# Neural Networks predictive modeling for Football Betting

John Sibony[a], Abdelfatah Tlemsani [a], Youssef Hamchi[a], Monika Gjergji[a], Evgueny Shurmanov[b], Marius Cristian Frunza[b,c]

[a]*Dauphine University, Place du Marechal de Lattre de Tassigny, 75016, Paris, France*
[b]*Ural Federal University(UrFU), 620002, 19 Mira street, Ekaterinburg, Russia*
[c]*Schwarzthal Tech,231b Business Design Centre, 52 Upper Street, Islington, London, United Kingdom, N1 0QH*

## Abstract

The aim of this paper is to explore neural network based modeling strategies for football betting. An neural network model and a challenger model based on a traditional econometric approach(Dixon and Coles (1997)) were estimated on data from five national leagues results including France, Spain, Italy, Germany, England) Our results show that the Neural network based model has better predictive accuracy compared with the traditional econometric models. Betting strategies were implemented using prediction outputs generated with both econometric and neural networks models. The latter provide with better return over investment. Nevertheless, both approaches lead to losses in the long run.

*Keywords:* Sport betting football, Machine Learning, Neural Networks, Statistical Models

## 1. Introduction

Football is the most popular sport and as such, it has been known for drawing strong and often extreme emotions out of its millions of devoted fans. Thus, it is unsurprising that with football's ever-growing popularity, betting activities have also flourished exponentially, especially in recent years with the rise of online betting. According to Hing (2014), the possibility of online gambling on football matches plays a key role in the rise of popularity that football has experienced as of late, particularly in regions that have not historically exhibited a strong relationship with the sport, such

*Email addresses:* john.sibony@hotmail.fr (John Sibony ), abdelfatah.tlemsani@gmail.com (Abdelfatah Tlemsani )

*Preprint submitted to Elsevier* *July 24, 2020*

Electronic copy available at: https://ssrn.com/abstract=3655700

as the United States and China.

A combination of deregulated gambling and online betting has led to what The Guardian calls the "gamblification" of watching football. According to a recent study by Dr. Darragh McGee, fans indulging in excessive football betting have suffered "dire consequences" in its wake with McGee (2018) concluding that betting is "far from being the knowledge-based", risk-free activity it is marketed as.

Indeed, the reasons behind betting losses are inevitably linked to bettors making decisions based on their own pre-conceived beliefs which are not based on any strong statistical evidence as the outcome of football matches is ultimately characterized by numerous uncertainties. A large portion of this outcome cannot be explained by measurable factors. While sports betting is not based purely on mathematics, many statistical models are used to predict final match outcomes which have become increasingly more sophisticated in order to improve prediction accuracy.

The betting industry generates billions in revenues, thriving despite the enforcement of restrictive legislation in several countries. This industry is composed of a large number of participants on a global scale. The betting market in the United Kingdom only was worth GBP 365 million in 2011. Betting markets provide a platform for bookmakers to list their odds for all possible outcomes of sporting events, here, football matches. A single football match will have separate markets for betting on the match result, the number of goals scored, the goal scorers and the victory margin. Bookmakers sometimes offer dozens of betting markets for high-profile fixtures. Some of the most popular bookmakers for sporting events are: Bet365, William Hill, Ladbrokes, Paddy Power, Betfred, BetVictor, betway, Unibet etc.

According to Sauer (1998), the betting market is considered efficient due to the fact that both parties, bettors and bookmakers, are interested in turning a profit. However, several studies have determined potential biases in this market such as *bettor sentiment* (Levitt (2004)) among which optimism/perception bias and loyalty bias, and home - away teams bias (Paul and Weinbach, 2009). This market holds great importance for academic research, in large part due to the ever-growing volume of placed bets and bookmakers which consistently brings new changes to the sports betting business and sports events themselves.

Electronic copy available at: https://ssrn.com/abstract=3655700

The *overround*, or the price of the odds, is a key factor in the betting market. In a fixed-odds setting, it refers to the profit margin of bookmakers wherein the odds are set such that the probability of the event occurring, i.e. the inverse sum of the odds, is strictly greater than one. The entry of new companies in the market has consistently influenced the overround which has in turn changed the structure of the market itself. In addition, the cross evolution of both the number of bookmakers and the overround has also been a topic of much interest to researchers due to the potential arbitrage opportunities it has introduced to the betting market, as first shown in the study of Pope and Peel (1989). For instance, this phenomenon has given rise to *surebets* which refer to the variation of odds across bookmakers and which offer the possibility of placing riskless bets, i.e. always making a profit regardless of the outcome. Consequently, the decrease of the values of the overround, implying reduced margins, combined with more dispersed odds due to the growing number of market participants, should generate more arbitrage opportunities in the betting market (Frunza (2015),Frunza (2014),Gomez-Gonzalez and Del Corral (2018) ).

This article enriches the academic literature about sport betting modelling and develops the topic of neural networks-based methodologies. The remainder of our paper is structured as follows:

- **Section 2** reviews extensively the current sport betting models and present some of the traditional as well as the newer models using a Machine Learning framework on the prediction of betting odds.
- **Section 3** presents and describes the data we have used as well as our chosen methodology.
- **Sections 4** describes our methodology in detail by first implementing a simple martingale strategy and then take an advanced Machine Learning approach, more specifically through recurrent neural networks.
- **Sections 5** discussed the results of our study.
- **Section 6** concludes on our findings as well as suggest potential new areas for research in this particular market.

Electronic copy available at: https://ssrn.com/abstract=3655700

## 2. Sport betting models

This section reviews the existing models on the prediction of football match outcomes, covering traditional models such as the ones based on goal scores and ordered logistic regressions as well as more recent literature on predictive betting models using a more advanced approach with a Machine learning framework.

### *2.1. Goal models*

Predictive models of football match outcomes in extant literature can be divided into two main categories. First, we have the *goal models* which initially forecast the amount of goals scored per team which through which the final match outcome is indirectly predicted.

One such model was first proposed by Maher (1982) which forecasts the number of goals scored by both the home and the away team by using Poisson distributions. The number of goals scored by the home team against the away team was denoted as the variable $X_{ij}$, where $i$ and $j$ are respectively the home and away teams while the variable $Y_{ij}$ represents the number of goals conceded by team i to team j. The variables $X_{ij}$ and $Y_{ij}$ are assumed to be independent. In addition to the attacking and defending parameters for both teams, another parameter is included which represents the home team's advantage.

Maher then assumes that each team $i$ has an attacking strength denoted as the parameter $\alpha_i$ as well as a defending strength denoted as $\beta_i$. A high value of $\alpha_i$ indicates that the team $i$ scores many goals while a low value of $\beta_i$ indicates that the team $i$ tends concede a small number of goals. In addition to the attacking and defending parameters for both teams, another parameter is included which represents the home team's advantage. This parameter is denoted as $k$ and is assumed to be the same for all teams. The **Maher** model determines that the outcome of the game between the home team $i$ and the away team $j$ follows a probability distribution formulated as:

$$P(X_{ij} = x, Y_{ij} = y)|\alpha, \beta, k) = Poisson(x|k\alpha_i\beta_j) \;*\; Poisson(y|k\alpha_j\beta_i) \tag{1}$$

where $Poisson(x|\lambda)$ denotes the Poisson probability distribution function with parameter $\lambda$ evaluated at $x$. Maher then found maximum likelihood estimators for each of the parameters in his model.

Electronic copy available at: https://ssrn.com/abstract=3655700

There have been two main criticisms towards the **Maher** model. First, if X follows a Poisson distribution then that would imply that the first two central moments of the distribution, i.e. the mean and the variance, are equal: $\mathbb{E}[X] = Var(X)$. However, in many studies it has been empirically found that the variance tends to be larger than the expectation, hence violating the Poisson distribution assumption.

In order to remove this violation, an extension of the initial **Maher** model was proposed by Karlis and Ntzoufras (2009) by adapting the model to a Bayesian setting. We first determine:
$(X|\Lambda = \lambda) \sim Poisson(\lambda)$ where $\Lambda$ is a random variable following a Gamma distribution. By the law of total variance we would then have: $Var(X) = \mathbb{E}[\Lambda] + Var(\Lambda) > \mathbb{E}[\Lambda] = \mathbb{E}[X]$, an assumption which is in line with the authors' empirical findings.

Furthermore, two additional extensions to the independence assumption of the **Maher** model were proposed by Dixon and Coles (1997). This model is able to generate ex ante match outcome probabilities. They defined the match outcome probabilities in a bivariate Poisson model, in order to account for the existing correlation between goals scored by each team, such that:

$$P(X_{ij} = x, Y_{ij} = y)|\alpha, \beta, k) = \tau(x, y) \; * \; Poisson(x|k\alpha_i\beta_j) \; * \; Poisson(y|k\alpha_j\beta_i) \quad (2)$$

where $\tau(x, y)$ is a function that makes low-scoring draws such as 0-0 or 1-1 more probable and the outcomes 1-0 and 0-1 more probable than in *Maher*'s model.

Secondly, they lift the assumption on the constancy of the attacking and defending parameters of each team as it does not correspond to the reality of football matches throughout seasons. A potential unavailability of certain football players either due to injury or suspension as well as changes made in the transfer window or in a team's game strategy can lead to fluctuations in the offensive and defensive abilities of the team, i.e. the parameters $\alpha$ and $\beta$. Hence, Dixon and Coles proposed a time-dependent model where team $i$'s attacking and defending abilities at time $t$ are respectively represented by the parameters $\alpha_i^t$ and $\beta_i^t$.

Rue and Salvesen (2000) improved the latter extension by defining a random walk

Electronic copy available at: https://ssrn.com/abstract=3655700

model with $\alpha_i^t | \alpha_i^{t'} \sim N(\alpha_i^{t'}, \frac{t-t'}{\tau}, \sigma^2)$ and similarly so for the defensive strength parameter. The parameter $\tau$ represents the 'strength of memory' and is estimated from real data. Thus, they assume that the time varying offensive and defensive abilities of a team follow a random path over time. In addition, they use Bayesian methods in order to continuously update the prior estimates of these parameters when new information on match results is received. They use Markov-Chain Monte Carlo iterative simulation methods for inference.

Crowder et al. (2002) develop a more flexible computational procedure for updating the team strength parameters. Other researchers have analyzed the impact of certain specific factors on match outcome predictions. Barnett and Hilditch (1993) investigate whether artificial playing surfaces first introduced and subsequently abandoned by football clubs during the 1990s, provided additional benefits to the home team's advantage. Additionally, Ridder et al. (1994) show that player dismissals have had a negative effect on match results for teams competing with fewer than 11 players. Clarke and Norman (1995) apply a number of non-parametric techniques to analyze and identify the effect of home advantage on match results. Dixon and Robinson (1998) analyze variations in the scoring rates of the home and away teams during the course of a match. These scoring rates depend upon the number of minutes elapsed, but also upon which (if either) team is leading at the time.

*Maher*'s framework can also be adapted to a Skellam distribution. In this particular case, the win, loss or draw result of a football match is modelled as the difference between the number of goals scored and those conceded by each team. On the other hand, the difference between the number of home team goals X and away team ones Y is defined as a team's margin victory. If we define Z as Z = X + Y where Z follows a Skellam distribution with intensities $\lambda_1$ and $\lambda_2$, we have the following probability function:

$$P_{sk}(Z = z|\lambda_1, \lambda_2) = e^{-(\lambda_1+\lambda_2)}(\frac{\lambda_1}{\lambda_2})^{z/2} I_{|z|}(2\sqrt{\lambda_1\lambda_2}) \tag{3}$$

where $I_{|z|}(.)$ is a modified Bessel function of order $|z|$.

Karlis and Ntzoufras (2009) showed that the underlying Poisson assumption of the original Maher model could be replaced by a Skellam distribution which could be considered as a distribution defined on integers. When modeling football match

Electronic copy available at: https://ssrn.com/abstract=3655700

outcomes, $\lambda_1$ corresponds to number of goals scored by the home team and $\lambda_2$ the number of goals scored by the away team. Therefore, during a match between home team $i$ and away team $j$ we can define the victory margin variable $Z$ as: $Z = X_{ijt} - Y_{ijt}$. Next, we use the maximum log-likelihood function to estimate the intensities $\hat{\lambda_1}$ and $\hat{\lambda_2}$. Then we are able to compute the respective probabilities for each of the three outcomes (win, draw, loss), which are given by:

$$\begin{aligned}
P(Z_{ij} > 0) &= \sum_{z=1}^{\infty} P_{sk}(Z = z|\hat{\lambda_1}, \hat{\lambda_2}) \quad (home) \\
P(Z_{ij} = 0) &= \sum_{z=1}^{\infty} P_{sk}(0|\hat{\lambda_1}, \hat{\lambda_2}) \qquad\quad (draw) \\
P(Z_{ij} < 0) &= \sum_{z=1}^{\infty} P_{sk}(Z = z|\hat{\lambda_1}, \hat{\lambda_2}) \quad (away)
\end{aligned} \tag{4}$$

### 2.2. Ordered Logistic regressions (Toto Models)

We review ordered logistic regression models(*toto models*) which directly predict match outcomes, unlike goal models. *Toto models* are also known as discrete choice regression models. In addition to being a more simple approach compared to goal models, they possess the advantage of avoiding the interdependence between the goal scores of the home and away teams by modeling match results directly. These models predict whether a football match is won by home or away team or whether the final result is a draw. For the estimation of *toto models*, the logistic regression method is often used.

Forrest and Simmons (2002) analyze the predictive quality of match results forecasts made by newspaper tipsters, as well as the performance of the pools panel by providing hypothetical results for postponed matches. Kuypers (2000) utilizes a variety of explanatory variables drawn from current-season match results in order to estimate an *ex ante* forecasting model. Audas et al. (2002) use a similar regression model to examine the impact of managerial change on English football match results data since the 1970s.

Koning (2000) proposed the use of an ordered probit regression model with static team strengths in order to analyze match results *ex post*, as part of a more extensive analysis on the competitive balance in Dutch football. The ordered logistic regression method used by Koning is described as follows:

Electronic copy available at: https://ssrn.com/abstract=3655700

There are three possible outcomes for each football match: win, loss or draw. The ordered logistic method uses two threshold values to separate these three outcomes, thus making the outcome a binary category at the determined thresholds.

As we know, both Poisson distributions and logistic regressions are Generalized Linear Models (GLM) whose main objective is the estimation of the parameter vector $\beta$. In the Generalized Linear Model (GLM) we have a link function $g(\mu_i)$ such that: $g(\mu_i) = \mu_i = x_i^T\beta$ with $i = 1, ..., n$.
Consequently, we have that $\mu_i = E(Y_i)$. When considering a Poisson distribution, the link function is equal to: $g(\mu_i) = log(\mu_i)$ and in the case of logistic regressions, we have: $g(\mu_i) = log(\frac{\mu_i}{1-\mu_i})$.
The estimator of $\beta$ is found through the maximum log-likelihood function:

$$l(\theta) = \sum_{i=1}^{n}(\frac{y\theta_i - b(\theta_i)}{\phi/A_i} + c(y, \phi/A_i)) \tag{5}$$

For $J$ outcomes, in the case of an ordered logit regression, we add the thresholds $\alpha$ thus modifying the previous link function as follows:

$$g(\mu_i) = Logit[P(Y_i < j)] = \alpha_j - x_i^T\beta \tag{6}$$

with $i = 1, ..., n$ and $j = 1, ..., J-1$
Since the cumulative probabilities are increasing and we have $P(Y_i) = 1$ we only have to model $J-1$ probabilities. If we order the 3 outcome categories as follows: away team win (1), draw (2) and home team win (3), we can compute the estimated log odds of categories 1 and 2 as follows:

$$\begin{aligned} Logit[P(Y_i < 1)] &= \alpha_1 - x_i^T\beta \quad (away) \\ Logit[P(Y_i < 2)] &= \alpha_2 - x_i^T\beta \quad (draw) \\ Logit[P(Y_i < 3)] &= \alpha_3 - x_i^T\beta \quad (home) \end{aligned} \tag{7}$$

We can then derive the probabilities for each category which corresponds to the inverse logit functions:

$$P(Y_i < j) = \frac{exp(\alpha_j - x_i^T\beta)}{1 + exp(\alpha_j - x_i^T\beta)} \tag{8}$$

Based on this model, Graham and Stott (2008) then compared the estimated probabilities of their own dynamic probit regression models on English leagues from August

Electronic copy available at: https://ssrn.com/abstract=3655700

2004 to November 2006 with the probabilities obtained from bookmaker William Hill. The predicted probabilities of William Hill outperformed the researchers' model, albeit not with a very significant difference. Similarly, Goddard and Asimakopoulos (2004) also use a static ordered probit regression model to forecast English league match football results where the estimated probabilities of this model could compete with those predicted by bookmakers. This paper was the first to find the predictive quality of past match results data as well as other explanatory factors among which the significance of the match for championship, the geographical distance between the home and away teams' hometowns etc. This paper concludes that ordered probit models are more flexible to predict match outcomes than team scores forecasting models. Their model has been used to test the weak-form efficiency of prices quoted by bookmakers for fixed-odds betting during the 1999-2000 seasons. These tests show that probit regression models contain information on match outcomes that has not been captured by the odds fixed by bookmakers.

On the other hand, Forrest et al. (2010) used static ordered logit regression models. Both of these studies introduced a selection of covariates through these static order regression models. More recently, Cattelan et al. (2013) introduced a semi-dynamic Bradley-Terry model in which team strengths are modeled as exponentially weighted moving average processes.

In an earlier study, Fahrmeir and Tutz (1994) proposed an ordered logit non-Gaussian state space model that defines team strengths as a random walk process. The model parameters are estimated by using a Kalman filter as well as recursive estimation methods. Similarly, the dynamic cumulative link model of Knorr-Held (2000) has been applied to German Bundesliga data, where the estimations of the model are performed through an extended Kalman filter and smoother. Finally, Hvattum and Arntzen (2010) propose an ordered logit model in which team strengths are updated over time by using an "Elo rating system".
Comparing the two model categories, Snyder (2013) shows in a study that the betting profit of several strategies is considerably higher for *toto models* than for goal models. Thus, it is more efficient to concentrate on models that directly predict match outcomes.

Electronic copy available at: https://ssrn.com/abstract=3655700

In the next subsection, we will take a look at some of the more complex models on match betting which use a Machine Learning framework. Incidentally, the latter is the methodology we have chosen to implement in our paper as well.

### *2.3. Predictive Betting Models with Machine Learning*

Machine learning (ML) is the scientific study of algorithms and statistical models used by computer systems and it is considered as a subcategory of artificial intelligence. In a machine learning framework, several specific tasks are efficiently performed without using explicit instructions, among which classification and prediction. Due to the great strides made on these two areas, ML is often used to build models on sports results prediction. Even club managers and owners themselves are interested in developing classification models in order to create effective game strategies. These models rely on numerous game factors, such as the results of historical matches, player performance indicators, and opposition information.

Per Witten and Frank (2002), the classification task is related to the prediction of a target variable or class by using previously undetected data (data processing), a task which is completed by creating a classification model that is used to predict the value of the class of test data. Sports prediction is considered as a classification problem with one class to predict (the 3 possible outcomes: win, lose, or draw). Once a prediction result for the match is obtained, you still have to decide whether to bet on a match, by taking into consideration the bookmaker's odds as well. A recent study by Bunker and Thabtah (2019) analyzes existing ML literature, particularly focusing on the application of Artificial Neural Networks (ANNs) for sports outcome predictions.

ANNs are the ML approach that is most commonly used for sports betting prediction models in an ML framework. An ANN contains interconnected components (neurons) which transform a set of inputs into a desired output. In the ANN algorithm, weights are non-linearly adjusted by hidden neurons. These continuous weight changes contribute to the final prediction of the model and its accuracy. The ANN model is built following the training dataset processing. Among the advantages of the ANN method are its flexibility related to the definition of class variables (whether it is the probability of victory or they represent home and away goals in the case of two

Electronic copy available at: https://ssrn.com/abstract=3655700

classes).

Most of the related work on this topic has been done by gambling organizations for the profit of bookmakers. However, due to the public nature of sports results data, academic researchers have also built prediction models using Machine Learning methods.

Tax and Joustra (2015) used Dutch football data from the past 13 years to predict the results of football matches. They compared a model with betting odds alone to a hybrid model with betting odds and other match features (public data model). They emphasized that cross-validation is not appropriate for sports prediction because of the time-ordered nature of the data. They used several methods such as a PCA analysis, sequential forward selection and correlation-based feature subset selection. Among the nine classification algorithms used for their study, the highest performing classifiers on the full feature set were naÃ¯ve Bayes (used with a 3-component PCA) and the ANN (used with a 3 or 7-component PCA), both achieving a classification accuracy of 54.7%. The authors concluded that there was not a statistically significant difference between the betting odds model and the public data model by using the McNemar test. However, they recommended the use of the betting odds model as a predictor of match outcomes.

Joseph et al. (2006) used Bayesian Nets to predict the results of Tottenham Hotspur over the period of 1995-1997. However, their model displayed many flaws and inconsistencies: it relies upon trends from a specific time period and is not extendable to later seasons, and they report wide variations in accuracy, ranging between 38% and 59%. Nevertheless, they were able to provide useful elements on model comparison and feature selection.

Timmaraju et al. (2013) focused on building a highly accurate system to be trained and tested with one season of data. Since their parameters were affected by such small datasets, their accuracies rose to 60% with a Radial Basis Function - Support Vector Machine model (RBF-SVM).

Ulmer et al. (2013) predicted the results of football matches in the English Premier League (EPL) using machine learning algorithms. They relied heavily upon the previ-

Electronic copy available at: https://ssrn.com/abstract=3655700

ous work of Rue and Salvesen (2000) who proposed a Bayesian linear time-dependent model to predict football results.

From historical data, they created a feature set which included gameday data and current team performance (form). Using these feature data they created five different classifiers: Linear from stochastic gradient descent, Naive Bayes, Hidden Markov Model, Support Vector Machine (SVM), and Random Forest. Their prediction was in one of three classes for each game: win, draw, or loss. Their error analysis focused on improving hyperparameters and class imbalances, which they approached by using grid searches and ROC curve analysis respectively. Some of the challenges they encountered were the limited amount of data as well as the randomness of the data. The latter was confirmed by an entropy value (a randomness measure) of 0.7 where a value of 1 signifies pure randomness of data. Due to the large number of upsets found in football match outcomes, i.e. when minor league teams win over major league ones, the number of outlier events is high, thus making football match results very difficult to predict.

Each of the classification models they used tended to under-predict draws. When they attempted to weight this class more, they found that improved accuracy of the draw results deteriorated that of the other two classes, wins and losses. The authors suggest that the failure to predict draws stems from the fact that draws are the least likely result to occur. The authors conclude that while their model was relatively successful, it still lags behind leading industry methods. They add that the models would perform even worse with a more limited dataset. Finally, they state that their model was only able to accurately predict that Manchester United would win the 2012-2013 EPL season.

Bunker and Thabtah (2017) propose a new ML framework called the Sport Result Prediction or SRP CRISP-DM framework which focuses on result prediction for all team sports. The proposed methodology consists of 6 main steps: domain understanding, data understanding, data preparation and feature extraction, modelling, model evaluation and finally model deployment.

In the following section we will implement the Dixon and COles model and this new methodology based on neural networks for sports outcome predictions, as it improved

Electronic copy available at: https://ssrn.com/abstract=3655700

the accuracy of ML models on this domain.

## 3. Dataset

The raw dataset used to test the models explored in this paper was retrieved from *www.indatabet.com website* encompassing results until the $1^{st}$ of April, 2019. It is composed of the first division of 5 countries (France, Spain, Italy, Germany, England) for a total of 163 different teams with seasons ranging from 2010/2011 to 2017/2018. We are given historical data for the results of the match (home goals and away goals), the date of the match, the name of the teams and the odds, from 2 bookmakers: Pinnacle and Bet365.
It is noteworthy to mention that the odds are pre-match opening.

The test sample for implementing the models is represented by results from the season of 2017/2018. The validation set of the Dixon's model to find $\xi$ consists of the matches of the 4 previous months before the test set. The validation set of the Neural Network's model is the season of 2016/2017.

To use these data in testing our models, it was necessary first to carry out some preliminary treatments:

1. Translate dates in digital format
2. Remove duplicates
3. Correct the names of the teams so that there are no duplicates (ex: "Paris SG" and "Paris SG (FRA)")
4. Assign an ID to each team
5. Delete rows with missing information
6. Removing all arbitrage opportunities assuming they were data errors

Translating dates in digital format is not really complicated, Excel already has a system of conversion, 1 being the 01/01/1900). To remove duplicates, we only had to use the Excel function. Correcting the names of the teams was a bite more tricky. It was necessary to use the data conversion function, and use " (" as a separator, so we get a first column with the name of the team without the country code. To assign an ID per team, we simply rank the teams in alphabetical order, and assign 1 to the first. In total, we are left with over 2000 teams. To simplify the dataset, we removed the following fields: FIFA ID of the league, League, Season, 1st half score only, 2nd half

Electronic copy available at: https://ssrn.com/abstract=3655700

score only, Odds of the total number of goals (greater or less than 2.5 goals). Table 1 presents an example of records from the cleaned dataset.

| Game | Team ID | | Final score | | Pinnacle odds | | | Bet365 odds | | | Goal spread |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Date | Home | Away | Home Team | Away Team | H | D | A | H | D | A | (H-A) |
| 2019-04-01 | 1102 | 1425 | 2 | 1 | 1.74 | 4.03 | 4.58 | 1.90 | 3.60 | 3.75 | 1 |
| 2019-03-31 | 5 | 918 | 1 | 1 | 2.31 | 3.34 | 3.09 | 1.90 | 3.60 | 3.75 | 0 |
| 2019-03-31 | 18 | 1024 | 0 | 3 | 3.64 | 3.45 | 2.11 | 3.40 | 3.75 | 2.00 | -3 |
| 2019-03-31 | 30 | 1363 | 4 | 0 | 1.37 | 4.23 | 8.91 | 1.44 | 3.80 | 9.00 | 4 |
| 2019-03-31 | 46 | 1437 | 3 | 1 | 1.76 | 4.18 | 4.29 | 1.70 | 4.20 | 4.20 | 2 |

Table 1: Example of records from the cleaned dataset

In Figures below, several graphs are presented and allow to better understand the data. Figure 1, we see that overall, home team is more likely to win scoring an average of 1.46 goals against 1.14 for the away team. The comparison between Figures 2 and 3 allows us to quickly observe that the odds proposed by Pinnacle are better than those of Bet365.

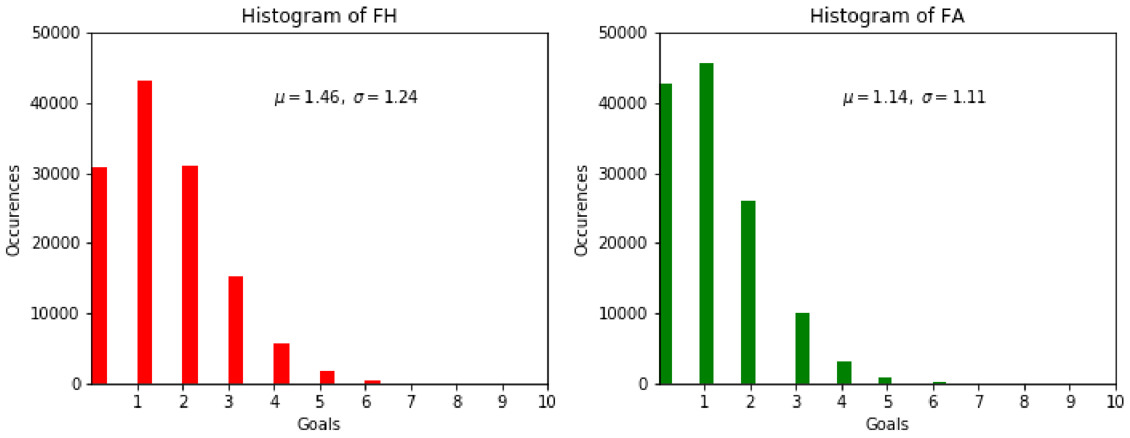

Figure 1: Histogram of Goals

Finally, Figure 4 tells us that the higher the odds (rare event), the more the expectations of bookmakers are heterogeneous and vice versa.

## 4. Model implementation

In this section we present the detailed methodology and Python implementation of the Dixon and Coles model and the adaptation of the neural-networks based model.

Electronic copy available at: https://ssrn.com/abstract=3655700

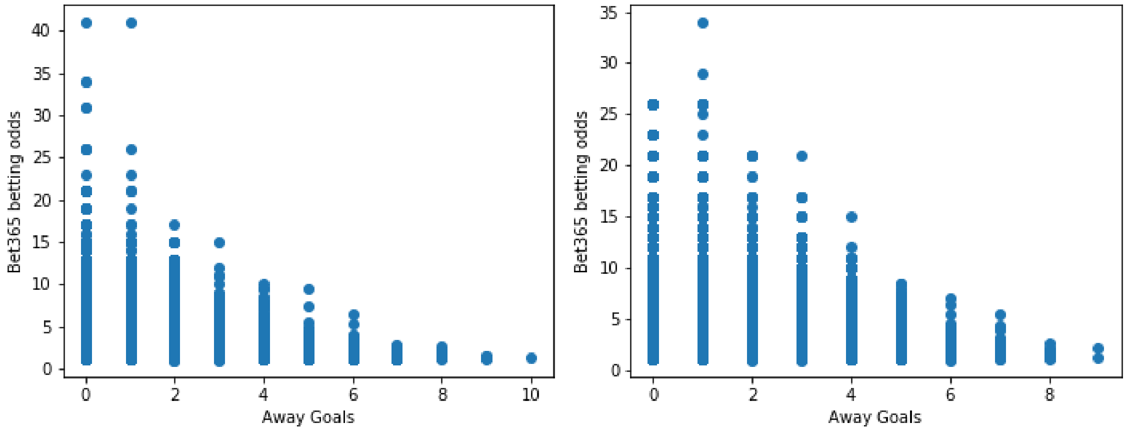

Figure 2: Scatter plot of goals vs Bet365 odds

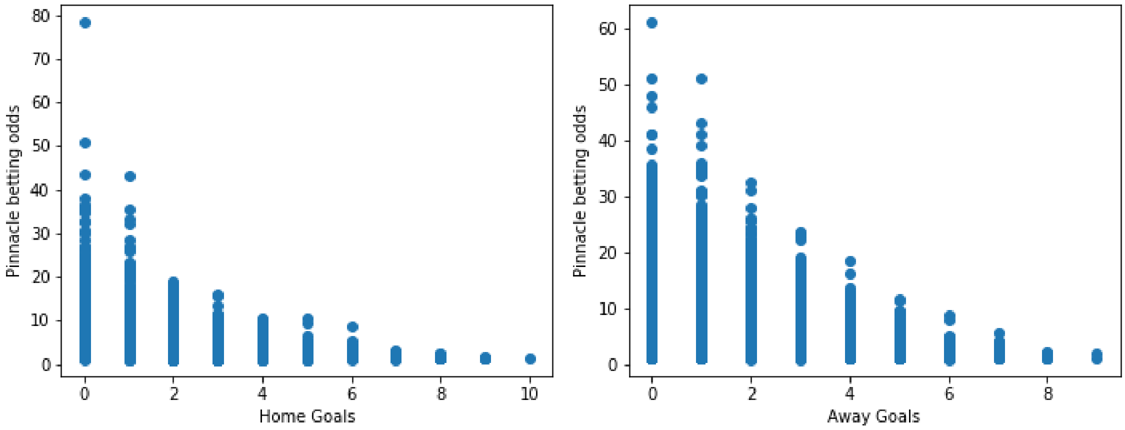

Figure 3: Scatter plot of goals vs Pinnacle odds

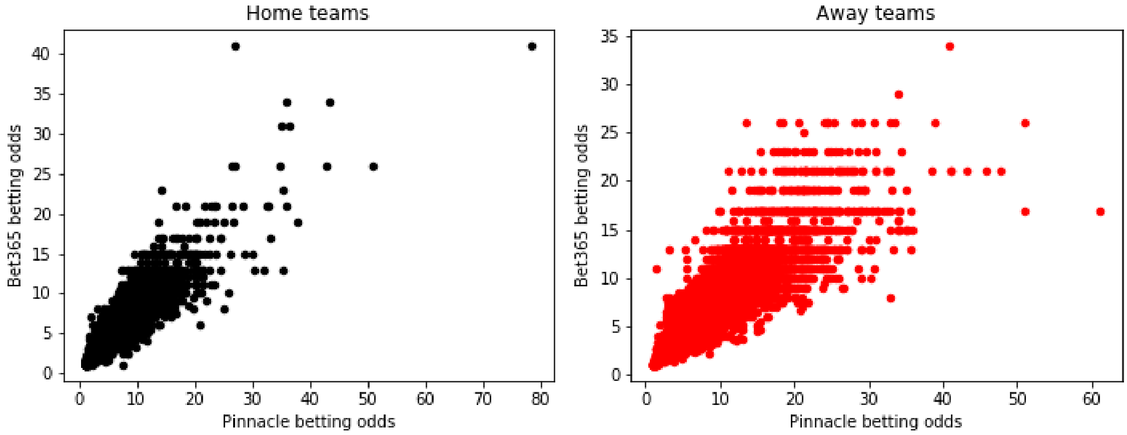

Figure 4: Scatter plot of Pinnacle vs Bet365 odds

Electronic copy available at: https://ssrn.com/abstract=3655700

### *4.1. Dixon and Coles*

Let denote by $X_{i,j}$ the number of goals scored by home team i against away team j, and $Y_{i,j}$ the number of goals scored by away team $j$ against home team $i$. In Maher's framework, $X \sim Poisson(\alpha_i\beta_j\gamma)$ and $Y \sim Poisson(\alpha_j\beta_i)$, where $\alpha_k$, $\beta_k$ and $\gamma$ are respectively the attack parameter of team $k$, the defense parameter of team k, and the home advantage parameter. The issue of a match is then defined by:

$$\mathbb{P}(X_{i,j} = x, y_{i,j} = y) = \mathbb{P}(X_{i,j} = x)\mathbb{P}(y_{i,j} = y) = \frac{e^{-\alpha_i\beta_j\gamma}(\alpha_i\beta_j\gamma)^x}{x!}\frac{e^{-\alpha_j\beta_i}(\alpha_j\beta_i)^y}{y!}$$

Dixon and Coles(Dixon and Coles (1997)) proposed to enhance the model by introducing dependence between $X$ and $Y$, and by weighting the matches according to a time decay function.

Let define

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

the correlation function of a match with x home goals and y away goals, where $\lambda$ and $\mu$ are respectively home and away parameter of the home and away Poisson distribution; and

$$\Phi(t) = e^{-\xi t}$$

the weighting time decay function that output a weight for a match that has been disputed t days ago. Current match will have a one unit weight whereas older matches will have a weight close to zero.

The issue of a match is then defined by :

$$\mathbb{P}(X_{i,j} = x, y_{i,j} = y) = (\tau_{\alpha_i\beta_j\gamma,\alpha_j\beta_i}(x,y)\frac{e^{-\alpha_i\beta_j\gamma}(\alpha_i\beta_j\gamma)^x}{x!}\frac{e^{-\alpha_j\beta_i}(\alpha_j\beta_i)^y}{y!})^{\Phi(t)}$$

and the log-likelihood $L(\alpha,\beta,\gamma,\rho,\xi)$ is defined by:

$$L(\alpha,\beta,\gamma,\rho,\xi) = \sum_{k=1}^{n} e^{-\xi(t-t(k))}[\log\tau_{\alpha_i\beta_j\gamma,\alpha_j\beta_i}(x_k,y_k) + x_k\log\left(\alpha_{i(k)}\beta_{j(k)}\gamma\right) + y_k\log\left(\alpha_{j(k)}\beta_{i(k)}\right)$$
$$- \alpha_{i(k)}\beta_{j(k)}\gamma - \alpha_{j(k)}\beta_{i(k)} - \log(x_k!) - \log(y_k!)]$$

Electronic copy available at: https://ssrn.com/abstract=3655700

where n is the number of matches on the dataset, $x_k$ and $y_k$ the number of home and away scores during the kth match, $i(k)$ and $j(k)$ the id of the home and away team during the kth match, and t-t(k) the number of days between current date t and the date t(k) of the kth match.

We first estimate the exogenous parameter $\xi$ by cross validation. Let define by G, a grid of 11 linearly spaced values between 0 and 0.005, the possible values taken by $\xi$. The optimal $\xi^*$ is the $\xi \in G$ maximizing :

$$S(\xi) = \sum_{k=1}^{M} S_k(\xi)$$

$$S_k(\xi) = \sum_{l=1}^{M_k} \log(p_{l,k}(\alpha_k^*, \beta_k^*, \gamma_k^*, \rho_k^*, \xi))$$

where M is the number of days on the validation set, $M_k$ is the number of match on the kth day of the validation set, $\alpha_k^*, \beta_k^*, \gamma_k^*, \rho_k^*$ are the optimal maximum likelihood estimators find on the whole dataset before the kth day, and $p_{l,k}$ is the probability of the true result (home win, draw or away win) returned by the model on the lth match of the kth day (i.e if the home team i win against away team j, then $p_{l,k}^* = \sum_{x>y} \mathbb{P}(X_{i,j} = x, y_{i,j} = y)$).

Finally, for each day t on the test set and using the optimal $\xi^*$ find previously, the model is retrained to find the optimal maximum likelihood estimators $\alpha_t^*, \beta_t^*, \gamma_t^*, \rho_t^*$ of $L(\alpha, \beta, \gamma, \rho, \xi^*)$ which will be used to predict the result of the matches on day t.

The Limited-memory BFGS optimization algorithm has been used to infer the optimal parameters. We also computed manually the derivatives of the log likelihood function to improve inference speed. On the validation set (last 4 months of season 2016-2017), we obtained $\xi^* = 0.004$. Each day of the test set (season 2017-2018), we obtained the remaining parameters : attack, defense, home advantage and correlation parameter.

Electronic copy available at: https://ssrn.com/abstract=3655700

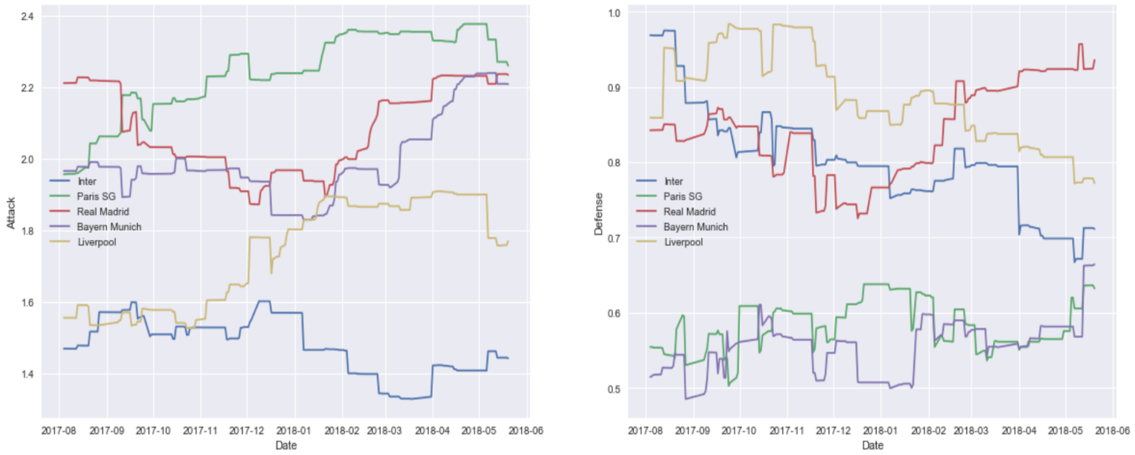

Figure 5: Example of attack (left) and defense (right) parameter

### *4.2. Neural Network*

In this section, we will try to implement Deep Learning methodologies using Neural Networks and test its efficiency when predicting football games' outcome. Clearly, neural networks and the mathematical theory behind it are beyond the scope of this course and it is therefore not relevant to go too deep into the details. However, we deem necessary to explain in a few sentences the way they function in order to at least have a graphical representation of what it actually is, and also to be able to distinguish some different architectures relevant to this project as they will be used. Neural networks are composed of a series of layers. Each layer takes in the outputs of the previous layer to calculate its value. The general equation is:

$$X_i = f_i(W_i X_{i-1} + h_i) \tag{9}$$

where $X_i$ and $X_{i_1}$ are the values of the current layer and the previous layer, $W_i$ represent the weights, $h_i$ the bias, and $f$ a non-linear function called activation function. In fact, research has shown that the use of non-linear functions between layers allows any neural network to move closer to any continuous function. The sigmoÃ¯d function is one of the most known activation functions $\sigma(x) = \frac{1}{1+\exp(-x)}$ with the rectified identity function **ReLU** $f(x) = max(0; x)$. $W_i$ and $h_i$ are the parameters of the model, they are initialized randomly before being learned during the training of the neural network by minimizing a pre-defined cost function.

Electronic copy available at: https://ssrn.com/abstract=3655700

### *4.2.1. Neural network training*

Training a neural network is a key optimization problem in which we use stochastic descent techniques in order to find the best parameters. During the training, the data is transmitted to the neural network in “batches”, and a passage both forward and backward through the neural network of the learning data set is called an “epoch”. The training is done using the algorithm of back-propagation: the data are first transmitted forward. We start with the layer input and each layer calculates its values using the previous layer until the output. Once we have reached the exit, we compare the calculated output with the desired one and we obtain a measured error using a loss function. Then we use this error and the chain rule to apply a gradient descent algorithm to update weights and minimize this mistake in order to obtain an output closer to reality.

### *4.2.2. Classical architecture*

*Dense layer*

The most common layer is the dense layer (also called fully connected). In this type of layer, each neuron is bound to all the neurons of the previous layer. This leads to a very large number of parameters: the number of parameters for this layer is equal to one plus number of neurons, multiplied by the number of neurons in the previous layer.

*Recurrent layer*

When dealing with data presented in a sequential way, for example images in a video or words in a sentence, recurrent neural networks are a common technique to use. The idea is that the recurrent layer receives the elements of the sequence one by one while keeping an internal state of the past elements. So, the prediction does not rest solely on the last element, but on all the previous elements and their order. For that reason, we thought it could be a good idea to try this type of architecture. However, we will later on explain why we encountered difficulties implementing this model and decided to sit this one out. Again, the number of parameters is very low compared to a conventional dense layer.

A problem encountered by recurrent layers is the problem of gradient propagation.

Electronic copy available at: https://ssrn.com/abstract=3655700

The calculation of the degradation through several layers or stages of a sequence is done by multiplying the gradient. Multiplying by values greater than 1 or between 1 and 0 results in more than one problems, respectively: the explosive gradient and the vanishing gradient. These problems prevent the neural network from learning the proper weights. One solution is to use the LSTM layers (Long short-term memory). They constitute an improved version of the recurrent layer. They have better control of the gradient using a gateway and a door of oblivion. An LSTM cell is defined in 3 steps:

- Input gate : as for a RNN cell, we apply a hyperbolic tangent transformation on the linear combination of the current score $x_t$ and the previous output $h_t$. We will however additionally multiply this transformation by a sigmoÃ¯d transformation of these inputs to bring more or less importance to the components of the first information.

- Forget gate : this step is the most important since it is the one that determines the information of the cell and thus the current state and who is in charge of transmitting it. She denies herself to help information from the previous cell, again updated by a sigmoÃ¯d transformation, as well as the input gate by adding the old interesting information (forget gate) with the new information (input gate).

- Output gate : once the information is captured in the previous step, a hyperbolic tangent transformation is applied to it, that we will once again modify in order to keep the relevant information to characterize the next state ($h_t$) and to predict the next score ($y_t$).

They are the ones we tried to use at first but we soon decided to give up on that idea given our data structure which did not allow the score prediction by football team. In fact, the occurrences of games is rather not stable over the period we have for each team and too many parameters are to be taken into account so that the regression actually makes sense. As a recurrent neural networks (RNN) feeds from all its past elements to predict the next one, it was not ideal in our case. Having better data with individual structure and an adjusted frequency of games could have strongly oriented us towards "successfully" implementing a RNN.

Electronic copy available at: https://ssrn.com/abstract=3655700

We trained Neural to predict directly the results of the matches (win home, draw, loss home). The architecture consists of 3 hidden layers with respectively 60, 20 and 10 neurons. The loss metric is the categorical crossentropy with a softmax activation function on the output layer. For each observation (=match), the inputs are the number of days since the match has been disputed, the home/draw/away odds of the bookmakers, the number of goals for the home and away team, and the ids of the home and away team of the match.

The model is defined by a set P of different parameters with their corresponding window of possible values:

- n: for each matchs team A vs B, the n previous matches of team A and team B are considered as inputs. Only the number of goals and the ids of the teams of the 2n previous matches are considered as inputs (n=0,1,2,3).
- Polynomial feature: whether or not second order polynomial feature with interaction is applied (only done if the accuracy is improved by more than 1% compared to a model without polynomial features).
- Exponential date: whether or not exponential time decay function ($f(x) = e^{-0.01x}$) is applied on the input date.
- Frequency: the frequency to re-train the Neural Network (every 1year, 3months, 1month, 1week).
- Dropout: whether or not a 20% dropout of is applied on each hidden layer.
- Neural Network parameters: activation function (relu, sigmoid, tanh, selu, exponential, softplus), optimizer (Adamax, SGD, Adagrad, Adadelta), learning rate (0.001, 0.01, 0.1, 0.2, 0.3), batch size (32, 64, 128, 256, 512).

We started with a set of default parameters and then for each parameters $p \in P$ we successively updated the parameters one by one in the order listed above: the previous optimal parameter found are used to find to optimal parameter of p. We applied the following process to find the optimal parameter among the possible values of p.

Electronic copy available at: https://ssrn.com/abstract=3655700

Let define by $t_1, ..., t_M$ the dates to re-train the Neural Network (with $t_1$ first date of the season). For each re-training date $t_i$, we defined a set $A_{t_i}$ consisting of the matches in $[t_i, t_{i+1}]$ (with $t_{M+1}$ the last date of the season), a set $B_{t_i}$ consisting of the 400 previous matches (=20% of the matches on a season) before $A_{t_i}$ and a set $C_{t_i}$ consisting of the whole dataset before $B_{t_i}$. The model is train on $C_{t_i}$ with 500 epochs and we choose the model on the epoch with the highest accuracy on $B_{t_i}$. Then we predict the outcomes on $A_{t_i}$ with the previous chosen model. Since $A_{t_i}$, i=1...M form a partition of the validation set, we have a prediction on each day of the set and we can compute the average accuracy on the whole validation set. We finally choose the parameter with the highest accuracy.

When the optimal parameters P has been found, we applied again the previous process on the test set (season to predict), using the optimal parameters, and we obtain the prediction for the season.

Starting with a default model, we found successively the optimal parameters on the validation set (season 2016-2017). More precisely, the process has been applied 30 times for each parameters to decrease the variability on the optimal parameters found. Once the optimal parameters found, we also applied 30 times the process on the test set to decrease the variability of the predictions (season 2017-2018). Table 2 shows the features of the optimal neural network model

| Model | n | Poly | Exp | Frequency | Dropout | Act. fct | Opt. | LR | Batch |
|---|---|---|---|---|---|---|---|---|---|
| Default | 0 | False | True | 1Y | False | Re, Sg, Sg | Adamax | 0.002 | 128 |
| Optimal | 3 | False | True | 3M | False | Re, Re, Re | Adagrad | 0.01 | 64 |

Table 2: Features of the optimal neural network model

### *4.3. Implementation accuracy*

We displayed in Figure 6 and Figure 6 the probabilities produced by the Neural Network model and the Dixon-Coles model against the probabilities of the bookmakers for each possible outcome (win, draw, loss). The probabilities of the bookmakers are the scaled one (to remove their overround profit). We took the average probabilities of the bookmakers Pinnacle and Bet365 for each matches.

Electronic copy available at: https://ssrn.com/abstract=3655700

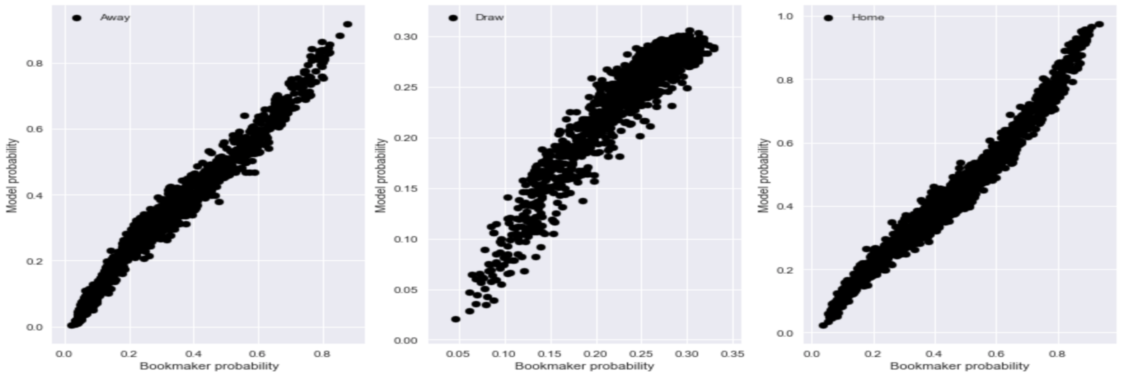

Figure 6: Neural network vs bookmaker probabilities of Away (left), Draw (center) and Home (right)

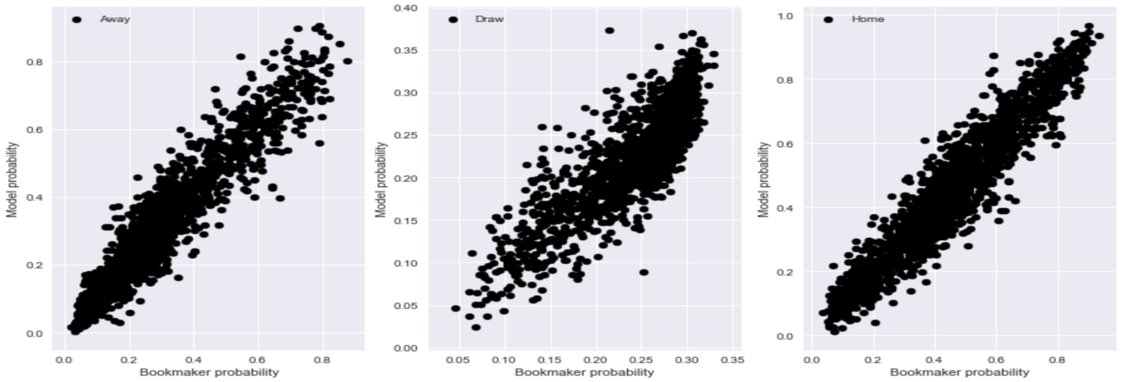

Figure 7: Dixon and Coles vs bookmaker probabilities of Away (left), Draw (center) and Home (right)

We computed the confusion matrix for the Neural Network model(Figure 8) and the Dixon-Coles model(Figure 9). The Neural Network accuracy and F1 macro score is 0.5454 and 0.4655. The Dixon and Coles accuracy and F1 macro score is 0.5387 and 0.4627.

## 5. Sport Betting as Investment strategy

In this section we expand the two model implemnted above to explore whter they ca produce sustainable investment strategies. Thus, we implemented 3 different portfolio strategies using the predictions on the test set (season 2017-2018). The portfolios start with an initial capital of 1$ and for each match we invest 1/1000 of the current capital.

- Strategy 1: For each match, we invest on the highest of the 3 probabilities outcomes of the model (win, draw, loss).

Electronic copy available at: https://ssrn.com/abstract=3655700

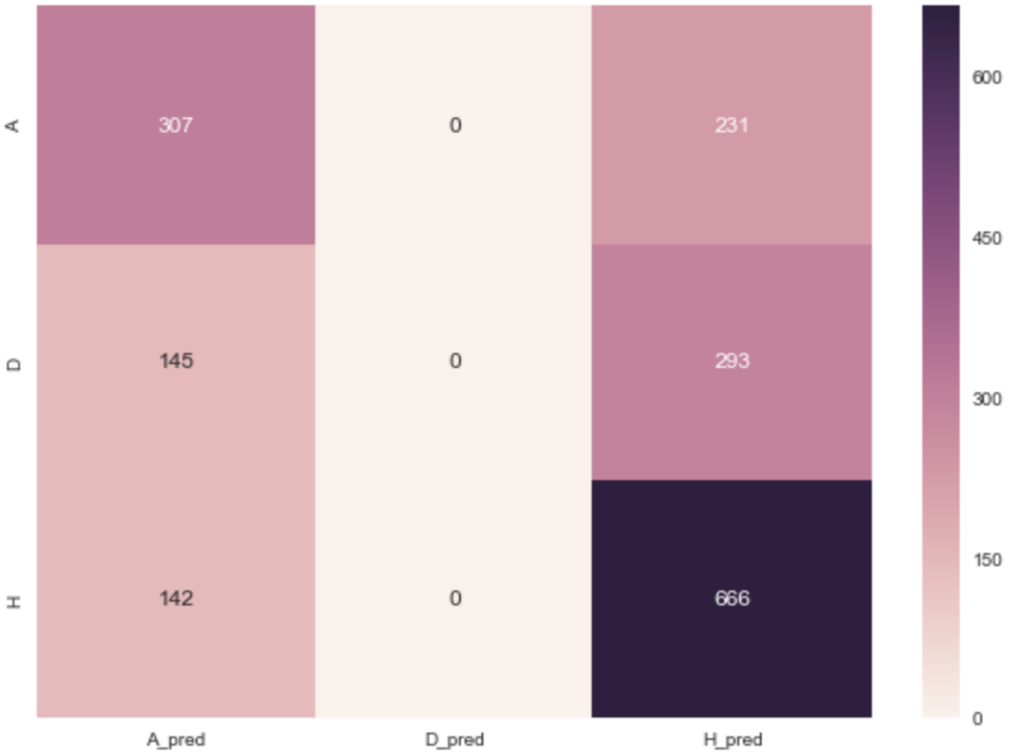

Figure 8: Confusion matrix for Neural Network model

- Strategy 2: For each match, we invest on the 3 outcomes of the model (win, draw loss) weighted by the probabilities.

- Strategy 3: For each match, we invest on the outcomes for which the ratio probability of the model to (unscaled) the average probabilities of the bookmakers is higher than a threshold t ($t > 1$).

The investment Strategy 1(Figure 10) and Strategy 2(Figure 11) are not profitable for both models. However, the Neural Network model is always better than the Dixon and Coles model in terms of Sharpe ratio..

Electronic copy available at: https://ssrn.com/abstract=3655700

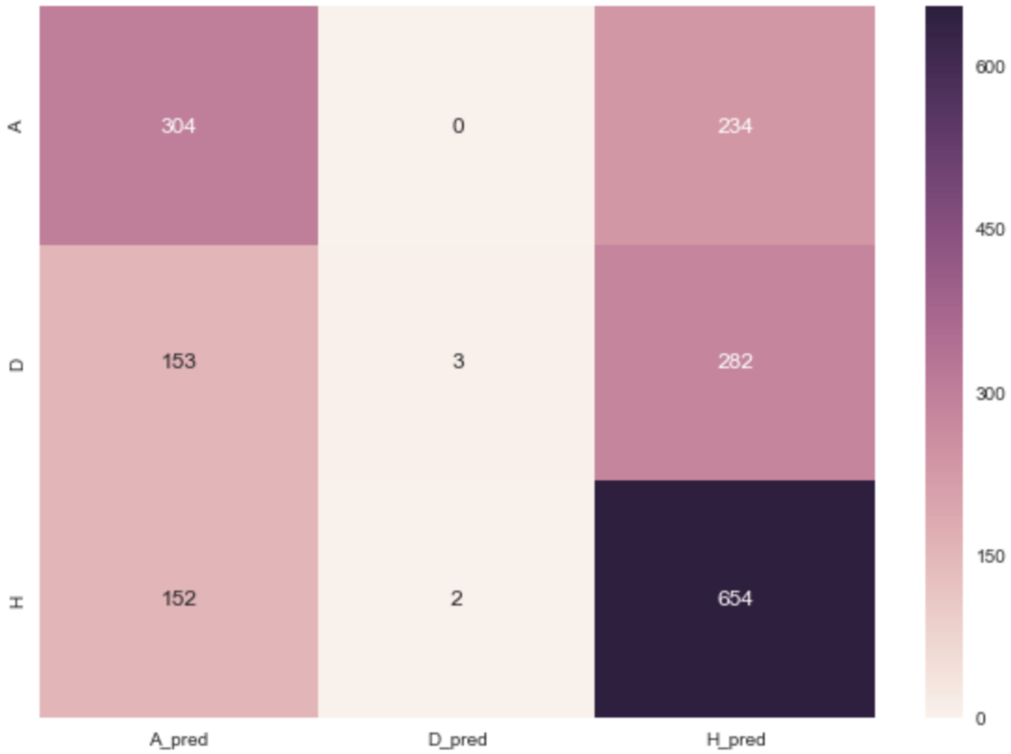

Figure 9: Confusion matrix for Dixon and Coles model

The Strategy 3 is is more profitable in terms of absolute returns for the Neural Network model 12 compared to the Dixon Coles model which deliver a negative performance(12). For both models the number of bets decreases with the threshold level. The Strategy 3 (Figure 14) is profitable for the Neural Network model for a threshold between t=1 and t=1.08 with a maximum yearly return of 8% for t=1.05.

## 6. Conclusion

In this article, we implemented the well-known statistical model of Dixon and Coles (Dixon and Coles (1997)) as a benchmark, as well as a more advanced Neural Network model. The goal of this paper was to compare the performance of the Neural Network model against the benchmark and to perform a betting strategy that is profitable on the long run (period of 1 season).

We found that the accuracy of the Neural Network model is slightly higher than the benchmark model. In addition, the probabilities produced by the Neural Network

Electronic copy available at: https://ssrn.com/abstract=3655700

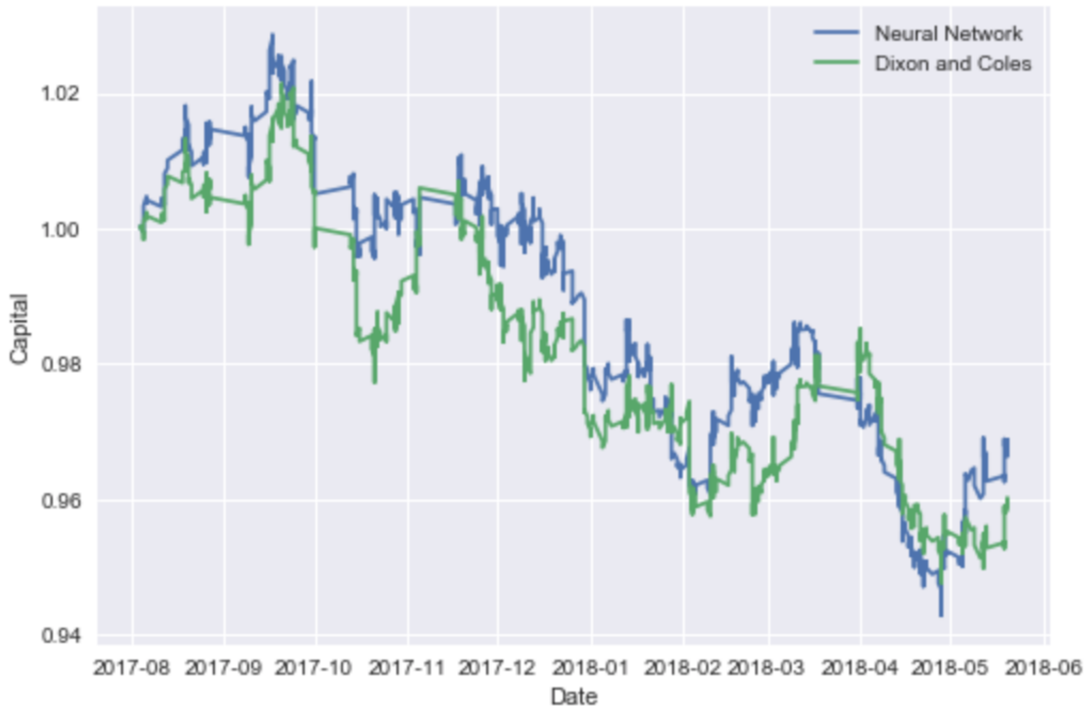

Figure 10: Model performance benchmark when applied to the investment strategy no. 1

model (for the three outcomes: win, draw and loss) are more in line with the bookmakers probabilities compared to the benchmark model. Indeed, unlike the model benchmark, the Neural Network model take as an input the odds of the bookmakers.

The Neural Netowork-based prediction model outperforms the benchmark model on the 3 different strategies implemented, although the two most basic strategies are not profitable. The Strategy 3 consists of investing on an outcome (win, draw, loss) only if the probability given by the model on that outcome is larger than the average probabilities of the bookmakers. We found that the overall portfolio return over a season using Strategy 3 for the Neural Network model can reach a target of 8%. Thus, by learning from the bookmakers odds, the Neural Network is able to outperform the probabilities of the bookmakers and their overround price to make profit on the long run.

Electronic copy available at: https://ssrn.com/abstract=3655700

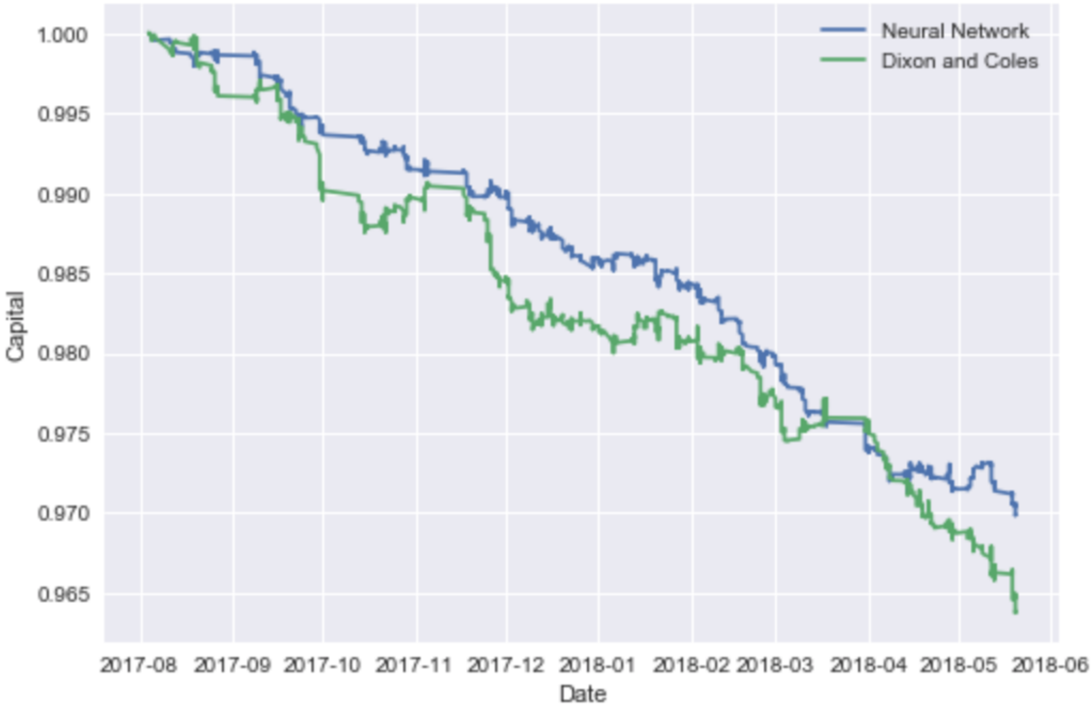

Figure 11: Model performance benchmark when applied to the investment strategy no. 2

## References

Audas, R., Dobson, S., Goddard, J., 2002. The impact of managerial change on team performance in professional sports. Journal of Economics and Business 54, 633–650.

Barnett, V., Hilditch, S., 1993. The effect of an artificial pitch surface on home team performance in football (soccer). Journal of the Royal Statistical Society: Series A (Statistics in Society) 156, 39–50.

Bunker, R.P., Thabtah, F., 2017. Applied computing and informatics .

Bunker, R.P., Thabtah, F., 2019. A machine learning framework for sport result prediction. Applied computing and informatics 15, 27–33.

Cattelan, M., Varin, C., Firth, D., 2013. Dynamic bradley–terry modelling of sports tournaments. Journal of the Royal Statistical Society: Series C (Applied Statistics) 62, 135–150.

Electronic copy available at: https://ssrn.com/abstract=3655700

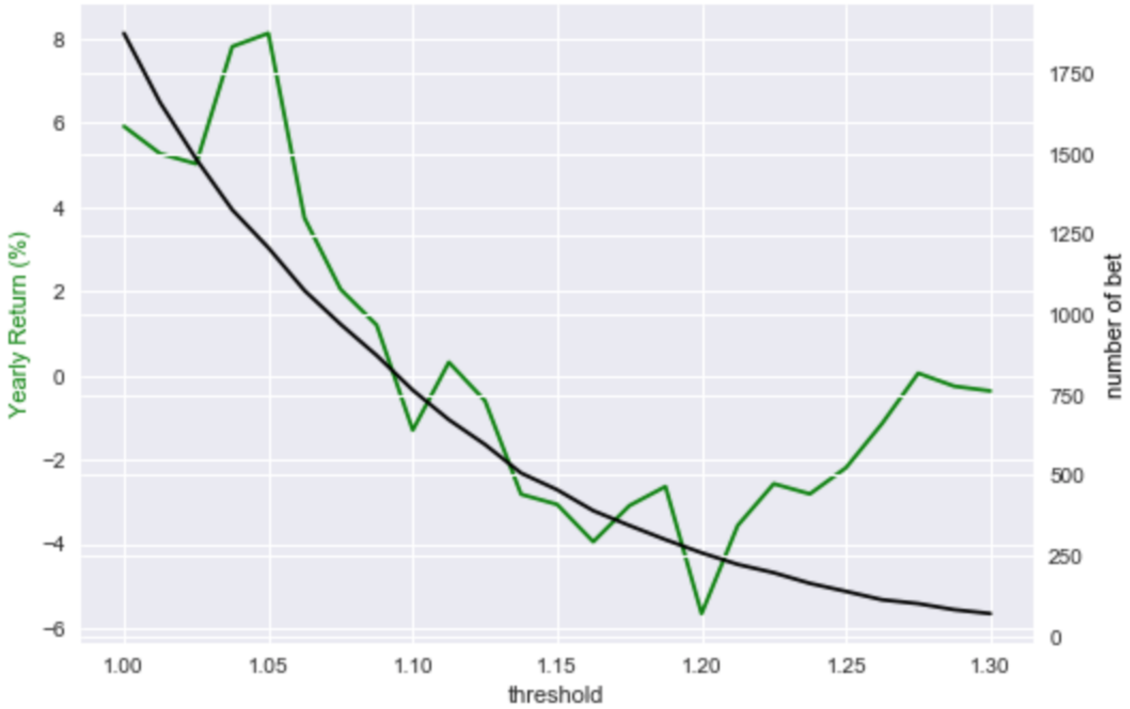

Figure 12: Yearly return (green) of investment Strategy 3 for the Neural Network model for different thresholds

Clarke, S.R., Norman, J.M., 1995. Home ground advantage of individual clubs in english soccer. Journal of the Royal Statistical Society: Series D (The Statistician) 44, 509–521.

Crowder, M., Dixon, M., Ledford, A., Robinson, M., 2002. Dynamic modelling and prediction of english football league matches for betting. Journal of the Royal Statistical Society: Series D (The Statistician) 51, 157–168.

Dixon, M., Robinson, M., 1998. A birth process model for association football matches. Journal of the Royal Statistical Society: Series D (The Statistician) 47, 523–538.

Dixon, M.J., Coles, S.G., 1997. Modelling association football scores and inefficiencies in the football betting market. Journal of the Royal Statistical Society: Series C (Applied Statistics) 46, 265–280.

Fahrmeir, L., Tutz, G., 1994. Dynamic stochastic models for time-dependent ordered

Electronic copy available at: https://ssrn.com/abstract=3655700

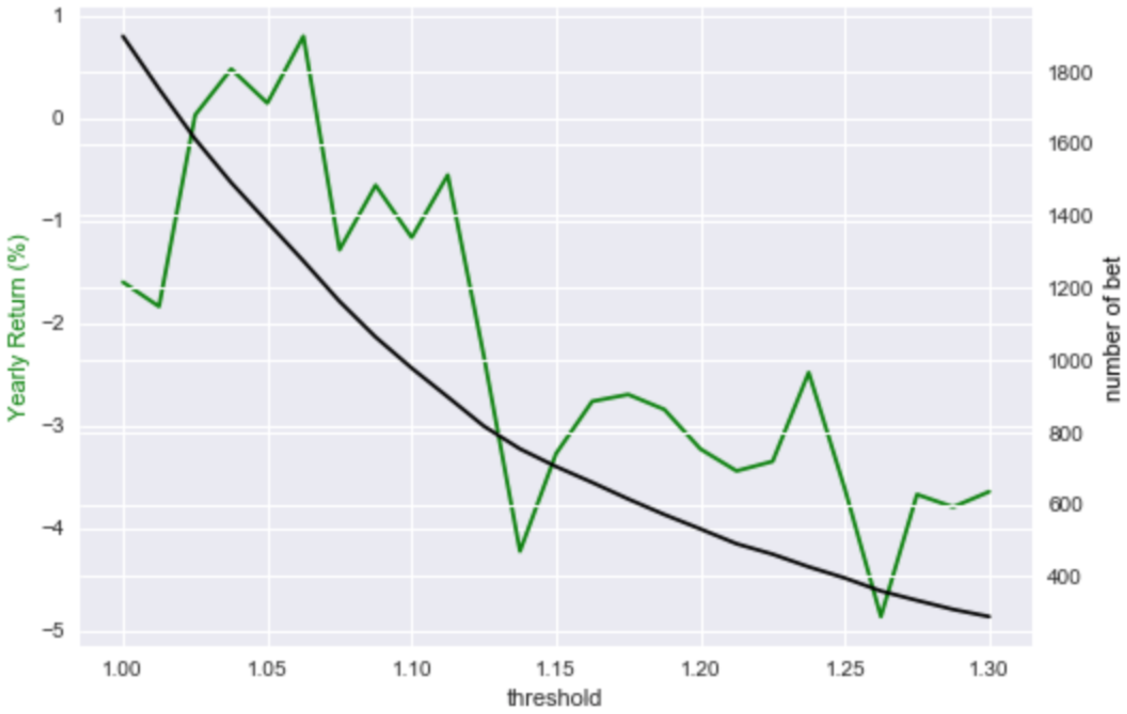

Figure 13: Yearly return (green) of Strategy 3 for Dixon-Coles model for different thresholds

paired comparison systems. Journal of the American Statistical Association 89, 1438–1449.

Forrest, D., Gulley, O.D., Simmons, R., 2010. The relationship between betting and lottery play. Economic Inquiry 48, 26–38.

Forrest, D., Simmons, R., 2002. Outcome uncertainty and attendance demand in sport: the case of english soccer. Journal of the Royal Statistical Society: Series D (The Statistician) 51, 229–241.

Frunza, M., 2014. Assessment of the link between the betting industry and financial crime: Application to association football. Available at SSRN 2528223 .

Frunza, M.C., 2015. Solving modern crime in financial markets: Analytics and case studies. Academic Press.

Goddard, J., Asimakopoulos, I., 2004. Forecasting football results and the efficiency of fixed-odds betting. Journal of Forecasting 23, 51–66.

Electronic copy available at: https://ssrn.com/abstract=3655700

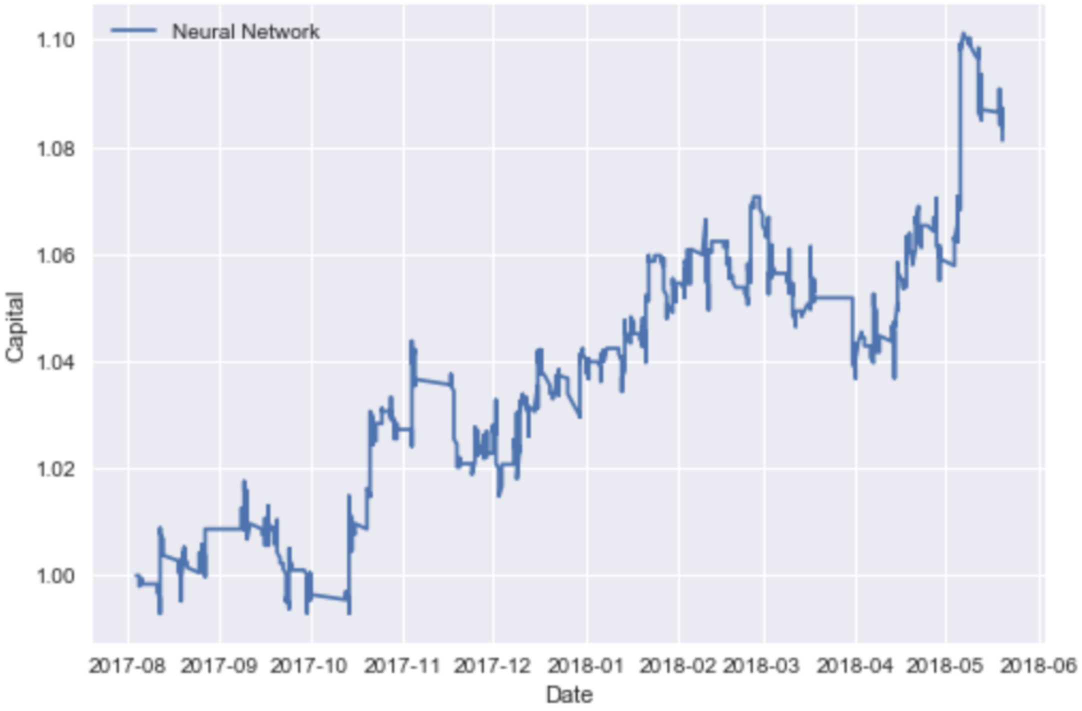

Figure 14: Strategy 3 for the Neural Network model with optimal t=1.05

Gomez-Gonzalez, C., Del Corral, J., 2018. The betting market over time: overround and surebets in european football. Economics and Business Letters 7, 129–136.

Graham, I., Stott, H., 2008. Predicting bookmaker odds and efficiency for uk football. Applied Economics 40, 99–109.

Hing, N., 2014. Sports betting and advertising. Australian Gambling Research Centre, Australian Institute of Family Studies.

Hvattum, L.M., Arntzen, H., 2010. Using elo ratings for match result prediction in association football. International Journal of forecasting 26, 460–470.

Joseph, A., Fenton, N.E., Neil, M., 2006. Predicting football results using bayesian nets and other machine learning techniques. Knowledge-Based Systems 19, 544–553.

Karlis, D., Ntzoufras, I., 2009. Bayesian modelling of football outcomes: using the skellam's distribution for the goal difference. IMA Journal of Management Mathematics 20, 133–145.

Electronic copy available at: https://ssrn.com/abstract=3655700

Knorr-Held, L., 2000. Dynamic rating of sports teams. Journal of the Royal Statistical Society: Series D (The Statistician) 49, 261–276.

Koning, R.H., 2000. Balance in competition in dutch soccer. Journal of the Royal Statistical Society: Series D (The Statistician) 49, 419–431.

Kuypers, T., 2000. Information and efficiency: an empirical study of a fixed odds betting market. Applied Economics 32, 1353–1363.

Levitt, S.D., 2004. Why are gambling markets organised so differently from financial markets?*. The Economic Journal 114, 223–246.

Maher, M.J., 1982. Modelling association football scores. Statistica Neerlandica 36, 109–118.

McGee, D., 2018. Beyond the betting shop: Youth, masculinity and the growth of online sports gambling, in: Youth Gambling Forum: Goldsmiths, University of London.

Pope, P.F., Peel, D.A., 1989. Information, prices and efficiency in a fixed-odds betting market. Economica , 323–341.

Ridder, G., Cramer, J.S., Hopstaken, P., 1994. Down to ten: estimating the effect of a red card in soccer. Journal of the American Statistical Association 89, 1124–1127.

Rue, H., Salvesen, O., 2000. Prediction and retrospective analysis of soccer matches in a league. Journal of the Royal Statistical Society: Series D (The Statistician) 49, 399–418.

Sauer, R.D., 1998. The economics of wagering markets. Journal of economic Literature 36, 2021–2064.

Snyder, J.A.L., 2013. What actually wins soccer matches: Prediction of the 2011-2012 premier league for fun and profit. University of Washington .

Tax, N., Joustra, Y., 2015. Predicting the dutch football competition using public data: A machine learning approach. Transactions on knowledge and data engineering 10, 1–13.

Electronic copy available at: https://ssrn.com/abstract=3655700

Timmaraju, A.S., Palnitkar, A., Khanna, V., 2013. Game on! predicting english premier league match outcomes.

Ulmer, B., Fernandez, M., Peterson, M., 2013. Predicting soccer match results in the English Premier League. Ph.D. thesis. Ph. D. thesis, Doctoral dissertation, Ph. D. dissertation, Stanford.

Witten, I.H., Frank, E., 2002. Data mining: practical machine learning tools and techniques with java implementations. Acm Sigmod Record 31, 76–77.

Electronic copy available at: https://ssrn.com/abstract=3655700