Advances and Applications in Statistics

http://www.pphmj.com
http://dx.doi.org/10.17654/0972361723026
Volume 87, Number 1, 2023, Pages 13-27 **P-ISSN: 0972-3617**

THE EFFECTS OF DECOMPOSITION OF THE GOALS SCORED IN CLASSIFYING THE OUTCOMES OF FIVE ENGLISH PREMIER LEAGUE SEASONS USING MACHINE LEARNING MODELS

Tomilayo P. Iyiola, Hilary I. Okagbue*, Adedayo F. Adedotun and Toluwalase J. Akingbade

Department of Mathematics
Covenant University
Ota, Nigeria
e-mail: hilary.okagbue@covenantuniversity.edu.ng

Abstract

The English Premier League (EPL) is one of the best football championships in the world and thus, data generated from it is highly sought after by users of football data. One of the uses of the data is in the prediction of outcome of the league matches. This paper applies four machine learning (ML) models in classifying the outcome (home

Received: September 3, 2022; Revised: October 19, 2022; Accepted: December 5, 2022
2020 Mathematics Subject Classification: 62R07, 68T01, 68U01.
Keywords and phrases: algorithm, classification, cross-validation, English Premier League, feature selection, machine learning, statistics.

*Corresponding author

How to cite this article: Tomilayo P. Iyiola, Hilary I. Okagbue, Adedayo F. Adedotun and Toluwalase J. Akingbade, The effects of decomposition of the goals scored in classifying the outcomes of five English Premier League seasons using machine learning models, Advances and Applications in Statistics 87(1) (2023), 13-27. http://dx.doi.org/10.17654/0972361723026

Published Online: April 15, 2023

win, draw, and away win) of five consecutive seasons of EPL using only six independent variables. Information Gain Ratio (IGR) and ReliefF were the feature selection algorithms that reduced the independent variables from 16 to 6. Spearman rank correlation gave a high significant positive correlation between the ranks of the 2 feature selection algorithms. The Kruskal-Wallis H test indicated that there is a significant difference in the dependent variable between the different Seasons (Chi-square = 15.36, Degrees of freedom = 4, P = 0.004). Adaptive boosting (AB), gradient boosting (GB), logistic regression (LR) and random forests (RF) were used in the classification of the outcome using the six independent variables and the performance metrics showed a perfect classification in almost all the models. This paper concluded that the knowledge of the number of goals scored by the home and away teams, and the number of Goals scored by home and away teams in the first half and second half are all that is needed to correctly classify the outcomes of the English Premier League (EPL). Secondly, the knowledge of the own goals and goals scored by penalty, and yellow and red cards conceded by the home or away teams is not necessarily needed in the determination or prediction of the outcomes of the EPL.

Introduction

Nowadays predictive models, classification, or prediction of results in any sport have become popular in artificial intelligence community, and particularly, English Premier League (EPL) in football gains much attention in the past few years. Predictions of the outcome of football are not only the aim of various researches in this emerging research domain. Researchers have also considered the prediction of market values of players in different football clubs and leagues [1] or the prediction of instances or possibilities of injuries [2] or spotting or identification of new talents in a football game [3] or tactical decision prediction to outperform the opponents [4]. The stochastic nature of the data generated from football is also a contributing factor in the continuous flow of research activities in that area [5].

There are three main approaches to predicting the outcomes of football matches: statistical approaches, machine learning approaches, and the Bayesian approaches. There appears to be a general shift from the use of statistical methods to Bayesian and machine learning, largely because the latter can handle a vast amount of multidimensional data without loss of information [6]. The choice of ranking score is responsible for many of the instances of loss of information in the use of ML in the classification of outcomes [7]. Statistical methods are further divided into two, regression and probability density fitting. Some examples of regression are the use of the bivariate Poisson regression in predicting home advantage [8], and the use of a multinomial logit model to model the impact of scoring first on the outcome of matches in the Chinese Football Super League [9]. The Bayesian approach utilizes the concept of conditional probability to compute the odds associated with prediction of football outcomes [10]. The odds provide the needed insight on the degree of effect of different variables in classification or prediction. Bayesian methods assume that the presence of prior and post distribution of outcomes and as such, forecast can be made before the start of a given football league as seen in the forecasting of Association Football match outcomes [11].

In the process of mining data from football, data are first obtained mostly through different web scrapping methods [12]. The availability of a vast amount of football data and robust data analytic tools have led to an unprecedented increase in research interest and opportunities derivable from a deeper understanding of football data [13]. Even data obtained using global positioning systems of players' on-field attributes or features can be used [14]. Several features which often characterize the data are investigated and the mining learning algorithms have the computational capacity to reveal the prominent or leading features that ultimately lead to accurate classification of the odds of the football team to win, lose or draw from any mined data [15]. Available research attests that prediction of outcome (win or draw or lose) using ML models gives more accurate results than predicting the actual scores [16]. Knowledge of the best features that predict the outcomes are

discovered from the various ML classifiers. The hidden patterns that could not be easily observed are laid bare to the astonishment of researchers and benefactors such as betting organizations, football coaches and managers, media, and scouts [17]. The features also known as variables could be data relating to in-game statistics like goals scored, yellow cards, penalties, number of throw-ins, offside, and so on [18] or off-game statistics such as weather, jersey colors, club values, and so on.

Most of the early works are the simple classification of football outcomes using ML models without cross-validation [19]. Nowadays, the data are divided into training and test, and testing is done using the test data (cross-validation) to improve the reliability of the classification/prediction and to guard against unrealistic predictions [20]. Apart from regression and classification, clustering algorithms have been deployed as seen in the clustering of FIFA World cup results [21] and Fuzzy C-mean clustering in differentiating between home and away teams' corners [22]. Advanced deep learning ML models and evolutionary computational algorithms have been used too. An example is the use of Particle Swarm Optimization (PSO) in the prediction of the value of football players in transfer markets [23]. Other football competitions have been considered such as the UEFA Champions League [24], the Brazilian football championship [25], and the Belgian soccer league [26]. Recently, the combination of deep learning methods and ensemble models appears to yield better results, especially in the presence of numerous attributes or features such as goals conceded, shooting precision, half-time results and scoring first [27]. Multilayer Perceptron (MLP) and Dense Neural Network (DNN) proposed by [28] are the deep learning models applied in the classification of the outcomes of the English Premier League. Despite the advantages of the use of ML methods, issues with the detailed workings of the algorithms, ethics, interpretability, and more recently, the absence of benchmarking platforms make comparisons, difficult [29]. That is, the independent variables considered in classifying outcomes differ, and hence, a consensus on the results may not be possible. In addition, interpretation is limited to only the data analyzed and no general comments

about the global outcome could be obtained. Nevertheless, the adoption of the ML as methodology has helped deepen understanding of different aspects of research as seen in [30-35].

The aim of this paper is to classify the outcome (home win, away win and draw) of five English Premier League seasons using sixteen independent variables. The feature selection algorithm reduced the variables to six and four ML models were used. This research is a clear departure from the adoption of the use of home advantage and goals in predicting football outcomes. The adoption of feature selection methods will help to reduce the variables and improve the predictive power of the algorithms which are yet to be reported in this context.

Methodology

Data

The data for the five seasons were obtained from the following websites: livescores.com, soccerway.com, allfootball.com, and sofascores.com. Also, the data was verified by the official website of the English Premier League (www.premierleague.com), which retrieves data from OPTA, whose data reliability ranges from 0.92 to 0.94.

Variables in the data

The dependent (target) variable is the outcome (home win, draw and away win). The data are nominal. The sixteen independent variables are listed in Table 1.

Table 1. The 16 independent variables

S/N	Second half	Acronym
1	Number of goals scored by the home team	NOG(H)
2	Number of goals scored by away team	NOG(A)
3	Goals scored by home team in the first half	GIFH(H)
4	Goals scored by away team in the first half	GIFH(A)
5	Goals scored by home team in the second half	GISH(H)
6	Goals scored by away team in the second half	GISH(A)
7	Own goal by home team	OG(H)
8	Own goal by away team	OG(A)
9	Number of goals scored by substitutes of home team	NGSS(H)
10	Number of goals scored by substitutes of away team	NGSS(A)
11	Goals by penalty for home team	GBP(H)
12	Goals by penalty for away team	GBP(A)
13	Yellow card conceded by home Team	YELLOW(H)
14	Yellow card conceded by away Team	YELLOW(A)
15	Red card conceded by home Team	RED(H)
16	Red card conceded by away Team	RED(A)

The variables are mostly from the decomposition of the total goals scored that births the respective outcomes.

Software used in the data analysis

Microsoft EXCEL, SPSS, R and Knowledge Extraction Based on Evolutionary Learning (KEEL) were used in this research for data extraction, frequency analysis, median test, feature selection and classification.

Statistical analysis

Frequency analysis was restricted only to the outcome for the five seasons. Kruskal Wallis test was used in the test of equality of the median outcome. Spearman rank correlation was used to test agreement between the two feature selection algorithms.

Machine learning

Information Gain Ratio (IGR) and ReliefF are Feature selection algorithms while data sampler (which is an algorithm for splitting data) was used to divide the data into training (70%) and test (30%). Adaptive boosting (AB), gradient boosting (GB), logistic regression (LR) and random forests (RF) are the adopted classification models. Area under curve (AUC), classification accuracy (CA), F1, precision and recall were the evaluation metrics.

Result

The frequency distribution of the outcome

The frequency distribution of the outcomes for the five EPL seasons is presented in Table 2. Home win is the modal observation except in the 2020/2021 season. The effect of the COVID-19 could be culpable and it appears that playing under closed door erodes the undue effect of home advantage. Generally, the home win is often more than the away win, which in turn is more than the draws.

Table 2. The frequency distribution of the outcomes for the five EPL seasons

	Season	Home win	Draw	Away win
X1	2016/2017	188	83	109
X2	2017/2018	172	99	109
X3	2018/2019	181	71	128
X4	2019/2020	172	92	116
X5	2020/2021	144	83	153

Test of equality of median outcome

Kruskal-wallis test was used to test the equality of median of the outcomes (HW, DR, AW) of the five (5) seasons. The null hypothesis, in this case, is that the median of the outcomes is equal (p-value > 0.05), while the alternative is that the outcomes have a median that is significantly different

from each other. The post-hoc analysis will be used when the null hypothesis is rejected. The outcome of each match was categorized into three variables (1 = HW, 2 = DR, 3 = AW). The Kruskal-Wallis H test indicated that there is a significant difference in the dependent variable between the different Seasons (Chi-square = 15.36, Degrees of freedom = 4, $P = 0.004$) with a mean rank score of 903.22 for the 2016/2017 season, 930.27 for 2017/2018 season, 941.13 for 2018/2019 season, 939.88 for 2019/2020 season and 1038 for 2020/2021 season. The post-hoc analysis was conducted using the Dunn's test and a Bonferroni corrected alpha of 0.005. The results presented as a *Z*-statistic and *p*-value indicated that the median ranks of SOME pairs of the seasons are significantly different as shown in Table 3.

Table 3. The result of the Kruskal Wallis Test

Pair	Z	*p*-value
X1-X2	0.7317	0.4644
X1-X3	1.0254	0.3052
X1-X4	0.9915	0.3214
X1-X5	3.6454	0.000267*
X2-X3	0.2937	0.769
X2-X4	0.2598	0.759
X2-X5	2.9137	0.003527*
X3-X4	0.03384	0.973
X3-X5	2.62	0.008793*
X4-X5	2.6539	0.007958*

*p-value < 0.05

The median of the pairs X1-X5 (2016/2017 and 2020/2021), X2-X5 (2017/2018 and 2020/2021), X3-X5 (2018/2019 and 2020/2021) and X4-X5 (2019/2020 and 2020/2021) are significantly different since their respective *p*-values are less than 0.05.

Feature selection

Feature selection is applied to rank the order of importance of the 16 independent variables that are to be used in classifying the outcomes of the

matches. It is similar to the stepwise regression where redundant variables are removed from the regression model. Information Gain Ratio (IGR) and ReliefF were feature selection methods used to reduce the number of variables based on some scores. Independent variables with high scores of IGR and ReliefF were chosen. The feature selection methods attempt to find the relevance of the independent variables with respect to the outcome which is the target or dependent variable. The ranks of the variables for the five seasons are presented in Tables 4 and 5.

Table 4. Ranking of independent variables (IGR)

Variable	X1	X2	X3	X4	X5
NOG(H)	1	2	1	1	1
NOG(A)	2	1	2	2	2
GIFH(H)	3	5	5	6	6
GIFH(A)	4	4	6	5	7
GISH(H)	5	6	3	4	3
GISH(A)	6	3	4	3	4
OG(H)	10	8	13	14	5
OG(A)	8	13	8	13	8
NGSS(H)	12	11	9	12	13
NGSS(A)	7	9	11	7	12
GBP(H)	9	12	12	8	10
GBP(A)	16	7	7	9	9
YELLOW(H)	14	14	14	15	16
YELLOW(A)	15	15	16	16	15
RED(H)	13	10	10	11	14
RED(A)	11	16	15	10	11

Table 5. Ranking of independent variables (ReliefF)

	X1	X2	X3	X4	X5
NOG(H)	1	1	1	1	3
NOG(A)	2	2	2	3	2
GIFH(H)	3	5	4	5	5
GIFH(A)	6	4	6	6	6
GISH(H)	5	3	3	4	1
GISH(A)	4	6	5	2	4
OG(H)	11	10	11	14	11
OG(A)	14	11	12	7	8
NGSS(H)	12	8	14	11	13
NGSS(A)	9	14	9	10	14
GBP(H)	7	7	8	9	7
GBP(A)	16	9	7	8	12
YELLOW(H)	15	16	13	16	16
YELLOW(A)	8	15	16	15	15
RED(H)	10	13	10	13	10
RED(A)	13	12	15	12	9

The two feature selection methods outputted different but similar statistics arranged in ranks. The measure of agreement is needed to ensure that a single rank (1 to 6) can be used in the classification. The remaining nine variables are discarded and were not used in the classification. The agreement between the two feature selection methods could not be computed using Cohen's Kappa or other inter-rater agreement methods because the raters, in this case, are not categorical. Spearman rank correlation (SRC) is used instead to determine if there is a significant association or relationship between the ranks of the two feature selection methods.

The null hypothesis is that no correlation exists $(p > 0.05)$ while the alternate hypothesis is that a significant association exists $(p < 0.05)$. The result is presented in Table 6 which showed that the two feature selection methods yielded almost the same values.

Table 6. Test of agreement between IGR and ReliefF

Season	SRC	*p*-value
2016/2017	0.829	< 0.0001
2017/2018	0.824	< 0.0001
2018/2019	0.900	< 0.0001
2019/2020	0.909	< 0.0001
2020/2021	0.871	< 0.0001

Classification

The six variables recommended via feature selection are; NOG(H), NOG(A), GIFH(H), GIFH(A), GISH(H) and GISH(A). The remaining ten variables are redundant and hence, contribute minimal to the classification.

Cross-validation was done by dividing the data using the data sampler, into training (70%) and testing (30%) and evaluation was done on then test data. This reduced the population from 380 to 266 per season. Adaptive boosting (AB), gradient boosting (GB), logistic regression (LR) and random forests (RF) were used in the classification the outcome using the six independent variables and the performance metrics that measured the performance of the models are presented in Table 7.

Table 7. Evaluation metrics of the four classification models

Season	Model	AUC	CA	F1	Precision	Recall
X1	RF	1	0.992	0.992	0.993	0.992
	LR	1	1	1	1	1
	GB	1	1	1	1	1
	AB	1	1	1	1	1
X2	RF	1	1	1	1	1
	LR	1	1	1	1	1
	GB	1	1	1	1	1
	AB	1	1	1	1	1
X3	RF	1	0.996	0.996	0.996	0.996
	LR	1	1	1	1	1
	GB	1	1	1	1	1
	AB	1	1	1	1	1

X4	RF	1	1	1	1	1
	LR	1	1	1	1	1
	GB	1	1	1	1	1
	AB	1	1	1	1	1
X5	RF	1	0.992	0.992	0.993	0.992
	LR	1	1	1	1	1
	GB	1	1	1	1	1
	AB	1	1	1	1	1

In 2017/2018 and 2019/2020, the four models gave a perfect classification of the outcomes with zero misclassification.

Conclusion

This paper has succeeded in decomposing the goal variables prominently used in classifying the outcomes of football matches into more variables that perfectly classify the outcomes of the English Premier League (EPL) results of five seasons. Feature selection algorithms outputted fewer independent variables that predicted the outcome. ML models were efficient in correctly classifying the outcome using the fewer variables recommended via feature selection. In summary, the paper makes the following conclusions.

(1) The knowledge of the number of goals scored by the home and away teams, and the number of Goals scored by home and away teams in the first half and second half are all that is needed to correctly classify the outcomes of the English Premier League (EPL).

(2) The knowledge of the own goals and goals scored by penalty, and yellow and red cards conceded by the home or away teams is not necessarily needed in the determination or prediction of the EPL.

Acknowledgements

The research was fully sponsored by Covenant University Centre for Research, Innovation and Discovery (CUCRID), Covenant University, Ota, Nigeria.

The authors thank the anonymous referees for their valuable suggestions and comments which led to the improvement of the presentation of the paper.

References

[1] V. S. Arrul, P. Subramanian and R. Mafas, Predicting the football players' market value using neural network model: a data-driven approach, ICDCECE. 2022. https://doi.org/10.1109/ICDCECE53908.2022.9792681.

[2] A. Majumdar, R. Bakirov, D. Hodges, S. Scott and T. Rees, Machine learning for understanding and predicting injuries in football, Sports Med. Open 8(1) (2022), Art. 79.

[3] S. Jain, E. Tiwari and P. Sardar, Soccer result prediction using deep learning and neural networks, Lect. Notes Data Engine. Commun. Technol. 57 (2021), 697-707.

[4] R. Beal, T. J. Norman and S. D. Ramchurn, Artificial intelligence for team sports: A survey. Knowl. Engine. Review (2019), e28. https://doi.org/10.1017/S0269888919000225.

[5] U. Haruna, J. Z. Maitama, M. Mohammed and R. G. Raj, Predicting the outcomes of football matches using machine learning approach, Commun. Comp. Info. Sci. 1547 (2022), 92-104.

[6] S. K. Andrews, K. L. Narayanan, K. Balasubadra and M. S. Josephine, Analysis on sports data match result prediction using machine learning libraries, J. Physics: Conf. Series, 1964(4) (2021), Art. 042085.

[7] E. Wheatcroft, Evaluating probabilistic forecasts of football matches: the case against the ranked probability score, J. Quant. Analy. Sports 17(4) (2021), 273-287.

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

[9] 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, Front. Psych. 12 (2021), Art. 662708.

[10] N. Razali, A. Mustapha, N. Mustapha and F. M. Clemente, A Bayesian approach for major European football league match prediction, Int. J. Nonlinear Anal. Appl. 12 (2021), 971-980.

[11] A. C. Constantinou, N. E. Fenton and M. Neil, Profiting from an inefficient association football gambling market: Prediction, risk and uncertainty using Bayesian networks, Knowl. Based Syst. 50 (2013), 60-86.

[12] L. Carloni, A. De Angelis, G. Sansonetti and A. Micarelli, A machine learning approach to football match result prediction, Commun. Comp. Info. Sci. 1420 (2021), 473-480.

[13] I. B. da Costa, L. B. Marinho and C. E. S. Pires, Forecasting football results and exploiting betting markets: the case of "both teams to score", Int. J. Forecasting 38(3) (2022), 895-909.

[14] A. Cortez, A. Trigo and N. Loureiro, Football match line-up prediction based on physiological variables: a machine learning approach, Computers 11(3) (2022), Art. 40.

[15] A. Ranjan, V. Kumar, D. Malhotra, R. Jain and P. Nagrath, Predicting the result of English premier league matches, Lect. Notes Netw. Syst. 203 (2021), 435-446.

[16] R. Nestoruk and G. Słowiński, Prediction of football games results, CEUR Workshop Proc., 2951, 2021, pp. 156-165.

[17] P. Xenopoulos and C. Silva, Graph neural networks to predict sports outcomes, Proc. IEEE Int. Conf. on Big Data, 2021, pp. 1757-1763.

[18] C. Pipatchatchawal and S. Phimoltares, Predicting football match result using fusion-based classification models, 18th Int. Joint Conf. on Comp. Sci. and Software Engine. Cybernet. Human Beings 2021, Art. 9493837.

[19] A. M. Sánchez Gálvez, R. Álvarez González, S. Sánchez Gálvez and M. Anzures García, Model to predict the result of a soccer match based on the number of goals scored by a single team, Computacion y Sistemas 26(1) (2022), 295-302.

[20] J. Fahey-Gilmour, J. Heasman, B. Rogalski, B. Dawson and P. Peeling, Can elite Australian football player's game performance be predicted? Int. J. Comp. Sci. Sport 20(1) (2021), 55-78.

[21] Y. Bai and X. Zhang, Prediction model of football world cup championship based on machine learning and mobile algorithm, Mobile Information Systems 2021 (2021), Art. 1875060.

[22] J. Yadav, Fuzzy C-mean clustering based soccer result analysis, Comm. Comp. Info. Sci. 1572 (2022), 3-14.

[23] I. Behravan and S. M. Razavi, A novel machine learning method for estimating football players' value in the transfer market, Soft Computing 25(3) (2021), 2499-2511.

[24] Y. W. Syaifudin and P. Puspitaningayu, Predicting winner of football match using analytical hierarchy process: an analysis based on previous matches data, In Int. Conf. on Data Analy. Bus. Industry, 2021, pp. 47-52.

[25] 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, J. Physical Educ. 32(1) (2022), e3254.

[26] 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), Art. 2378.

[27] E. Filiz, Evaluation of match results of five successful football clubs with ensemble learning algorithms, Res. Quart. Exer. Sport 2022. https://doi.org/10.1080/02701367.2022.2053647.

[28] M. Muszaidi, A. B. Mustapha, S. Ismail and N. Razali, Deep Learning Approach for football match classification of English Premier League (EPL) based on full-time results, Springer Proc. Physics 273 (2022), 339-350.

[29] R. Bunker and T. Susnjak, The application of machine learning techniques for predicting match results in team sport: a review, J. Artificial Intel. Res. 73 (2022), 1285-1322.

[30] 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 J. Scientometrics Info. Manag. 16(1) (2022), 117-124.

[31] H. I. Okagbue, E. M. Akhmetshin and J. A. Teixeira da Silva, Distinct clusters of cite score and percentiles in top 1000 journals in Scopus, COLLNET J. Scientometrics Info. Manag. 15(1) (2021), 133-143.

[32] H. I. Okagbue, P. E. Oguntunde, P. I. Adamu and O. A. Adejumo, Unique clusters of patterns of breast cancer survivorship, Health Technol. 12(2) (2022), 365-384.

[33] 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 Technol. 11(4) (2021), 887-893.

[34] 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 Technol. 11 (2021), 23-37.

[35] C. O. Iroham, S. Misra, O. C. Emebo and H. I. Okagbue, Predictive rental values model for low-income earners in slums: the case of Ijora, Nigeria. Int. J. Constr. Manag. (2021). https://doi.org/10.1080/15623599.2021.1975021.