Journal of Informatics and Web Engineering

Vol. 4 No. 3 (October 2025) eISSN: 2821-370X

Developing A Predictive Model for Football Players' Market Value Using Machine Learning

Muhammad Afif Jazimin Idris[1*], Sew Lai Ng[2]

[1,2]Faculty of Computing and Informatics, Multimedia University, Jalan Multimedia, 63100 Cyberjaya, Malaysia

corresponding author: (1211103419@student.mmu.edu.my; ORCiD: 0009-0002-6999-0199)

Abstract - Football is the world's most popular sport, and evaluating the market value of players is crucial for clubs and managers in making informed decisions regarding transfers, contracts, and financial planning. This study aims to develop a predictive model to estimate the market value of football players using machine learning (ML) algorithms and real-life statistics performance data from the top five European leagues such as English Premier League, Italian Serie A, Spanish La Liga, German Bundesliga, and French Ligue 1 between the 2017/18 and 2019/20 seasons. By reviewing past research, various ML methods such as Random Forest, LightGBM, XGBoost, and Gradient Boosting Decision Tree (GBDT) are developed. Data preprocessing techniques, including data cleaning, feature selection, feature encoding, splitting, and standardization, are applied to ensure data quality and consistency. To tune the hyperparameter of the models, RandomizedSearchCV is applied alongside cross validation. The model evaluation is conducted using regression metrics such as mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R²), to determine the most accurate model. The best-performing model is further utilised to analyse the correlation between the features and market value, offering insights into the key features that significantly impact the market value for each position.

Keywords—Machine Learning, Football, Market Value, Correlation Coefficient, Key Features

Received: 09 Feb 2025; Accepted: 25 May 2025; Published: 16 October 2025

1. INTRODUCTION

Football, also known as soccer in some countries, is the world's most popular sport. In football, two teams will compete in a match to get the ball into the opposing team's goal, and which team that scores the most goals wins. There will be 11 players for a team on the pitch, which consists of a goalkeeper and ten outfield players to play as either defender, midfielder or attacker position. To win a match in football, not only do the players have to work on maintaining good performance, but the managers of the team also must play their role. Managers need to scout and analyse players from other clubs to have one good team by signing contracts with new players. Different managers have different ways of signing new players. Some managers decide to focus on hiring star players, and some decide to focus on hiring young talented players, which can reduce the expenditures of signing players [1].

MULTIMEDIA UNIVERSITY

Journal of Informatics and Web Engineering
https://doi.org/10.33093/jiwe.2025.4.3.12
© Universiti Telekom Sdn Bhd.
Published by MMU Press. URL: https://journals.mmupress.com/jiwe

In professional football, it is crucial for clubs, agents and analysts to understand what influences a player's market value. Instead of using more advanced ML methods, football researchers are more likely to use traditional linear regression methods for evaluating the market value [2]. However, only linear relation between a dependent feature and independent features can be analysed, which is not applicable to real-life situation [2]. Instead of using linear regression methods, there are many ML methods that have been used in past research to predict the market value of football players such as random forest [2], and multi linear regression [3]. Furthermore, existing research relies on FIFA football video game datasets as their primary source of player attributes and performance features [2-5]. Although much research has shown that well-performing models can be developed with the use of football video game data, real-life features such as goals scored, and duels won are not included. On top of that, since the attributes of the players are influenced by assessments from game developers, it is possible that some features of a player in the video games dataset are partially based on the transfer fee that has been paid by a football club [4].

To estimate the market value of football players, several performance features can be used [3]. In addition, market value is also influenced by external factors such as age, height, and nationality [3]. Key performance features differ by position; for example, midfielders show big differences in touches and passes of the ball compared to the other positions because midfielders play an important role in organizing and attacking phase, while strikers show big differences in shots, shots on targets, and dribbles due to their role to score a goal [6]. Aside from the performance features, the key features of external factors also differ by position. Height helps strikers and defenders in their heading ability to score or prevent a goal, and the age of the players reflects their experience and ability [3]. It was found out in research where European players are overrated, which receive more fees than non-European players [3]. Due to these variations from performance features and external factors features, the model needs to generate results that highlight the key features influencing market value based on the player's position.

The aims of this study are:

- To identify the most accurate ML method for predicting football players' market value using real-world performance statistics.
- To determine the key features that significantly influence market value based on the player's position.

This study contributes to the development of a robust model for valuation, which can be applied to other sports where key features play a crucial role.

2. LITERATURE REVIEW

2.1 ML in Sport Analytics

ML is yet another technological advancement that completely alters the approach of how sports analysts. One of the critical applications of ML in sports analytics is injury prediction. Using models that utilise deep learning algorithms, researchers have been able to build models that estimate the risk of injury by assessing workload metrics, movement patterns, and other physiological factors. Research in [7] utilised deep learning techniques for early injury warning systems. This could help a sport team to manage their financial costs due to the inability of injured players to play any match.

Next application of ML is tactical analysis. Graph neural networks (GNNs) were utilised in [8] for predicting game results and helping in decision-making in volleyball. By capturing complex interactions between players and teams, ML models provide coaches with deeper insights into the best plans, counter strategies, and player positioning

ML also helps with player scouting and talent discovery. Traditionally, recruiting scouts depends on objective metrics and subjective assessments, which are likely to vary widely amongst assessors. Chmait and Westerbeek [9] claimed that ML algorithms can evaluate a player's fitness with the team dynamics, forecast future performance, and identify hidden talent by examining enormous volumes of player performance data. This helps managers and scouts to have a better idea of who they can hire, reducing the negative risks associated with onboarding new players.

Sports analytics has advanced to a new level with the use of ML to forecast athletes' market values. ML models could make assessments that help clubs make well-informed decisions during transfer negotiations by evaluating large datasets that include player performance attributes and external factors. Multiple linear regression (MLR) models, decision trees, and ensemble methods were used in [10] to forecast Major League Baseball (MLB) player salaries based on contract history, age, injury history, and player statistics (e.g., batting average, home runs, and earned run average for pitchers).

Research was conducted in [11] for cricket using ML method to forecast the Indian Premier League (IPL) players' auction prices during the IPL auction. The dataset contains historical performance metrics such as runs scored, balls faced, innings played, wickets taken, and matches played. ML models, including decision trees and K-nearest neighbours (KNN), were used to forecast players' selling prices. Section 2.4 will cover more details on the use of ML to forecast football players' market values.

2.2. Key Performance Features Based on Football Players' Positions

Research in [12] used ML models to examine performance characteristics from 864 Qatar Star League (QSL) games played between 2012 and 2019 in their study, ML Models Reveal Key Performance Metrics of Football Players to Win Matches in Qatar Stars League. The research identified specific key performance features for each football position. For example, the key performance features of forward positions are shots on target and distance covered at high speed, while for defender positions are tackles made and interceptions. The player positions were separated into more specific positions in research conducted in [5]. For example, there are different categories for forward positions, which are central forward, left forward, right forward, left winger, right winger, left striker, and right striker.

2.3. External Factors Influencing Football Players' Market Value

Research in [13] examined external factors that affect football players' market value in their research on Market Value Prediction of Football Players. First, the external factor influencing market value is the player's age. Due to their longevity in the professional football industry and their potential for growth, younger players typically have higher market values. The research also discovers that a player's market value is significantly influenced by their position. In general, forward players are more valuable than defensive players. For instance, Thibaut Courtois, the highest paid goalkeeper player, received 60 million euros, while Kylian Mbappe, the highest paid forward player, received 180 million euros. According to [3], the players' nationalities have an impact on their market value as well. It was discovered that non-European players are underrated, and European players are overrated. Hence, the market value of European players is typically higher than non-European players.

2.4. ML in Market Value Prediction in Football

One ML algorithm that researchers use to forecast the market value of football players is MLR. Multiple learning regression was employed as an improvement from the baseline model, which is simple linear regression (SLR) [3]. The accuracy of this algorithm is higher than the baseline model because as SLR evaluates the relationship between a single dependent and a single independent variable, MLR evaluates the relationship between a single dependent with multiple independent variables. However, MLR is not able to handle non-linear data because it still assumes a linear relationship between variables. This study uses the FIFA 20 football video game as its dataset while also considering the other factors like player popularity, performance, and attributes that affect football players' market value. The decision tree and random forest, which are non-linear models were also used in this research due to having these non-linear factors. Using Analysis of Variance (ANOVA) and Pearson correlation, the data was analysed to determine the key features impacting the market value. Features that do not improve predictive accuracy are not used. According to the study's findings, the random forest algorithm gives the most accurate results with the lowest RMSE.

Jana and Hemalatha [5] used ridge regression, stepwise regression, and particle swarm optimisation (PSO) in their research. FIFA 19 football video games were utilised as the dataset for this research. Similar with the research conducted in [3], correlation analysis was used to find key features. However, this research used correlation analysis for every football position (e.g., left striker), which resulted in different key features for each position. PSO was utilised to optimise player scores by simulating match scenarios to obtain benchmark scores. PSO performs well in handling multi-dimensional search spaces and identifying global optimal solutions, which makes it suitable for modelling complex interactions in player attributes. However, PSO can be sensitive to parameter settings (e.g., inertia weights) and the performance may worsen when the search space is overly constrained. The stepwise regression and ridge regression, enhanced with smoothing splines were then applied as different models to predict overall player scores and the market value. After the two models were compared, the stepwise regression model has better accuracy with the lowest RMSE.

This research predicted the market value of the players based on the attributes of the players without considering any external factor (e.g., age of the players), which plays a crucial role affecting the market value in professional football.

Laros [4] investigated the prediction of transfer fees of football players using SLR as the baseline, MLR, support vector regression (SVR), and random forest in his research. The research also relied on FIFA football video game dataset from 2015 to 2021. By expanding the dataset across multiple seasons, the author aims to achieve more comprehensive statistical analysis rather than limiting the evaluation to player attributes from a single season. To avoid overfitting, K-Fold cross-validation was used to tune the hyperparameters without using any optimization method, after the data was standardised using Scikit-learns' StandardScaler. Football players were categorised based on their positions to understand the attributes influencing the market value per position. MLR was used as a baseline to analyse linear relationship between multiple independent variables with a dependent variable, transfer fee. Due to limitations in handling non-linear relationships, another model which implements SVR was deployed to improve the prediction by using hyperplanes to fit data points within defined boundaries, but it requires careful hyperparameter tuning, which can be computationally expensive. However, random forest model outperforms SVR with better accuracy.

Research in [14] used optimised ensemble learning approach in predicting football players' market value. A dataset from FIFA 22 video game was obtained and used in this research. Both attributes on the pitch and external factors were taken into consideration to predict the market value. Correlation analysis was applied to identify the features that correlate with the players' market value. The author implemented GBDT, SLR, lasso, elastic net, and kernel ridge regression as the baseline models for this research. The author found out that the GBDT model has outperformed the other baseline models. Another two models, which implement LightGBM and XGBoost, were developed. The LightGBM model, which possesses high efficiency (e.g., fast training while maintaining high performance) compared to GBDT and XGBoost, is optimised using Bayesian optimization with tree-structured Parzen estimator (TPE). Since Bayesian optimization excels in handling high-dimensional data efficiently, it can find better hyperparameters in a short time, but it increases computational complexity and may require specialised resources for practical applications. As result, the optimised model outperforms baseline regression models and non-optimised gradient boosting models in prediction accuracy.

3. DATA AND METHOD

In this study, all processes are performed using a local computer with Windows operating system. Figure 1 shows the flowchart of the study that highlights the processes involved.

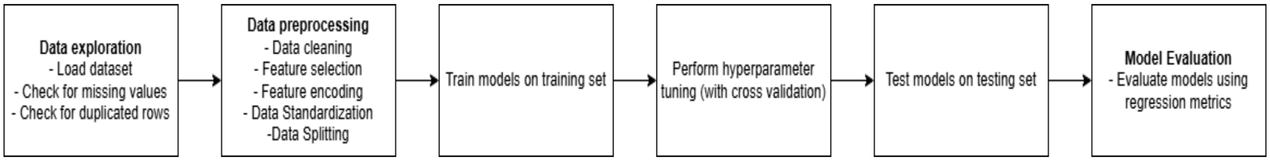
Data exploration
- Load dataset
- Check for missing values
- Check for duplicated rows
Data preprocessing
- Data cleaning
- Feature selection
- Feature encoding
- Data Standardization
-Data Splitting
Train models on training set
Perform hyperparameter tuning (with cross validation)
Test models on testing set
Model Evaluation
- Evaluate models using regression metrics

Figure 1. Flowchart of the Study

Since football analysis is a widely explored field, such datasets can be found on the internet. A dataset titled 'European Soccer Dataset: Season 2017–2020' provided by Alessia Simone in CSV file from Kaggle [15] website is chosen to be used in this study. The dataset contains real-life statistics performance of the football players, along with their attributes (e.g., age, height, nationality, league, and club) and market value from season 17/18 to season 19/20 from five European leagues. The market value of the players given in the dataset is the value after the end of the season. The dataset is chosen for its extensive coverage of features based on real-life statistics performance and usability rating of 10 by Kaggle. Most of the datasets that use real-life statistics are either have limited attributes, which are basic features such as goals scored, assists, or minutes played, or low usability rating. Other datasets with detailed features utilise football video game data. Since this research focuses on analysing the real-life statistics performance features, these datasets are not chosen.

Data exploration is crucial for understanding the dataset's structure and the types of data. The features of the dataset are identified alongside their respective data types. The chosen dataset contains 46 different features. With the use of Pandas via Python, the sum of missing values and duplicated rows per attribute is determined. Table 1 shows the list of the attributes in the dataset and the data types gotten with their respective descriptions.

Table 1. List of Features in The Dataset with Description and Data Type

No.	**Feature**	**Description**	**Data Type**
1	squad	Teams name	String
2	Season	Season	Date
3	Pts	The total points the player's team earned in the league	Integer
4	GF	Number of goals scored by the player's team	Integer
5	GA	Number of goals conceded by the player's team	Integer
6	attendance	Number of attendees at the team's matches	Integer
7	CL	If the player's team played in Champions League (a prestigious European tournament)	Boolean
8	WinCL	If the player's team won in Champions League	Boolean
9	CLBestScorer	If the player has been the top scorer in Champions League	Boolean
10	MP	Number of teams' matches played	Integer
11	W	Number of teams' matches won	Integer
12	D	Number of teams' matches draws	Integer
13	L	Number of teams' matches losses	Integer
14	player	Players' first and last name	String
15	age	Age of each player	Integer
16	height	Height of each player (cm)	Integer
17	nationality	Nation where the player was born	String
18	value	Value of the player in football market (€)	Floating
19	position	Player's position on the pitch	String
20	league	League's name	String
21	LgRK	Team's league ranking	Integer
22	games	Number of matches played by the player	Integer
23	games starts	Number of matches where the player started	Integer
24	minutes	Number of minutes played	Integer
25	ball recoveries	The number of times the player regained possession of the ball from the opposing team	Integer
26	cards yellow	Number of yellow cards (warnings) received by the player	Integer
27	passes completed	The number of successful passes the player made to teammates	Integer
28	fouls	The number of times the player committed fouls	Integer
29	fouled	The number of times the player was fouled by opponents	Integer
30	offsides	The number of times the player was caught offside	Integer
31	own goals	The number of goals accidentally scored by the player into their own team's net	Integer
32	crosses	The number of passes the player made across the field to teammates in attacking positions	Integer
33	shots total	The total number of shots the player attempted	Integer
34	goals	The total number of goals the player scored	Integer
35	assists	The number of passes the player made that led directly to goals	Integer
36	pens won	The number of penalties kicks the player won for their team	Integer
37	touches	The total number of times the player touched the ball during matches	Integer
38	dribbles completed	The number of times the player successfully moved past opponents with the ball	Integer
39	sca	The number of actions the player made that led to a shot attempt	Integer
40	gca	The number of actions the player made that directly led to a goal being scored	Integer
41	tackles	The number of times the player successfully stopped an opponent from advancing with the ball	Integer
42	blocks	The number of times the player blocked a shot, pass, or cross made by opponent	Integer

43	pressures	The number of times the player applied pressure to an opponent with the ball	Integer
44	shots on target against	Number of shots on target received by goalkeeper	Integer
45	saves	The number of shots on target that successfully stopped by goalkeeper	Integer
46	goals against gk	Number of goals conceded by goalkeeper	Integer

For data preprocessing, the processes are divided into data cleaning, feature selection, feature encoding, data standardization and data splitting. For data cleaning, rows containing missing values are removed because removing the missing values is better than filling them with mode or mean as it may affect the prediction accuracy due to potential bias [16]. The duplicated rows are also removed. The process is simple and time saving. This dataset seems to not contain any duplicated row. The number of rows before and after data cleaning are identified to ensure that the cleaning process has been applied correctly.

Next for feature selection, team-related features are removed because these features provide information about the team as a whole, as this study focuses on the individual players. However, although ‘League’ and ‘Squad’ features are also team-related features, they are not removed as these two features are crucial for identifying the league and team associated with each player. In addition, ‘player’ and ‘Season’ features are not included as these two features only serve as an identifier of the players and which season the statistics belong to. Table 2 shows all features that are kept and removed.

Table 2. List of Features that are Kept and Removed

Kept Feature	Removed Feature
Squad, age, height, nationality, value, position, league, games, games starts, minutes, ball, recoveries, cards yellow, passes completed Fouls, fouled, offsides, own goals, crosses shots total, goals, assists, pens won, touches dribbles completed, sca, gca, tackles, blocks pressures, shots on target against, saves, goals against gk	Pts, GF, GA, attendance, MP, W, D, L, Lg Rk, Season, Player, CL, WinCL, CLBestScorer

For feature encoding, since there are features/columns in string type, they cannot directly be used in statistics model [2]. Hence, in this section, ‘position’, ‘league’, ‘squad’, and ‘nationality’ features are encoded using label encoding instead of one-hot encoding since the unique categories in the string type features/columns are in a large number. This can increase computational efficiency. For instance, the feature encoding of string type features/columns used is shown in Table 3. There are four positions, five leagues, 121 squads, and 103 nationalities. The value assigned to the values is based on the alphabetic order. In the next process, which is data splitting and data standardization, the data is split into training and testing set. 80% of the data is split into the training set, and 20% of the data is split into the testing set. Research conducted in [17] concluded that having more data in the training set leads to higher accuracy of the model. Once split, the data is standardised for training set and testing set to ensure that features are on the same scale. For example, the age of the players ranges from 18 to 45, while market value ranges from thousands to millions. Data standardization is applied after data splitting to avoid data leakage from the testing set into the training set [18]. This is because mean and standard deviation will be calculated using both the training and testing set if applied before data splitting.

Table 3. Feature Encoding of String Type Features/ Columns

Position	League	Squad	Nationality
Defender -> 0	Bundesliga -> 0	Alavés -> 0	Albania -> 0
Forward -> 1	La Liga -> 1	Amiens -> 1	Algeria -> 1
Goalkeeper -> 2	Ligue 1 -> 2	Angers -> 2	Angola -> 2
Midfield -> 3	Premier League -> 3	Arsenal -> 3	Argentina -> 3
	Serie A -> 4	Aston Villa -> 4	Armenia -> 4
		……	…….

In the model development process, models are developed to predict market value. One of the models acts as the baseline model, which serves as a reference. In other words, the other models are expected to surpass the performance of the baseline. Due to its simplicity compared to the other models that are being used in this study but still reliable, the Random Forest model is chosen as the baseline model. ML models that are used in this study, which have also been used in past related research are LightGBM, XGBoost, and GBDT.

Next, the hyperparameter of the models is tuned to find the best hyperparameters configuration, improving the model's predictive accuracy [19]. Since the models involve using large hyperparameter space, RandomizedSearchCV is used as the optimization method as it consumes less time to optimise all the models. Using optimization method is also considered more practical than manual tuning for hyperparameter selection by saving time from manually testing numerous combinations. To ensure that the models perform well on unseen data, K-fold Cross Validation is applied alongside the optimization method. The testing set is split into k folds, train on k-1 folds and validate on the remaining fold. This process will be repeated k times, when all parts have been left out. By applying cross validation within the optimization process, the best hyperparameters are selected based on average performance across all folds.

Table 4 shows the inputs of hyperparameter that are being tuned by RandomizedSearchCV for all developed models.

Table 4. Input of Hyperparameter of All Models

	Random Forest	**LightGBM**	**XGBoost**	**GBDT**
n_estimators	50, 100, 150, 200	50, 100, 150, 200	50, 100, 150, 200	50, 100, 150, 200
max_depth	3, 5, 10, 20	3, 5, 10, 20	3, 5, 10, 20	3, 5, 10, 20
min_samples_split	2, 5, 10, 20	-	-	2, 5, 10, 20
min_samples_leaf	1, 5, 10, 20	-	-	1, 5, 10, 20
max_features	'sqrt', 'log2'	-	-	'sqrt', 'log2'
bootstrap	True, False	-	-	-
learning_rate	-	0.01, 0.05, 0.1, 0.3	0.01, 0.05, 0.1, 0.3	0.01, 0.05, 0.1, 0.3
subsample	-	0.8, 1.0	0.8, 1.0	0.8, 1.0
colsample_bytree	-	0.8, 1.0	0.8, 1.0	-
num_leaves	-	20, 31, 40, 50	-	-
min_child_samples	-	5, 10, 20, 30	-	-
gamma	-	-	0, 0.1, 0.2	-
min_child_weight	-	-	1, 3, 5, 10	-
reg_alpha	-	0, 0.01, 0.1, 1	0, 0.01, 0.1, 1	-
reg_lambda	-	0.1, 0.5, 1, 2	0.1, 0.5, 1, 2	-

Some hyperparameters are algorithm-specific and may not apply to all models. For example, parameters like gamma, reg_alpha, reg_lambda, and min_child_weight are specific to boosting algorithms such as XGBoost and LightGBM.

4. RESULTS AND DISCUSSIONS

After developing all models, model evaluation process is executed to identify which model is the most accurate to predict the market value of the football players. In the model evaluation process, all the models developed are compared by using regression metrics such as MAE, mean squared error (MSE), RMSE, and R^2. Table 5 shows the results obtained from all models for their evaluation.

Based on the model evaluation, LightGBM has outperformed all models with the lowest MAE, RMSE and the highest R^2. This means that LightGBM is the most suitable model for predicting the market value in this study. LightGBM uses leaf-wise tree growth, allowing higher complexity by building much deeper trees compared to level-wise models such as XGBoost and GBDT but it is more prone to overfit [20]. However, applying hyperparameter tuning on the model reduces overfitting, making LightGBM the most accurate model in this study. Next, key features that significantly impact the player's market value based on the player's position are determined by using LightGBM model. To achieve this, correlation coefficients between the features and the market value based on position are calculated, and the resulting values are sorted to highlight features with the highest and lowest correlation to the market value. This visualization eases the interpretation of features with positive or negative impact on the market value.

Table 5. Model Evaluation

No.	Model	MAE	RMSE	R²
1	Random Forest	0.37	0.64	0.51
2	LightGBM	0.28	0.51	0.69
3	XGBoost	0.31	0.53	0.67
4	GBDT	0.30	0.53	0.66

Figure 2 shows the correlation coefficient for all positions with market value.

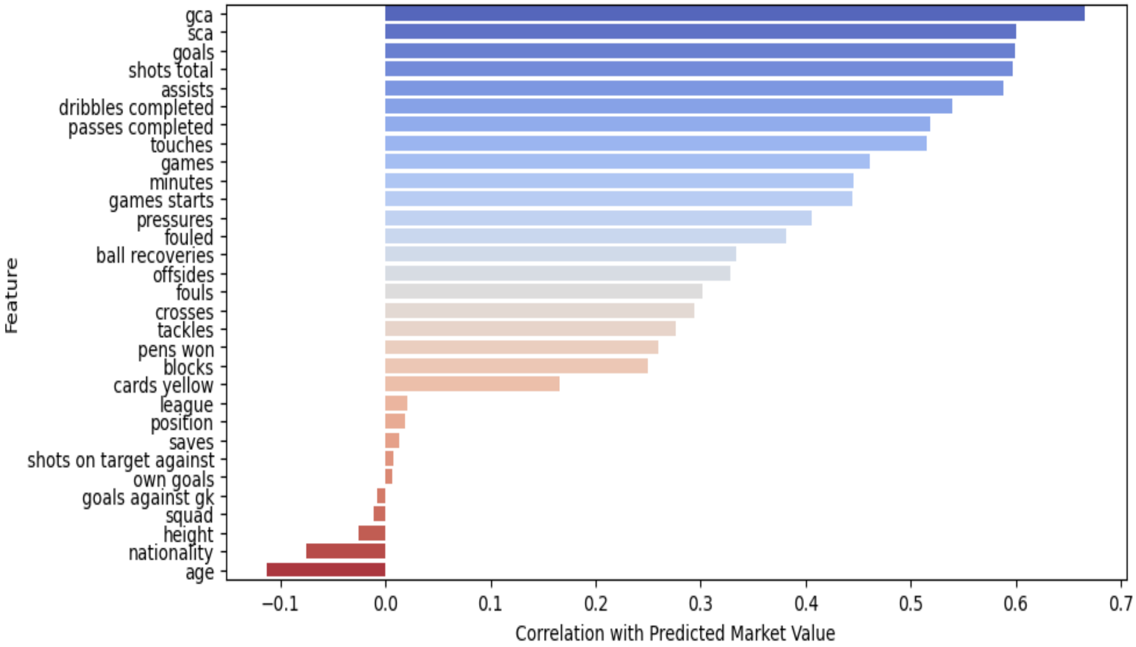

Figure 2. Correlation Coefficients for All Positions with Market Value (Based on LightGBM Predictions)

Figure 3 to Figure 6 show the correlation coefficients for specific positions with market value. Features at the top indicate the most positive correlation with the market value, meanwhile features at the bottom indicate the most negative correlation with the market value, Gca, or the number of actions the players made that directly led to a goal being scored gives the most positive impact towards the market value for all position and forward position as the chance of scoring the goal is very important, especially for forward players since that is their main role in football. As for defender players, midfield players, and goalkeeper players, passes completed have the most positive correlation with the market value. Since these players are far away from the opponent's goal, their role is to pass the ball forward to the forward players in order for them to get goals. This is not called gca hence, passes completed have the most position correlation with the market value for these positions. Age has the most negative correlation for overall positions, defender players, midfield players, and goalkeeper players. This means that as the players get older, the market value of the players decreases due to their performance getting worse. However, the forward position has height as the most negative correlation with the market value. It is noticed that many forward players are still shining even get older.

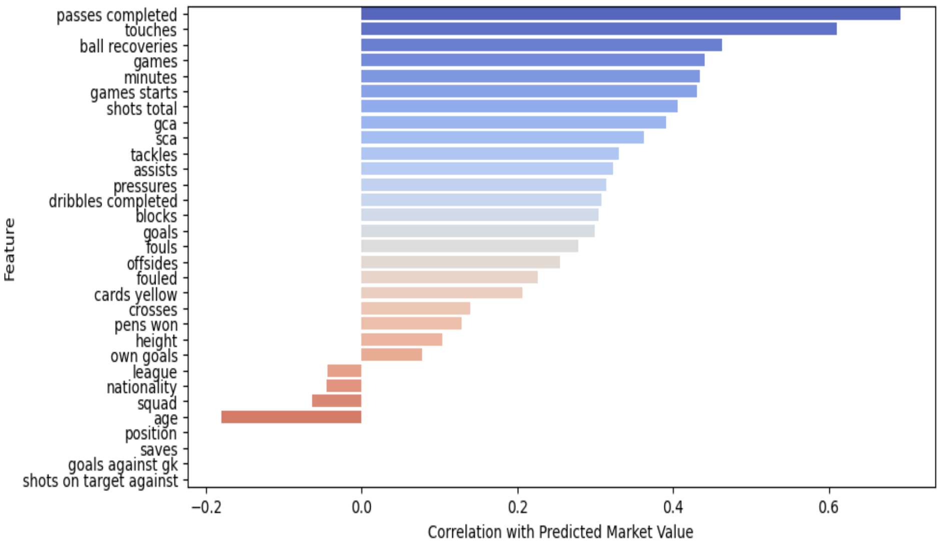

Figure 3. Correlation Coefficients for Defender Players with Market Value (Based on LightGBM Predictions)

.

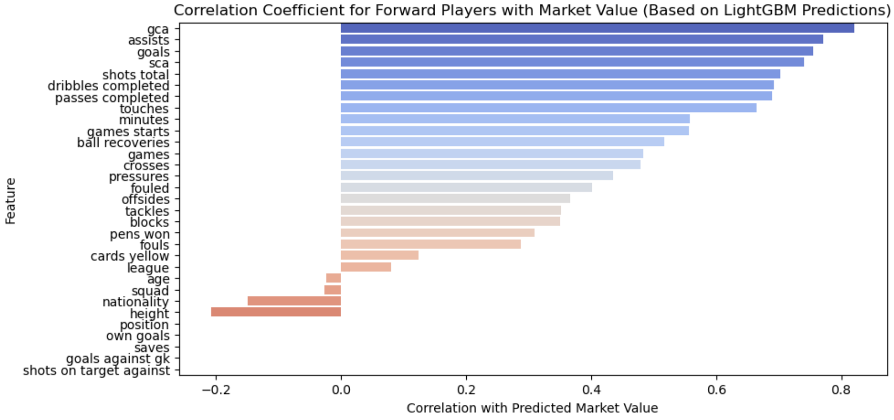

Figure 4. Correlation Coefficients for Forward Players with Market Value (Based on LightGBM Predictions)

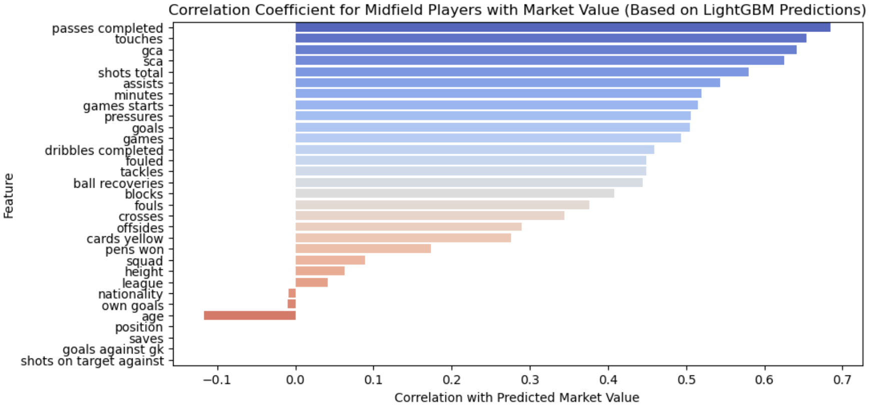

Figure 5. Correlation Coefficients for Midfield Players with Market Value (Based on LightGBM Predictions)

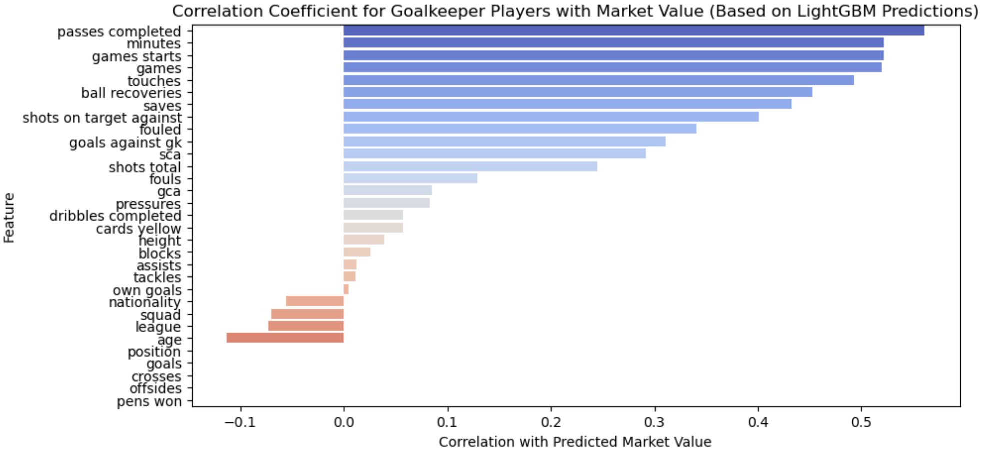

Figure 6. Correlation Coefficients for Goalkeeper Players with Market Value (Based on LightGBM Predictions)

5. CONCLUSION

This study focuses on identifying the most accurate method for predicting football players’ market value using real-world performance statistics, as well as determining the key features that significantly influence market value based on the player’s position. This study utilises statistics of real-life statistical performance of football players. Real-life statistical performance is based on actual situations on the pitch. On the other hand, FIFA video game statistics are simplified and may not capture the complexities of a player’s performance in real matches [4]. For instance, while FIFA video game assigns a shooting ability score on a scale of 1-99, real-life statistics evaluate shooting effectiveness based on key features such as shot on target, and goals scored. models such as Random Forest, XGBoost, LightGBM, and GBDT are used to develop the predictive model with RandomizedSearchCV to tune the hyperparameters.

Based on the findings, LightGBM model is the most accurate method for predicting football players’ market value as it has less error than the other models and the key features that significantly influence market value based on the player’s position have been determined. From the visualization in the discussion, it is best to say that key features differ for each football position.

For future study, some improvements could be made to improve the accuracy of the prediction by including more external factors such as media coverage, player injury history, and contract length. It could also be useful to add more hyperparameters for hyperparameter tuning. While this study focuses on standard hyperparameters, expanding the search to include other hyperparameters may result in a more accurate model, and improve the predictive performance.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for the suggestions to improve the paper.

FUNDING STATEMENT

The authors received no funding from any party for the research and publication of this article.

AUTHOR CONTRIBUTIONS

Muhammad Afif Jazimin Idris: Conceptualization, Data Curation, Methodology, Validation, Writing – Original Draft Preparation;
Sew Lai Ng: Supervision, Writing – Review & Editing.

CONFLICT OF INTERESTS

No conflict of interests were disclosed.

ETHICS STATEMENTS

Informed consent was obtained from all participants involved in this study for data collection. The platform's data redistribution policies were complied with. Our study follows The Committee of Publication Ethics (COPE). https://publicationethics.org/.

REFERENCES

[1] I. Behravan, and S. M. Razavi, "A novel machine learning method for estimating football players" value in the transfer market," *Soft Computing*, 2020, doi: 10.1007/s00500-020-05319-3.

[2] C. Li, S. Kampakis, and P. Treleaven, "Machine learning modeling to evaluate the value of football players," *arXiv.org*, 2022, doi: 10.48550/arXiv.2207.11361.

[3] M. A. Al-Asadi, and S. Tasdemir, "Predict the value of football players using FIFA video game data and machine learning techniques," IEEE *Access*, vol. 10, pp. 22631-22645, 2022, doi: 10.1109/access.2022.3154767.

[4] G. P. K. Laros, "Predicting transfer value of professional football players based on player skills and characteristics using multiple linear regression, support vector regression, and random forest regression," *Tilburg University*, 2020.

[5] J. Almulla, and T. Alam, "Machine learning models reveal key performance metrics of football players to win matches in Qatar Stars League," *IEEE Access*, vol. 8, pp. 213695–213705, 2020, doi: 10.1109/access.2020.3038601.

[6] Q. Yi., M. Gomez-Ruano, H. Liu, S. Zhang, B. Gao, F. Wunderlich, and D. Memmert, "Evaluation of the technical performance of football players in the UEFA champions league," *International Journal of Environmental Research and Public Health*, vol. 17, no. 2, pp. 604, 2020, doi: 10.3390/ijerph17020604.

[7] W. R. Johnson, A. Mian, D. G. Lloyd, and J. A. Alderson, "On-field player workload exposure and knee injury risk monitoring via deep learning," *Journal of Biomechanics*, vol. 93, pp. 185–193, 2019, doi: 10.1016/j.jbiomech.2019.07.002.

[8] R. Tracy, H. Xia, A. Rasla, Y.-F. Wang, and A. Singh, "Graph encoding and neural network approaches for volleyball analytics: From game outcome to individual play predictions," *arXiv.org*, 2023, doi: 10.48550/arXiv.2308.11142.

[9] N. Chmait and H. Westerbeek, "Artificial Intelligence and machine learning in sport research: An introduction for non-data scientists," *Frontiers in Sports and Active Living*, vol. 3, pp. 682287, 2021, doi: 10.3389/fspor.2021.682287.

[10] H. Al-Shari, Y. A. Saleh, and Alper Odabas, "Comparison of gradient boosting decision tree algorithms for CPU performance," *Erciyes Medical Journal*, vol. 37, pp. 157–168, 2021.

[11] J. Prathuri, A. Kulkarni, A. Kamath, A. Menon, P. Dhatwalia, and D. Rishabh, "Prediction of player price in IPL auction using machine learning regression algorithms", *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pp. 1-6, 2020, doi: 10.1109/conecct50063.2020.9198668.

[12] A. Jana, and S. Hemalatha, "Football player performance analysis using particle swarm optimization and player value calculation using regression," *Journal of Physics: Conference Series*, vol. 1911, no. 1, pp. 012011, 2021, doi: 10.1088/1742-6596/1911/1/012011.

[13] M. Elahi, S. Pandey, and S. S. Malhi, "Market value prediction of football players," *SSRN Electronic Journal*, 2024, doi: 10.2139/ssrn.4485449.

[14] H. Lee, B. A. Tama, and M. Cha, “Prediction of football player value using bayesian ensemble approach,” *arXiv.org*, 2022, doi: 10.48550/arXiv.2206.13246.

[15] Alessia, “European soccer dataset,” Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/alessiasimone/european-soccer-dataset-season-20172020.

[16] N. Tamboli, “Tackling missing value in dataset,” *Analytics Vidhya*, 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/.

[17] M. S. Jalani, H. Ng, T. T. V. Yap, and V. T . Goh, “Performance of Sentiment Classification on Tweets of Clothing Brands”, *Journal of Informatics and Web Engineering*, vol. 1, no. 1, pp. 16–22, Mar. 2022, doi: 10.33093/jiwe.2022.1.1.2.

[18] S. B. Pinjosovsky, “Normalize data before or after split of training and testing data?,” *Medium*, 2023. [Online]. Available: https://medium.com/@spinjosovsky/normalize-data-before-or-after-split-of-training-and-testing-data-7b8005f81e26.

[19] C. M. Chituru, S.-B. Ho, and I. Chai, “Diabetes Risk Prediction using Shapley Additive Explanations for Feature Engineering”, *Journal of Informatics and Web Engineering*, vol. 4, no. 2, pp. 18–35, Jun. 2025, doi: 10.33093/jiwe.2025.4.2.2.

[20] C. Lee, P. Hsu, M. Cheng, J. Leu, N. Xu, and B. Kan, "Using machine learning to predict salaries of major league baseball players", Lecture Notes in Computer Science, pp. 28-33, 2021, doi: 10.1007/978-3-030-79463-7_3.

BIOGRAPHIES OF AUTHORS

	Muhammad Afif Jazimin Idris is a final year student in Bachelor of Science Computer (Hons.) in Data Science in Multimedia University, Cyberjaya, Malaysia. He is passionate about studying data and analytics. He can be contacted at email: 1211103419@student.mmu.edu.my.
	Sew Lai Ng is an Assistant Professor in the Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Malaysia. Her research interests lie in the development and application of predictive modeling, statistical and econometric techniques, time series analysis, and risk management. She focuses on leveraging advanced modeling approaches to understand complex systems and inform decision-making in various domains. She can be contacted at email: slng@mmu.edu.my.