Advances and Applications in Statistics

http://www.pphmj.com
http://dx.doi.org/10.17654/0972361722080
Volume 82, 2022, Pages 53-64 **P-ISSN: 0972-3617**

USE OF THE FIRST AND SECOND HALVES RESULTS TO CLASSIFY THE FINAL OUTCOME OF ENGLISH PREMIER LEAGUE MATCHES

Tomilayo P. Iyiola, Hilary I. Okagbue and Oluwole A. Odetunmibi

Department of Mathematics
Covenant University
Ota, Nigeria

Abstract

English premier league (EPL) is one of the top leagues in Europe and any analysis of data generated from the league is highly sought after by fans, betters, coach, managers and scouters. The paper applied four machine learning models in classifying the outcome of five seasons using the results of the first and the second halves. Each half and the final outcome were made up of just three data points, namely, home win (HW), draw (DR) and away win (AW). Home win is the most frequent followed by AW and DR in descending order. There is no significant relationship between the results of the two halves. On the other hand, there are significant relationships between the first

Received: August 8, 2022; Accepted: September 15, 2022
2020 Mathematics Subject Classification: 62P99.
Keywords and phrases: adaptive boosting, chi-square test, English premier league, football, gradient boosting, linear regression, machine learning, random forests.

How to cite this article: Tomilayo P. Iyiola, Hilary I. Okagbue and Oluwole A. Odetunmibi, Use of the first and second halves results to classify the final outcome of English premier league matches, Advances and Applications in Statistics 82 (2022), 53-64.
http://dx.doi.org/10.17654/0972361722080

Published Online: November 3, 2022

half and the outcome and also, between the second half and the outcome. Random forests (RF), gradient boosting (GB) and adaptive boosting (AB) yielded better results than the logistic regression (LR). Generally, the accuracy averaged over 90 percent with few misclassifications. Implementation of the research in a decision support system is highly recommended.

Introduction

Football is a very popular sport, and the English premier league (EPL) is one of the top five leagues in Europe, with a fan base across the globe. EPL's popularity has also encouraged enormous betting, television licenses, and other socioeconomic benefits. The probabilistic nature of the outcomes of football games in general and EPL, in particular, have made the art of predicting the results exciting. The prediction benefits betting companies and bookmakers, media, fans [1], scouts for sourcing new players, and football managers and coaches for studying opponents' weaknesses, tactics, and strategies for winning matches [2].

Three approaches exist for predicting football outcomes: statistics, Bayesian, and machine learning (ML) [3]. All the aspects could be in-game and pre-game features. The enormous data generated from matches could be in the form of player performance metrics, goals scored, match statistics, injuries, and other related data [4]. Statistical approach is the use of probability models and regression models. Examples of statistical approaches are the use of the Weibull model [5] or Poisson [6] or multinomial logit [7] models to model the number of goals scored by the home and away teams or the instances of home advantage or probability of winning a football match given the teams to score first. In the same vein, design of experiment has been utilized too [8]. The Bayesian approach uses the Bayes theorem to compute probabilities that measure the degree of occurrence of events. This approach has been applied in predicting the outcome of the five (5) major European Football Leagues [9]. Restricting the analysis to predict a season of EPL [10] and a particular football club [11] yielded better accuracy.

ML approach is the most popular because it allows the development of predictive capability with unprecedented accuracy [12]. The vast amount of data predicts using complex models and algorithms to unravel hidden patterns that will assist in predicting different aspects of the game [13]. Aside from the data, the vast amount of monetary rewards available in betting continues to inspire researchers to apply ML models in developing predictive models in this context [14]. Moreover, ML is one of the fastest growing research methodologies which has been applied in different areas of research, such as health informatics [15-17] and scientometrics or research evaluation [18, 19], with excellent results. However, the issue of interpretability and ethics remains the primary concern in the adoptability or adaptability of the method. Apart from the use of ML in predicting football outcomes, it has also been used in different aspects of football, such as the prediction of the final team position in a football league [20] and; the prediction of the actual market values of football players [21] and prediction of injuries [22].

A recent review [23] showed that there had been a shift in the use of artificial neural networks (ANN) to other ML models in predicting the outcome of football matches. The application of various ML models gives rise to different performance metrics, as reported by different authors. Examples of ML models that have been used yielded different accuracies. Gradient boosting gave accuracy of 89.6% in [24] and *k*-nearest neighbor (83.95%) [25].

This paper used the ML approach is predicting the outcome of football matches of five seasons of EPL using the results of each half of the respective individual matches. This approach has not been featured in scientific literature as those available prefer to use many variables as independent variables in the prediction. Their rationale is to increase prediction accuracy and flexibility in handling a vast amount of data obtained mainly via data scrapping methods [26]. However, using only two independent variables, in this case, reduces the inherent difficulty in ML interpretability while increasing accuracy and reducing parsimony. Again,

this paper adopts the use of only three data values (win, draw, lose) as prediction is more accurate than attempting to predict the actual football scores [27].

Materials and Methods

The data was obtained from various online sources. The data is the outcome (home win, draw, away win) of five seasons of EPL, from 2016/2017 to 2020/2021 season. Each season contains 380 matches, and the results were grouped into three: first half, second half and final result. The actual scores were converted into the three groups: HW, DR and AW, representing home win, draw and away win, respectively.

There are only two independent variables (first half and second half) results and the final result, the target or the dependent variable.

Three hypotheses were crafted to establish the relationship between the pairs of first half, second half and the outcome. The chi-square test of independence was chosen because the data is nominal.

Four ML models which are often used in classification are used. They are random forests (RF), logistic regression (LR), gradient boosting (GB), and adaptive boosting (AB). Before the data was passed through the models, cross-validation was done by dividing the data into training data (70%) and testing data (30%).

Four performance metrics were used to assess the accuracy of the classification. The evaluation metrics are area under curve (AUC), classification accuracy (CA), F1, precision, and recall. Confusion matrix was used to determine the instances of correct and incorrect classifications.

Result

The frequency analysis of the first half, second half and outcome of the five EPL seasons is presented in Table 1.

Table 1. Frequency of the first and second halves and outcome

Season	First half			Second half			Outcome		
	HW	DR	AW	HW	DR	AW	HW	DR	AW
2016/2017	137	151	92	158	124	98	188	83	109
2017/2018	125	160	95	156	132	92	173	99	108
2018/2019	126	148	106	152	116	112	181	71	128
2019/2020	138	139	103	146	117	117	172	92	116
2020/2021	121	153	106	115	140	125	144	83	153

It can be seen from Table 1 that the outcomes of the first half are not necessarily the same as the second half. For instance, in the 2018/2019 season, 126 matches were HW in the first half. In the second half, it increased to 152 and 182 as the final outcome.

First half. There were more draws in the first half than home win and away win in each of the five seasons. In this case, the 2017/2018 season had more draws in the first half and the 2018/2019 season had the least draws. Noticeable, away win had the lowest number in the first half for each of the five seasons considered.

Second half. There were more home wins in the second half for the first four seasons except the 2020/2021 season, which had more draws. Also, there was a growing decline in the number of home wins in the second half of the five seasons.

Outcome. The final outcome had more home wins for the first four seasons and the last season had away wins as its highest number. The least ranked outcome is the number of draws across all five seasons.

Generally, it could be seen that the 2016/2017 season had more draws in the first half, more home wins in the second half, and more home wins in the final outcome. Similarly, this trend is the same for the 2017/2018, 2018/2019 and 2019/2020 season(s). For the 2020/2021 season, there were more draws in the first half, more draws in the second half and finally more away wins in the final outcome, which is a deviation from the trend for the final outcome for the first four seasons.

To confirm whether there are relationships between each pair of the trio (first half, second half, and outcome), three hypotheses will be defined and answered using the chi-square test of hypothesis.

The first hypothesis. The null hypothesis is that there is no significant relationship between the outcomes of the English premier league's first half and second half $(p > 0.05)$. The alternate hypothesis is that there is a significant relationship between the two halves $(p < 0.05)$.

The second hypothesis. The null hypothesis is that there is no significant relationship between the outcomes of the first half and the final outcome of the English premier league $(p > 0.05)$. The alternate hypothesis is that there is a significant relationship between the two $(p < 0.05)$.

The third hypothesis. The null hypothesis is that there is no significant relationship between the outcomes of the second half and the final outcome of the English premier league $(p > 0.05)$. The alternate hypothesis is that there is a significant relationship between the two $(p < 0.05)$.

The result of the three hypotheses is summarized in Table 2.

Table 2. Chi-square values of the pairs of the first and second halves and outcome

Season	First and second halves		First half and outcome		Second half and outcome	
2016/2017	9.293	0.054	**174.643**	**<0.0001**	**222.638**	**<0.0001**
2017/2018	6.732	0.151	**168.192**	**<0.0001**	**243.993**	**<0.0001**
2018/2019	**13.022**	**0.011**	**166.749**	**<0.0001**	**232.393**	**<0.0001**
2019/2020	4.736	0.315	**143.816**	**<0.0001**	**217.561**	**<0.0001**
2020/2021	6.015	0.198	**169.891**	**<0.0001**	**218.265**	**<0.0001**

Table 2 shows the result of the chi-square test of the independence of the pairs.

For the first hypothesis, the chi-square test confirmed that there is no significant relationship between the first and second halves for the

2016/2017, 2017/2018, 2019/2020, and 2020/2021 season(s), while there is a significant relationship between these pairs in the 2018/2019 season.

The test showed a significant relationship between the first half and the outcome across the five seasons. In this case, the alternative hypothesis for the second hypothesis is true for all seasons and the null hypothesis is false.

Finally, the values in Table 2 confirm that the alternate hypothesis for the third hypothesis is true. Hence, there is a significant relationship between the second half and the outcome across all five seasons.

The final analysis applies four ML models to classify the outcome by the first and second halves. The following codes are used as presented in Table 3.

Table 3. Codes for the variables

First half	Second half	Outcome
Home win (1)	Home win (1)	Home win (1)
Draw (2)	Draw (2)	Draw (2)
Away win (3)	Away win (3)	Away win (3)

The same codes were used throughout for the three variables. HW is coded as one (1), draw is coded as two (2), and away win is coded as three (3).

Data sampler (which is an algorithm for splitting data) was used to divide the data into training (70%) and test (30%). Cross-validation is done to guide against overfitting. Also, the parameter settings of the Orange software used in the analysis remain unchanged. The application of cross-validation reduced the data from 380 matches to 266 matches. Subsequently, testing was done on test data.

The four ML models (LR, GB, RF and AB) were able to classify the outcome using the outcomes of the first half and second half. Also, performance metrics were used to assess the precision of the machine learning models for the classification of outcomes using the two independent variables. The results were obtained for the five EPL seasons, and the

confusion matrix was used to output the instances of correct and incorrect classifications, as shown in Table 4.

Table 4. Performance metrics of the ML models for the five EPL seasons

2016/17	AUC	CA	F1	Precision	Recall	Correct	Incorrect
RF	0.993	0.944	0.946	0.955	0.9444	251	15
LR	0.992	0.932	0.933	0.935	0.932	248	18
GB	0.992	0.944	0.946	0.955	0.944	251	15
AB	0.985	0.944	0.946	0.955	0.944	251	15
2017/18	AUC	CA	F1	Precision	Recall	Correct	Incorrect
RF	0.985	0.917	0.920	0.932	0.917	244	22
LR	0.983	0.898	0.900	0.905	0.898	239	27
GB	0.983	0.917	0.920	0.932	0.917	244	22
AB	0.946	0.917	0.920	0.932	0.917	244	22
2018/19	AUC	CA	F1	Precision	Recall	Correct	Incorrect
RF	0.988	0.925	0.929	0.944	0.925	246	20
LR	0.986	0.917	0.917	0.918	0.917	244	22
GB	0.988	0.925	0.929	0.944	0.925	246	20
AB	0.959	0.925	0.929	0.944	0.925	246	20
2019/20	AUC	CA	F1	Precision	Recall	Correct	Incorrect
RF	0.982	0.906	0.910	0.928	0.906	241	25
LR	0.982	0.895	0.895	0.899	0.895	238	28
GB	0.982	0.906	0.910	0.928	0.906	241	25
AB	0.970	0.906	0.910	0.928	0.906	241	25
2020/21	AUC	CA	F1	Precision	Recall	Correct	Incorrect
RF	0.991	0.929	0.932	0.947	0.929	247	19
LR	0.991	0.891	0.884	0.902	0.891	237	29
GB	0.991	0.929	0.932	0.947	0.929	247	19
AB	0.991	0.929	0.932	0.947	0.929	247	19

Generally, the results for RF, GB and AB are nearly the same for the five seasons. The three models generally performed better than the LR, with fewer misclassifications than the latter.

Conclusion

The paper applied four machine learning models in classifying the outcome of five EPL seasons using the individual results of the first and second halves. Statistically, in all the seasons except one, the home wins are more than the away wins, which are equally higher than the draws. The performance metrics for the classification average over 90 percent and the chi-square test of independence showed that the first half results are statistically different from the second half. Numerous users of football data will benefit immensely from this research.

Acknowledgment

The efforts of the anonymous reviewers are greatly appreciated. Covenant University sponsored the research through CUCRID.

References

[1] P. Xenopoulos and C. Silva, Graph neural networks to predict sports outcomes, Proceedings, IEEE International Conference on Big Data, Big Data, 2021, pp. 1757-1763. https://doi.org/10.1109/BigData52589.2021.9671833.

[2] E. Filiz, Evaluation of match results of five successful football clubs with ensemble learning algorithms, Research Quarterly for Exercise and Sport (2022). https://doi.org/10.1080/02701367.2022.2053647.

[3] A. Ranjan, V. Kumar, D. Malhotra, R. Jain and P. Nagrath, Predicting the result of English premier league matches, Lecture Notes in Networks and Systems 203 (2021), 435-446. 10.1007/978-981-16-0733-2_30.

[4] S. Jain, E. Tiwari and P. Sardar, Soccer result prediction using deep learning and neural networks, Lecture Notes on Data Engineering and Communications Technologies 57 (2021), 697-707. https://doi.org/10.1007/978-981-15-9509-7_57.

[5] G. Boshnakov, T. Kharrat and I. G. McHale, A bivariate Weibull count model for forecasting association football scores, International Journal of Forecasting 33(2) (2017), 458-466. https://doi.org/10.1016/j.ijforecast.2016.11.006.

[6] L. S. Benz and M. J. Lopez, Estimating the change in soccer's home advantage during the Covid-19 pandemic using bivariate Poisson regression, AStA Advances in Statistical Analysis (2021). https://doi.org/10.1007/s10182-021-00413-9.

[7] T. Liu, A. García-de-Alcaraz, H. Wang, P. Hu and Q. Chen, Impact of scoring first on match outcome in the Chinese Football Super League, Frontiers in Psychology 12 (2021), 662-708. https://doi.org/10.3389/fpsyg.2021.662708.

[8] F. Liu, Y. Shi and L. Najjar, Application of design of experiment method for sports results prediction, Procedia Computer Science 122 (2017), 720-726. https://doi.org/10.1016/j.procs.2017.11.429.

[9] N. Razali, A. Mustapha, N. Mustapha and F. M. Clemente, A Bayesian approach for major European football league match prediction, International Journal of Nonlinear Analysis and Applications 12 (2021), 971-980. https://doi.org/10.22075/IJNAA.2021.5544.

[10] A. C. Constantinou, N. E. Fenton and M. Neil, Profiting from an inefficient association football gambling market: prediction, risk and uncertainty using Bayesian networks, Knowledge-Based Systems 50 (2013), 60-86. https://doi.org/10.1016/j.knosys.2013.05.008.

[11] A. Joseph, N. E. Fenton and M. Neil, Predicting football results using Bayesian nets and other machine learning techniques, Knowledge-Based Systems 19(7) (2006), 544-553. https://doi.org/10.1016/j.knosys.2006.04.011.

[12] R. Baboota and H. Kaur, Predictive analysis and modelling football results using machine learning approach for English Premier League, International Journal of Forecasting 35(2) (2019), 741-755. https://doi.org/10.1016/j.ijforecast.2018.01.003.

[13] S. K. Andrews, K. L. Narayanan, K. Balasubadra and M. S. Josephine, Analysis on sports data match result prediction using machine learning libraries, Journal of Physics: Conference Series 1964(4) (2021), 042-085. https://doi.org/10.1088/1742-6596/1964/4/042085.

[14] R. P. Bunker and F. Thabtah, A machine learning framework for sport result prediction, Applied Computing and Informatics 15(1) (2019), 27-33. https://doi.org/10.1016/j.aci.2017.09.005.

[15] H. I. Okagbue, P. E. Oguntunde, P. I. Adamu and O. A. Adejumo, Unique clusters of patterns of breast cancer survivorship, Health and Technology 12(2) (2022), 365-384. https://doi.org/10.1007/s12553-021-00637-4.

[16] H. I. Okagbue, P. I. Adamu, P. E. Oguntunde, E. C. M. Obasi and O. A. Odetunmibi, Machine learning prediction of breast cancer survival using age, sex, length of stay, mode of diagnosis and location of cancer, Health and Technology 11(4) (2021), 887-893. https://doi.org/10.1007/s12553-021-00572-4.

[17] H. I. Okagbue, P. E. Oguntunde, E. C. M. Obasi, P. I. Adamu and A. A. Opanuga, Diagnosing malaria from some symptoms: a machine learning approach and public health implications, Health and Technology 11 (2021), 23-37. https://doi.org/10.1007/s12553-020-00488-5.

[18] H. I. Okagbue, C. A. Nzeadibe and J. A. Teixeira da Silva, Predicting access mode of multidisciplinary and library and information sciences journals using machine learning, COLLNET Journal of Scientometrics and Information Management 16(1) (2022), 117-124. https://doi.org/10.1080/09737766.2021.2009745.

[19] H. I. Okagbue, E. M. Akhmetshin and J. A. Teixeira da Silva, Distinct clusters of CiteScore and percentiles in top 1000 journals in Scopus, COLLNET Journal of Scientometrics and Information Management 15(1) (2021), 133-143. https://doi.org/10.1080/09737766.2021.1934604.

[20] M. Kleina, M. N. D. Santos, T. N. D. Santos, M. A. M. Marques and W. D. A. Silva, Artificial intelligence techniques applied to predict teams position of the Brazilian football championship, Journal of Physical Education 32(1) (2022), e3254. https://doi.org/10.4025/jphyseduc.v32i1.3254.

[21] V. S. Arrul, P. Subramanian and R. Mafas, Predicting the football players' market value using neural network model: a data-driven approach, IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics, ICDCECE, 2022. https://doi.org/10.1109/ICDCECE53908.2022.9792681.

[22] A. Majumdar, R. Bakirov, D. Hodges, S. Scott and T. Rees, Machine learning for understanding and predicting injuries in football, Sports Medicine-Open 8(1) (2022), Article 73. https://doi.org/10.1186/s40798-022-00465-4.

[23] R. Bunker and T. Susnjak, The application of machine learning techniques for predicting match results in team sport: a review, Journal of Artificial Intelligence Research 73 (2022), 1285-1322. https://doi.org/10.1613/jair.1.13509.

[24] Y. Geurkink, J. Boone, S. Verstockt and J. G. Bourgois, Machine learning-based identification of the strongest predictive variables of winning and losing in Belgian professional soccer, Appl. Sci. 11(5) (2021), 2378. https://doi.org/10.3390/app11052378.

[25] U. Haruna, J. Z. Maitama, M. Mohammed and R. G. Raj, Predicting the outcomes of football matches using machine learning approach, Communications in Computer and Information Science 1547 (2022), 92-104. https://doi.org/10.1007/978-3-030-95630-1_7.

[26] L. Carloni, A. De Angelis, G. Sansonetti and A. Micarelli, A machine learning approach to football match result prediction, Communications in Computer and Information Science 1420 (2021), 473-480. https://doi.org/10.1007/978-3-030-78642-7_63.

[27] R. Nestoruk and G. Słowiński, Prediction of football games results, CEUR Workshop Proceedings 2951 (2021), 156-165.