An Artificial Neural Network-based Prediction Model for Underdog Teams in NBA Matches

Paolo Giuliodori

University of Camerino, School of Science and Technology,
Computer Science Division, Italy
paolo.giuliodori@unicam.it

Abstract. In this work, we present an artificial neural network-based prediction model for underdog teams in NBA matches (ANNUT). We describe the steps of our supervised algorithm, starting from data acquisition to prediction selection. We talk about prediction selection because the final stage of our model is represented by a filtration phase. In this phase, the outputs returned from the neural network are evaluated according to how the events are quoted on one of the most famous bookmakers. Experimental results prove that the model is able to select with a certain accuracy winning teams. In particular, it reaches excellent results when we restrict the selection among underdogs (teams which probably will not win). Furthermore, we show that a significant sports prediction model cannot ignore bookmaker's odds.

Keywords: sports prediction model, artificial neural network, bookmaker's odd analysis, supervised classification, basketball

1 Introduction

The world of sports betting is a real jungle. There exists a huge number of bookmakers and prediction model for every sport. In this paper, we deal with predictions of outcome for basketball events. We consider the most famous basketball league, the National Basketball Association (NBA) played in North America. The worth of NBA is its incredible amount of matches. In fact, during the regular season, each team plays 82 games, 41 each home and away. Furthermore, most of those matches end with a small winning gap, this makes the NBA league one of the most exciting league in all sports. Due to this success, bookmakers and bettors follow NBA with extreme interest. These are some reasons why we choose NBA, in addition, we can count on a well-updated statistics database provided directly by NBA.
In this work, we present a prediction model based on training and application of artificial neural networks (ANNs) with the final aim to predict if home or away team is going to win the match. In detail, we see why all the predictions are not the same. In fact, we add another ingredient: bookmaker's odd. We describe all needed steps of analysis and the experimental results. Using classical techniques,

we train the ANN with data related to the last regular season ended in April 2017 and we show its performance. Moreover, we test our ANNUT model on data concerning the last part of the NBA regular season showing its worth also in practice i.e., money gained by betting on these predictions. We show how our model, according to experimental results, is able to select underdog teams which have an underestimated odd (according to the ANN output).
The paper is organized in the following way, in Section II, we report a brief background dealing with neural network and machine learning in general. Furthermore, we provide a description of sports betting. In Section III, we describe the state of art and similar works. In Section IV, we explain in details the phase of our analysis starting from data acquisition to outputs. In Section V, we present the experimental results applying our prediction model to matches of the last part of the regular season. Finally, in Section VI, we conclude the paper also with some possible future works.

2 Background

In this section, we are going to introduce briefly what is machine learning, in particular, data classification. Furthermore, we provide a description of the main actors involved in a betting scenario.

2.1 Machine Learning: Data Classification

Machine learning [8] is the subfield of computer science, it is considered the ability of a computing system to learn without being explicitly programmed. In machine learning, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations whose category membership is known (supervised learning [7]).
Artificial neural networks (ANNs) [5] are computing systems inspired by the biological neural networks that constitute animal brains. Such systems learn (progressively improve performance) to do tasks by considering examples, generally without task-specific programming. For example, in image recognition, they might learn to identify images that contain cats by analysing example images that have been manually labelled as "cat" or "no cat" and using the analytic results to identify cats in other images.
In particular, a feed-forward neural network is an artificial neural network wherein connections between the units do not form a cycle. In this network, the information moves in only one direction, forward, from the input nodes, through the hidden nodes (if any) and to the output nodes. It is common to use a backpropagation method [5] as a part of algorithms that optimize the performance of the network by adjusting the weights. This approach calculates the gradient of the loss function with respect to the weights in an artificial neural network.

2.2 Sports Betting

Sports betting is the activity of predicting sports results and placing a wager on the outcome. The frequency of sports bet upon varies by culture, with the vast majority of bets being placed on association football, American football, basketball, baseball, hockey, track cycling, auto racing, mixed martial arts and boxing at both the amateur and professional levels. Sports betting can also extend to non-athletic events, such as reality show contests and political elections, and non-human contests such as horse racing, greyhound racing and illegal, underground dog fighting.
The bookmaker functions as a market maker for sports wagers, most of which have a binary outcome: a team either wins or loses. The bookmaker accepts both wagers and maintains a spread which will ensure a profit regardless of the outcome of the wager (bookmaker's fee). In other words, bookmakers have a fixed income for every event, in fact, they underestimate every possible outcome of the event, i.e. they calculate the odd according to a higher probability than the expected one, meaning a lower odd for the outcome.
Odds for different outcomes in single bet are presented either in European format (decimal odds), UK format (fractional odds), or American format (moneyline odds). European format (decimal odds) are used in continental Europe, Canada, and Australia. They are the ratio of the full payout to the stake, in a decimal format. Decimal odds of 2.00 are an even bet with a theoretical implied probability of 50% (without considering bookmaker's fee).

3 Related Work

In the state of art concerning NBA prediction models we can find several works, for instance in [9], authors present a comparison between NBA and the National College Athletics Association Basketball (NCAAB). They evaluate the implementations of the multilayer perceptrons, random forest, and Naive Bayes classifiers. They used fewer variables than our model without considering every match as a single record and they do not take into account the distinction between home and road team. Furthermore, they do not consider the bookmaker's expectation of events.
In [4], authors formalize the problem of predicting NBA game results as a classification problem and apply the principle of Maximum Entropy to construct an NBA Maximum Entropy (NBAME) model that fits discrete statistics for NBA games, and then predict the outcomes of NBA playoffs using the model.
In journal paper [6], authors propose two network-based models to predict the behaviour of teams in sports leagues. This represents a different approach from our model, in fact, they do not start from team statistics or other parameters but they model sports leagues as networks of players and teams where the only information available is the work relationships among them.

4 Model Description

In the following section, we describe the analysis process flow. Our model is essentially composed of four phases. First of all, we have to collect data. Then, in the filtering phase, we process this raw dataset by removing irrelevant features. After that, we train the ANN deploying the resulting dataset as training source and we evaluate the performance of the generated network. Finally, we compare the ANN output, computed on a new input set of data referring to future events, to the bookmaker's odd of those events. In Fig. 2 we present a diagram describing the ANNUT model.

GP	Games Played	**FT%**	Free Throw Percentage
W	Wins	**OREB**	Offensive Rebounds
L	Losses	**DREB**	Defensive Rebounds
WIN%	Win Percentage	**REB**	Rebounds
MIN	Minutes Played	**AST**	Assists
FGM	Field Goalds Made	**TOV**	Turnovers
FGA	Field Goals Attempted	**STL**	Steals
FG%	Field Goal Percentage	**BLK**	Blocks
3PM	3 Point Field Goals Made	**BLKA**	Blocked Field Goal Attempts
3PA	3 Point Field Goals Attempted	**PF**	Personal Fouls
3P%	3 Point Field Goals Percentage	**PFD**	Personal Fouls Drawn
FTM	Free Throws Made	**PTS**	Points
FTA	Free Throws Attempted	**+/-**	Plus Minus

Table 1: List of variables provided by the NBA database.

4.1 Data Acquisition

In our model, every record represents a match. In every row, we insert home team statistics and road (away) team statistics. Statistics refer to previous matches played during the season until the day of the match. It is important to note that we take into account only home matches for the home team and only away matches for the away team. This is because we assume that teams have a slighltly different behaviour playing home or away. Data are extracted directly from the official NBA site [3]. In table 1, we can see all the features provided by NBA data source.
The NBA calendar is full of matches and it is common to have that two teams have a different number of played matches in the season. For this reason, to have a balanced dataset we normalize our data according to the actual played

minutes, considering in this way also overtime periods. Thus, we have values per minute, for instance: *points scored per minute*. The two chosen categories for classification are *win home* and *win road*, they are boolean variables set to 1 if home team or road team won the match. Data refer to the second half of the season allowing us to have a set of data based on a significant amount of already played matches.

4.2 Data Filtering

Once prepared the raw dataset, the next step is the filtering phase. We decided to select a subset of available variables. In details, we select ten features which describe almost completely the characteristics of a team. The selected features are shown in Tab. 2. We choose to remove first of all the statistics concerning win or loss matches because they do not describe playing styles of teams. Then, we remove the number of shots, we evaluate that having the success shot percentage and the points scored it is sufficient to have a consistent description of how much and how a team scores points. Furthermore, we remove foul statistics because they do not improve the network performances. Finally, we consider other variables (as the number of turnovers, blocks, etc.) giving information on the ability of teams mostly concerning defence playing. We get records composed of twenty fields, ten fields for each team.

PTS	Points	**AST**	Assists
FG%	Field Goal Percentage	**TOV**	Turnovers
3P%	3 Point Field Goals Percentage	**STL**	Steals
FT%	Free Throw Percentage	**BLK**	Blocks
REB	Rebounds	**+/-**	Plus Minus

Table 2: List of selected variables provided by the NBA database.

4.3 Neural Network Training Phase

Our ANNUT model is based on the deployment of an ANN for pattern recognition. Our aim is to predict which team will probably win the match, home or road. In order to build the network, we deploy the "Neural Network" tool provided by Matlab [2]. The algorithm implemented by the tool is based on a two-layer feed-forward network, with sigmoid hidden and output neurons. It is used to classify vectors into specified target categories. The network is trained with scaled conjugate gradient backpropagation. The training set is composed of 566 matches, played between late December 2016 and late March 2017. In the training phase, we select the following split of input data: 396 matches for training, 141 matches for validation, 28 matches for testing. We follow a classical neural network training approach by using also validation data to select the best performing network. Furthermore, we do not select a bigger training

dataset because we note that using a bigger one will train a network with worse or at least equal performance.
In order to evaluate the performance of the generated network, we show the receiver operating characteristic (ROC) curve presented in Fig. 1. By looking at the area under the curve ROC (AUROC) of Tab.3 (computed constructing trapezoids under the curve), we have a measure of the predictive accuracy of the model. We can see that the neural network has good prediction capabilities, in fact, we have to consider that we are into a non-deterministic context (sports events) in which we have a huge amount of involved variables. Our AUROC values are compared in Tab. 3 with classifiers presented in [4]. This table shows a significant gap between our ANNUT model and other approaches, underlying the good results reached by our trained network. In detail, we compare our ANNUT model (ANNUT^H for classification of the win of home team and ANNUT^R for the road team) with Naive Bayes (NB), logistic regression (LR), backpropagation neural networks (BP-ANN), random forest (RF) and NBAME model. At this point, the trained ANN can be used to classify new data and we can go further with the last analysis step.

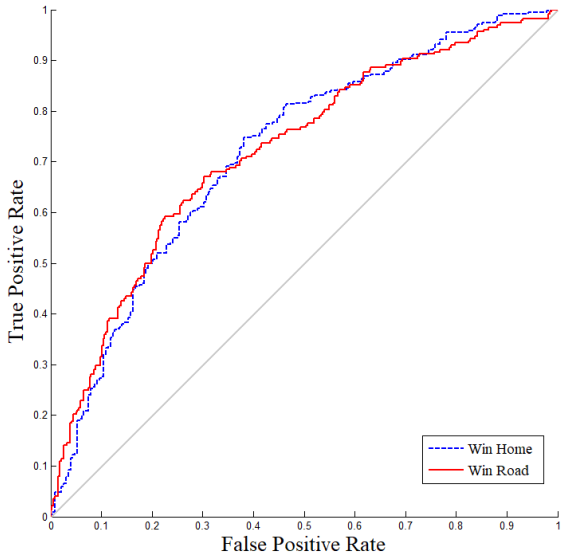

Fig. 1: The resulting ROC curve of the trained ANN.

NB	**LR**	**BP-ANN**	**RF**	**NBAME**	**ANNUT^H**	**ANNUT^R**
0.5575	0.5705	0.5635	0.5597	0.5853	0.7183	0.7195

Table 3: Comparison of Area under the ROC curve values for the two categories of our ANNUT model and the other classifiers presented in [4].

4.4 Comparison with Bookmaker's Odd

Now, we want to introduce a different perspective in order to evaluate the output of the network. Other approaches take simply the output from the network as a prediction, instead of in our model we evaluate network output in comparison with betting odds (in decimal format) of the match. In other words, we compare the real normalized value returned by the ANN (included between zero and one) with the implied bookmaker's probabilities. In fact, bookmakers select their odds according to the expectation of an event (implied probability) and the market. Starting from bookmaker's quotation, we can compute this probability of an event p by the equation:

$$p = \frac{1}{q \cdot c} \tag{1}$$

where q is the bookmaker's odd for that event and c is a fixed parameter indicating the bookmaker's commission (the average bookmaker's gain). The value p should return the implied probability computed by the bookmaker, i.e. the bookmaker's expectation on that outcome. For this reason, we consider also bookmaker's commission c. Once computed the value p for the event, we compare it to the normalized ANN output o. If the difference between values p and o exceeds a predefined threshold t, we choose this prediction. In other words, it means that our expectation of that event is greater at least of a factor t than the bookmaker's one.

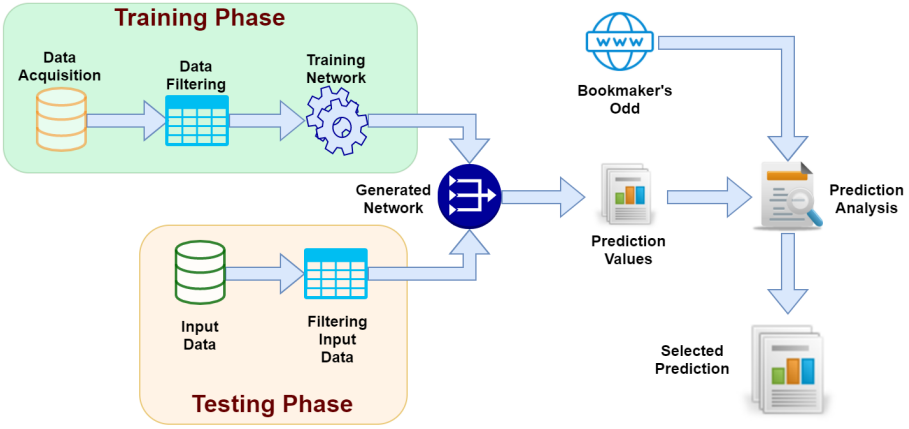

Fig. 2: Schema of the ANNUT model.

5 Experimental Results

We are now able to present the experimental results. We build the network with a dataset composed of 566 matches extracted from the official NBA database and odds are extracted from one of the most famous bookmakers, Bet365 [1]. After training the ANN, we apply our ANNUT model to another set of matches referring to the last part of the regular season 2016/2017, the first two weeks of April. This test dataset has 117 records (matches), in which every record always contains statistics for home and away team. For what concerns the Eq. 1 the parameter c is setted to 1.041, meaning a bookmaker's fee of $4,1\%$. This value comes out from the computation of the average commission of ten similar events quoted on Bet365. Furthermore, the selection threshold is equal to 0.10. So, it results that for every normalized output o if $o > p + 0.10$, we keep this prediction otherwise we discard it. Thus, we are interested not only in the pure prediction but in some way we are looking for a quality prediction. In fact, our aim is to design a profitable model because our perspective is to gain money in betting (or at least to not lose it). After this consideration, if we want to have a profitable model, we cannot neglect which is the bookmaker's odd of that event.
in order to evaluate our ANNUT model, accuracy was used as a performance measure, and it was calculated by the following formula:

$$\text{Accuracy} = \frac{\text{number of correct predictions}}{\text{number of predictions}} \cdot 100 \tag{2}$$

In the Tab. 4, we expose our results according to our prediction schema. In the first row, we consider every range of odd and we collect 69 event predictions over 117 considered matches, their winning rate, average odd and the balance (computed assuming that we bet one unit per prediction). In next rows, we restrict the selected predictions according to the odd, for instance in the second row we consider only events listed with an odd greater than 1.99. The winning percentage must be evaluated considering that our model selects events with high odds, i.e., low probability but high profitability. In order to better show the performance of our model, we compute also the implied probability according to the odds of the selected matches. The implied probability is exactly the value p computed by Eq.1 with $c = 1.041$ and q is the average odd. Tab. 4 shows that the accuracy of our model has a good margin over the implied probability, especially in the case of odds greater than 2.99 where we have a significant percentage of correct predictions with respect to the implied probability.
In the last two rows, we can see a consistent gain in units. We can interpret this result as the capability of our model to predict the win of an underdog team. Considering only odds between 3.00 and 10.00, our model selects 26 events with an average odd of 4.63. 10 out of 26 events are winning predictions with a winning average odd of 4.89. Assuming to bet one unit for each event, we have a total balance of +22.89 units. Taking into account matches referring to the last row of Tab. 4 and assuming that we have a base bet amount of 10.00 € on each match, in Fig. 3 we present the trend of our balance.

Odd	# Events	Accuracy	Av. Odd	Imp. Prob.	+/-
no restrictions	69	36.23%	3.39	27.36%	+5.02
$x > 1.99$	44	29.55%	4.48	21.44%	+12.55
$x > 2.99$	29	34.48%	5.53	17.37%	+19.89
$2.99 < x < 10.01$	26	38.46%	4.63	20.74%	+22.89

Table 4: Prediction results according to different odd levels.

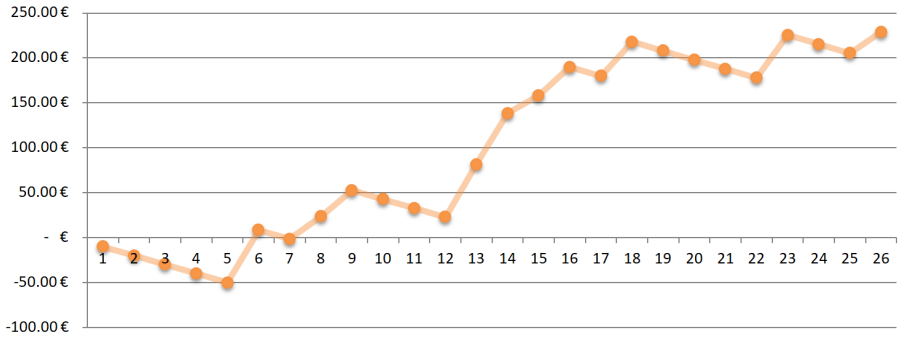

Fig. 3: The balance trend considering a base bet of 10.00 € per match with odds between 2.99 and 10.01.

6 Conclusion & Future Work

In conclusion, in this paper, we describe the ANNUT model showing its profitability on real events. The ANNUT model reveals good performance in term of ANN ROC curve but also concerning the selection of winning underdog teams which lead us to interesting money earnings showing his effectiveness into practice.
The interpretation of the success of our framework can be related to the consensus phenomenon which shows what percentage of the general betting public is on each side of the game. Thus, in order to balance the market, bookmakers change odds according to the sentiment of bettors. Considering this scenario, our ANNUT model shows its ability in spotting undervalued teams.
There are several future works whom we can add, for instance, we can make an additional analysis on how to choose the threshold t. Furthermore, one weak point is represented by the fact that the model does not consider missing players. This can lead to a wrong prediction if one or more relevant players will not probably play the game because odd will very high for that team.

References

1. Bet365 sports bookmaker, https://mobile.bet365.com/
2. Matlab neural network toolbox, https://it.mathworks.com/products/neural-network.html
3. Nba statistics team database, http://stats.nba.com/teams/traditional/#!?Season=2016-17&SeasonType=Regular%20Season&PerMode=Totals&sort=W_PCT&dir=-1
4. Cheng, G., Zhang, Z., Kyebambe, M.N., Kimbugwe, N.: Predicting the outcome of NBA playoffs based on the maximum entropy principle. Entropy 18(12), 450 (2016), https://doi.org/10.3390/e18120450
5. Cilimkovic, M.: Neural networks and back propagation algorithm http://dataminingmasters.com/uploads/studentProjects/NeuralNetworks.pdf
6. de Melo, P.O.S.V., Almeida, V.A.F., Loureiro, A.A.F., Faloutsos, C.: Forecasting in the NBA and other team sports: Network effects in action. TKDD 6(3), 13:1–13:27 (2012), http://doi.acm.org/10.1145/2362383.2362387
7. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of Machine Learning. The MIT Press (2012)
8. Munoz, A.: Machine learning and optimization. URL: https://www. cims. nyu. edu/˜ munoz/files/ml_optimization. pdf [accessed 2016-03-02][WebCite Cache ID 6fiLfZvnG] (2014)
9. Zimmermann, A.: Basketball predictions in the NCAAB and NBA: similarities and differences. Statistical Analysis and Data Mining 9(5), 350–364 (2016), https://doi.org/10.1002/sam.11319