IMPROVING THE ACCURACY OF MACHINE LEARNING MODELS FOR PREDICTING OUTCOMES IN VIRTUAL SPORTS SYSTEMS

Georgi Iliev, Todor Atanasov

Abstract. *This study explores the use of machine learning (ML) models to analyze and predict players' performance in fantasy football. Using publicly available data, the impact of various factors such as value per minute, points from the bonus points system (BPS), minutes played, and others on players' total points is analyzed. Effective ensemble ML models are built using Classification and Regression Trees (CART) and Random Forests algorithms. To improve model accuracy, a CART model with an optimized number of variables is used as a feature selection method. The results of the methodology, including the reduction of the dataset, demonstrate an increase in the model's prediction accuracy.*

Key words: Football Analytics, Fantasy Premier League (FPL), Random Forests, Feature Selection, Model Accuracy, Ensemble Model.

Introduction

Machine learning (ML) is a core paradigm of artificial intelligence (AI). Its methods are used for analysis, pattern extraction, and the prediction of the unknown or future state of the studied system through data processing. In recent years, in the field of football, specific frameworks, architectures, and ML-based statistical methods have been actively developed and applied. Predicting the performance of teams and players in football leagues is a challenging task related to team composition, training strategies, match tactics, budget management, and other factors. The availability of large datasets from various public sources for the English Premier League (EPL), Fantasy Premier League (FPL), and others provides an opportunity for effective use of football analytics. The construction and evaluation of the models is carried out on the basis of known variables (factors) for different past periods in the careers of players, teams, coaches, their respective financial indicators, etc. Among the purposes of prediction are, for example, evaluating players' performance metrics (passes, goals, and fouls), analyzing and forecasting football match results, and others [1].

This study focuses on predicting players' performance using data for Fantasy Premier League (FPL). FPL is a popular virtual football game with over 11 million fans worldwide by the end of 2025 [2]. Players register as virtual football managers and get the chance to build their dream team for a single season. They oversee all the team's activities – from lineup selection and match tactics to budget management – and compete based on the real performance of professional footballers in the EPL [3].

The basic ML methods for sport and football analytics are Artificial Neural Networks (ANN), Decision Trees (DT), Random Forests (RF), Support Vector Machines (SVM), eXtreme Gradient Boosting (XGBoost), Deep Learning, Long Short Term Memory (LSTM), Light Gradient Boosting (LightGBM), and others [1, 4]. Moya et al. discussed the integration of ML techniques in professional football, addressing analyses of player and team performance, and match outcome prediction in a recent publication [5]. The authors of [6] set out to predict the outcomes of matches between two EPL football teams using a large dataset with over 30 variables. To this end, the results of ML models built with eight algorithms – Naive Bayes (NB), K-nearest neighbors (KNN), RF, SVM, C5.0 (decision trees), Xgboost, Multinomial Logistic Regression (MLR), and ANN were evaluated and compared. The prediction accuracy of all selected final models is over 62%, with RF achieving the best result (65.26%), followed by XGBoost and SVM (63.95%). Several LSTM and convolutional neural network (CNN) models are built in [3] to predict the players' total points in the FPL. Among these, the LSTM models proved to be more accurate across different test scenarios. Tamimi and Tran used various algorithms, including Ridge Regression, LightGBM, one-dimensional CNN, and LSTM, to forecast players' performance in FPL. By generating numerous new features from additional sentiment and news datasets, the classification and regression models achieved better predictions [7]. Bangdiwala et al. compared three machine learning models built with Linear Regression, DT, and RF to predict the number of points that each footballer would earn during the season [8]. Minimal errors were obtained with the RF model. Other applications of ML in the field of FPL include service prediction [9], improving lineup selection [10], and strategy development [11], among others.

The purpose of this work is to investigate some flaws in the application of supervised ML for football predictions. Despite the availability of numerous statistics (goals, assists, expected goals, expected assists, minutes played, etc.), there are systematic errors in forecasting minutes played, accounting for fixture difficulty, and reporting unexpected events (injuries, red cards, rotations). The influence of various factors on forecast accuracy and the effectiveness of careful data preprocessing through feature engineering, including feature selection, has been studied. Several ensemble learning models using CART and RF, with and without feature selection, have been built and evaluated, and the accuracy of the predictions obtained has been compared. It has been demonstrated that in the second case, the proposed approach can reduce the root mean squared error of the models by 10–20% compared to the baseline models.

Data and research methodology

To build the models, we used freely available data from FPL on players' performance for the 2025 season [12]. Our goal is to determine the influence of various variables (features) that impact the total points these players earn during the season, based on the provided historical data. This variable is labeled *total_points*. We consider it as a target variable, depending on the 17 available variables. The initial dataset contains information about 746 footballers with no missing data. Of these, 307 have 0 total points, so their data were removed as noise. As a result, the size of the initial dataset was reduced to $n = 439$.

The modeling process employs the ML methods $CART$ and RF [13, 14]. The research methodology includes the following steps: (1) Building and assessing $CART$ models with and without feature selection and identify the most important independent variables in the models; (2) constructing and evaluating RF models created with different feature combinations; (3) comparing the predictions of *total_points* from the selected models; (4) determining the most valuable features influencing players' performance; (5) interpreting the results.

Model performance was evaluated using the following statistical indicators to compare the values of the target variable Y and the model's prediction variable P: root mean squared error ($RMSE$), mean absolute percentage error ($MAPE$), and coefficient of determination R^2, defined by the expressions:

$$RMSE = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(P_k - Y_k\right)^2}, \quad MAPE = \frac{1}{n}\sum_{k=1}^{n}\left|\frac{P_k - Y_k}{Y_k}\right|,$$
$$R^2 = \frac{\left[\sum_{k=1}^{n}\left(P_k - \bar{P}\right)\left(Y_k - \bar{Y}\right)\right]^2}{\sum_{k=1}^{n}\left(P_k - \bar{P}\right)^2 \cdot \sum_{k=1}^{n}\left(Y_k - \bar{Y}\right)^2}. \tag{1.1}$$

In (1.1), Y_k, $\bar{Y}$ are the values and the mean of Y, respectively, and P_k, $\bar{P}$ are the model predicted values and their mean, respectively, n is the sample size of the used dataset. As the best of the created models, we will select the one with minimal errors and R^2 close to 1.

Building and evaluating *CART* models

Multiple CART models were built by varying their hyperparameters as follows: number of observations in parent and child nodes $(j1, j2)$, tree depth, type of cross-validation (CV), number of features, and others. With the reduced dataset size of $n = 439$ observations and dependent variable *total_points*, two models – $C1$ and $C2$ were selected. They were built with parameters $(j1, j2) = (10, 5)$, tree depth up to 10 levels, and trained with the usual 10-fold CV. With 17 independent variables, the number of possible combinations is enormous. For example, the number of possible combinations of 10 out of 17 features is 19448. In practice, instead of trying all these possibilities, it is sufficient to apply recursive sequential elimination of the variable with the highest variable importance. To find the best model with 10 variables according to measures (1), we use an integrated algorithm in the SPM software [15]. Model $C1$ is built with all the initial 17 features, while model $C2$ uses the ten variables extracted by SPM.

The statistics of the two CART models, $C1$ and $C2$, are shown in Table 1. Model $C2$ slightly outperforms $C1$ with fewer minor errors and almost the same value of the coefficient of determination R2. This coefficient indicates that both CART models explain 97.8% of the data with a 95% probability. We can assume that the selected 10 features are basic, and the remaining seven can be ignored. Table 2 shows the independent variables in the analyses and the corresponding relative importance values for the presented models. There are no significant differences among the first 7–8 variables, except for Minutes, which was dropped due to the feature selection procedure.

Table 1. Prediction statistics of selected built models in the study

Model	Statistics			
	Number of features	RMSE	MAPE	R^2
*C*1	17	2.13799	0.05914	0.97781
*C*2	10	2.13669	**0.05462**	0.97783
*R*1_100	17	1.41776	0.08906	0.99246
*R*2_200	17	1.44195	0.08725	0.99215
***R*3_100**	**10**	**1.32297**	**0.06559**	**0.99291**
*R*4_100	9	1.40239	0.06997	0.99160

Results from *RF* models

Different RF models were built by varying the following hyperparameters: $ntree$ – number of trees in the RF model ensemble, $mmin$ – minimum node size in a terminal node of each $CART$ tree in the ensemble, $ntry$ – number of features randomly selected from the total number when splitting each branch. Table 1 gives the statistics of the three selected best RF models. They were built with $mmin = 5$ and $ntry = 3$, and trained on a standard random out-of-bag (OOB) test sample for RF. Of these models, the first two – $R1\_100$ and $R2\_200$ have $ntree = 100$ and 200 trees, respectively, and were built with all 17 features. From Table 1, it is seen that the $R1\_100$ model with 100 trees has better statistics than $R2\_200$ for all three measures (1). From Table 2, it is observed that there is a non-significant difference in the ranking of the relative importance of the 17 features.

The next model, $R3\_100$, was built with the same hyperparameters as $R1\_100$ and the 10 features of model $C2$. From all the models in Table 1, it is seen that this model outperforms all the others with the smallest error, $RMSE$ 1.32297, and matches the data 99.3%. Also, it has the smallest error, $MAPE = 0.06559$ (6.6%), of the other RF models. For comparison, model $R4\_100$ was also built using the first nine variables from model $R1\_100$, sorted by size and with importance > 5. From Table 1, it is seen that this model has weaker statistics than $R3\_100$. In addition, Figure 1 shows the correspondence of the relative importance of the main features in models $R1\_100$ and $R3\_100$.

Feature	Models					
	C1	*C2*	*R1_100*	*R2_200*	*R3_100*	*R4_100*
BPS	100	100	100	92.68	63.27	94.39
Value_per_m	98.39	99.05	99.70	100	100	100
Influence	86.62	86.76	46.65	33.46	23.42	19.85
Minutes	69.91	-	25.40	25.20	-	5.93
Bonus	70.25	70.29	25.40	19.70	7.40	7.47
Selected_by_percent	12.84	13.11	20.10	20.36	11.87	14.52
Clean_sheets	5.97	61.63	18.97	14.82	5.03	3.77
ICT_index	76.62	76.76	13.11	16.01	8.43	6.38
Goals_scored	2.85	3.65	10.54	6.55	0.99	2.70
Threat	4.23	9.47	4.22	2.87	1.38	-
Now_cost	2.87	4.22	2.38	3.17	3.92	-
Assists	0.33	-	2.33	2.24	-	-
Element_type$	0.48	-	1.20	1.25	-	-
Goals_conceded	0.30	-	1.08	1.95	-	-
Creativity	2.05	-	0.95	0.64	-	-
Yellow _cards	0.03	-	0.26	0.24	-	-
Red_cards	0.00	-	0.02	0.01	-	-

Table 2. Feature importance in constructed models

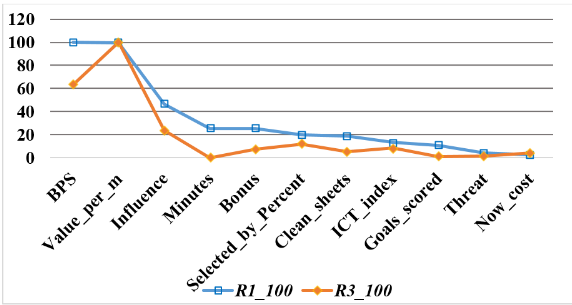

Figure 1. Comparison of the relative importance of the first sorted 11 features in model $R1\_100$ and model $R3\_100$

By pre-ranking the influence of the 17 variables on the target *total_points* using a $CART$ model, the variables with low importance were identified (see Table 2). After removing the features with the lowest influence, the RF models showed improved performance on the three selected statistical criteria (1).

Overall, the $RMSE$ of model $R3\_100$ is 7% lower than the $RMSE$ of model $R1\_100$ and 8% lower than the same error of model $R2\_200$. For the dimensionless error $MAPE$ of $R3\_100$, this reduction is more significant – by 26% and 24%, respectively, compared to the $MAPE$ of models $R1\_100$ and $R2\_200$.

Interpreting the results

Our main task in modeling was to demonstrate how to identify the most important players' characteristics to accurately predict their total points for the season (variable *total_points*). For this purpose, two high-performance ML methods were used – $CART$ and $RandomForests$. Both techniques allow us to determine the relative importance of each variable in the model. With a large number of features, the question arises of rejecting some to achieve a good enough prediction. In this study, we used $CART$ models for feature selection. It turned out that the 10 variables extracted in model $C2$ (column 3 of Table 2, used with RF) achieved the best statistics with model $R3\_100$ (see Table 1). The variables with the most significant contribution in the models are: *BPS*, *Value_per_m*, *Influence*, *Bonus*, *Minutes*, *Selected_by_percent*. Non-significant features are: *Assists*, *Element_type$*, *Goals_conceded*, *Creativity*, *Yellow_cards*, and *Red_cards*.

Conclusions and future work

Following the results, we can formulate the following conclusions:

The Random Forests models achieve very high accuracy ($R^2 \approx 0.99$ on 439 active players). Removing noise (players with zero total points) makes the model more stable and reliable for real predictions. Some aggregated indicators (*Value_per_m*, *BPS*, *Influence*) turn out to be more informative than raw statistics such as goals and assists.

The combination of multiple trees in the RF ensemble provides stability and high predictive ability. Variables such as yellow and red cards and position type have negligible importance in the model. Aggregated indicators are much more informative.

Some possible directions for future work are as follows:

- Gameweek predictions (time model instead of seasonal amount) for more dynamic transfer planning;
- Classification model to detect "explosive" players (15+ points per round) to identify differentiators;
- Integration of additional variables – injuries, schedule, team tactics,

opponent form;

- Comparison with real-world models on platforms such as FPL Hub / Fantasy Football Scout for validation;
- Creation of an API for real-time integration with Fantasy Premier League data.

Acknowledgments

This study is supported by the project FP25-FMI-010 "Innovative interdisciplinary research in informatics, mathematics and educational pedagogy" at the Paisii Hilendarski University of Plovdiv.

References

[1] R. Beal, T. Norman, S. Ramchurn, Artificial intelligence for team sports: a survey, *The Knowledge Engineering Review*, 2019, Vol. 34, p.e28. ISSN: 0269-8889, https://doi.org/10.1017/S0269888919000225

[2] Fantasy Premier League (FPL), https://fantasy.premierleague.com/ (last visited on December 5, 2025).

[3] A. Lombu, I. Paputungan, C. Dewa, Predicting fantasy premier league points using convolutional neural network and long short term memory, *Jurnal Teknik Informatika*, 2024, Vol. 5, No. 1, pp. 263–272, ISSN: 2723-3863, https://doi.org/10.52436/1.jutif.2024.5.1.1792

[4] R. Bunker, T. Susnjak, The application of machine learning techniques for predicting match results in team sport: A review, *Journal of Artificial Intelligence Research*, 2022, Vol. 73, pp. 1285–1322, ISSN: 1076-9757, https://doi.org/10.1613/jair.1.13509

[5] D. Moya, C. Tipantuña, G. Villa, X. Calderón-Hinojosa, B. Rivadeneira, R. Álvarez, Machine learning applied to professional football: Performance improvement and results prediction, *Machine Learning and Knowledge Extraction*, 2025, Vol. 7, art. 85, ISSN: 2504-4990, https://doi.org/10.3390/make7030085

[6] F. Rodrigues, Â. Pinto, Prediction of football match results with machine learning, *Procedia Computer Science*, 2022, Vol. 204, pp. 463–470, Online ISSN: 1877-0509, https://doi.org/10.1016/j.procs.2022.08.057

[7] M. Tamimi, T. Tran, Players performance prediction for Fantasy Premier League, using transformer-based sentiment analysis on news and statistical data, *International Journal of Computer Science in Sport*, 2025, Vol. 24, No. 1, ISSN: 1684-4769, https://doi.org/10.2478/ijcss-2025-0008

[8] M. Bangdiwala, R. Choudhari, A. Hegde, A. Salunke, Using ML models to predict points in Fantasy Premier League, *2nd Asian Conference on Innovation in Technology (ASIANCON)*, Pune, India. Aug 26–28, 2022, pp. 1–6, IEEE, ISBN: 978-1-6654-6851-0/22/

[9] D. Groos, OpenFPL: An open-source forecasting method rivaling state-of-the-art Fantasy Premier League services, *arXiv preprint*, 2025, https://arxiv.org/abs/2508.09992v1

[10] D. Ramezani, T. Dinh, Data-driven team selection in Fantasy Premier League using integer programming and predictive modeling approach, *arXiv preprint*, 2025, https://arxiv.org/abs/2505.02170

[11] E. Notelid, T. Östlund, *Enhancing Fantasy Premier League strategies through machine learning and large language models*, Faculty of Science and Technology, Uppsala University, 2025

[12] V. Anand, Fantasy Premier League Historical Dataset, www.github.com/vaastav/Fantasy-Premier-League/blob/master/data/2025-26/cleaned_players.csv (last visited on 22 November 2025)

[13] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Chapman and Hall/CRC, 2017

[14] L. Breiman, Random Forests, *Machine Learning*, 2001, Vol. 45, No. 1, pp. 5–32, https://doi.org/10.1023/A:1010933404324

[15] Salford Predictive Modeler (SPM), https://www.minitab.com/en-us/products/spm/ (last visited on 5 December 2025)

Georgi Iliev[1], Todor Atanasov[1]
[1] Paisii Hilendarski University of Plovdiv,
Faculty of Mathematics and Informatics,
236 Bulgaria Blvd., 4027 Plovdiv, Bulgaria
Corresponding author: iliev86@gmail.com