UNIVERSITÄT OSNABRÜCK

Institute of Cognitive Science

Bachelor Thesis

Prediction of Association Football Results Using Bayesian Neural Networks with Financial and Performance-Based Data

Ferdinand Schlatt

August 17, 2017

First supervisor: Prof. Dr. Gordon Pipa
Second supervisor: Leon Sütfeld, M.Sc.

Contents

1. Introduction

The prediction of association football results is, due to the high variance in outcomes of a game, a difficult task. The result of a game can depend on a large range of factors, from weather conditions and psychological aspects, to team form and quality. Therefore, a subset of adequate predictors needs to be chosen to model the outcome using reasonable resources. A prominent trend in the world's best leagues has given rise to a new category of factors. Through the increasing commercialization of football, financial aspects are playing a large role in how well teams place in the final standings.[1, pp.54–58] This trend reflects itself for example in the transfer fees, in which of the top 25 record transactions, 21 have happened in the last six years.[2] It can also be seen in the revenues of top football clubs. From 2012 to 2016 the total revenue of the five biggest European leagues rose from 20€ billion to 25€ billion.[3] This poses the question if a team's financial status holds any information about its strength and if this information could be used to predict match results or to further improve models which use more classical predictors.

A second current trend, which has shown excellent performance in prediction tasks, is the use of artificial neural networks (ANNs). Recent technological advancements have allowed for unparalleled performance on a multitude of tasks. Notably, with the help of ANNs, Poker and Go, two of the most complex perfect and imperfect information games, are played by artificial intelligence agents on a higher than expert level.[4, 5] In general, ANNs are capable of learning complex non-linear functions based on data examples.[6, pp.225–256] Training data is passed through a network of non-linear units, called neurons, and the output is compared to a desired target. The error is computed using a loss function and propagated backwards through the network, updating the weights of the neurons. After training converges and the error reaches a minimum, new data can be passed through the network. In the case of a prediction task, probabilities for a set of categories can be obtained for a new set of inputs.

On top of this, recent successful attempts have been made at combining Bayesian probabilistic techniques and ANNs efficiently.[7, 8] Instead of using point fixed real-valued network parameters, the parameters underlie prior distributions. In principle, Bayesian techniques can be used on any ANN structure, but for prediction tasks the most simple feed-forward network, the multilayer perceptron (MLP), suffices. The Bayesian variant of an MLP will from now on be referred to as a Bayesian neural network (BNN) and grants multiple advantages over typical ANN approaches. These include among others, implicit regularization of the network weights and uncertainty of network weights and outputs.[9] Since only a limited amount of football data is available, the former property is useful in preventing over fitting and the latter offers a natural uncertainty on predictions.

Both of the aforementioned trends are used in this thesis. BNNs are adopted to

predict the results of matches of the English Premier League (EPL), with results denoting the exclusive outcomes of a home team win, draw or away team win. This is done using three different data sets. The first consisting of classical predictors based on past results of the teams, the second consisting of financial predictors and the third is a combination of both. To compare the predictive performance of the models, a slew of different statistical and economic scores are employed. In addition, they are compared to a prominent model from past literature.

A number of hypotheses were established before testing the models. Performance data describes team strength more directly, in comparison to financial data, since it explicitly captures the strength exhibited in past games. The former is therefore expected to perform better when predicting the results of a game. Furthermore, since the model using the combined data set has the same and additional data at its disposal, it should feature equal or better performance than the performance and financial models. In comparison to models from past literature, the BNN models should perform better by incorporating more data and being able to learn more complex functions. Lastly, when betting against bookmaker odds, it is assumed that no profitable strategy can be generated. This assumption is grounded on the fact that bookmaker odds are highly optimized to gain as much profit as possible and are imbalanced. This means betting on all outcomes guarantees a loss and generates a risk free profit for the bookmakers, even with sub-optimal probabilities. This imbalance is commonly referred to as overround.

In this thesis, first previous literature about football and sports predictions, as well as efficiency of bookmaker odds will be reviewed. A description of the data acquisition, formatting and structure follows. Next, an overview of the different performance metrics is given, together with a description of the comparison model. Afterwards, the BNN is described in more detail and the specific architectures for the different models are reported. This is followed by an evaluation of the performances of all models, using the different performance measures. This is combined with an analysis of the output predictions and uncertainty and the results are discussed. Lastly, a conclusion and overview is given to summarize the thesis.

2. Literature Review

In past literature, many different approaches have been used in an attempt to predict sports results. Bouilier and Stekler use seedings and logistic regression to predict league and tournament play in professional tennis and college basketball.[10] Stefani utilized a least squares prediction system based on self computed ratings.[11] He was able to improve on this by accounting for win margins and implementing a correction factor, handling the home field advantage.[12] The system is used to predict American college football outcomes on a weekly basis and the knockout round of the 1978 World Cup of Football. More recently, McCabe and Trevathan use MLPs with backpropagation and root mean squared errors to predict the results of games in Australian rules football, rugby and football.[13] Different performance measures of a team are used as input to compute a single confidence value, representing the win likelihood of a team. For a single game, both values are computed for both teams and the team with the highest value chosen as the prediction.

Regarding predictions of association football, early notable academic attempts were made by Maher.[14] A Poisson model is used to model the number of goals scored in a game. Multiple distributions are fitted per team, with the distribution means implicitly representing the team's home and away attack and defense strengths. An extension to this approach is used by Dixon and Cole to generate predictions for match results.[15] The model allows for the inclusion of multiple divisions and incorporates time dependent dynamic fluctuations of team strengths. Bookmaker odds are used to evaluate the model performance. A significant advantage over the bookmaker probabilities is shown, but due to an 11% overround the model was unable to prove profitable in betting. Goddard and Asimakopoulos propose an ordered logistic regression model and show the significant influence on model performance of multiple non-trivial factors.[16] These factors include for example the geographical distance between home grounds, or the importance of a game regarding relegation or championship for a team. Again, to evaluate the model it is pitted against bookmaker odds, outperforming them in one of two seasons of English professional football leagues. Albeit, not to a statistical significant margin. Using an ordered logistic regression with elo values as explanatory variables, Hvattum and Arntzen are able to outperform the previously mentioned model.[17] Elo ratings are based on preceding results and updated after every match. A larger rating gain is achieved if a team wins against a higher rated opponent, opposed to winning against a lower rated opponent. The elo system is extended to also take the win margin into account. Again, it was not possible to obtain a profitable betting strategy against bookmaker odds using data from the English league system. Lastly, Baio and Blangiardo use a hierarchical Bayesian model to again predict the number of goals scored in a game.[18] The Italian Serie A is the league of choice and the model is shown to be on par with the bivariate Poisson model by Maher when

predicting final league standings.

As bookmaker odds are considered to be benchmarks in football prediction, numerous investigations have been made into the efficiency and exploitability of betting markets. Cain, Law and Peel are able to show that, alike to the horse racing betting market, the football betting market also exhibits a favorite long-shot bias.[19] In games with an extreme favorite against an underdog, frequently bookmakers offer fair bets on the favorite and unfair bets on the underdog. Potential profitable actions could therefore be to bet on the favorite. Similar results are obtained by Levitt.[20] Bookmakers utilize the systematic biases of bettors to provide theoretically sub-optimal, but profit maximizing odds. Against the majority of bettors this is a profitable strategy, but a prediction system yielding better probabilities will be able to exploit it. An example of a bettor bias is given by Forrest and Simmons.[21] The number of active supporters of a club significantly influence the odds setting of bookmakers in Spanish and Scottish football. A disproportionate amount of bets is placed on teams with large fan bases, leading bookmakers to bias the odds. These findings show that bookmaker odds are sub-optimal, but the creation of a profitable betting strategy remains a difficult task.

3. Data

An abundance of large data sets, containing statistics of past games of major football leagues can be found online. Financial data about football clubs, on the other hand is more scarce. The most accessible league in this respect is the EPL, perhaps due to the unconstrained investment policy. For this reason, all collected data is from the EPL from the 2005/2006 up to the 2016/2017 season. The 2016/2017 season is used for evaluation, while all other seasons make up the training sets. In the following, an overview of the different explanatory variables in the different data sets is given. Also bookmaker odds are discussed and the different data sources are mentioned.

3.1 Performance-Based

Four different values are used to attempt to capture the strength of a team: points, elo value, goals scored and goals conceded. It should be noted that the points as well as goals scored and conceded are reset after every season. To estimate these values for the first game of a season, the values from the last matchweek of the previous season are used. Due to this reset, the points and goals could wrongly reflect the actual strength of a team at the beginning of a season. If for example a weak team has an exceptionally good start, playing and winning against two also weak teams, it will have unconventionally many points and be overrated. Given the matchweek number, a model could theoretically weight these predictors less, early in the season. Therefore, the matchweek number is included as a final predictor. If this effect can be measured in model performance is reviewed in section 6.

3.1.1 Points

In most professional football league settings, winning a game grants three points, drawing grants one point and losing zero points. Points are over the course of a season and teams ranked by their number of points. Rank is also a possible performance measure, but points are preferred as they have additional informational value. A small difference in rank does not necessarily mean a small difference in team strength. If for example one team dominates the league, this team will have significantly higher points than other teams, reflecting their superiority. On the other hand, a large difference in rank does not imply significantly different team strengths. Multiple teams could have the same amount of points, expressing a close to equal team strength, which is not reflected by the rank. In other words, points are interval scaled and therefore better reflect the performance of a team, compared to the ordinal scaled rank. So as to remove the dependence on the matchweek, the points are divided by the number of games played to obtain the points per game.

3.1.2 Elo Values

Elo values were originally developed by Elo to rank and assess chess players.[22] As aforementioned, Hvattum and Arntzen have successfully shown that they are efficient and effective predictors for association football game results. To compute the values in this thesis, the extension incorporating win margins developed by Hvattum and Arntzen is used.[17]

After every game, elo values are updated for both teams. This is done by first computing the expected score of both teams. Assuming r_0^H and r_0^A are the current ratings of the home and away teams, the expected scores s^H and s^A are computed by:

$$s^H = \frac{1}{1 + c^{(r_0^A - r_0^H)/d}} \quad s^A = 1 - s^H = \frac{1}{1 + c^{(r_0^H - r_0^A)/d}}.$$

Rating updates for both teams are then obtained using the actual scores a^H and a^A with

$$a^H = \begin{cases} 1 & \text{if home team won} \\ 0.5 & \text{if result is draw} \\ 0 & \text{if away team won} \end{cases} \qquad a^A = 1 - a^H.$$

The new ratings r_1^H and r_1^A for the home and away teams are computed using the absolute goal difference g by

$$r_1^H = r_0^H + k(1+g)(a^H - s^H) \quad r_1^A = r_0^A + k(1+g)(a^A - s^A).$$

The two parameters c and d are used for scaling and set to $c = 10$ and $d = 400$, while k influences the rate at which the elo values change. It is set to $k = 10$. The beginning elo values are initialized to 1500 and the 2003/2004 and 2004/2005 seasons used to obtain proper starting values for training.

As at the end of a season, three teams are relegated from the EPL and three promoted from the League Championship, three teams' elo values need to estimated at the beginning of a season. A rating update is a zero sum operation, hence the average value of all teams always stays at 1500. To guarantee this property is sustained, the three relegated teams' values are summed up and averaged. The three promoted teams are then assigned these values, under the assumption that the promoted and relegated teams have close to equal team strength. It can be argued that this is an inaccurate method of initializing these values, as getting promoted into a league can be considered more difficult than staying in the league. Therefore, relegated teams should be weaker than their corresponding promoted teams. An underestimation of the quality of promoted teams would then reflect itself in an early increase of their elo values. However, the mean elo change of all promoted teams in their first five league games is fairly low at a gain of 1.809. Consequently, the estimation using the relegated teams values is reasonable. See Figure 3.1 for a plot of the course of elo values for three chosen EPL teams.

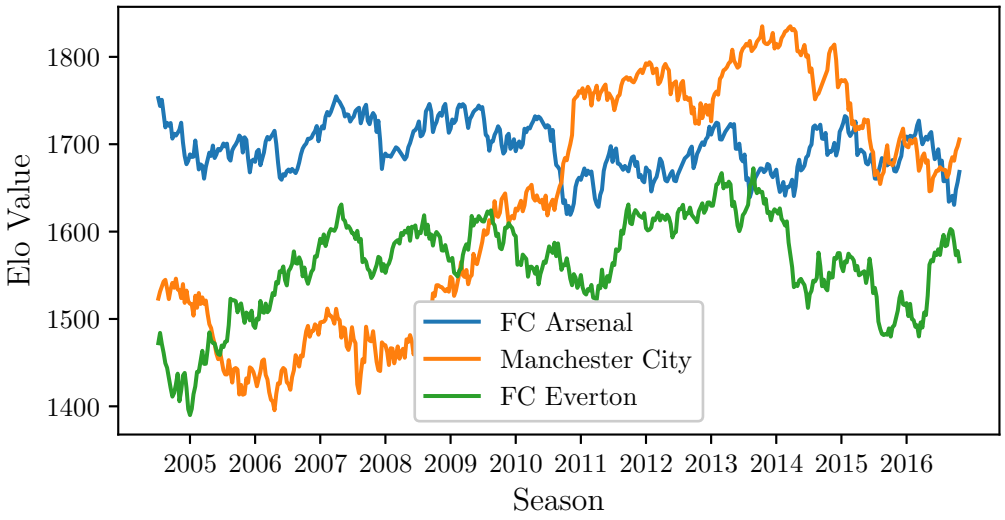

Figure 3.1: Elo values of FC Arsenal, Manchester City and FC Everton over the course of 10 EPL seasons.

3.1.3 Goals Scored and Conceded

Team strength can be divided into different properties, with a very intuitive way being into offensive and defensive strength. To capture these directly the average scored and conceded goals per game are considered. A good team should naturally not only score many goals, but also concede few, while bad teams should show opposite behavior. Differentiation between offensive and defensive strengths is hence only sensible, if teams exist which exhibit an abnormal high offensive and low defensive strength or vice versa. The correlation between the average goals scored and conceded of all EPL teams from all matchweeks and teams between the 2003/2004 and 2016/2017 does support the mentioned intuition at -0.492. However, a linear regression on the data reveals a high spread around the regression line, with a subsequent R-squared value of only 24.1% (Figure 3.2a). Therefore, the trend of good teams scoring many and conceding few goals and vice versa, explains only little of the actual variance found in the data. Differentiating between offensive and defensive strength using the scored and conceded goals of a team is subsequently sensible.

3.1.4 Form

A final performance measure which was considered is the form of a team. For this, fluctuations of team performance in recent games are considered. Naturally, one assumes a team that has a run and won multiple games in a row will play well again. To capture this property, a weighted average on the performance measures of recent games can be used, weighting further past games less than recent ones.

To measure the performance of a team in one game, the difference between elo values before and after the game can be used. The elo change compares the expected to the actual score and therefore further differentiates between results other than win, draw or loss. If for example, a weak team draws against a strong team, this will be considered a good performance and the weak team will increase its value.

The elo change between two consecutive games can therefore be used, to test if a form property exists. The correlation between elo changes of 5280 game pairs is extremely low and even slightly negative at −0.040. On top of this, plotting the data shows a close to normal distribution of values (Figure 3.2b). Since a form effect should be greatest between two consecutive games, it can hence be considered as extremely minimal and is therefore not regarded in the model.

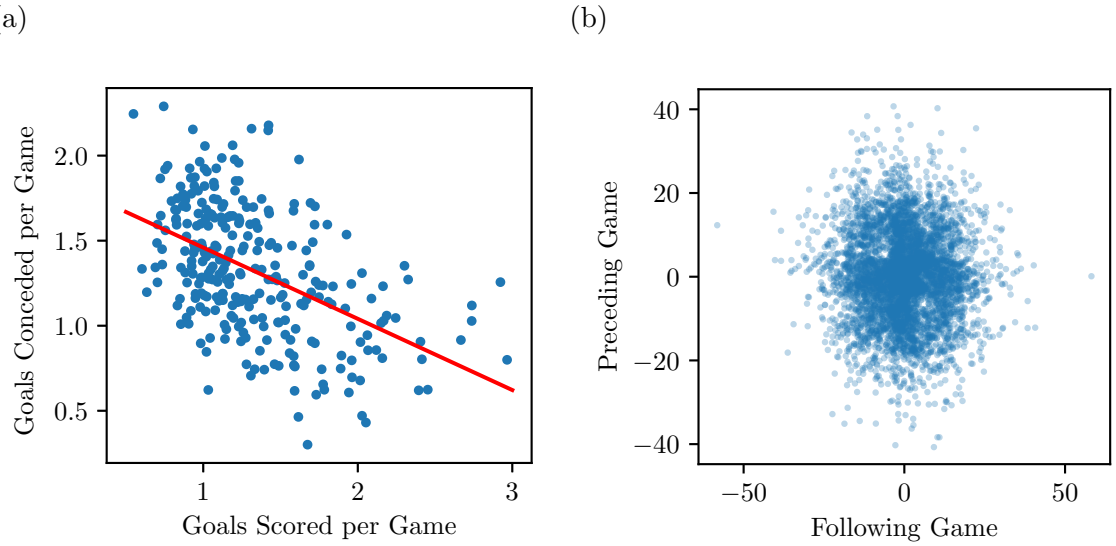

Figure 3.2: (a) Conceded goals per game plotted against scored goals per game of all teams from the EPL from 2003/2004 to 2016/2017 with the regression line. (b) Elo changes between two consecutive preceding and following games plotted against each other over of a total of 5200 EPL game pairs.

3.2 Financially-Based

The financial data set consists of three values per team and for similar reasons as the performance data set, also the matchweek number. These are the transfer expenditures, transfer income and the market value of a team. Contrary to the performance based values, the financial values are not updated on a game to game basis, but are static and set at the beginning of a season. This is on the one hand, due to, to the scarcity of data which is rarely updated and on the other, to the transfer window, which only allows for player transfers once every half a year. This could lead to poor model performance during the end of the season, when the data is potentially outdated and does not capture the team strength well. If this effect occurs will be reviewed in section 6.

3.2.1 Transfer Expenditures and Income

During a transfer period, teams are allowed to buy, sell, lend and trade players with the essential goal to strengthen the team. Large sums of money are paid for individual players, such that the amount paid for a player can be considered as a gauge for the player's quality. The transfer expenditures and income before the

beginning of a season therefore give insight into the quality of players that have left and joined the team. A large transfer surplus could thus mean that a team sold multiple good players, without signing new ones to replace them. However, this by no means is a guarantee that team quality has decreased. For example, youth players and free agents are not considered in the transfer values and transfer fees are on average 30% higher for offensive players.[1, p. 77] Still, transfer expenditures and income can be considered as guidelines to the quality increase and decrease of a team.

3.2.2 Market Values

The market value of a player is not an empirical value, but a subjective estimate based on expert knowledge of how much the player is worth. A guideline is given by actual transfer fees of players, while many different factors influence a players worth. These are for example, the amount of experience a player has gained, how well he is performing or if he is prone to injuries or not. Summing up the values of all players in a team then grants the team value. The relationship between market values and points at the end of a season of a team is not linear but logarithmic. The same increase in market value in a team with an already high market value will on average result in a smaller points increase compared to a team with a lower initial market value. This routes from the fact that there exists a maximum amount of points that can be achieved and the difficulty of achieving a high amount of points increases exponentially towards the maximum. Still, as the correlation between points of a team at the end of a season and their corresponding market value before the season is high at 0.758, the team market value is a good indicator of team quality. The data is plotted with a fitted logarithmic curve for visualization in Figure 3.3.

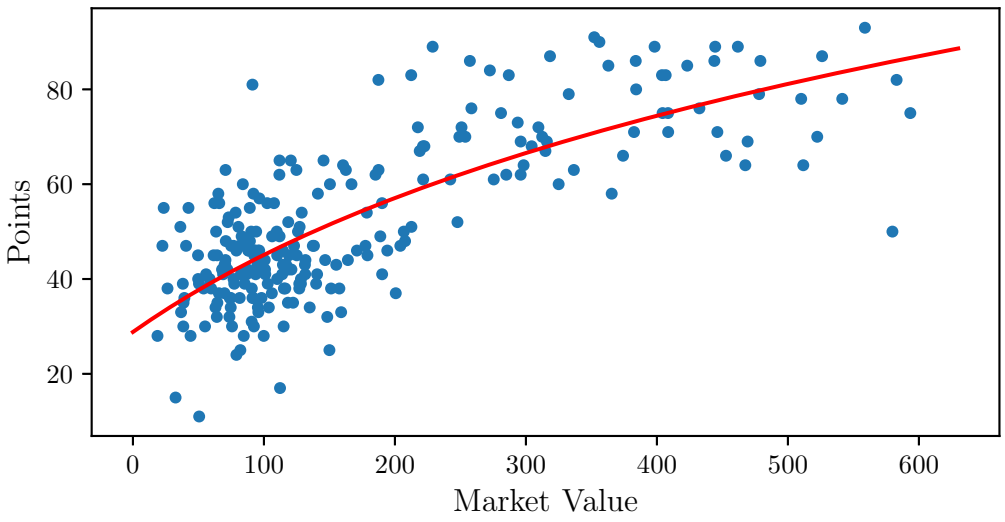

Figure 3.3: Points at the end of a season plotted against the market value at the beginning of a season of all teams from all EPL season from 2005/2006 to 2016/2017. The red line is a curve of the form $f(x) = \ln(xa + b)c + d$ fitted using the non-linear least squares method.

3.2.3 Wage Bill and Relative Attendance

Two further financial variables which were considered are the wage bill and attendance per game of a team. The rationale for the wage bill being that more valuable players earn more money, with the added effect that players which are paid more money are more satisfied and play better. Unfortunately, teams rarely make their wage bill public, making the data scarce. On top of this, different data sources post different contradicting values. Therefore, a correct and valid data set could not confidently be created.

Attendance values on the other hand are easily accessible. The average attendance per game relative to stadium size is an indicator of team quality, assuming better teams attract more visitors. However, attendance values across all teams were extremely high, with a mean relative attendance of 91%. Attendance values therefore give no informational value about the quality of a team and both predictors were discarded.

3.3 Data Set Size

As aforementioned, all data is taken from the EPL from the 2005/2006 up to the 2016/2017 season, with the last season used as test data. One season consists of 38 matchweeks with 10 games per matchweek. That results in 380 games per season, with maximum of 3800 games in the training data set. In the performance data set, the points and goals for the first game are estimated using the last game of the previous season. As three teams are promoted at the beginning of every season, no values can be estimated for these. All three games of the first matchweek involving these three teams therefore have to be disregarded for the performance data set. The performance and also the combined training sets are subsequently comprised of 3770 games while the financial training set is able to take advantage of all 3800 games. The validation set is the same for all three models, consisting of data from the 2016/2017 season with a size of 377 games, where the three games of the promoted teams have been removed.

3.4 Bookmaker Odds

Many different forms of odds are used around the world. The preferred representation on the European market is the decimal odd. In this representation, the posted value is the amount a bettor would win, multiplied by his wager. For example, if a bettor would place a 5 unit bet on a posted odd of 1.75 and win, the payout would equal $1.75 \cdot 5 = 8.75$. A decimal odd o can easily be converted into its corresponding probability p by computing its multiplicative inverse

$$p_o = \frac{1}{o}.$$

This allows for an easy comparison with model probabilities, after considering the overround. It ensures betting on all outcomes is always a losing action for the bettor and even sub-optimal odds will generate a positive revenue for the bookmaker. An

overround affects the odds such that the corresponding probabilities of all outcomes sum up to a value larger than one. Therefore, the probabilities need to be scaled to be able to use them in the comparison. Scaling is done by dividing all odds of a game by their sum. Lastly, odds were collected and averaged over 16 different bookmakers, ensuring them to be unbiased towards any specific bookmaker. The resulting average overround over all odds from the 2016/2017 season is 3.31%.

3.5 Data Sources

All EPL game data, which was used to build the performance data set, was gathered from a website containing reliable information about a multitude of professional leagues.[23] Market values are available on a website which features a large community of registered users, that update the market value of players regularly based on different criteria. Transfer expenditures and income can also be found on this site.[2] Lastly, all bookmaker odds data was collected from an odds archive, which records the final odds set by multiple bookmakers shortly before the beginning of a game.[24]

4. Comparison Methods

To be able to evaluate and compare the performance of the models, different score functions are used. Also, by comparing the models with the logistic regression model by Hvattum and Arntzen, they are put into context with best known model from previous literature.[17] An overview of different statistical functions which can and have been used to evaluate models predicting association football results is given by Constantinou.[25] The score functions are separated into two groups. One considers all outcomes of a game and the other considers only the observed outcome of a game. An observed outcome is the actual outcome which was the result of a game. One function is chosen from both groups, from the first group the Brier score and from the second the geometrical mean. These are then used to statistically compare the models with one another and against the bookmaker odds. As a third and more economic measure, the betting advantage or disadvantage against the bookmaker odds is considered, using a predefined betting strategy. The different score functions and comparison model are examined in the following.

4.1 Brier Score

The Brier score was first defined by Brier for the evaluation of weather forecasting models.[26] It essentially is the mean squared error of the predicted probabilities with their actual outcomes. It not only takes the error of the outcome that took place into account, but also the error of the outcomes that did not take place. For example in a football game, if the home team won, not only the error made in the home team win prediction but also the error for the draw and away team win predictions are considered. The Brier score B is defined for multi-category events by

$$B = \frac{1}{N} \sum_{t=1}^{N} \sum_{i=1}^{C} (p_{ti} - e_{ti})^2.$$

where N is the number of events, C the number of event categories, p_{ti} the probability predicted and e_{ti} the outcome at event t for category i. For a single football game, with prediction probabilities of $[0.5, 0.3, 0.2]$ for a home team win, draw and away team win, in which the home team won, the Brier score would be

$$(0.5 - 1)^2 + (0.3 - 0)^2 + (0.2 - 0)^2 = 0.38.$$

Therefore, the lower the Brier score the better the prediction model.

4.2 Geometric Mean

The geometric mean, in contrast to the arithmetic mean, uses the product in a measure of central tendency. It can be used as a performance measure by averaging over the probabilities predicted for the observed outcome of a game. Therefore, it only takes the observed outcome of a game into account. For a number of N games and a predicted probability of p_i for the outcome of game i, the geometric mean is defined as

$$G = (\prod_{i=1}^{N} p_i)^{\frac{1}{N}}.$$

This results in a normalized pseudo-likelihood of the predictions and consequently, a higher geometric mean is the preferable score.

4.3 Betting Dis-/Advantage

In section 3.4, the structure of decimal odds is discussed. By multiplying a decimal odd o_{gr} with its corresponding probability p_{gr} for a game $g \in G$ and result $r \in R$, one obtains the expected return k_{gr}, when betting a one unit wager on that result. If a model generates better probabilities than a bookmaker, on average, betting on the outcome of a game with the highest return $h_g = \arg\max_{r \in R} k_{gr}$ will result in a profit. This holds if the bookmaker does not employ an overround, otherwise the maximum expected return $k_g > 1$ must hold.

The Kelly criterion is a criterion developed to calculate the optimal amount of money that should be placed on a bet, given the odd and probability of a result.[27] It is frequently used by professional bettors and was also used to generate a betting strategy by Hvattum and Arntzen with the predictions generated by their elo-based model.[17] For a betting scenario in which the bookmaker uses an overround, it is defined as $b = (op - 1)/(o - 1)$ for a posted decimal odd o and computed probability p. The result b is the fraction of the budget one should optimally place, to achieve exponential growth. Using the Kelly criterion, a betting strategy can be constructed. A bet of size $b_g = ((k_{gh_g} - 1)/(o_{gh_g} - 1))10$ is placed on the result h_g of a game, assuming a fixed budget size of 10 units per game. If h_g is equal to the actual result of the game a_g, the profit is the bet sizing multiplied by the odd, $w_g = b_g o_{gh_g}$. If the results are not equal, the wager is lost, $w_g = -b_g$. Additionally, if a game does not have any profitable outcome on which to bet, no bet is placed.

Using this betting strategy, the betting advantage A of a model can be computed. Formally, it is defined as

$$A = \frac{1}{N} \sum_{g=1}^{N} w_g \qquad w_g = \begin{cases} 0 & \text{if } k_g < 1 \\ b_g o_{gh_g} & \text{if } h_g = a_g \\ -b_g & \text{else} \end{cases},$$

with h_g, k_g and b_g defined as

$$h_g = \arg\max_{r \in R}(o_{gr} p_{gr}) \quad k_g = \max_{r \in R}(o_{gr} p_{gr}) \quad b_g = ((k_{gh_g} - 1)/(o_{gh_g} - 1))10.$$

With this strategy, only one bet per game is allowed, so as to keep the betting strategy as simple as possible. Commonly, bookmakers also allow for bets to be placed on multiple results of a game and also to combine bets between games to increase the odds. However, these possibilities are not considered. The model scoring the highest betting advantage then subsequently has the economic advantage over the others.

4.4 Elo Regression

To put the models into context with previous models, the logistic regression model based on win margin optimized elo values and posited by Hvattum and Arnzten is used (ELOReg).[17] They used an ordered logistic regression, but since the outcomes of a game have no intrinsic ordering, a plain multinomial logistic regression was implemented in this study. Changing the model architecture lead to a sizable performance increase compared to the findings of the paper. Still, ELOReg is suitable as a comparison model by providing a scope of the performance of models from past literature. Finally, the model was trained using the ELO values from the 2005/2006 season up to the 2015/2016 season.

5. Bayesian Neural Networks

BNNs are a variant of ANNs, in which the parameter values of a network underlie a prior distribution. Training of BNNs is inherently different from ANN training, since no longer single values, but distributions need to be learned. On top of this, due to the nonlinear, multi-layer structure of BNNs, an exact evaluation of the posterior distributions of the network parameters is impossible. Instead, variational inference is used to approximate the posterior distributions using approximating variational distributions.[6, pp.277–284] In this chapter, a short explanation of variational inference will be given, along with an overview of the adopted network architecture, how the networks were trained and lastly, how a network is sampled to obtain a prediction.

5.1 Variational Inference

In a Bayesian model, a prior distribution is posited on a set of latent random variables θ. In inference, the posterior distribution $p(\theta|x)$ needs to computed, given observations x_n of a random variable X. As mentioned, in the case of BNNs and other large models, this computation is intractable. Therefore, a variational distribution $q(\theta|\phi)$, with $\phi \in \Phi$ as parameters, is used as an approximation. To ensure that the variational distribution estimates the true posterior distribution as well as possible, the distance between them needs to be minimized. This is done by minimizing the KL divergence,

$$\phi' = \arg\min_{\phi \in \Phi} KL(q(\theta|\phi) \,||\, p(\theta|x)). \tag{5.1}$$

After which $q(\theta|\phi')$ can be used as an approximation of the posterior. Furthermore, equation 5.1 can be transformed into

$$ELBO(\phi) = \mathbb{E}_q[\log p(x, \theta)] - \mathbb{E}_q[\log q(\theta|\phi)], \tag{5.2}$$

where instead of minimizing the KL divergence, the evidence lower bound (ELBO) is maximized. Maximizing the ELBO is insofar an improvement, that now the intractable posterior $p(\theta|x)$ is no longer required and a numerical solution can be computed.

Specifically, an extension to variational inference, automatic differentiation variational inference (ADVI) is used to train the models. It alleviates the problem of choosing a suitable variational distribution q, by mapping the model into a space where the latent variables are unconstrained.[7]

5.2 Network Architecture

Different network architectures were tested to attempt to achieve optimal performance. In the end, all three models used the same architecture, except for the size of the input layer. A common problem in training the models, was that it would often converge into a local minimum, such that a model would only learn the distribution of outcomes of the training data set. This would reflect itself in close to constant predictions of $[0.45, 0.27, 0.28]$ for a home team win, draw and away team win respectively, irrespective of the input. Only for small network sizes, was a variable output and therefore improved predictions obtained.

In total, one *linear* input layer, three hidden layers with a *tanh* activation function and one $softmax$ output layer were used. The number of input nodes was equal to the input dimensions of the respective data set, while each hidden layer contained five and the output layer three nodes, one for each outcome. The $softmax$ function used in the output layer guarantees an output vector that sums up to one, while all entries range between zero and one. Finally, each nodes' weight and bias means were initialized randomly from the standard normal distribution, with the standard deviations initialized randomly between zero and one.

5.3 Training

Due to the aforementioned effect, that training commonly did not converge to provide variable predictions, multiple networks had to be trained. After evaluation, the best of 100 networks, in respect to the Brier score and geometric mean, were chosen as the model network. Batch training was used to train a single network, with 250 samples per batch over a total of 6000 iterations of ADVI. For a single batch, 250 samples were chosen at random from the training data set. See Figure 5.1 for a plot of the convergence of ELBO values of one training procedure. After circa 4500 iterations, the values converge to around -4100, using a learning rate of 0.01.

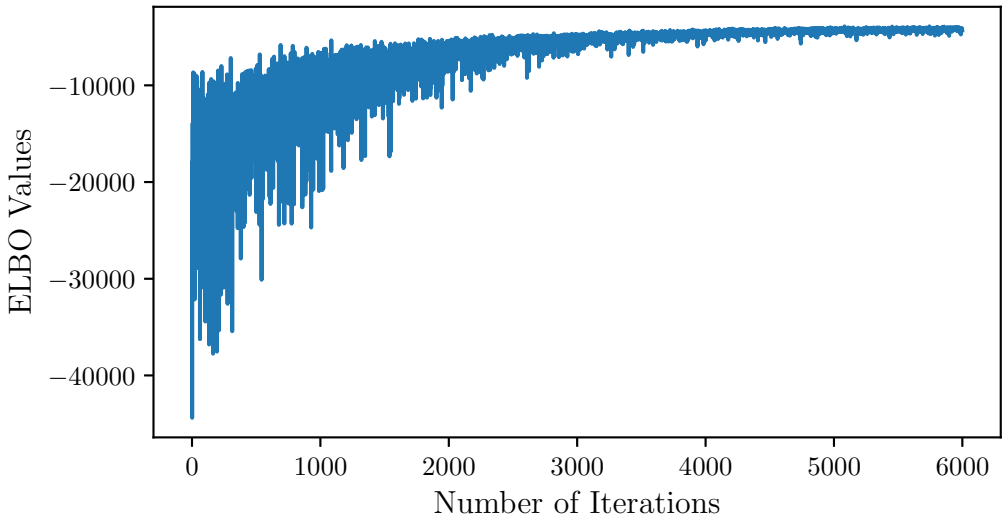

Figure 5.1: ELBO values of one training procedure plotted against the number of iterations.

5.4 Sampling

By taking a sample from all network weight and bias distributions, one obtains a single MLP with real valued parameters and can avoid the extensive process of solving the network for new data from the test data set. To obtain an estimation of the actual BNN output, a data sample can be fed through the sampled network. Multiple MLPs can then be sampled and used to obtain an output set of predictions for a single game as input. By computing the mean and standard deviations of the results, an increasingly better estimation of the actual output is obtained, the more MLP samples are taken from the BNN. In total, 500 samples were taken of each model network. The means of the outputs for a single game then culminated in the probabilities ascribed to the three different results, with the standard deviation representing a degree of uncertainty of the probability.

6. Evaluation

Through the usage of BNNs not only a probability mean, but also a standard deviation, interpretable as uncertainty, was obtained for a prediction of a result of a game. This allows for a further layer of analysis of model differences, on top of the performance measures. As the ELOReg model does not provide any uncertainty measures, it was omitted from any uncertainty analysis. In the following, first the results of the different performance measures are given. Afterwards, the output means and standard deviations are analyzed and finally all results combined and discussed.

6.1 Performance Measures

As mentioned, the 2016/2017 EPL season was used to evaluate the performance of all models using the Brier score, geometrical mean and betting advantage. Table 6.1 summarizes the observed results. All significance tests were conducted two-tailed and using a 95% confidence interval. To perform the t-test for the geometric mean, the logarithm of the geometric mean and geometric standard deviation was used.

To begin with, the statistical measures, Brier score and geometric mean, are inspected. As expected, the Brier score and geometric mean both provided very similar results, with a model scoring well in one, also scoring well in the other. When comparing the models to the bookmaker odds, all of them performed worse by a significant margin. Even the best model, the financial model, featured a substantial difference of 1.26% for the Brier score and 0.72% for the geometric mean, compared to the odds' performance. The differences between the models among themselves were more moderate. When ranked, the financial model featured the best scores, followed by the combined and ELOReg models. The worst performance recorded was made by the performance model. Still, no model presented significantly different results from the others. Hence, no model can confidently be proclaimed better than the others, on the basis of the Brier score and geometric mean alone.

Model	Brier Score	Geometric Mean	Betting Advantage
Performance	0.5572 (5)	0.3895 (5)	0.2561 (2)
Financial	0.5433 (2)	0.3978 (2)	0.2626 (1)
Combined	0.5489 (3)	0.3946 (3)	0.1321 (4)
ELOReg	0.5496 (4)	0.3942 (4)	0.1885 (3)
Odds	0.5307 (1)	0.4050 (1)	

Table 6.1: Overview of the performance measures for all models and the bookmaker odds. The rank in the respective measures is given in parentheses.

The betting advantage, on the other hand, presents a different result. All models were capable of scoring a positive revenue, with the financial model again reaching the highest score. Surprisingly, the performance model followed close behind, while the ELOReg and lastly the combined model were not able to score as well by a considerable margin. Even though the financial and performance model achieved very similar betting advantages, only the financial model showed to be significantly different from zero and therefore confidently profitable. Yet again, a statistical difference between any of the models can not be reported.

Further insight is gained by regarding the cumulative sum of bets made by a model over the progression of a season, plotted in Figure 6.1. Four distinct trajectories can be analyzed. The performance model tended to struggle in the beginning, but achieved large profits during the middle of the season to overtake the other models. The financial model on other hand, featured a more consistent and steady incline over the entire season. So did the combined model, though with a less steep slope. Lastly, the ELOReg model, alike to the performance model, lost a large portion of its bets in the beginning, but was able to make consistent profits by the middle of the season until the end. In the end, the largest net gain was scored by the financial model at 99.26 units, closely followed by the performance model at 96.80. The ELOReg model achieved a moderate 71.27, while the combined model scored half of the financial model at 49.93.

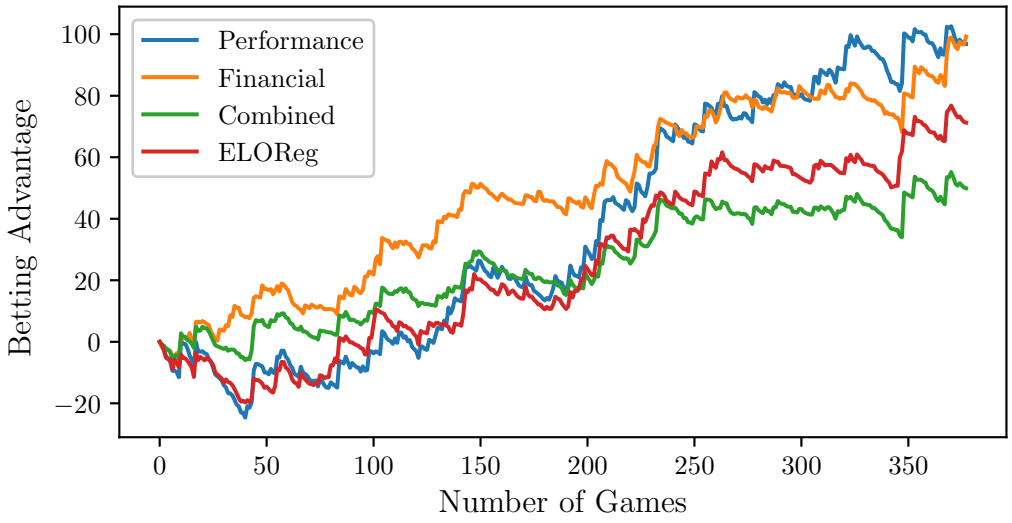

Figure 6.1: Cumulative betting advantages plotted against the number of games of the three different models over the 2016/2017 EPL season.

6.2 Output Analysis

The analysis of the descriptive statistics of the prediction means and standard deviations also holds valuable information about model differences. For example, tendencies towards a certain result or higher uncertainty can be used to explain why a model performs worse or better than the others. To begin with, the mean prediction probabilities and mean standard deviations for every result are considered (Table

6.2). On average, of all models the performance model predicted a home team win with the lowest probability and the combined model with the highest. The exact reverse holds true for the away team win probabilities, where from the combined over the financial and the ELOReg to the performance model, the respective mean probabilities rose. The mean probabilities for a draw were close to equal for all four models, with the combined model weighting it 1% less. Compared to the observed distribution of results in the test data set of $[0.4933, 0.2202, 0.2865]$, the combined model achieved the closest fit.

When examining the standard deviation means over the different results, only subtle differences are found. Overall, the financial model featured the least uncertainty over all results, with only marginal differences between the performance and combined model. Striking was the extremely small uncertainty for the draw probability for all models, compared to the other results. The difference in uncertainty between a home team and away team win was comparably small, with the uncertainty for the home team win on average circa 1% higher.

Model	HTW	DRW	ATW
Probability Means			
Performance	0.4439	0.2539	0.3022
Financial	0.4671	0.2577	0.2753
Combined	0.4860	0.2461	0.2679
ELOReg	0.4629	0.2586	0.2785
Standard Deviation Means			
Performance	0.0792	0.0300	0.0662
Financial	0.0519	0.0271	0.0466
Combined	0.0794	0.0364	0.0670

Table 6.2: Table of probability and standard deviation means of the predictions of all models from the 2016/2017 EPL season. HTW, DRW and ATW denote home team win, draw and away team win respectively.

Lastly, the relationship between single predictions and their respective uncertainty is investigated. In Figure 6.2, the uncertainty of a prediction of a result is plotted against the corresponding probability. For all three results and all three models, an arch structure can be identified. In the case of extreme probabilities for an outcome, the uncertainty of a model becomes minimal, while for more moderate probabilities the uncertainty rises. Especially the performance model features a tight trajectory along its arch structure, with the financial model clustering tightly at the extreme ends of result probabilities, but exhibiting the largest spread in the center. Also worth noting are the breadths of the combined and financial arches, compared to the performance arches. In all three different results, the combined and financial models feature a larger range of probabilities. All of these properties hint at the fact that the performance model features a limited variability in comparison to the other models.

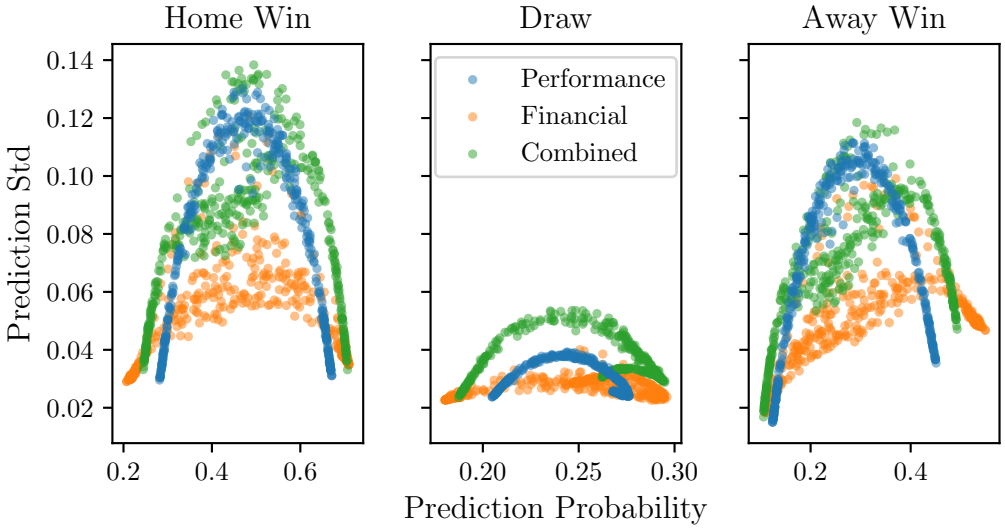

Figure 6.2: Standard deviation plotted against the probability of a result of a game of the 2016/2017 EPL season for all three models.

6.3 Discussion

To begin with, Johnstone is able to show that, to maximize future profits, prediction models should be ranked by statistical scores and not economically.[28] This means a model achieving better statistical scores should be used, instead of a model which granted a larger profit profit in the past. The underlying rationale is that financial evaluation is commonly grounded on a volatile market, which leads to highly variable profits on the basis of chance factors. A model which provides accurate probabilities is hence more likely to profit than another, which showed a high profit due partially to chance on a volatile market. Therefore, the results from the betting advantage should be handled and considered with care.

Even though the different evaluation methods revealed partially contradicting results, an overarching conclusion across all of them can be drawn. With the worst statistical scores of all models, the performance model was unable to perform as well as the financial model. The financial model on the other hand, accomplished the highest score of all models in not only the statistical but also the economic analysis. Therefore, financial football data is highly effective at encoding information about team strength and is at least on par with a subset of classical predictors based on past performances of a team.

The interpretation of why the performance model performs worse than the financial and combined models is supported by the analysis of Figure 6.2. On the one hand, it is less capable of deciding a clear favorite and underdog in games with low uncertainty, based on the lacking breadth in the range of probabilities. On the other hand, in games between equally matched teams it features higher, but static uncertainty. This means it is unable to differentiate between for example, two games in which in the first, two equally matched but worse teams play against one another, and in the other two better, but also equally matched teams square off. Due to both of these factors, slightly sub-optimal probabilities are obtained, which reflect themselves greatly in the Brier score and geometric mean.

Going into more detail, by examining the minimal Brier score and geometric mean differences between the combined and financial model, one could infer that the informational value of the financial data fully overlaps the informational value of the performance data. Through the financial data explaining any effects the performance data can explain and more, the combined model completely relies on the financial data and achieves close to equal performance. When regarding the betting advantage though, this is evidently not the case. If the combined model were basing all predictions to a large degree only on the financial data, the betting advantage would likely also be very similar, which is not the case. On top of this, the graph of the cumulative bets in Figure 6.1 of both the performance and financial models feature substantially different trajectories. Therefore, the encoded information of both data sets will likely intersect, but not overlap.

If the combined model can then benefit from the information of both data sets, this raises the question of why the combined model, even though showing good statistical scores, performs comparably bad in betting. Multiple reasons could weigh in to explain this effect. First and foremost, is the chance factor involved in betting and in football games in general. This is connected to the aforementioned findings of Johnstone. The combined model could potentially have come out unlucky in this particular set of bets and will on average score better. Secondly, even though the odds were averaged over 16 different prominent bookmakers, a completely unbiased set cannot be obtained. Tendencies and biases in the odds are the reason why a model is able to exploit them and achieve a positive return. Table 6.2 shows the different models also exhibit biases towards certain results, making therefore some models better at exploiting a certain set of odds than others. Thus, the combined model is unable to turn good statistical scores into a highly profitable strategy, by being unable to exploit the odds optimally.

A further interesting point is the discussion of how a model is able to create a profitable betting strategy against bookmaker odds, even though it produces statistically worse probabilities and an overround of 3.31% is used by the bookmaker. For this, the betting behavior of the different models is analyzed. To score a profit, optimally a large portion of bets needs to be won. Counter intuitively, the win percentages of all models was relatively low, ranging between 26.17% and 28.18%, compared to the accuracy scores of up to 60% when predicting the winner. However, by using the Kelly criterion, large bets are only placed on highly profitable games, in which a large discrepancy between the computed probabilities and the bookmaker probabilities is found. On top of this, by placing low stake but high risk bets, a model is able to score large profits when it is correct and lose little when it is not. The ELOReg model showed the lowest average odd in which bets were placed at 4.61, equaling a probability of 21.69%. This probability is extremely low, since the largest portion of bets was placed on the home team and the financial model featured the lowest probability for this result at 20.70%. Hence, the models are able to profit against the bookmaker odds by placing highly profitable longshot bets on high odds.

This conflicts with the results found by Cain, Law and Peel.[19] They report a bias of bookmaker odds towards the favorite, but this discrepancy is likely due to the change in betting markets since those findings. At that time, when betting on results, bookmakers required the combination of multiple bets. A bias towards favorites reduces the average size of all odds, thereby reducing the multiplicative

effect of combining longshot bets and the risk of the bookmaker losing large sums in a single bet.

Lastly, the intuitively perceived disadvantage of the financial data being static across an entire season, turns out be an advantage. Observing the Brier score over each match independently shows that all models indicate a slight but insignificant performance increase over the course of a season. Therefore, the assumption that the financial data becomes dated during the end of the season and therefore no longer describes the strength of a team well, could not be shown to hold. At the same time, the performance model also did not show a significantly different performance between the first five and the following matchweeks. Poor estimation of team strength at the beginning of a season for the performance data could subsequently also not be shown. The financial data thus has insofar an advantage, that it is not dependent on how a team has played in the past and hence is not susceptible to volatile changes, which are unconnected to the team strength. These changes arise from the numerous chance factors involved in a football game, resulting in game outcomes that do not coincide with the actual team strengths. By estimating the team strength at the beginning of a season, the financial data is not influenced by random fluctuations and, incorporated in a model, better predicts the outcomes of football games.

7. Conclusion

Despite few models from past literature being able to profit when betting against bookmaker odds, some improvements can and some acknowledgments have to be made to the posited models. Especially the feature extraction and selection leaves room for improvement. Only a small and limited subset of predictors was considered. Specifically for the performance data set, an abundance of statistics are recorded, such as scoring chances, ball possession and fouls committed, which were not considered in the models. These may hold additional valuable information about team strength, aiding to close the gap between the financial and performance model and further increase the performance of the combined model to match bookmaker odds.

On top of this, even though considerations and tests were conducted for investigating the influence and relevance of predictors, conclusive evidence of how and to what degree a model is using a certain predictor is not gained. This is due to the black box nature of ANNs. In spite of all advantages, parameters of ANNs commonly yield no information about the actual weighting of an input. By using an entirely different model this problem can be alleviated. Other models exist which provide relevant information about the effects of predictors, but do not scale well for large models. However, since only small BNN sizes were necessary to achieve maximum performance, large scale models are not required and the usage of a more transparent model is sensible.

A further shortcoming of the BNNs, and ANNs in general, can be seen in the statistical performance measures. Theoretically, since the performance model can take advantage of the elo values and even further information, it should at least be on par with the ELOReg model. Similarly, the combined model's performance would equal or surpass the performance of all other models, if training would converge in the global minimum, given that it has the most amount of data at its disposal. There in lies the problem, that training ANNs is commonly a trial and error process and can require a lengthy process of fine tuning. In this case, even after extensive testing it was not possible to achieve the theoretically possible results.

Finally, improvements to the betting strategy are considered. Based on the knowledge of the odd and probability of a result, the Kelly criterion computes the optimal bet sizing for exponential growth on a variable budget. This variable budget results in erratic and volatile behavior, in which large wins but also large losses can develop in quick succession. In the proposed betting strategy, a fixed budget of 10 units per game was set to achieve more stable behavior. A more stable behavior could potentially be favorable over a profit maximizing one, as it ensures a long term steady growth, instead of a possibly large win or loss. In this respect, a natural extension, that is made available by the usage of BNNs, is the integration of uncertainty. For higher risk bets, i.e. betting on a result with a large odd and small probability, the Kelly criterion proposes small bet sizes. The uncertainty of

a probability could be used to influence the bet sizing in a similar way. For high uncertainty of a probability, bet sizes should be kept small, as betting on a result with low certainty is risky. For probabilities with a high certainty on the other hand, bet sizes can be increased. A further method to ensure steady behavior would be to replace the Kelly criterion by a function prioritizing long term growth. Using approaches such as reinforcement learning, such a function could be learned. In addition, more alternative betting options such as combinations of bets and betting on multiple outcomes could be included, broadening the range of betting options.

Concluding, it can be said that, through the usage of financial data, the proposed models perform exceptionally well compared to models from past literature. As the market of the major football leagues continue to grow, the informational value and importance of financial predictors is unlikely to decrease. The inclusion of financial data in models for football prediction will thus continue to be sensible. Still, classical predictors based on past performances of teams are not obsolete and also hold valuable information not encoded in financial variables. Lastly, potential for improvement to achieve even better predictions of association football remains. Either by including more data, analyzing model behavior more thoroughly or using superior model architectures, statistically better predictions than bookmaker odds still remain to be achieved.

Bibliography

[1] A. Heuer. *Der perfekte Tipp: Statistik des Fußballspiels*. John Wiley & Sons, 2013.

[2] www.transfermarkt.co.uk. *Transfer Market Data*. 2017. URL: `https://www.transfermarkt.co.uk/` (visited on 01/07/2017).

[3] Deloitte. *Annual Review of Football Finance: Reboot*. 2016.

[4] D. Silver et al. "Mastering the game of Go with deep neural networks and tree search". In: *Nature* 529.7587 (2016), pp. 484–489.

[5] M. Moravčík et al. "DeepStack: Expert-Level Artificial Intelligence in No-Limit Poker". In: *arXiv preprint arXiv:1701.01724* (2017).

[6] C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006, pp. 225–256.

[7] A. Kucukelbir et al. "Automatic differentiation variational inference". In: *arXiv preprint arXiv:1603.00788* (2016).

[8] H. Wang and D.-Y. Yeung. "Towards Bayesian deep learning: A framework and some existing methods". In: *IEEE Transactions on Knowledge and Data Engineering* 28.12 (2016), pp. 3395–3408.

[9] Y. Gal and Z. Ghahramani. "On modern deep learning and variational inference". In: *Advances in Approximate Bayesian Inference workshop, NIPS*. 2015.

[10] B. L. Boulier and H. O. Stekler. "Are sports seedings good predictors?: an evaluation". In: *International Journal of Forecasting* 15.1 (Feb. 1999), pp. 83–91.

[11] R. T. Stefani. "Football and basketball predictions using least squares". In: *IEEE Transactions on systems, man, and cybernetics* 7 (1977), pp. 117–121.

[12] R. T. Stefani. "Improved least squares football, basketball, and soccer predictions". In: *IEEE Transactions on Systems, Man, and Cybernetics* 10 (1980), pp. 116–123.

[13] A. McCabe and J. Trevathan. "Artificial intelligence in sports prediction". In: *Information Technology: New Generations, 2008. ITNG 2008. Fifth International Conference on*. IEEE. 2008, pp. 1194–1197.

[14] M. J. Maher. "Modelling association football scores". In: *Statistica Neerlandica* 36.3 (1982), pp. 109–118. ISSN: 1467-9574.

[15] M. J. Dixon and S. G. Coles. "Modelling Association Football Scores and Inefficiencies in the Football Betting Market". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 46.2 (1997), pp. 265–280.

[16] I. Asimakopoulos and J. Goddard. "Forecasting football results and the efficiency of fixed-odds betting". In: *Journal of Forecasting* 23.1 (2004), pp. 51–66.

[17] L. M. Hvattum and H. Arntzen. "Using ELO ratings for match result prediction in association football". In: *International Journal of forecasting* 26.3 (2010), pp. 460–470.

[18] G. Baio and M. Blangiardo. "Bayesian hierarchical model for the prediction of football results". In: *Journal of Applied Statistics* 37.2 (2010), pp. 253–264.

[19] M. Cain, D. Law and D. Peel. "The Favourite-Longshot Bias and Market Efficiency in UK Football betting". In: *Scottish Journal of Political Economy* 47.1 (2000), pp. 25–36.

[20] S. Levitt. "Why are gambling markets organised so differently from financial markets?*". In: *The Economic Journal* (Jan. 2004).

[21] D. Forrest and R. Simmons. "Sentiment in the betting market on Spanish football". In: *Applied Economics* 40.1 (2008), pp. 119–126.

[22] A. E. Elo. *The rating of chessplayers, past and present*. Arco Pub., 1978.

[23] www.fussballdaten.de. *Football Fixture Data*. 2017. URL: `https://www.fussballdaten.de/england` (visited on 01/07/2017).

[24] www.oddsportal.com. *Football Odds Data*. 2017. URL: `http://www.oddsportal.com/soccer/england/premier-league-2016-2017/results/` (visited on 01/07/2017).

[25] A. C. Constantinou and N. E. Fenton. "Evaluating the Predictive Accuracy of Association Football Forecasting Systems". In: *Technical Report* (2010).

[26] G. W. Brier. "Verification of forecasts expressed in terms of probability". In: *Monthly weather review* 78.1 (1950), pp. 1–3.

[27] J. L. Kelly. "A new interpretation of information rate". In: *Bell Labs Technical Journal* 35.4 (1956), pp. 917–926.

[28] D. Johnstone. "Economic Darwinism: Who has the Best Probabilities?" In: *Theory and Decision* 62.1 (2007), pp. 47–96.

Declaration of Authorship

I hereby certify that the work presented here is, to the best of my knowledge and belief, original and the result of my own investigations, except as acknowledged, and has not been submitted, either in part or whole, for a degree at this or any other university.

Osnabrück, August 3, 2023