Analyzing data from football matches utilizing deep neural networks with the aim of predicting the outcome of future games in order to establish if profit can be made by strategically placing bets based upon those predictions. Matches from the English Premier League. Italian Serie A, Spanish La Liga, Liga Portugal and French Ligue 1 have been analyzed.

Bets.ai Betting Research Company, G. P. Fialho

Abstract

Using artificial neural networks, this system aims to predict results of confrontations between two teams in the form of a home win, an away win or a draw, to determine if there are likely to be more than 2.5 goals or less than 2.5 goals scored during the match, if both teams will score, if any of the teams will score more than 1.5 goals or if they will win by at a difference of at least two goals and to verify if it is possible to profit from strategically placing bets using the obtained predictions.

Statistics were collected from approximately 24,000 matches across 5 different leagues: English Premier League, Italian Serie A, Spanish La Liga, French Ligue 1 and Liga Portugal. The data includes information on teams, (individual) players, numbers of goals scored, numbers of shots on target, numbers of corner kicks and some other (but no biometric data). A system for rating players and teams was developed in order to enhance the model and the (resulting) dataset analyzed using feature engineering to improve the data by filtering.

A deep neural network incorporating multiple layers was built with the aim of predicting the outcome of the matches. Predictions for 1,564 matches from 2021 and 2022 seasons achieved an accuracy of up to 81.5% using past historical data for high certainty matches.

In addition, a betting strategy was developed using the model and a simulation based on the predictions generated from the last two months of the 2021-2022 season. At the end of the simulation, the initial balance had increased by 333%.

To conclude, the results clearly demonstrated that it is indeed possible to predict the outcome of football matches and to use those predictions to make profit in the sports betting industry.

Introduction

Technology is a proven and essential part of our day-to-day life which is already having a radical impact on many areas - the betting industry being one of them. Significant progress has been made in recent years in the field of machine learning techniques and we have seen a considerable increase in its accuracy. Artificial neural networks are now outperforming humans in many areas of activity including that of forecasting which has seen constant progress and increases in accuracy that are only expected to improve from year to year.

Nowadays, a punter has a tremendous amount of data available online providing guidance for his journey through the world of betting. For example, there are TV programs, newspaper articles and many websites that present data and insights along with debates over who will win a specific match, the World Cup, the Premier League, or the UEFA Champions League etc. If a punter is serious about wanting to improve their income by making money from the betting industry, it is now pretty much essential for them to make use of data-based tools derived from AI technology.

The main objective of our work is to use the branch of machine learning, more specifically deep neural networks, to predict the outcomes of football matches from English Premier League, Italian Serie A, Spain La Liga and French Ligue 1 by combining both technology and statistical data and then verify if it's possible to use those predictions to make profit by placing bets on betting platforms.

Machine learning for football is a technique employed within the field of artificial intelligence which not only aims to highlight statistics and offer reports but that also aims to predict the results of future matches based on past data. Bets.ai aids bettors to an even greater extent by offering valuable additional information using a simple but rich presentation that even includes predictions for some matches. We can confidently say that artificial intelligence is changing the field of sports result prediction and improving the chances for punters to win.

Literature review

Constant technological advancement has allowed artificial intelligence to develop and become involved in many fields, frequently revolutionizing them and making AI very popular among many entrepreneurs and computer scientists. We will now discuss some projects and the conclusions drawn by people who attempted to use artificial intelligence to predict the outcome of football matches.

Rue and Salvesen in 2000 [1] created a Bayesian dynamic generalized linear model to predict football matches from the English Premier League and Spanish Division 1 and to make betting simulations.

They extracted time dependent skills and employed algorithms to predict matches based on previous results. The dataset was composed of more than 3800 matches from 1993 to 1997. As a result, a final cumulative return of 54% from Spanish Division 1 and 40% from the English Premier League was obtained from which they concluded that the presented model seemed to capture most of the information contained in the match results and to provide reasonable predictions.

Michael Purucker conducted research on predicting the results of sports matches [2] and made an artificial neural network to predict matches from the National Football League (NFL). The dataset consisted of statistics such as time of possession, yards gained and turnover margin from the first eight rounds of the league. The model achieved an accuracy of 61% compared with 72% from the domain experts.

Kahn [3] continued his work and was able to increase the accuracy. He gathered more statistical data from a greater number of matches compared to the previous paper. A total of 208 matches from the 2003 season were used combined with new features such as an away team indicator and a home team indicator. He used 192 matches for the training dataset and the last 2 weeks of the season as a testing set. As a result, the network achieved an accuracy of 75%.
Ben Ulmer and Matthew Fernandez, in 2014, predicted results of football matches from the English Premier League using artificial intelligence models [4]. They reported certain challenges when building the model including a lack of data and its randomness. They used features such as streak, ratings for each team and whether a team was playing at home or away. In addition, they implemented and fine-tuned 8 different models and compared them. The best model was a support vector machine with an error rate of 0.48.
On the other hand, they found that the models drastically under-predicted draws, with the lowest accuracy of 1% correctly predicting just 3 draws out of 294. This problem caught our attention, and we took it into particular account when building our prediction models. As we can see from the

studies, predicting the outcome of a football match with a relatively high accuracy is not an easy task. In fact, specific models are needed: a large set of historical data, current data analysis, pre-processing techniques, and computational methods for training.

Dataset

A relevant dataset filled with quality information is the foundation of any efficient machine learning model and the pathway through which artificial neural networks will detect patterns.

The quality of the dataset is directly associated with the accuracy of the neural network. In our case, for the dataset to be of quality, it must contain important information for the prediction and a positive relationship between the amount of information, the number of samples and the number of classes that the model predicts.

Generally, the number of samples must be greater than the number of variables in a factor of "x%" that can vary according to the purpose of the network (for complex problems this factor must be high). This is because the neural network needs to learn the importance of each variable to the result - the more information, the more difficult this task and therefore a greater number of samples is essential. There must also be a number "Y" of independent samples for each class, where Y can be tens, hundreds, or thousands depending on the complexity of the problem, so that the algorithm is able to calculate all the parameters that influence the choice of a particular class.

The dataset includes statistics from 24,143 matches from the following leagues: English Premier League; France Ligue 1; Italy Serie A; Liga Portugal; Spain La Liga. All of them have the same characteristics incorporating information about both players and teams plus a ranking system by points. All collected statistics from teams are presented in Table 1 and from players individually, in Table 2.

League	**Date**	**Team**	**Goals scored**	**Goals suffered**
Total shots	**Possession**	**Pass success (%)**	**Dribbles won**	**Aerials won**
Tackle success (%)	**Dispossessed**	**Total corners**	**Shots on Target**	**Dribbles Success (%)**
Fouls Committed	**Total Tackle success**	**Total Passes**	**Interceptions**	**Clearances**
Blocks	**Total Touches**	**Clean Sheets**		

Table 1 – ***collected team's statistics***

League	**Player's Name**	**Date**	**Position**	**Total Shots**
Shots on Target	**Total Key Passes**	**Pass Success (%)**	**Total Aerials Balls Won**	**Total Touches**
Total Interceptions	**Total Tackles Won**	**Total Clearances**	**Committed Fouls**	**Total Passes**
Total Crosses	**Total Accurate Crosses**	**Total Long Ball Passes**	**Long Ball Pass Accuracy**	**Total Passes Through Ball**

Passes Through Ball Accurate	**Total Red Cards**	**Total Yellow Cards**	**Goals**	**Total Minutes Played**
Number of great chances missed	**Total Assists**	**Total Errors that lead to enemy's goal**	**Total Last Man Tackle**	**Total Goalkeeper Interventions**

Table 2 – ***collected player's statistics***

For each match, 22 items of statistical data were collected for each of the two teams along with 29 items for every player. Multiplying these values by the 24,193 matches analyzed, a dataset composed initially of 16,499,666 data points was created.

In order to maximize the network's ability to identify patterns and improve its performance, new features were created using mathematical equations and manipulations of this (just described) dataset. Such feature creation is a key step as the data may well contain information that can only become meaningful in a transformed state [5].

New Features

Using the dataset, new features were created to improve the capacity of the neural network to learn and to find patterns and thereby increase its accuracy. Firstly, a ranking system was built for both players and teams.

The system allocates points to generate a score where the higher the score, the better the ranking. For each match, an attack and defense score was calculated for both home and away teams using relevant statistics combined with mathematical equations. Some data is directly proportional, and the statistic therefore increases the number of ranking points - other data is inversely proportional and decreases the number of points.

For example, statistics that are directly proportional for the attack rating system are: goals scored, shots on target, possession, number of corners, number of successful dribbles. Similarly, for the defense score these statistics are: number of interceptions, number of successful tackles, number of blocks and goalkeeper interventions.

Furthermore, each statistic is weighted to reflect the importance of its data more accurately to the rating system (attack or defense). For example, the number of goals scored is an essential piece of information about a team's attack power which is more important than, let us say, the number of corners or shots on target. It is therefore weighted to have a greater impact on the ranking score. Table 3 below shows the top 5 teams with the best attacks in the English Premier League of 2021-2022 season based on this rating system.

Position	Team	Score
1	**Manchester City**	**123.68**
2	**Liverpool**	**117.95**
3	**Chelsea**	**101.01**
4	**Tottenham**	**95.19**
5	**Leicester**	**87.49**

Table 3 – ***top 5 attack teams of England Premier League based on our rating system***

A similar process was applied to the players but using a different approach: an individual player's score was generated by combining their attacking and defensive performances whilst taking into consideration both their statistics during a match and their position on the field.

For example, for a defensive player, their defense rating is more important than the attack rating and so has a higher weighting on their final score even though their offensive capabilities were taken into consideration.

Table 4 shows the top 5 players with the highest rated performances in the English Premier League 2021-2022 season using the system.

Position	Name	Team	Score
1	**Heung-Min Son**	**Tottenham**	**119.17**
2	**Mohamed Salah**	**Liverpool**	**118.20**
3	**Cristiano Ronaldo**	**Manchester Utd**	**104.01**
4	**Harry Kane**	**Tottenham**	**103.82**
5	**Sadio Mané**	**Liverpool**	**101.46**

Table 4 – ***top 5 England Premier League's attack players based on our rating system***

In addition to the rating system, new data was generated using the dataset statistics to improve the model's capacity to learn. For example, a feature that we call ''goals expectation'' was calculated based on the combined ratings of both players and teams, focusing upon their attack vs defense performances. This new feature represents the expectation that a team will score goals.

Moreover, using the same idea of transforming the data into a more impactful form having a greater correlation to the outcome, these new features were created: precision, corner power, creation power, attack players power, defense players power, mid field players power.

Furthermore, some valuable additional information, the Pirating and Elo ratings, were calculated and incorporated into the dataset as they have already proved to be highly effective in predicting the outcome of football matches using machine learning [5, 6].

Even though we can make a prediction for every single match from the five leagues being considered, we have decided to release on the website only those predictions which exceed a certain confidence value.

Also, purely for reasons related to popularity and audience and despite having achieved results which we consider to be extraordinarily good, the Liga Portugal will not feature on the Bets.ai website in the first year. We are thinking of adding it together with the German Bundesliga 1 starting with the 2023-2024 season.

Pre-process

The data we use (see Table 1 and Table 2) to establish the team and player ratings is collected from finished matches. A model intended to predict future events does not include such data (for example: possession, number of shots on target, goals etc.) and must predict the outcome of a future match based on a combination of factors that includes historical data.

The only up to date element is represented by the lineups, the system formation, the specific teams involved in the match and the particular league where the match is being played. As can be seen in Tables 1 and 2, there are several areas of additional statistical information that will not be available until the end of a game. Therefore, we will use techniques that are able to interpret past data to predict future data.

There are several already proven methods that can perform this calculation for predicting football match results based on previous match results including the simple average [6], a weighted average [7] or the exponential average [8]. All of these different approaches were assessed using different models, compared, and the most effective one chosen.

Separate home and away data

The dataset consists of 45% home team wins, 29% wins of teams playing away from home and 26% of draws. These values are in line with the reality of the sport and clearly demonstrate the existence of a superiority for teams playing at home [9].

Below, Figure 1 illustrates this difference in concentration.

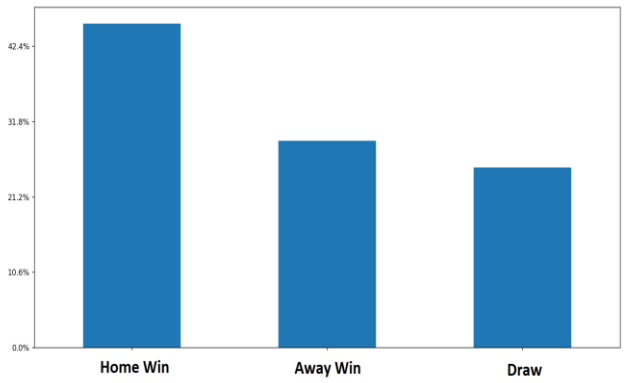

Figure 1 - Graph of numbers of matches by results - respectively a home team win, an away team win and a draw

Furthermore, figures 2, 3 and 4 show a concentration of ball possession, shots on target and goals scored for teams playing both at home and away. The blue line represents an identity line - if there were no advantage for a team playing at home, we would expect the distribution of values to be scattered on or around this line. However, as we can see, they sit to the right and below, clearly reflecting the superiority of the home team.

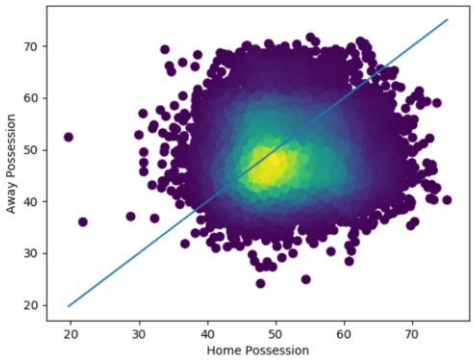

Figure 2 – Home and Away ball possession concentration

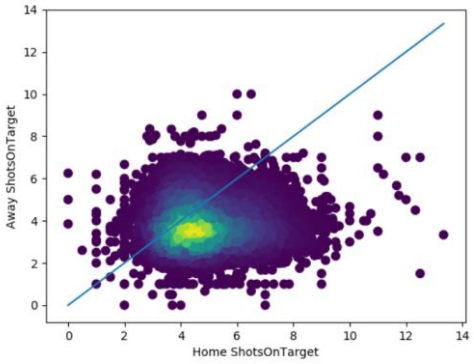

Figure 3 – Home and Away Shots on Target Concentration

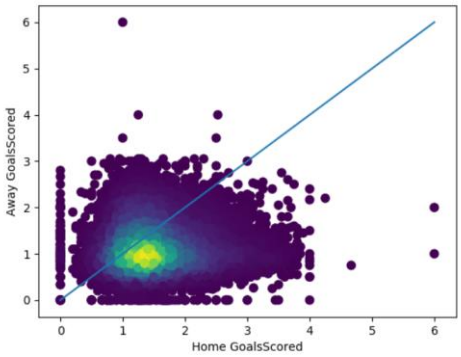

Figure 4 – Home and Away Goals Scored Concentration

In addition to organizing in a temporal way, to account for this home field advantage, the data has been calculated separately depending on whether a team is playing at home or away.

For a team playing at home, the statistical data is derived from those matches played most recently at home - equally, for a team playing away from home, the data is calculated using recent matches played away from home.

Feature Scaling

Feature scaling is necessary to ensure that all entries are within a comparable range [10]. For example, considering a neural network with say two inputs, x1 and x2 - if x1 varies from 0 to 0.5 and x2 from 0 to 1000, a change in x1 of 0.5 equates to a 100% change whereas a change in x2 of 0.5 would be equivalent to a change of just 0.05%.

Feature scaling is therefore absolutely necessary for the correct functioning of the model since the statistical data making up the dataset contain differing (and sometimes wide) variations in the range of values. Two techniques are used: normalization and standardization. The results are compared, and the best is then selected.

Encode categorical data

The dataset contains categorical data which corresponds to the league in which the football match is being played. As artificial neural networks are unable to support categorical data as an input, the technique of the binary vector “One-hot Encoding” is used [11].

The league feature is transformed into vectors of binary numbers with each column representing a specific league. For example, for a game in Serie A (which corresponds to the second column of the binary vector), the value attributed to this column will be 1 and the remaining columns 0.

Machine Learning Model

The model aims to represent a real problem in the form of a mathematical expression. The predictability of a particular phenomenon depends upon finding a pattern amongst the variables that encompass it. Football involves many variables and, being a sport played by human beings, the behavior of the players affects the outcome.

The relationship between these variables can often give rise to surprising and unexpected results. The problem is of high complexity and the neural network needs to be structured in a manner that can support this.

For instance, if a deep neural network with multiple hidden layers is used, these extra layers will enable the network to identify complex, nonlinear relationships between the variables [12].

The network output is composed by three binary values: the first for the victory of the home team, the second for a draw and the third for the victory of the away team. As previously mentioned, the results of the matches are unbalanced in that there are more home team wins than draws and/or away team wins. To compensate for this, a weight function is applied to each class (home team win, draw, away team win) so that it is proportional to sample numbers. The error function also becomes weighted and the degree of weighting for each sample is specified [13]. Consequently, by distributing the model’s output, its ability to learn will be improved [14].

The network's training dataset (24k matches) was divided as follows: 80% (19k) were used as training dataset, 10% (2,4k) were used as validation and the remaining 10% as test dataset. The dates on which the games took place are arranged in ascending order - from the training set through to validation and testing.

Feature Engineering

It is important to establish the value of the data. This can be determined by observing the increase in error level for each model after the data has been randomly permuted since the relationship between that variable and the result will have been cut.

Data is considered important if after randomly 'shuffling' its values, the error of the model increases as this demonstrates the model's dependence on that variable to make the prediction. Conversely, if the model's error does not change or even decreases after the values have been randomly permuted, the data is considered unimportant as it was not required to make the prediction and the feature is therefore not needed.

This technique was introduced by Breiman [15] and a model for calculating the importance of data for a forecast model subsequently proposed by Fisher, Rudin and Dominici [16]. The feature importance has been calculated for all of the features from the dataset and variables that are not important to the model have been removed.

Results

The model was used to predict the results of 1,564 matches from the 2021-2022 English Premier League, Liga Portugal, Spain La Liga, France Ligue 1, Italy Serie A with a combined accuracy of 50.8%.

However, where the deep neural network predicted a home win, draw or away win with a certainty of more than 60%, the accuracy increased to 69.27% and where the certainty was more than 70%, accuracy increased to 78.46%. For a certainty of more than 80%, accuracy increased to 83.3%.

The degree of certainty represents the anticipated accuracy of the model's predictions. For example, if a prediction indicates 69% for the home team to win, 5% for the away team to win and 26% for a draw, then the model has a 69% certainty that the home team will win (some matches have a high randomness factor whereas other matches are more predictable).

In addition, there are variations in results across different leagues with some leagues bring more predictable than others. For example, table 5 below shows the accuracy of all leagues for a certainty of more than 60%.

League	Accuracy
Liga Portugal	81.57%
Italy Serie A	74.4%
England Premier League	72.9%
France Ligue 1	70.8%
Spain La Liga	67.5%

Table 5 – Accuracy of predictions with more than 60% certainty

To verify whether or not it is possible to win money using the model, a betting simulation was made. The experiment consisted of placing bets based only on the deep neural network's predictions and using a betting strategy for matches selected from the last 2 months of the 2021-2022 season. The betting strategy is explained below.

Betting Simulation

The intended goal is to double the initial balance. For example, to take an opening balance of 1000$ and to increase it to 2000$ as quickly as possible, withdraw the profit and then to start over. For this simulation, we will use only predictions having odds between 1.50 and 1.99.

In addition, the stake is to be dynamic. The first stake will be 20% of the initial balance and then following each win, the stake will rise by 10% of the initial balance. For example, for an initial balance of 1000$, the first bet will be 200$ - if a win is achieved, the second bet will be 300$, the third bet 400$ and so on following each win. Note that the stake grows only after a win - following a loss, the bet does not increase and remains at its current level. Should the suggested stake ever become higher than the available balance, the entire amount will be used to go 'all in'.

This strategy was implemented for predictions obtained in the last 2 months of the 2021-2022 season. The results of this strategy using the machine learning model can be seen in Table 6.

Match				Predictions			Simulation			
Date	**Home Team**	**Away Team**	**Result**	**Home Win**	**Draw**	**Away Win**	**Odd**	**Bet**	**Profit**	**Balance**
2022-04-02	Celta Vigo	Real Madrid	1-2	0.091518	0.220327	**0.688155**	1,80	200$	160$	1160$
2022-04-03	Fiorentina	Empoli	1-0	**0.567488**	0.277117	0.155395	1,53	300$	159$	1319$
2022-04-03	West Ham	Everton	2-1	**0.591350**	0.276141	0.132509	1,75	400$	300$	1619$
2022-04-03	Tottenham	Newcastle	5-1	**0.649805**	0.242894	0.107301	1,52	500$	260$	1879$
2022-04-09	Southampton	Chelsea	0-6	0.113897	0.250882	**0.635221**	1,90	600$	540$	2419$
2022-04-09	Cagliari	Juventus	1-2	0.093170	0.225264	**0.681566**	1,62	200$	124$	1124$
2022-04-10	Napoli	Fiorentina	2-3	**0.603296**	0.257307	0.139397	1,75	300$	-	824$
2022-04-10	Torino	AC Milan	0-0	0.100431	0.232155	**0.667414**	1,85	300$	-	542$
2022-04-16	Lazio	Torino	1-1	**0.505226**	0.35853	0.136244	1,65	300$	-	242$
2022-04-17	Nice	Lorient	2-1	**0.504598**	0.252699	0.242703	1,66	242$	159$	401$
2022-04-20	Osasuna	Real Madrid	1-3	0.105388	0.242999	**0.651613**	1,80	342$	273$	674$
2022-04-23	St Etienne	Monaco	1-4	0.151476	0.291933	**0.556591**	1,70	442$	310$	984$
2022-04-24	Empoli	Napoli	3-2	0.270135	0.231821	**0.598044**	1,60	542$	-	442$
2022-04-25	Sassuolo	Juventus	1-2	0.107674	0.244795	**0.647531**	1,78	442$	345$	787$
2022-04-29	Sevilla	Cadiz	1-1	0.128334	0.266808	**0.604858**	1,55	542$	-	245$
2022-04-02	Newcastle	Liverpool	0-1	0.075207	0.199062	**0.725731**	1,50	245$	122$	367$
2022-04-02	Spezia	Lazio	3-4	0.156472	0.303554	**0.539974**	1,50	367$	183$	550$
2022-05-01	AC Milan	Fiorentina	1-0	**0.520573**	0.297527	0.181901	1,58	467$	271$	821$
2022-05-07	Torino	Napoli	0-1	0.099882	0.231266	**0.668852**	1,95	567$	519$	1340$
2022-05-07	Celta Vigo	Alaves	4-0	**0.517842**	0.279003	0.203155	1,80	667$	533$	1873$
2022-05-08	Norwich	West Ham	0-4	0.172388	0.301651	**0.525961**	1,70	767$	537$	2410$

2022-05-08	Lorient	Marseille	0-3	**0.508723**	0.297274	0.194003	1,90	200$	180$	1180$
2022-05-08	Verona	AC Milan	1-3	**0.531090**	0.298104	0.170806	1,65	300$	195$	1375$
2022-05-14	Udinese	Spezia	2-3	**0.503931**	0.303492	0.192577	1,65	400$	-	975$
2022-05-15	Napoli	Genoa	3-0	**0.693851**	0.194217	0.115395	1,66	400$	264$	1239$
2022-05-15	Betis	Granada	2-0	0.108468	0.24244	**0.649092**	1,53	500$	265$	1504$

Table 6 – testing the predictions with >= 50% chances

**When two or more predictions were available for matches played at the same hour, the prediction with the greatest chance of success was chosen.*

Table 6 shows the model's predictions for the home team to win, the away team to win and a draw. This value is between 0 and 1 and represents the probability for the event to happen - with the sum of all probabilities always adding up to 1. For example, in the first line the probability for Real Madrid to win at Celta Vigo is 0.688155 - equivalent to a 68.81% probability.

A bet is placed for the event having the highest probability. So, for the first match, a bet of 200$ was placed on Real Madrid to win the match vs Celta Vigo. The table shows the odds for the prediction with the highest probability, the actual result of the match and the profit plus balance after the match. When the updated balance exceeds double the initial balance, all profits are withdrawn, and the strategy repeated using the initial balance to recommence.

The initial balance doubled on two occasions during the course of the simulation representing a gross profit of 2829$ and an increase of almost 283% on the initial balance. The simulation ended on the penultimate match day with a balance of 1504$ – adding a further 50% of the initial balance. The total profit therefore amounted to 333% over just one and a half months

This result clearly shows that our strategy is extremely efficient and that neural deep networks can indeed be used to predict football matches and thus to make profit in the betting industry.

Conclusion

This work aimed to create a model for predicting the results of football matches using artificial intelligence and to verify if it is possible to use such a model for profit. Several steps and techniques were used to create the final machine learning model. Furthermore, the dataset was pre-processed to optimize the training of the network and new data generated to improve performance.

A deep neural network was created and a betting simulation was developed with a specific strategy. Based on the results obtained, we can conclude that it is possible to predict the outcome of football matches using artificial intelligence and to use those predictions to make a profit from the betting industry. There is potential to increase an initial balance by a factor of almost five over a three-month period.

References:

[1] Rue, H. and Salvesen, O. (2000) “Prediction and retrospective analysis of soccer matches in a league”. J. R. Stat. Soc. Ser.D (Stat.) 49, 399–418.

[2] M. C. Purucker, (1996) "Neural network quarterbacking," in IEEE Potentials, vol. 15, no. 3, pp. 9-15, Aug.-Sept. 1996, doi: 10.1109/45.535226.

[3] Kahn, J. (2003). “Neural Network Prediction of NFL Football Games”. World Wide Web Electronic Publication ECE539 Fall 2003 December 19.

[4] Ulmer, B., & Fernández, M.P. (2014). “Predicting Soccer Match Results in the English Premier League”.

[5] Heaton, Jeff (2020), “An Empirical Analysis of Feature Engineering for Predictive Modeling” pp. 1-6, doi: 10.1109/SECON.2016.7506650.

[6] Buursma, D. (2011). “Predicting sports events from past results Towards effective betting on football matches”, University of Twente P.O. Box 217, 7500AE Enschede The Netherlands.

[7] Kumar, G. (2013). “Machine Learning for Soccer Analytics”, Thesis submitted for the degree of Master of Science in Artificial Intelligence, option Engineering and Computer Science.

[8] Dixon, M. and Coler, S. (1997) “Modelling association football scores and inefficiencies in the football betting market”. Journal of the Royal Statistical Society: Series C (Applied Statistics) 46.2 (1997), pp. 265–280.

[9] Pollard, R. (2008). “Home Advantage in Football: A Current Review of an Unsolved Puzzle”. The Open Sports Sciences Journal. 1. 10.2174/1875399X00801010012.

[10] Ahsan, M.M.; Mahmud, M.A.P.; Saha, P.K.; Gupta, K.D.; Siddique, Z. (2021) “Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance”. Technologies 2021, 9, 52. https://doi.org/10.3390/ technologies9030052

[11] Hancock, J. &Taghi, K. (2020). “Survey on categorical data for neural networks”. Journal of Big Data. 7. 10.1186/s40537-020-00305-w.

[12] Ogunmolu, Olalekan & Gu, Xuejun & Jiang, Steve & Gans, Nicholas. (2016). “Nonlinear Systems Identification Using Deep Dynamic Neural Networks”, arXiv:1610.01439v1 5 Oct 2016.

[13] Jung, S. Kwon, S. (2013). “Weighted error functions in artificial neural networks for improved wind energy potential estimation”, Aplied Energy, Volume 111, 2013, Pages 778-790 ISSN 0306-2619.

[14] Kibet, A. (2020), “Classification in Imbalanced Data Sets. Understanding and utilizing imbalanced data.” Towards Data Science, available at: https://towardsdatascience.com/classification-framework-for-imbalanced-data-9a7961354033

[15] Breiman, L. (2001) “Random Forests”. Machine Learning 45 (1). Springer: 5-32, 2001.

[16] Fisher, A., Rudin, C. and Dominici, F. (2018). “Model Class Reliance: Variable importance measures for any machine learning model class, from the ‘Rashomon’ perspective.” 2018, available at: http://arxiv.org/abs/1801.01489