Football League Ranking Prediction Using Machine Learning Regression Model

[1]Mohamed Rebbouj*, [1]Said Lotfi

[1]Sport Science Assessment and Physical Activity Didactic, Normal Higher School, Hassan II University, Casablanca, Morocco.
***:** Corresponding Author: **Mohamed Rebbouj.** E-mail: Mohamed.rebbouj@enscasa.ma

ABSTRACT

Background. Sport results predictive analysis is based on betting apps outcomes and has not yet been examined academically by concerned organizations in Morocco. **Objectives**. This study aims to predict a football national league ranking using a Machine Learning regression model with Elastic Net algorithm, where we determine the important features' weight on prediction. **Methods**. A dataset of historical scores of 8 standing teams since the 2009/2010 season was manually filled in and categorized into 9 columns: season, team, points, goal difference (+/-), matches played (M), matches won (W), matches drawn (D), matches lost (L), goals for (F) and goals against (A). Then preprocessed into Categorical data, categorical Hash, and numerical. **Results.** the machine learning analysis results in R^2 score = 0.999, NRMSE= 0.001 and Spearman correlation = 0.997. However, the predicted ranking was correct about 5 from 8 compared to the actual results till the 2021/2022 season. **Conclusion.** The Ranking prediction has been accurate by 75% in actual results compared to the regression analysis outcomes. This proves the quality of data needs to be more precise by including other parameters.

Keywords: Football Ranking, Machine Learning, Regression, Prediction.

INTRODUCTION

Football scores and results prediction has been the focus center of the tipster and betting market experts (1), and has become the more important center of interest for coaches, sports scientists, analysts, and performance specialists; to design the best practice, training, and competition tasks (2–4).
Therefore, researchers have begun applying mathematical formulas and statistics (5) to predict the outcomes, while machine learning and intelligent algorithms have become commonly used (6) and treating the football results as a classification problem with one class to predict (win, lose, or draw). But other researchers considered the problem a numerical value to predict in a regression model based on numerical analysis and values to predict specifically distance traveled (7) or the performance realized by athletes in jumping and throwing.
The sport results prediction problem lies in the data to gather, and the input features to consider impactful on the outcomes. Some researchers have focused on teams' historical data such as points of the team, goal difference, matches won, drawn, lost, goals for and goals against (8); while (9)

used more prediction criteria as a condition of the team in recent weeks and in the league, quality of the opponent in the last matches and week of match. More external features such as managerial change, fatigue, and club budget have been considered by (10) to predict the Dutch football competition, and a recent technique based on players rating scores related to their abilities on each team has resulted in a performing forecasting model (11–13) to predict the winner of the European champions league.

Regression analysis

Regression analysis (14) in machine learning is a type of supervised learning to determine the relationship between variables (features) with inputs and known outputs to predict (the team's scores in our case of study). the Elastic Net algorithm has gained significant attention in recent years due to its ability to handle high-dimensional datasets and address the limitations of traditional regression methods. And proven to be efficient in one-class classification machine learning analysis (15) and likewise used in human action recognition in real-time activity monitoring (16). The Elastic Net algorithm provides a powerful approach for regression analysis, combining the prediction ability for numerical values, and is usually used in sport performance studies(17)

MATERIALS AND METHODS

Data collection

A dataset has been manually transcribed into an Excel file from these two web sources:
www.footballdatabase.combhh
www.flashscore.com

The feature selection is based on the common data of teams: points, goal difference, total of matches played, matches won, matches drawn, matches lost, goals for, and goals against. These values have been recorded from the season 2009/2010 till 2021/2022 and we created a table with average scores; where we maintained only the 8 standing teams in all the seasons we collected as shown in the table below:

Table 1. **Average scores for the standing teams from the 2009/2010 season**

Club	P/12	+/-/12	M/12	W/12	D/12	L/12	F/12	A/12
WAC Casablanca	55.33	18.58	30.00	15.25	9.58	4.77	42.17	23.58
RCA	53.75	18.08	29.83	14.92	9.00	5.46	43.92	25.83
D.H. ElJadida	41.83	4.58	29.83	10.25	11.08	7.85	31.75	27.17
Hassania Agadir	39.08	-1.08	29.92	9.50	10.58	9.08	31.67	32.75
FAR Rabat	43.17	4.75	29.92	10.92	10.42	7.92	34.00	29.25
Moghreb Tétouan	41.50	1.83	29.92	10.33	10.50	8.38	31.58	29.75
FUS Rabat	44.08	4.83	29.83	11.17	10.58	7.46	30.92	26.08
Olympic Club de Safi	36.67	-5.83	29.83	8.67	10.67	9.69	29.00	34.83

P/12: mean points scores for 12 seasons. +/-: goal difference. M: matches played. W: matches won. D: matches drawn. L: matches lost. F: goals for. A: goals against.

Procedure and analysis

We run a machine learning job in Microsoft Azure Machine Learning Studio, where we uploaded the original dataset containing all the seasons as a csv file, with normalized root mean squared error (NRMSE) as a primary metric to evaluate the model. The data transformation process and the applied algorithms are shown in the figure below:

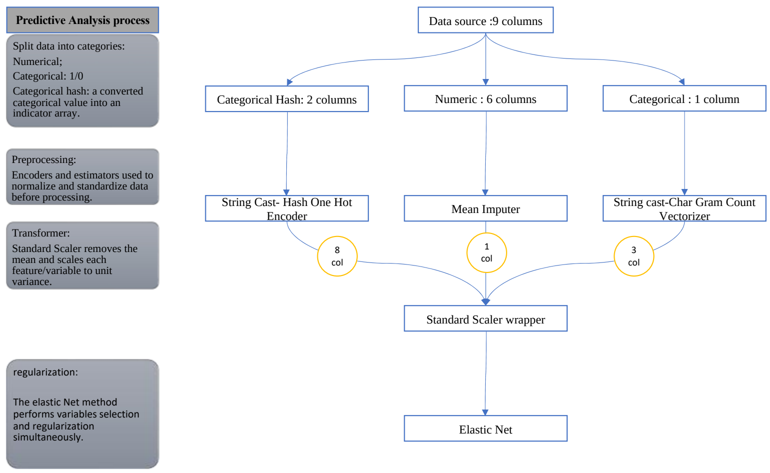

Figure. 1: Data preprocessing and engineering method for regression model with Elastic Net algorithm.

With this model, NRMSE is considered a scatter index, has a value of 0.001 closer to 0, and represents the best fitting model, and with $r^2=0,999$ as a metric, it indicates that the response variable can be perfectly explained without error by the predictor variable. Moreover, a perfect Spearman correlation of value +1 means a perfect association of rank, which is our case with a value of 0.997.

RESULTS

The regression analysis is meant to predict a numerical value as a target label depending on the features. In our case study, the mean feature by importance is the matches won historically during all the seasons. The figure below shows the aggregate feature importance on a scale from 0 to 4.

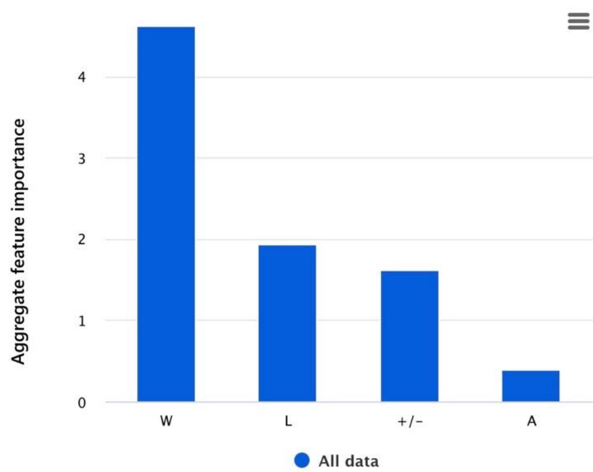

Figure 2. Aggregate feature importance for all the data in the dataset on a scale from 0 to 4.

All four features have different impacts on the ranking position at the end of the 2021/2022 season, where we can explore them individually in a datapoint chart with logarithmic scaling, with the position of importance by row in the dataset.

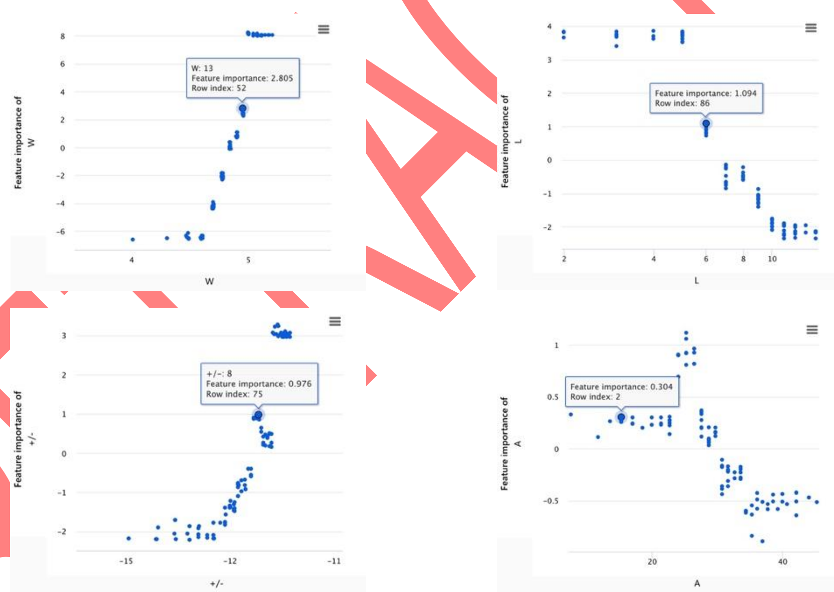

Figure 3. Top 4 features importance values with row index in the dataset.

When it comes to the matches won feature, the regression analysis allows us to compare the importance between two or more data points in different seasons. The figure below shows a comparison of two “W” feature importance in two different seasons.

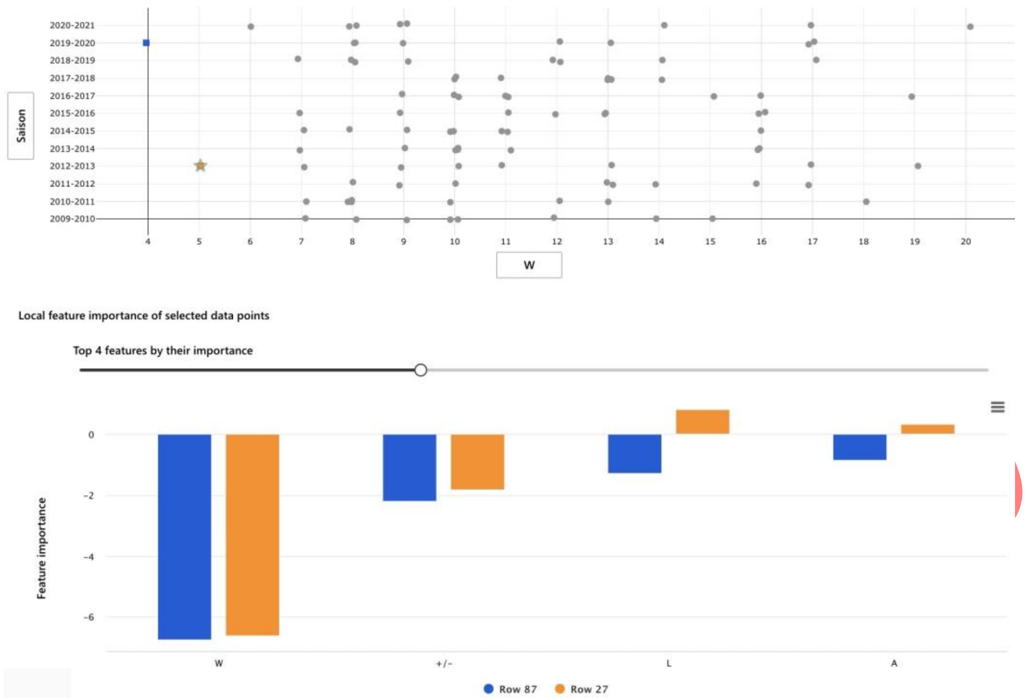

Figure 4. Matches' Won feature importance comparison of two different seasons.

Prediction

In a regression model, the prediction is a numerical value that is considered as a label. In our method of analysis, we have deployed an endpoint so we can define the team in a JSON file and obtain the values that we transcribed in a table for comparison with the actual outcomes. The tables below highlight the prediction for the season 2021/2022 with actual outcomes.

Table 2. Ranking of teams by 2021/2022 season: actual ranking.

Club	P	+/-	M	W	D	L	F	A	Actual 2021/2022
WAC Casablanca	63	23	29	19	6	4	45	22	1st
RCA	59	20	29	17	8	4	41	21	2nd
D.H. ElJadida	35	-9	29	8	11	10	31	40	9th
Hassania Agadir	33	-4	29	9	6	14	26	30	10th
FAR Rabat	45	8	29	12	9	8	37	29	3d
Moghreb Tétouan	0	0	0	0	0	0	0	0	out
FUS Rabat	40	3	29	10	10	9	32	29	5th
Olympic Club de Safi	38	2	29	9	11	9	29	27	7th

Table 3. Regression values and ranking prediction for the 2021/2022 season.

Club	Regression values	Predicted from 12 yrs	Actual 2021/2022
WAC Casablanca	123.49	1st	1st

RCA	119.37	2nd	2nd
D.H. ElJadida	98.44	4th	9th
Hassania Agadir	92.84	7th	10th
FAR Rabat	102.53	3d	3d
Moghreb Tétouan	69.21	out	out
FUS Rabat	96.86	5th	5th
Olympic Club de Safi	93.53	6th	7th

Table 4. Regression values and ranking prediction for the 2022/2023 season.

Club	**Regression values**	**Predicted from 13 yrs**
WAC Casablanca	116.1878453	1st
RCA	102.2777284	3rd
D.H. ElJadida	92.17679822	6th
Hassania Agadir	89.2303382	7th
FAR Rabat	111.6901356	2nd
FUS Rabat	97.28467347	4th
Olympic Club de Safi	92.84940406	5th

Based on the collected data till two weeks before the league championship ends, the predictive regression analysis shows the promising teams for standing where the 8th standing team will leave the first pro league (out) in Table 3 and doesn't appear in Table 4. Other teams have raised from the second league and dispute the standing with the remaining teams from the 2009/2010 season.

Botola Pro 2022/2023

Morocco

See past competitions

Standings

- Total
- Home
- Away

#	Club	P	+/-	M	W	D	L	F	A
1	FAR Rabat	67	31	30	20	7	3	50	19
2	Wydad Casablanca	66	26	30	19	9	2	47	21
3	FUS Rabat	55	20	30	15	10	5	36	16

#	Club	P	+/-	M	W	D	L	F	A
4	Olympic Club de Safi	47	6	30	12	11	7	34	28
5	RCA Raja Casablanca Athletic	44	5	30	11	11	8	31	26
6	RSB Berkane	44	2	30	11	11	8	31	29
7	Hassania Agadir	39	1	30	10	9	11	30	29
8	Union de Touarga	36	-6	30	9	9	12	34	40

DISCUSSION

Regression analysis turns out the most accurate way to predict ranking since the scores are numerical values based on historical data collected for 13 years. And the predicted values represent the highest scores for the top-ranking teams in quantitative order, which aims through this case study to understand the variables influencing a team's performance and its position in the league table. Furthermore, many other features can directly impact the team's performance as player attributes, match statistics in possession and attack efficacity, management change, and player physiological abilities after and before each match(18,19). Taking into consideration these facts, our dataset is built on historical data of the teams and has disregarded the players' and managers' contributions to the team's performance in strategies and tactics (20).

Prediction accuracy was significant in this study due to the quality of the data gathered and preprocessing method adopted. The predicted values are correct by 6 out of 8 taking into consideration the 7th actual position missed by one rank, which represents around 75% of accuracy. This difference could be impacted by changes made in team structure as players transfer, players injured and substitutions to take into consideration.

CONCLUSION

Sports outcomes prediction has become the most common tool for the actors working in this field, club managers, team's coach, tipsters, and bookies relay on the continuous data flow and real time analysis (21). In our case, we used the Moroccan national football league standing teams during 13 years since 2009/2010 season, based on their historical scores to predict the ranking in the end of 2021/2022 and 2022/2023 season. The last ranking prediction must be compared to the final results when the season ends. Moreover, players' features and teams' structure could create a powerful features selection to obtain more accurate results.

APPLICABLE REMARKS

- Teams could use the predictive analysis to plan their training sessions and tactics against the away teams.
- Relay on data about other teams to consider the strength and weakness position during a season.
- Explore more predictive analysis as classification to determine the winning probability before and during a match based on historical and real-time data.

CONFLICT OF INTEREST

The authors declare that no conflicts of interest could be perceived as interfering with the publication of this study.

REFERENCES

1. Hubáček O, Šourek G, Železný F. Exploiting sports-betting market using machine learning. Int J Forecast. 2019;35(2):783–96.
2. Leitner C, Zeileis A, Hornik K. Forecasting sports tournaments by ratings of (prob) abilities: A comparison for the EURO 2008. Int J Forecast. 2010;26(3):471–81.
3. Haghighat M, Rastegari H, Nourafza N, Branch N, Esfahan I. A review of data mining techniques for result prediction in sports. Advances in Computer Science: an International Journal. 2013;2(5):7–12.
4. Couceiro MS, Dias G, Araújo D, Davids K. The ARCANE project: how an ecological dynamics framework can enhance performance assessment and prediction in football. Sports Medicine. 2016;46(12):1781–6.
5. Kipp K, Warmenhoven J. Applications of regularized regression models in sports biomechanics research. Sports Biomech. 2022;1–19.
6. Delen D, Cogdell D, Kasap N. A comparative analysis of data mining methods in predicting NCAA bowl outcomes. Int J Forecast. 2012;28(2):543–52.
7. Maszczyk A, Gołaś A, Pietraszewski P, Roczniok R, Zając A, Stanula A. Application of neural and regression models in sports results prediction. Procedia-Social and Behavioral Sciences. 2014;117:482–7.
8. McCabe A, Trevathan J. Artificial intelligence in sports prediction. In: Fifth International Conference on Information Technology: New Generations (itng 2008). IEEE; 2008. p. 1194–7.
9. Arabzad SM, Tayebi Araghi ME, 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. 2014;1(3):159–79.
10. Tax N, Joustra Y. Predicting the Dutch football competition using public data: A machine learning approach. Transactions on knowledge and data engineering. 2015;10(10):1–13.
11. Holmes B, McHale IG. Forecasting football match results using a player rating based model. Int J Forecast. 2023;
12. Yeadon MR, Pain MTG. Fifty years of performance-related sports biomechanics research. J Biomech. 2023;111666.
13. Ricketts C, Malete L, Myers ND, Bateman AG, Bateman CJ. Sport bodies: An examination of positive body image, sport-confidence, and subjective sport performance in Jamaican athletes. Psychol Sport Exerc. 2023;67:102434.
14. Miočić J, Zekanović-Korona L, Hotti L. Importance of Regression Analysis in Sports Information Systems at Evaluation of Sports and Sports Associations. 2019;
15. Zhan W, Wang K, Cao J. Elastic-net based robust extreme learning machine for one-class classification. Signal Processing. 2023;211:109101.
16. Nasir IM, Raza M, Ulyah SM, Shah JH, Fitriyani NL, Syafrudin M. ENGA: Elastic Net-Based Genetic Algorithm for human action recognition. Expert Syst Appl. 2023;227:120311.

17. Saavedra-Garcia M, Matabuena M, Montero-Seoane A, Fernandez-Romero JJ. A new approach to study the relative age effect with the use of additive logistic regression models: A case of study of FIFA football tournaments (1908-2012). PLoS One. 2019;14(7):e0219757.
18. Cortez A, Trigo A, Loureiro N. Football Match Line-Up Prediction Based on Physiological Variables: A Machine Learning Approach. Computers. 2022;11(3):40.
19. Araújo D, Davids K, Hristovski R. The ecological dynamics of decision making in sport. Psychol Sport Exerc. 2006;7(6):653–76.
20. Hammoudeh A, Vanderplaetse B, Dupont S. Soccer captioning: dataset, transformer-based model, and triple-level evaluation. Procedia Comput Sci. 2022;210:104–11.
21. Hubáček O, Šourek G, Železný F. Exploiting sports-betting market using machine learning. Int J Forecast. 2019;35(2):783–96.