A Machine Learning Approach to Football Match Result Prediction

Luca Carloni, Andrea De Angelis, Giuseppe Sansonetti, and Alessandro Micarelli

Department of Engineering, Roma Tre University,
Via della Vasca Navale 79, 00146 Rome, Italy
ailab gs ansone @dia.uni roma3.it

Abstract This paper describes the design and implementation of predictive models for sports betting. Specifically, we focused on exploiting Machine Learning (ML) techniques to predict football match results. To this aim, we realized an architecture that operates in two phases. First, it extracts data from the Web through scraping techniques. Then, it gives the collected data in input to di erent ML algorithms. Experimental tests showed encouraging performance in terms of the Return on Investment (ROI) metric.

Keywords: Machine learning, Artificial neural networks, Sport bets

Introduction

Recent technological advances in Machine Learning (ML) [29] (e.g., Deep Learning [22, 14]) play a central role in our lives. ML designs the services of our cities [8], recommends which places [21] may be of interest to us (e.g., cultural heritage resources [23] or restaurants [3]) and how to reach them [9]. It suggests to us which news articles [6] or research papers [13] to read, which music artists and songs to listen to [19], which movies to watch [2], which products to buy [4], and even people to hang out with [11]. It is, therefore, natural that we turn to ML even when it comes to betting our money on sports events.

In this paper, we illustrate a forecasting system aimed to profit in the sports betting market using ML techniques. More specifically, we adopted Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Naïve Bayes, and Random Forest, as well as a four-layer Artificial Neural Network (ANN). Those ML techniques were compared with each other using the prediction accuracy as an evaluation metric. The comparative analysis we carried out led us to choose the ANN as the learning technique since it allowed us to achieve better results. Subsequently, through simulation and prediction trials, we assessed the system performance. Finally, the system was also tested in predicting results related to matches still to be played. The performance in terms of Return on Investment (ROI) has been encouraging, which motivates us to pursue our research activities in this area.

2 Related Work

Among the first approaches advanced in the research literature to predict the result of sports events is the one illustrated in [26], in which the author proposes the use of least squares to make predictions on football and basketball matches. Since then, there have been numerous contributions on the same topic. For instance, in [10] the authors argue that regression models allow for more accurate results than those provided by domain experts. Furthermore, human experts are unable to process the publicly available data efficiently. Loe elholz *et al.* [16] put forward the use of neural network models on a dataset of 620 games. One of the reasons why neural networks are so widely used in this domain lies in their flexibility in defining the class to be identified. For instance, in [1] the authors use two classes (home goals and away goals), while in [18] the class expresses the probability of winning. In [7], the authors describe a model, named *pi football*, that takes advantage of a Bayesian Network to predict the outcome of football matches belonging to the English Premier League (EPL). This model takes into account both objective and subjective information, also weighing the available data through degrees of uncertainty. However, apart from a few authoritative exceptions, the results obtained in the first decade of the 2000s are generally not particularly encouraging. In [12], Haghighat *et al.* suggest that this is due to the limited size of the datasets. They, therefore, propose to consider player-level statistics as well and to adopt more advanced ML models. The advent of Deep Learning (DL) has provided a new boost to the entire domain. For example, in [17] the author illustrates an experimental evaluation conducted on di erent DL models on US National Basketball Association (NBA) matches. Interestingly, the author shows that significant results in terms of profit could be attained only by integrating statistics data with features extracted by experts from video recordings. The authors of [15] describe a convolutional neural network-based approach for the prediction of the outcome of basketball games, where the convolutional layer allows the system to exploit player-level data. They also suggest that correlating the prediction of the results with the predictions made by bookmakers is not helpful for the accuracy of the final results. In [20], the authors illustrate a system based on a Multi-Layer Perceptron (MLP) for predicting the outcome of football matches and show that it outperforms approaches based on traditional ML algorithms like Support Vector Machine (SVM) and Random Forest. Tiwari *et al.* [28] propose a model for predicting the result of football matches that makes use of Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM). To increase the model performance, the authors consider any events during the match and their consequences on the final result. In [5], the authors study the potential provided by ANNs by thoroughly analyzing ANN-based approaches for predicting the outcome of sports events and identifying what they believe are the challenges still to be solved. Consequently, they propose a prediction system based on the CRISP-DM model [24]. In [27], the authors propose to use di erent ML techniques to predict the outcome of football matches based on features related to not only the game but also the players.

3 System Architecture

The overall architecture of the proposed system is shown in Figure 1. The system

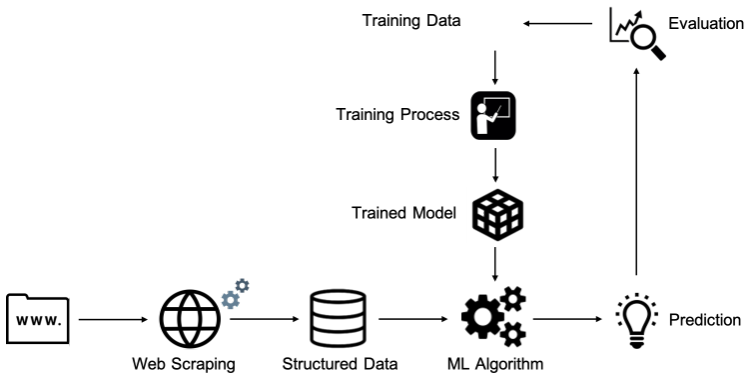
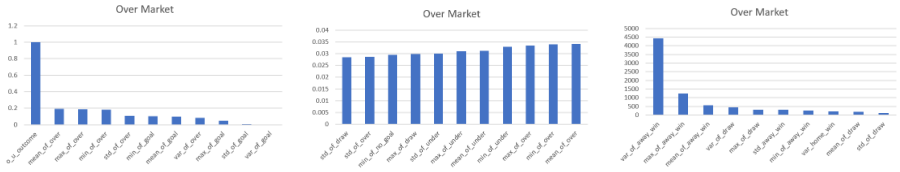


Fig 1 The overall system architecture.

works in two phases. In the first phase, the extraction of the betting odds is performed using a web scraping technique. These odds are related to the top betting markets such as *1X2*, *over under 2.5*, and *goal no goal*. Specifically, with the 1X2 market, we mean the type of bet that can be made to predict the outcome of a match, with the over-under 2.5 market we mean the market that takes into account the sum of the number of goals scored within a single match, with the goal-no-goal market we mean whether or not the match ended with at least one goal. In the second phase, the collected data is taken as input, and predictions are made through ML techniques. The web scraping operation was carried out from June to December 2020, through which we collected the following data:

Number of countries: 12;
Number of matches: 49,319;
Number of seasons: for the highest leagues all the seasons starting from 2008-2009 until 2019-2020, for lower leagues from 2010-2011 until 2018-2019;
Number of original features: 47;
Missing cells: 349,215.

This data was subjected to a data cleaning operation, which reduced the number of matches to 36,461. Before applying the Machine Learning techniques, it was essential to carry out a relevance analysis of the features available. The objective of this analysis is to identify the most relevant features and discard the least relevant ones, which could harm the forecasting model and, consequently, reduce its accuracy. Among the techniques used there are Univariate Selection, Feature Importance, and Correlation Matrix. The example graphs shown in Figure 2



(a) Correlation matrix (b) Feature importance (c) Univariate selection

Fig 2 Feature analysis results.

are related to the over-under 2.5 market. These analyzes allowed us to reduce the total number of features to 31, thereby improving performance in terms of accuracy as well as computational efficiency.

4 Experimental Evaluation

Once the data extraction, cleaning, and analysis procedures were completed, ML techniques were applied to the data just learned. This process occurred in two phases. Initially, we employed di erent ML techniques to understand which of them was the most efficient for that specific market. Once the best technique for that market had been identified, it was then reused for incoming matches. The di erent ML models were:

Logistic Regression;
K-Nearest Neighbors;
Support Vector Machine;
Naïve Bayes;
Random Forest;
Four-layer Artificial Neural Network.

Subsequently, through the use of simulation and prediction techniques, we wanted to evaluate the e ectiveness and efficiency of the system. Figure 3 shows the results in terms of accuracy for the over-under 2.5 market by applying the di erent ML techniques. It can be noted that ANN outperformed the other methods. The comparative analysis, therefore, led us to choose the ANN as the learning technique, since - as we have seen - it is more efficient than the other classification methods tested. Based on the previous results, we used the ANN as a model for the testing phase. The data was divided according to the classic 80-20 method: 80% of the data was used to train the network, while the remaining 20% was used as a test-set. We performed two experimental sessions. In the first session, we used the Return on Investment (ROI) as a metric to evaluate the probability of gaining a profit from an investment. ROI is defined as the ratio between the gain/loss realized by an investment and its initial cost. More specifically, we evaluated the system by monitoring the ROI trend as the threshold varied, which was set on the output of the neural network. It varied from the value of 0.9 up

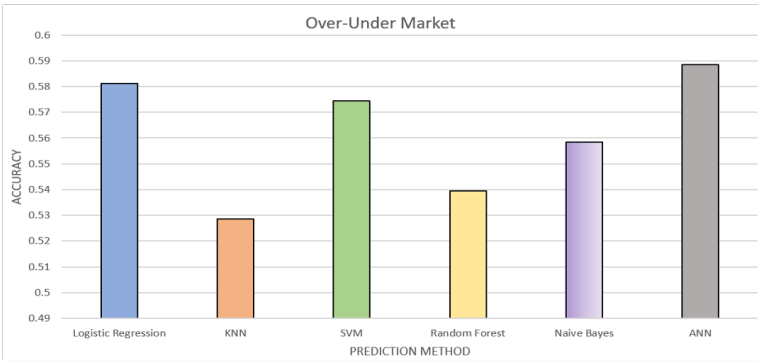


Fig 3 Comparative analysis results.

to 0.1: if the match was higher than the considered threshold, then it was considered for that market, else it was not taken into consideration. Figure 4 shows some example graphs. On the x-axis, there is the number of matches that are

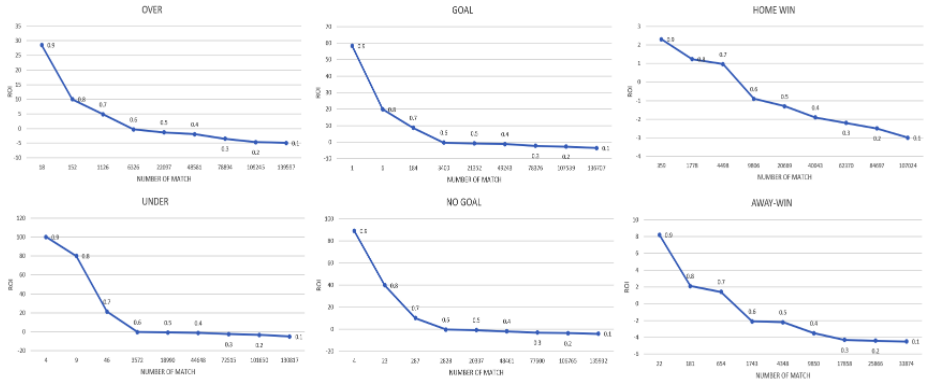


Fig 4 Results of the first experimental session.

higher than the aforementioned threshold, on the y-axis, the ROI value. As can be expected, the number of matches increases as the threshold value decreases, but this does not mean having a higher profit since the probability of winning is reduced. This session was repeated for each market that was scrapped. From the graphs it is possible to notice that there is a positive ROI value for high thresholds, generally starting from 0.6, 0.7. This is since for lower values we are in the area of complete randomness.

In the second experimental session, we wanted to compare the proposed approach with a baseline system that, di erently, always plays the lowest odd for the given market. Taking once again the over-under 2.5 market as an example,

this means that if we want to play the over for a game, the baseline system places credits on the over only if the betting odd is lower than that of the under, while our system places credits on the over only if the match is higher than the set threshold value. The obtained results are shown in Figure 5. We can note

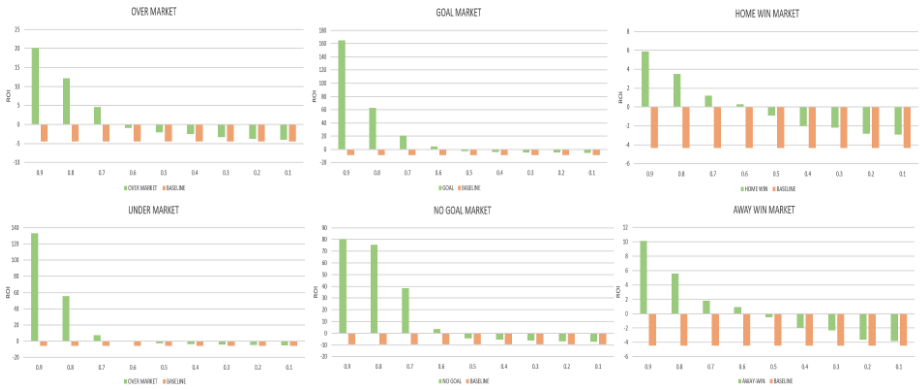


Fig 5 Results of the second experimental session.

how the baseline system has a negative ROI, which happens for every market. This information confirms that the baseline system approach does not lead to profits. We can also observe that, for low threshold values, our system does not lead to gains, but this is predictable. Taking a threshold value equal to 0.2 as an example, this means that event has a probability of success of 20%, which turns out to be a negligible probability. From the graphs, we can see that, di erently, for higher thresholds (i.e., for higher probabilities), the ROI value has negligible losses or, most of the time, goes positive. Below the 0.5 value included, it is normal to have a negative ROI since playing for such low thresholds is equivalent to relying on chance.

Once the test phase on matches already played had been completed, we tested the system performance even on matches still to be played. The considered matches were all those on the weekend from 4 to 6 December 2020. Overall, out of 89 matches selected from higher and lower leagues, we obtained 65 positive outcomes, 13 negative outcomes, while 9 were considered *NoBet* (i.e., looking at the di erent probabilities it was deemed not convenient to place credits on that match), or credits were placed but the game was not played for di erent reasons. Even after this test, the ROI was calculated: placing 10 credits on each game, we reached a gain of 999.7 credits out of 790 played, with a net gain of 209.7 and an overall ROI of 26.54%, that is, an encouraging final result.

5 Conclusions and Future Work

In conclusion, we first developed a web scraping system, which retains the opening and closing betting odds of football matches belonging to di erent national leagues. Then, we employed di erent prediction techniques with diverse classification algorithms. Finally, we carried out experimental trials aimed to show the efficiency and e ectiveness of the proposed system, also testing it on matches still to be played with good final results.

Among the possible future developments, there is undoubtedly the possibility of taking any temporal variation of odds for each archived match. This was not done as the scrapped betting sites did not report those variations for archived matches. Furthermore, for matches still to be played, it is possible to retain any change in odds up to the last available at the scraping time. We can also think of replacing the ANNs with Recurrent Neural Networks, which are very efficient in managing instances belonging to di erent time intervals. Finally, we can integrate the data extracted from online betting sites with user-generated content on social media. For example, in [25], the authors use tweet posts to increase the accuracy of a US National Football League (NFL) match outcome forecasting system.

References

1. Arabzad, S.M., Tayebi Araghi, M., Sadi-Nezhad, S., Ghofrani, N.: Football match results prediction using artificial neural networks; the case of Iran pro league. Journal of Applied Research on Industrial Engineering 1(3), 159–179 (2014)
2. Biancalana, C., Gasparetti, F., Micarelli, A., Miola, A., Sansonetti, G.: Context-aware movie recommendation based on signal processing and machine learning. In: Proceedings of the 2nd Challenge on Context-Aware Movie Recommendation. pp. 5–10. CAMRa '11, ACM, New York, NY, USA (2011)
3. Biancalana, C., Gasparetti, F., Micarelli, A., Sansonetti, G.: An approach to social recommendation for context-aware mobile services. ACM Trans. Intell. Syst. Technol. 4(1), 10:1–10:31 (Feb 2013)
4. Bologna, C., De Rosa, A.C., De Vivo, A., Gaeta, M., Sansonetti, G., Viserta, V.: Personality-based recommendation in e-commerce. In: CEUR Workshop Proceedings. vol. 997. CEUR-WS.org, Aachen, Germany (2013)
5. Bunker, R.P., Thabtah, F.: A machine learning framework for sport result prediction. Applied Computing and Informatics 15(1), 27–33 (2019)
6. Caldarelli, S., Gurini, D.F., Micarelli, A., Sansonetti, G.: A signal-based approach to news recommendation. In: CEUR Workshop Proceedings. vol. 1618. CEUR-WS.org, Aachen, Germany (2016)
7. Constantinou, A.C., Fenton, N., Neil, M.: pi-football: A bayesian network model for forecasting association football match outcomes. Knowledge-Based Systems 36, 322–339 (2012)
8. DAniello, G., Gaeta, M., Orciuoli, F., Sansonetti, G., Sorgente, F.: Knowledge-based smart city service system. Electronics (Switzerland) 9(6), 1–22 (2020)
9. Fogli, A., Sansonetti, G.: Exploiting semantics for context-aware itinerary recommendation. Personal and Ubiquitous Computing 23(2), 215–231 (Apr 2019)

10. Forrest, D., Simmons, R.: Forecasting sport: the behaviour and performance of football tipsters. International Journal of Forecasting 16(3), 317–331 (2000)
11. Gasparetti, F., Sansonetti, G., Micarelli, A.: Community detection in social recommender systems: a survey. Applied Intelligence (2020)
12. Haghighat, M., Rastegari, H., Nourafza, N.: A review of data mining techniques for result prediction in sports. Advances in Computer Science: an International Journal 2, 7–12 (2013)
13. Hassan, H.A.M., Sansonetti, G., Gasparetti, F., Micarelli, A.: Semantic-based tag recommendation in scientific bookmarking systems. In: Proc. of the 12th ACM Conf. on Recommender Systems. pp. 465–469. ACM, New York, NY, USA (2018)
14. Hassan, H.A.M., Sansonetti, G., Gasparetti, F., Micarelli, A., Beel, J.: Bert, elmo, USE and infersent sentence encoders: The panacea for research-paper recommendation? In: Tkalcic, M., Pera, S. (eds.) Proceedings of ACM RecSys 2019 Late-Breaking Results. vol. 2431, pp. 6–10. CEUR-WS.org (2019)
15. Hubácek, O., Sourek, G., Železný, F.: Exploiting sports-betting market using machine learning. International Journal of Forecasting 35(2), 783–796 (2019)
16. Loe elholz, B., Bednar, E., Bauer, K.W.: Predicting NBA Games Using Neural Networks. Journal of Quantitative Analysis in Sports 5(1), 1–17 (January 2009)
17. Maymin, P.Z.: Wage against the machine: A generalized deep-learning market test of dataset value. International Journal of Forecasting 35(2), 776–782 (2019)
18. McCabe, A., Trevathan, J.: Artificial intelligence in sports prediction. In: Proceedings of the Fifth International Conference on Information Technology: New Generations. pp. 1194–1197. ITNG '08, IEEE Computer Society, USA (2008)
19. Onori, M., Micarelli, A., Sansonetti, G.: A comparative analysis of personality-based music recommender systems. In: CEUR Workshop Proceedings. vol. 1680, pp. 55–59. CEUR-WS.org, Aachen, Germany (2016)
20. Rudrapal, D., Boro, S., Srivastava, J., Singh, S.: A deep learning approach to predict football match result. In: Computational Intelligence in Data Mining. pp. 93–99. Springer Singapore, Singapore (2020)
21. Sansonetti, G.: Point of interest recommendation based on social and linked open data. Personal and Ubiquitous Computing 23(2), 199–214 (Apr 2019)
22. Sansonetti, G., Gasparetti, F., D'Aniello, G., Micarelli, A.: Unreliable users detection in social media: Deep learning techniques for automatic detection. IEEE Access 8, 213154–213167 (2020)
23. Sansonetti, G., Gasparetti, F., Micarelli, A., Cena, F., Gena, C.: Enhancing cultural recommendations through social and linked open data. User Modeling and User-Adapted Interaction 29(1), 121–159 (March 2019)
24. Shearer, C.: The crisp-dm model: the new blueprint for data mining. Journal of data warehousing 5(4), 13–22 (2000)
25. Sinha, S., Dyer, C., Gimpel, K., Smith, N.A.: Predicting the NFL using twitter. In: Proceedings of the 2nd Workshop on Machine Learning and Data Mining for Sports Analytics co-located with ECML PKDD 2013. pp. 28–38 (2013)
26. Stefani, R.T.: Football and basketball predictions using least squares. IEEE Transactions on Systems, Man, and Cybernetics 7(2), 117–21 (1977)
27. Stübinger, J., Mangold, B., Knoll, J.: Machine learning in football betting: Prediction of match results based on player characteristics. App. Sciences 10(1) (2020)
28. Tiwari, E., Sardar, P., Jain, S.: Football match result prediction using neural networks and deep learning. In: Proceedings of ICRITO 2020. pp. 229–231 (2020)
29. Vaccaro, L., Sansonetti, G., Micarelli, A.: An empirical review of automated machine learning. Computers 10(1) (2021)