this case, most bets win. However, 'correct' we would win about four times for every five 'wrong' bets.

It should also be noted that the occurrence rate of the above two cases is likely to be affected by the well known phenomenon of the favourite longshot-bias observed by the markets.\footnote{The phenomenon whereby bettors have a preference in backing risky outcomes and hence, bookmakers offer more-than-fair odds to 'safe' outcomes, and less-than-fair odds to 'risky' outcomes. This phenomenon is not only observed in football but also in many different markets\cite{1,3,38,33,35,34,41,39,18,23,6}. Various theories exist, such as risk-loving behaviour, on why people are willing to bet on such uncertain propositions\cite{37,36}.}

Figs. 8 and 9 demonstrate the cumulative unit-based returns given $BP_3$ and $BP_4$, respectively. In both cases, considerably higher returns are generated relative to $BP_1$ and $BP_2$. In particular, the concluding balance of $BP_3$ at match instance 380 is \$180.34, whereas for $BP_2$, it is \$922.97. Since $BP_4$ is a replicative version of $BP_1$ (with the difference that stakes generated are based on the relative, rather than the absolute, discrepancy of model to market probabilities), it is normal for $BP_4$ to generate cumulative returns that are excessive versions of those of $BP_3$. The cumulative distributions in Figs. 8 and 9 show that $BP_4$ experienced a maximum loss of $-U43.65$ (81.63% less relative to its maximum profit of U237.57), whereas $BP_4$ experienced a maximum loss of $-U1066.33$ (14.54% less relative to its maximum profit of U1247.86). Further, $BP_4$ remained at a state of loss for a longer period throughout the season, whereas $BP_3$ remained at a state of loss for only a period of 11 match instances (out of 380). Table 4 presents the risk probability values for ending up with less than, or equal to, the specified concluding profit/loss balances according to the specified betting procedure, and Fig. B.1 presents the respective predicted probability density risk distributions.

4.1.1. Arbitrage opportunities and risk assessment

There are various ways to reduce our exposure to risk. In our case, a straightforward solution would be to take advantage of existing arbitrage opportunities and replace the betting procedure with arbitrage bets when such risk free match instances are exposed. In fact, 70 match instances (out of the 380) allowed for risk free returns for the season under study, where arbitrage betting guaranteed an average profit of 0.57% per such match instance with minimum and maximum risk free returns at 0.03% and 1.94% respectively. Figs. C.1, C.2, C.3 and C.4 demonstrate how the profit rate converges relative to an initialised bankroll on the basis of $BP_{3.1}$, $BP_{3.2}$, $BP_{3.3}$, and $BP_{3.4}$ (as described in Section 3.2). Table 4 and Fig. B.1 demonstrate the reduction in risk and uncertainty, when taking advantage of arbitrage instances, relative to the respective procedures of $BP_1$, $BP_2$, $BP_3$, and $BP_4$, which do not take advantage of such opportunities. As expected, due to the relatively high number of arbitrage instances the profitability is heavily dependent on the initialised bankroll. When an arbitrage opportunity is discovered the bet is equal to the value of the bankroll at that specific time. Bankrolls with sufficiently high initialised values (i.e. $\geq 1000$ or $\geq 10,000$ in this case) eventually overshadow the predictive performance of the model since generated returns converge towards the arbitrage profit rate.

4.2. Effectiveness of model components

Figs. 10–12 demonstrate the transitions of profitability at component levels 1, 2 and 3 given $BP_1$, $BP_2$, $BP_3$ and $BP_4$. We observe that the model component at level 2 (team form) generates profitability that is substantially superior to that of level 1, for all of the betting procedures. However, profitability is reduced at level 3 (team fatigue and motivation). We have therefore analysed the sub-parameters of that component in an attempt to investigate how they have negatively affected the performance of the model relative to market odds. Figs. D.1, D.2, D.3 and D.4 demonstrate the profitability of the model over procedures $BP_1$, $BP_2$, $BP_3$, and $BP_4$ when:

(a) we only consider match instances with evidence of fatigue (but no evidence of motivation);
(b) we only consider match instances with evidence of motivation (but no evidence of fatigue);
(c) we only consider match instances with evidence of both fatigue and motivation;
(d) we only consider match instances where neither evidence of fatigue nor evidence of motivation exist.

Assuming that we rank profitability-based performances from 1 to 4 (1 being best), the results suggest that evidence of fatigue provided the worse overall performance with resulting ranks of 3, 4, 4 and 4 under procedures $BP_1$, $BP_2$, $BP_3$ and $BP_4$ respectively.
This suggests that we have, most likely, overestimated the negative impact of fatigue for a team (i.e. the number of days gap since last competing match, the toughness of previous match, the involvement in European competitions, and player participation with their national team). On the other hand, motivation (whereby the quality of the input is predominantly dependent on the expert) provided performances with resulting ranks of 4, 1, 3 and 1 under the four respective betting procedures, and signs of improvement (relative to test (d)) in forecasting capability are observed only under two of the four betting procedures.

4.3. Team-based market inefficiency

The results reported in this section add further evidence of market inefficiency to an already extensive list, particularly in the presence of regular predetermined biases, arbitrage opportunities, as well as conflicting daily adjustments in published odds between firms [6]. We also considered a team-based profitability assessment (see Table 5), where the percentage values represent the returns \( U \) of a team relative to the returns over all teams based on the specified betting procedure.\(^{14}\)

Our results demonstrate notable differences in profitability for five out of the twenty teams. In particular, for match instances involving Liverpool, QPR, Arsenal and Newcastle our model generated notably higher returns relative to the overall team, whereas for match instances involving Chelsea our model generated notably lower returns. Fig. E.1 illustrates the team-based explicit returns throughout the season against market odds for the above five teams. Results show that:

(a) market odds overestimated the performances of Liverpool at a consistent rate, and particularly over the final third of the season (during which Liverpool accumulated only 10 points during their last 10 matches). This allowed our model to generate profitable returns during the specified period;

(b) as in (a), the same applies to Arsenal but to a lower extent. This allowed our model to generate profitable returns during the specified period;

(c) market odds underestimated the performances of Newcastle at a consistent rate, and particularly over the first half of the season. It is important to note that Newcastle finished at position 5 with 65 points after being promoted to the EPL only a season earlier. This allowed our model to generate profitable returns during the specified period;

(d) we do not consider that market odds underestimated performances of QPR at the absence of consistency and high uncertainty in returns; profit was generated due to a pair of match instances with excessive returns;

(e) our model overestimated the performances of Chelsea, particularly over the first two thirds of the season, at a consistent rate. This is highly likely to be due to Chelsea’s erratic performances under a new manager who was eventually sacked during that period. This led our model to generate unprofitable returns during the specified period. The returns over the final third of the season, during which Chelsea provided more consistent performances under a new manager, appear to be evened.

4.4. Performance comparison against the previously published BN model

Figs. F.1, F.2 and F.3 compare the unit-based cumulative returns over a period of 380 match instances (but for different seasons\(^{15}\)) between the two models. The results show that the new model generates superior returns under all of the betting procedures.\(^{16}\) In particular, for \( BP_1 \) and \( BP_2 \), the new model generated increased net-profit of 33.67% and 210.98% respectively. An interesting distinction between the two models (according to the first two betting procedures) is that the previous model provides higher profit rates but lower net-profit due to the significantly lower number of bets simulated (as discussed in Section 3.1, and Tables 3 and 6 verify this behaviour). Further, for scenarios \( BP_3 \) and \( BP_4 \) the new model generates respective net-profit that is 158.43% and 49.68% higher relative to respective returns from the previous model.

5. Concluding remarks

We have presented a Bayesian network (BN) model for forecasting football match outcomes that not only simplifies a previously publish BN model, but also provides improved forecasting capability. The model considers both objective and subjective information for prediction. The subjective information is important for prediction but is not captured in historical data. The model was used to generate the match forecasts for the EPL season 2011/2012, and forecasts were published online [29] prior to the start of each match.

For assessing the forecast capability of our model, we have introduced an array of betting procedures. These are variants of a standard betting methodology previously considered for assessing profitability by relevant published football forecast studies. A unit-based profitability assessment over all betting procedures demonstrates that:

(a) at level 2 (team form) the model component provided inferred match forecasts that were substantially superior to those generated at level 1 (which were solely based on historical performances);

(b) at level 3 (team fatigue and motivation) the model component failed to provide inferred match forecasts that were superior to those generated at level 2. This resulted in concluding match forecasts with inferior profitability relative to that of level 2, but still superior relative to that of level 1;

(c) a sub-component evaluation at level 3 revealed that we have overestimated the negative impact introduced by evidence of fatigue, and this should serve as a lesson-learned for relevant future models;

(d) despite the consequences of (b), the concluding profitability of our model was even superior to that generated by the previous successful and profitable model under all of the betting procedures;

(e) the predictive probability density distributions of unit-based returns showed that a bettor’s exposure to risk increases together with the substantial profitable returns that \( BP_3 \) and \( BP_4 \) provide over \( BP_1 \) and \( BP_2 \). However, we showed that one way a bettor may reduce his exposure to risk is by exploiting arbitrage opportunities which occur relatively frequently (70 out of the 380 match instances);

\(^{14}\) If for the specified betting procedure a team generates returns \( A \) which are equal to the returns \( B \) generated by all of the teams (overall), then team \( A \) is 100% related to set \( B \).

\(^{15}\) We compare the forecasting capability between the two models relative to market odds, where the old version was assessed over the EPL season 2010–2011, and the new version (presented in this paper) over the EPL season 2011–2012.

\(^{16}\) Following the discussion in Section 4.1, we have ignored the scenarios whereby the discrepancy levels of \( BP_1 \) and \( BP_2 \) are set to \( \geq 11\% \).
(f) a team-based profitability assessment revealed further market inefficiencies (to the already extensive list) whereby published odds are consistently biased towards the trademark rather than the performance of a team.

Evidently, the results of our study are critically dependent on the knowledge of the expert. Given that the subjective model inputs were provided by a member of the research team (who is a football fan but definitely not an expert of the EPL), it suggests that (a) subjective inputs can improve the forecasting capability of a model even if they are not submitted by a genuine expert who is a professional for the specified domain, and (b) if the model were to be used by genuine experts we would expect that the more informed expert inputs would lead to posterior beliefs that are even higher in both precision and confidence.

The results of this paper have demonstrated a number of benefits of using Bayesian networks: in particular they enable us to incorporate crucial subjective information easily and enhance our understanding of uncertainty and our exposure to the relevant risks involved.

Acknowledgements

We acknowledge the financial support by the Engineering and Physical Sciences Research Council (EPSRC) for funding this research and Agena Ltd for software support.

Appendix A. Cumulative Returns based on $BP_1$ and $BP_2$

Figs. A.1 and A.2.

Fig. A.1. Cumulative unit-based returns based on $BP_1$ and $BP_2$ according to the specified discrepancy level.
Fig. A.2. Cumulative unit-based returns based on $BP_1$ and $BP_2$ according to the specified discrepancy level.
Fig. B.1. Risk assessment of profit and loss based on the specified betting procedure.
Appendix C. Model performance when considering arbitrage opportunities

Figs. C.1, C.2, C.3 and C.4.

Fig. C.1. Cumulative unit-based returns based on $BP_{5.1}$ assuming no discrepancy restrictions (set to 0%) and according to the specified bankrolls prior to initialising the betting simulation.

Fig. C.2. Cumulative unit-based returns based on $BP_{5.2}$ assuming no discrepancy restrictions (set to 0%) and according to the specified bankrolls prior to initialising the betting simulation.

Fig. C.3. Cumulative unit-based returns based on $BP_{5.3}$ and according to the specified bankrolls prior to initialising the betting simulation.
Appendix D. Performance based on parameters of component level 3

Figs. D.1, D.2, D.3 and D.4.
Fig. D.3. Cumulative unit-based returns based on $\text{BP}_3$ for match instances with the specified evidence.

Fig. D.4. Cumulative unit-based returns based on $\text{BP}_4$ for match instances with the specified evidence.
Appendix E. Team-based efficiency

Fig. E.1. Team-based explicit returns against market odds throughout the EPL season.

Appendix F. Unit-based performance relative to the old model

Figs. F.1, F.2 and F.3.

Fig. F.1. Cumulative unit-based returns based on $BP_1$ and $BP_2$: a comparison between the new and the old model.
Fig. F.2. Cumulative unit-based returns based on $BP_3$: a comparison between the new and the old model.

Fig. F.3. Cumulative unit-based returns based on $BP_4$: a comparison between the new and the old model.

Appendix G. Description of model variables and actual examples of the BN model

Table G.1 and Figs. G.1,G.2,G.3,G.4 and G.5.

Table G.1

<table>
<thead>
<tr>
<th>Model component</th>
<th>Variable (node) name</th>
<th>Variable type</th>
<th>Observable/latent</th>
<th>Definition</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1r</td>
<td>Number of Wins</td>
<td>Integer interval</td>
<td>Observable</td>
<td>~ Binomial ($\frac{\text{NumberOfMatchesPlayed}}{p(\text{Win})}$)</td>
<td>Used for inferring $p(\text{Win})$. Same applies to “Number of Draws” and “Number of Loses”</td>
</tr>
<tr>
<td>Level 1r</td>
<td>Number of matches played</td>
<td>Integer interval</td>
<td>Definitional</td>
<td>Serves as hyperparameter $n$ for the variables: number of wins, draws, loses, and as a hyperparameter for number of residual matches</td>
<td></td>
</tr>
<tr>
<td>Level 1r</td>
<td>Current Points</td>
<td>Integer interval</td>
<td>Definitional</td>
<td>$\min \left( \frac{114.3 \times \text{NumberOfWins} \div \text{NumberOfDraws} }{\text{NumberOfResidualMatches}} \right)$</td>
<td>Represents the summation of wins, draws and losses. Similarly, the definition of “Number of residual matches” is 38 minus “Number of matches played”</td>
</tr>
<tr>
<td>Level 1r</td>
<td>$p$ (Win)</td>
<td>Continuous Interval -- Beta</td>
<td>Latent</td>
<td>$\sim \text{Beta}\left( 1 + \text{NumberOfWins}, 1 + 38 - \text{NumberOfWins} \right)$</td>
<td>Assumes prior $\sim\text{Beta}(1,1)$. Same applies to “$p$ (Draw)” and “$p$ (Lose)”</td>
</tr>
<tr>
<td>Level 1r</td>
<td>Expected Residual Points Difficulty of residual opponents ERP given opponent difficulty</td>
<td>Continuous Interval (Ranked)</td>
<td>Latent</td>
<td>$\min \left( \frac{114.7 \times \text{NumberOfResidualMatches} }{3 \times p(\text{Win}) + p(\text{Draw})} \right)$</td>
<td>Represents subjective indications</td>
</tr>
<tr>
<td>Level 1r</td>
<td>Team Strength (S)</td>
<td>Continuous Interval (Arithmetic)</td>
<td>Latent</td>
<td>$\pi_e$ as defined in Section 2.1</td>
<td></td>
</tr>
</tbody>
</table>

(continued on next page)
Table G.1 (continued)

<table>
<thead>
<tr>
<th>Model component</th>
<th>Variable (node) name</th>
<th>Variable type</th>
<th>Observable/latent</th>
<th>Definition</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1: Inconsistency (Variance)</td>
<td>Continuous Interval (→Uniform (a,b))</td>
<td>Latent</td>
<td>~Uniform (0,150)</td>
<td>~Uniform (0,114)</td>
<td></td>
</tr>
<tr>
<td>Level 1: Overall Performance (mean points)</td>
<td>Continuous Interval (→Uniform (a,b))</td>
<td>Latent</td>
<td>~Uniform (0,114)</td>
<td>The same applies to Seasons y2 to y5</td>
<td></td>
</tr>
<tr>
<td>Level 1: Season y1</td>
<td>Integer Interval (→TNormal (μ,σ²,a,b))</td>
<td>Observable</td>
<td>~TNormal (OverallPerformance, Inconsistency 0.001, 0.1)</td>
<td>Φ is measured outside of the BN (see Section 2.2)</td>
<td></td>
</tr>
<tr>
<td>Level 2 Form (F)</td>
<td>Continuous Interval (→TNormal (μ,σ²,a,b))</td>
<td>Observable</td>
<td>~TNormal(Φ,0.001,0,1)</td>
<td>Represent subjective indications</td>
<td></td>
</tr>
<tr>
<td>Level 2 Availability of players who resulted in current form (LA)</td>
<td>Continuous Interval (→TNormal (μ,σ²,a,b))</td>
<td>Observable</td>
<td>5 Ordered states from “Very Low” to “Very High”</td>
<td>Represent subjective indications</td>
<td></td>
</tr>
<tr>
<td>Level 2 Important players return (or new transfers) (LR)</td>
<td>Continuous Interval (→TNormal (μ,σ²,a,b))</td>
<td>Observable</td>
<td>4 Ordered states from “None” to “High”</td>
<td>Represent subjective indications</td>
<td></td>
</tr>
<tr>
<td>Level 2 Expected Form given player availability</td>
<td>Continuous Interval (→TNormal (μ,σ²,a,b))</td>
<td>Observable</td>
<td>Φ₄ as defined in Section 2.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2 Expected form given further important players</td>
<td>Continuous Interval (→TNormal (μ,σ²,a,b))</td>
<td>Observable</td>
<td>Φ₄ as defined in Section 2.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 3 Toughtness of previous match</td>
<td>Continuous Interval (→TNormal (μ,σ²,a,b))</td>
<td>Observable</td>
<td>5 Ordered states from “Very Low” to “Very High”</td>
<td>Represent subjective indications</td>
<td></td>
</tr>
<tr>
<td>Level 3 EU Match Involvement</td>
<td>Continuous Interval (→TNormal (μ,σ²,a,b))</td>
<td>Observable</td>
<td>6 Ordered states from “None” to “Very High”</td>
<td>Represent subjective indications</td>
<td></td>
</tr>
<tr>
<td>Level 3 National Team Involvement</td>
<td>Continuous Interval (→TNormal (μ,σ²,a,b))</td>
<td>Observable</td>
<td>5 Ordered states from “None” to “Very High”</td>
<td>Represent subjective indications</td>
<td></td>
</tr>
<tr>
<td>Level 3 Days Gap</td>
<td>Continuous Interval (→TNormal (μ,σ²,a,b))</td>
<td>Observable</td>
<td>5 Ordered states from “1–2” to “6+”</td>
<td>Represent subjective indications</td>
<td></td>
</tr>
<tr>
<td>Level 3 Motivation</td>
<td>Continuous Interval (→TNormal (μ,σ²,a,b))</td>
<td>Observable</td>
<td>5 ordered states from “Very Low” to “Very High”</td>
<td>Represent subjective indications</td>
<td></td>
</tr>
<tr>
<td>Level 3 Head To Head Bias</td>
<td>Continuous Interval (→TNormal (μ,σ²,a,b))</td>
<td>Observable</td>
<td>5 states: “HT Advantage” and “AT Advantage”</td>
<td>Represent subjective indications</td>
<td></td>
</tr>
<tr>
<td>Level 3 Prior Fatigue</td>
<td>Continuous Interval (→TNormal (μ,σ²,a,b))</td>
<td>Latent</td>
<td>Gᵢ as defined in Section 2.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 3 Expected Fatigue</td>
<td>Continuous Interval (→TNormal (μ,σ²,a,b))</td>
<td>Latent</td>
<td>Gᵢ as defined in Section 2.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 3 Fatigue and Motivation</td>
<td>Continuous Interval (→TNormal (μ,σ²,a,b))</td>
<td>Latent</td>
<td>G as defined in Section 2.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topology: Confidence in historical inconsistency</td>
<td>Continuous Interval (→TNormal (μ,σ²,a,b))</td>
<td>Latent</td>
<td>3 ordered states from “Low” to “High”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topology Team Strength (S) L1</td>
<td>Continuous Interval (→TNormal (μ,σ²,a,b))</td>
<td>Latent</td>
<td>S₂₁ as defined in Section 2.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topology Team Strength (S) L2</td>
<td>Continuous Interval (→TNormal (μ,σ²,a,b))</td>
<td>Latent</td>
<td>μ if (Φ &lt; 0.5) then: S₂₁ ((114 – S₁₂) × (0.5 – Φ)), else: S₂₁ – (S₁₂ × (Φ – 0.5)), σ² = 1 × ABS(Φ – 0.5) × 10, a = 0, b = 114</td>
<td>The same applies to: “Team Strength (S) L3”, where S₂₁ is replaced by S₂,2 and Φ is replaced by G</td>
<td></td>
</tr>
</tbody>
</table>
Fig G.1. A simplified representation of the overall Bayesian network model. An example based on the actual scenarios of the Arsenal vs. Liverpool EPL match, August 20th 2011. The observed outcome was A (0–2).
Fig. G.2. Level 1 Component (P): formulating S prior. An example with four actual scenarios based on Fulham, Man City, Wigan, and Man United data, as retrieved at gameweek 37 during season 2011/2012.
Fig. G.3. Level 1 Component (I): measuring a team’s historical inconsistency (V) based on league point totals of the five most recent seasons. An example with four actual scenarios based on Fulham, Man City, Wigan and Man United data for the five seasons preceding EPL 2011/2012.
Fig. G.4. Level 2 Component ($F$): measuring team form. An example with four scenarios (scenario 4 represents uncertain inputs whereby values follow predetermined subjective prior probabilities).
Fig. G.5. Component 3 (M): measuring fatigue and motivation. An example with four scenarios (scenario 4 represents uncertain inputs whereby values follow predetermined subjective prior probabilities).

References


