



ML Question1

05.25.2023

Hossein JourEbrahimian
G1

Overview

We would like to predict the market value (MV) of players using their information, such as age, number of caps, goals scored, and other relevant statistics. We would like to do this using an appropriate machine learning regression model.

To accomplish this, we gathered a dataset of player information and their corresponding market values from "Transfermarkt.com". The dataset that we collected was described in Phase 1. We completed this task previously.

In this phase, after choosing appropriate features from the collected data, we needed to preprocess and clean the data. This involved handling missing values, normalizing or standardizing the features, and encoding categorical variables. Once the dataset was cleaned and preprocessed, we were able to split it into training and testing sets to evaluate the performance of our model.

Finally, we chose an appropriate machine learning regression model, such as linear regression, decision tree regression, or random forest regression, and trained it on the training set. During training, the model learned the relationship between the input features and the target variable (i.e., market value). Once the model was trained, we evaluated its performance on the testing set and made predictions for new players.

Steps

1. Define the problem and the goals of the project.
2. Gather and preprocess the data.
3. Perform exploratory data analysis (EDA).
4. Split the data into training, validation, and test sets.
5. Select an appropriate machine learning algorithm or model.
6. Train the model on the training set and evaluate its performance on the validation set.
7. Perform feature engineering.
8. Tune the model's hyperparameters.
9. Evaluate the final model on the test set.

Literature review

1.




European Journal of Operational Research


Volume 306, Issue 1, 1 April 2023, Pages 389-399





Innovative Applications of O.R.

Estimating transfer fees of professional footballers using advanced performance metrics and machine learning

Ian G. McHale^a  , Benjamin Holmes^{a b}


Show more 

+ Add to Mendeley  Share  Cite

<https://doi.org/10.1016/j.ejor.2022.06.033> 

[Get rights and content](#) 

Under a Creative Commons [license](#) 

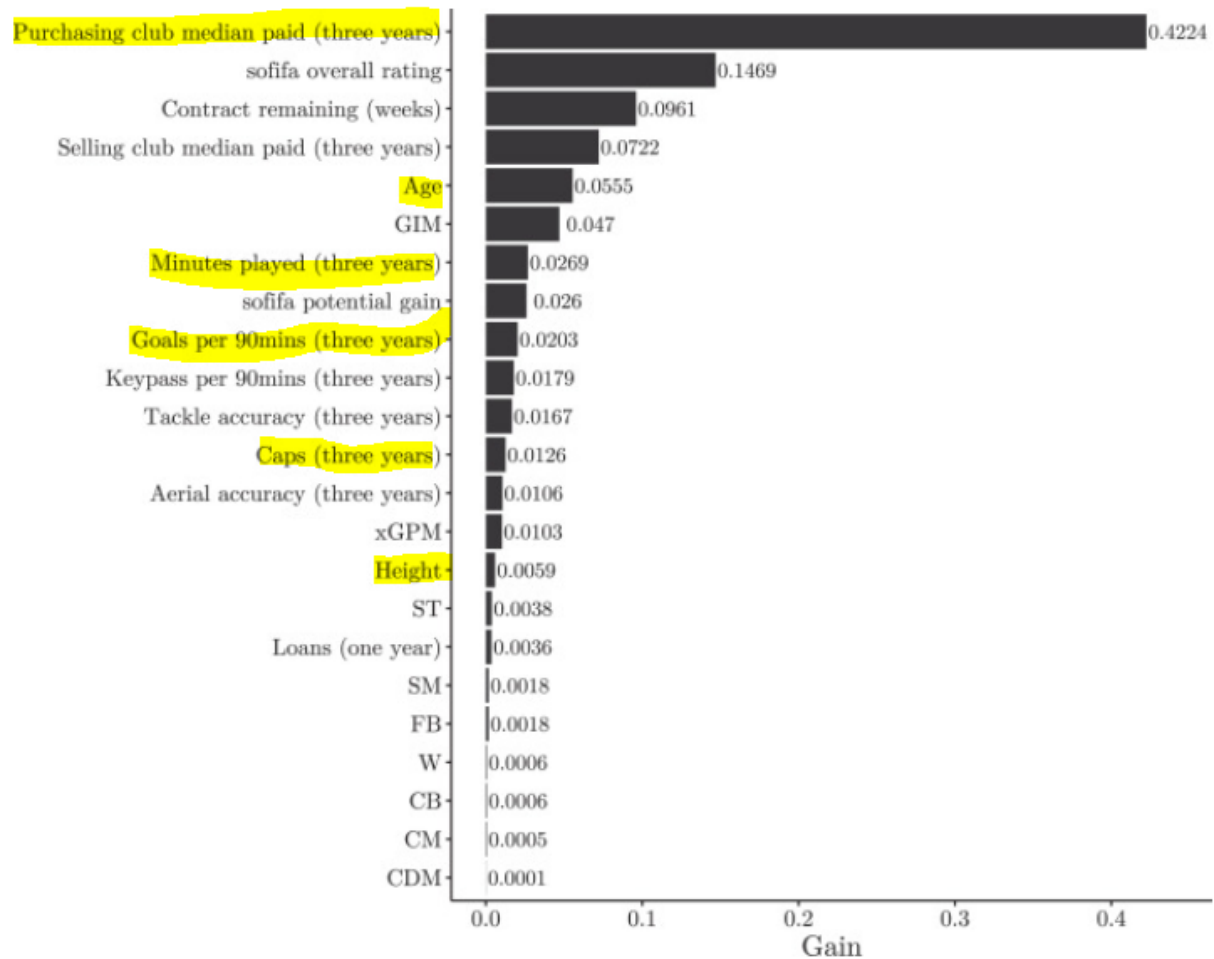
 [open access](#)

2.

Since the onrush of these economics papers, the topic has increasingly gained interest in the operational research literature and more recently, in the field of machine learning. An important distinction between papers in the literature is in the specifics of the data employed in studies. The data used for modelling varies in two dimensions. First, some authors choose crowd-sourced valuations of players as the dependent variable to be explained or predicted, not actual transfer fees. Second, some authors use subjective player ratings, not objective performance metrics, as the predictor variables.

Crowd-sourced player valuations have most commonly been taken from the transfermarkt.com website (TM). Members of the website offer their valuations of players and a panel of experts calculates a weighted average of the values to arrive at a single transfer value for each player. The panel of experts calculate the weights based on judging how accurately each member has valued players historically. Although the TM values have proved to be highly correlated with transfer fees (see, for example, Herm, Callsen-Bracker, & Kreis 2014), the two quantities are, of course, different. One is a subjective assessment of the valuation of the player made by a large crowd, whilst the other is an actual fee paid by one club to another for the services of a player. Indeed, Coates and Parshakov (2021) found that despite a high correlation with transfer fees, the TM valuations suffer from systematic bias in that adding simple descriptive statistics (such as goals scored) as a covariate in a model in which transfer fee is regressed

3.



Players

FIFA 23 MAY 26, 2023

Automate Endpoint Maintenance
Ensure mitigate security threats with proactive endpoint management solutions Endpoint Mgmt Software

Trending Added Updated Free On loan Removed Customized Create player Calculator Random

► COLUMNS SELECTED

► BASKET




SEARCH

Name

All Players


Continents

Nationality / Region

NAME	AGE	OVERA...	POTENTIAL	TEAM & CONTRACT	VALUE	WAGE	TOTAL S...
 K. Koné CM	21	77	86	Borussia Mönchengla... 2021 ~ 2025	€23.5M	€21K	2042
 K. Thuram CM CDM	21	79	85	Nice 2019 ~ 2025	€28M	€35K	2114
 B. Meijer LB LM LWB	19	73	86	Club Brugge 2022 ~ 2026	€7M	€10K	1930

Deep soccer analytics: learning an action-value function for evaluating soccer players

updates

Guiliang Liu¹ · Yudong Luo¹  · Oliver Schulte¹ · Tarak Kharrat²

Received: 12 September 2019 / Accepted: 10 July 2020 / Published online: 21 July 2020

© The Author(s), under exclusive licence to Springer Science+Business Media LLC, part of Springer Nature 2020

Abstract

Given the large pitch, numerous players, limited player turnovers, and sparse scoring, soccer is arguably the most challenging to analyze of all the major team sports. In this work, we develop a new approach to evaluating all types of soccer actions from play-by-play event data. Our approach utilizes a Deep Reinforcement Learning (DRL) model to learn an action-value Q-function. To our knowledge, this is the first action-value function based on DRL methods for a comprehensive set of soccer actions. Our neural architecture fits continuous game context signals and sequential features within a play with two stacked LSTM towers, one for the home team and one for the away team separately. To validate the model performance, we illustrate both temporal and spatial projections of the learned Q-function, and conduct a calibration experiment to study the data fit under different game contexts. Our novel soccer **Goal Impact Metric (GIM)** applies values from the learned Q-function, to measure a player's overall performance by the aggregate impact values of his actions over all the games in a season. To interpret the impact values, a mimic regression tree is built to find the game features that influence the values most. As an application of our GIM metric, we conduct a case study to rank players in the English Football League Championship. Empirical evaluation indicates GIM is a temporally stable metric, and its correlations with standard measures of soccer success are higher than that computed with other state-of-the-art soccer metrics.

4.

Table 8. Summary statistics of the predictions of our models on the out-of-sample test set.

Model	MAE	MAPE	R^2
xgbTree	3.60	67.47	0.77
xgbDART	3.64	68.90	0.76
glmer	4.11	69.05	0.74
glmnet	9.93	90.57	0.50
OLS	10.34	91.81	0.50

Features

Age & Nationality

Age is a crucial factor in predicting a player's market value, but nationality may also play a significant role. Many believe that European and South American players are more valuable than players of the same quality from Asia or Africa. To incorporate nationality as a feature in our machine learning model, we divided countries into four categories based on their typical player market values. We found that some countries are associated with higher player market values than others, even when players have similar levels of talent. Group 1 includes countries whose players are typically among the most valuable in the transfer market, such as Argentina, Brazil, and France. Group D includes all other countries whose players typically have lower market values than other groups. However, after training and evaluating our model, we decided to drop this feature and train the model again. We then evaluated the model's performance with and without the nationality feature to judge its effect on market value prediction. By doing so, we hope to better understand the influence of nationality on player market values and improve the accuracy and effectiveness of our model.

1. Group A:
 - Argentina
 - Brazil

- France
- England
- Germany
- Italy
- Portugal
- Spain

2. Group B:

- Belgium
- Croatia
- Netherlands
- Uruguay

3. Group C:

- Colombia
- Senegal
- Serbia
- Poland
- Egypt
- Mexico
- Nigeria
- Denmark
- Ghana
- Ivory Coast
- Cameroon
- Algeria
- Morocco
- Tunisia
- South Korea
- Japan
- Norway

4. Group D: All other countries not mentioned above.

Columns: age, citizenship, nationality

Caps

The number of national games that a player has played in and the number of goals they have scored in these matches can be one of the most important features for predicting the market value of a player.

Columns: international_goals, caps

Performance indicators at club level

Squad, Appearances, Goals, Goals by Penalty, Assists, Yellow cards, Red cards, Fixed in squad games, mean points per game, total time played, and, furthermore, for goalkeepers, clean sheets and goals conceded are some features that can show the performance of players in their club team.

Columns: total_squad, total_appearance, total_own_goal, total_sub_off, total_sub_on, total_yellow_card, total_second_yellow_card, total_red_card, total_penalty, total_minutes_per_goal, total_minutes_played, total_goal_conceded, total_clean_sheet, total_PPG

Performance indicators at UCL level

Intuitively, we believe that appearances, goals, assists, and minutes played in the UEFA Champions League are more important performance indicators than other club games. Therefore, we use these features to predict a player's market value at the Champions League level. After training and evaluating the model, we will create another model without these features to determine their effect on market value prediction.

Columns: total_squad, total_appearance, total_own_goal, total_sub_off, total_sub_on, total_yellow_card, total_second_yellow_card, total_red_card, total_penalty, total_minutes_per_goal, total_minutes_played, total_goal_conceded, total_clean_sheet, total_PPG

Physical parameters

Physical parameters such as height, weight, strength, agility, acceleration, jumping, and pace are essential features for a football player. However, it is unfortunate that by using 'Transfermarkt.com' we may not have access to any of these parameters except height. While height alone is not sufficient to describe all of a player's physical attributes, it is the only data that we have available. Therefore, we cannot afford to ignore this feature and must still incorporate it into our analysis.

Columns: height

Foot

The footedness, or preferred foot, of a football player can impact their playing style and value. Left-footed players may have an advantage in certain positions and may be more valuable due to their rarity and unique playing style. Players who are equally proficient with both feet, or ambidextrous, can be highly valuable in football due to their versatility and adaptability on the pitch. Ambidextrous players can use either foot to pass, shoot, and control the ball with equal ease, which allows them to play in multiple positions and adapt to different game situations. Examples of such players include Osman Dembele and Francesco Totti, who have demonstrated the value of being equally proficient with both feet in football.

Columns: foot

Position

The main position and other available positions of a football player can have a significant impact on their market value. A player who is versatile and capable of playing in multiple positions can be highly valuable to clubs, as they can fill gaps in the squad and provide tactical flexibility. This versatility can be particularly valuable for smaller clubs with limited budgets, as they can acquire fewer players to cover more positions.

However, a player's market value can also depend on their primary or main position. Players who are considered to be among the best in the world in their primary position can command very high transfer fees and salaries. For example, a world-class striker or goalkeeper may be more valuable than a versatile midfielder who can play in multiple positions.

Columns: main_position, available_positions

Current club

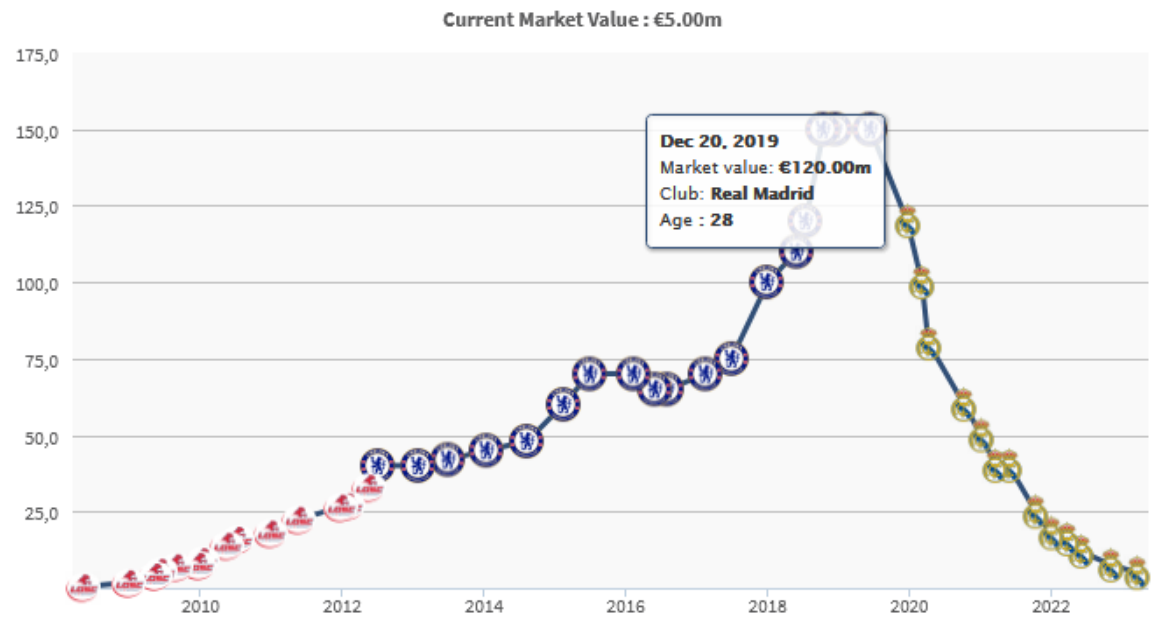
The current club of a football player can have a significant impact on their market value. Players who are currently playing for a successful and high-profile club may be more valuable in the transfer market than players who are playing for a less successful or lower-profile club. Playing for a high-profile club can increase a player's visibility, exposure, and perceived value. Additionally, the perceived strength and reputation of a player's current club can also impact their market value.

Columns: current_club_mv

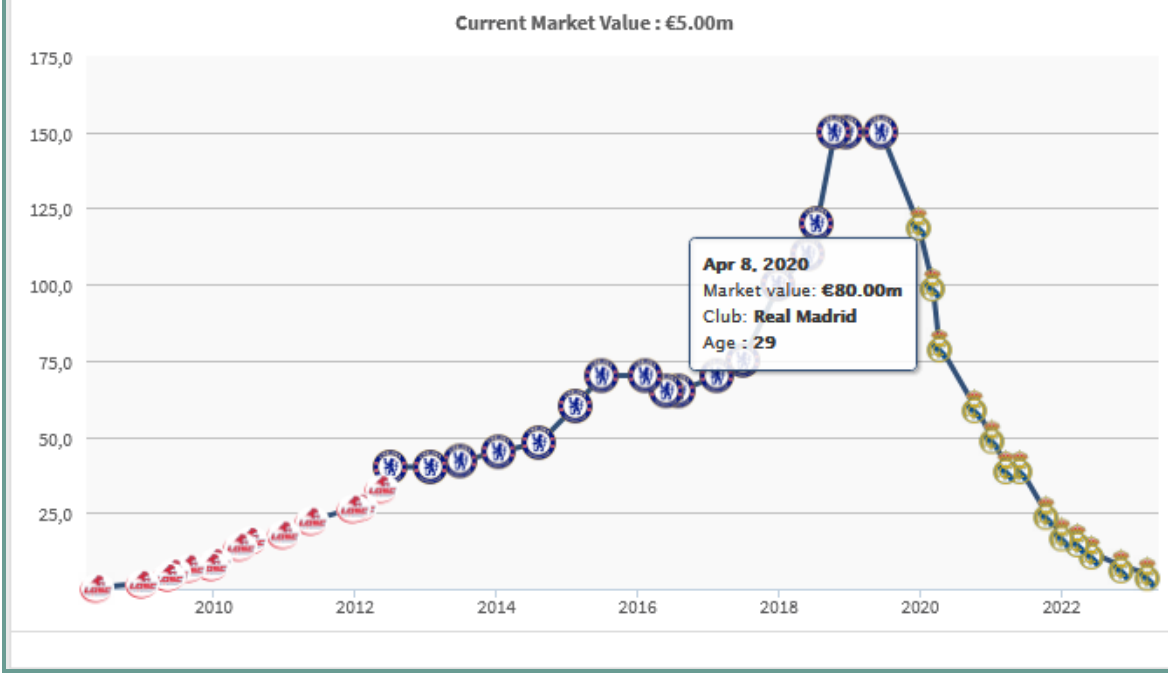
Columns: team_id

Previous transfers

MARKET VALUE OVER TIME



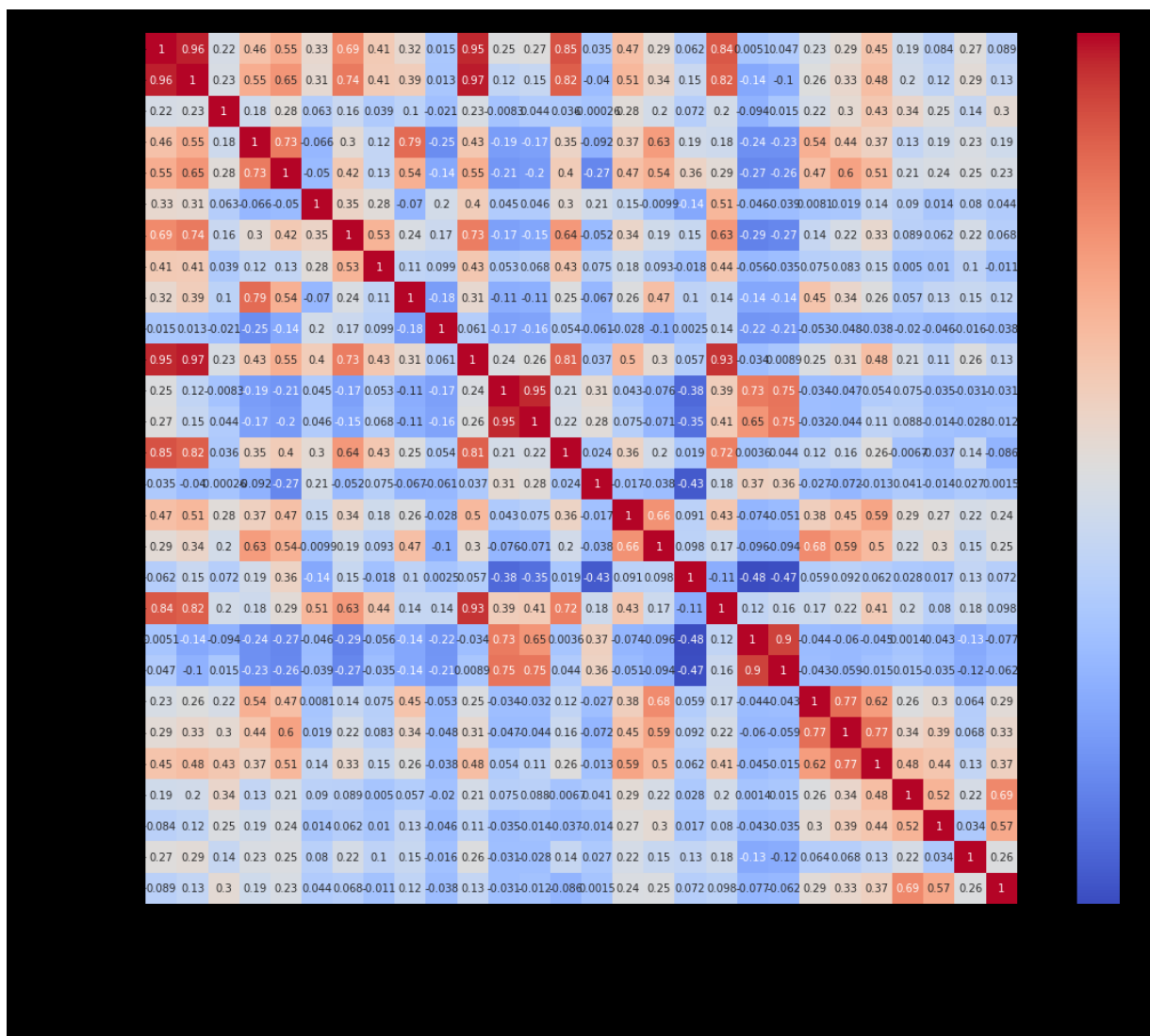
MARKET VALUE OVER TIME



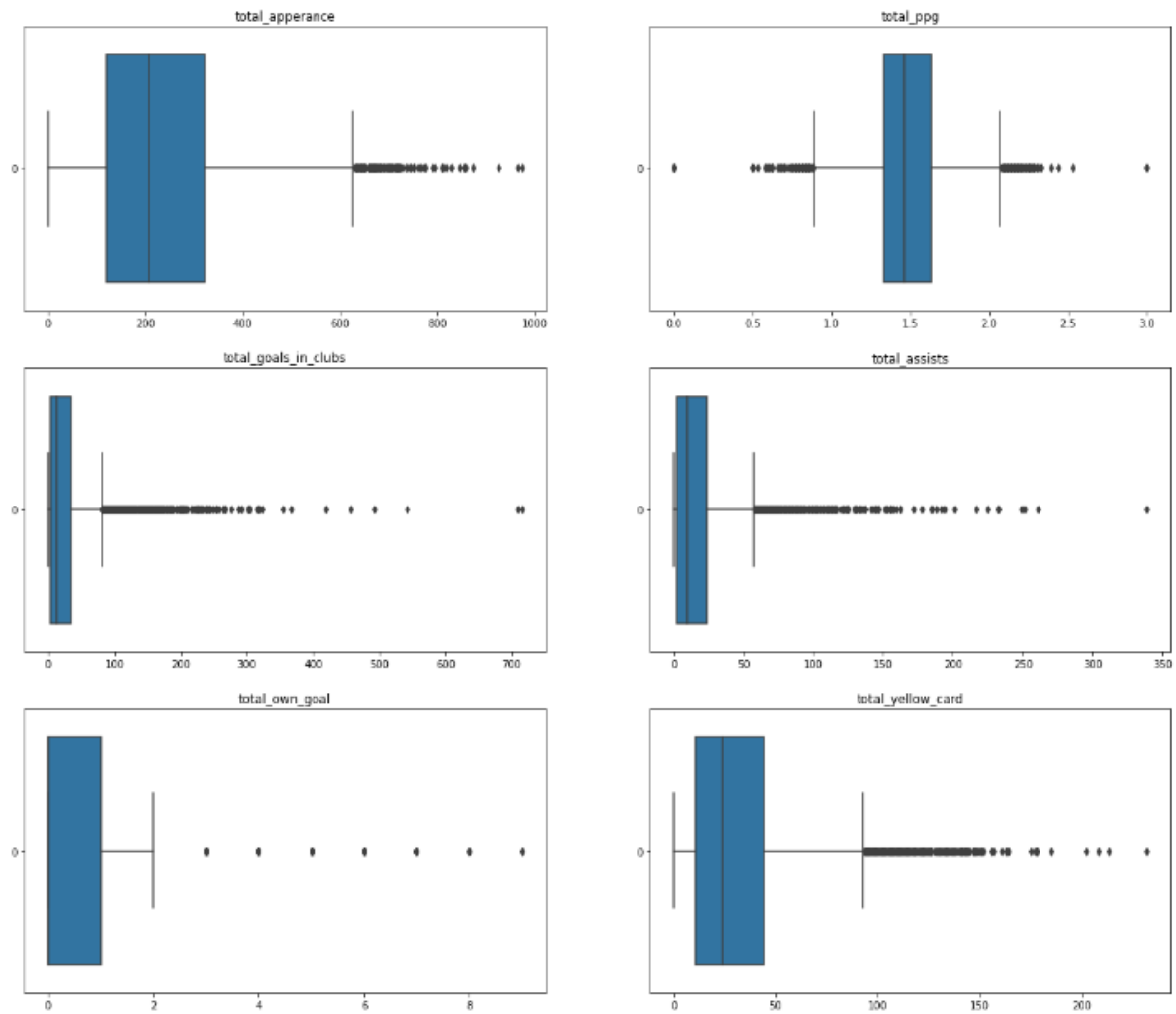
Columns: * Transfer table

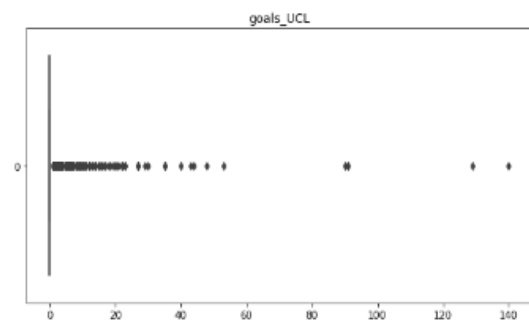
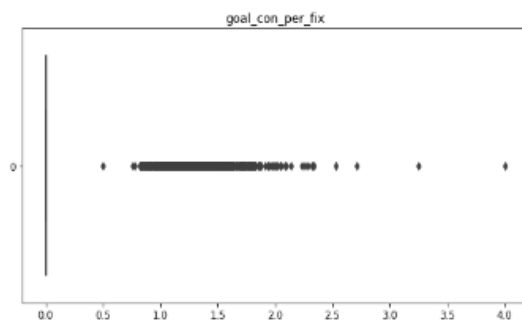
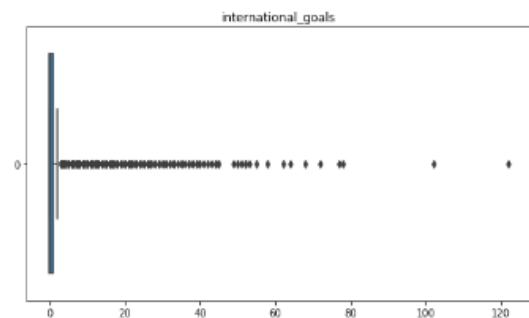
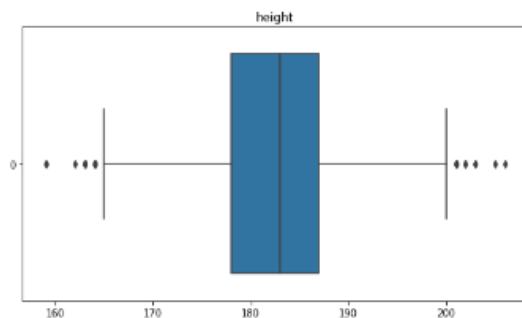
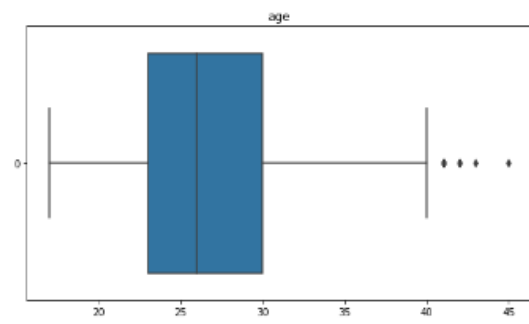
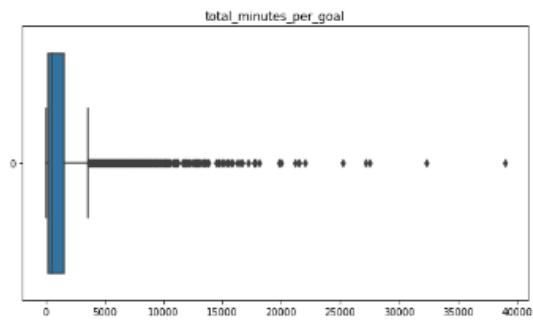
Total_appearance	/	total_ppg	/	total_goals_in_clubs	/	total_assists	/	total_own_goal
total_yellow_card	/	total_red_card	/	total_penalty	/	total_minutes_per_goal	/age	/height/
international_goals	/	goals_con_per_fix	/	goals_UCL	/	assists_UCL	/	appearance_UCL
Fee	/	year	/	current_market_value	/	nation_rank	/	pos_ATT
pos_DEE	/	pos_GK	/	pos_MID				

I. Correlation Matrix

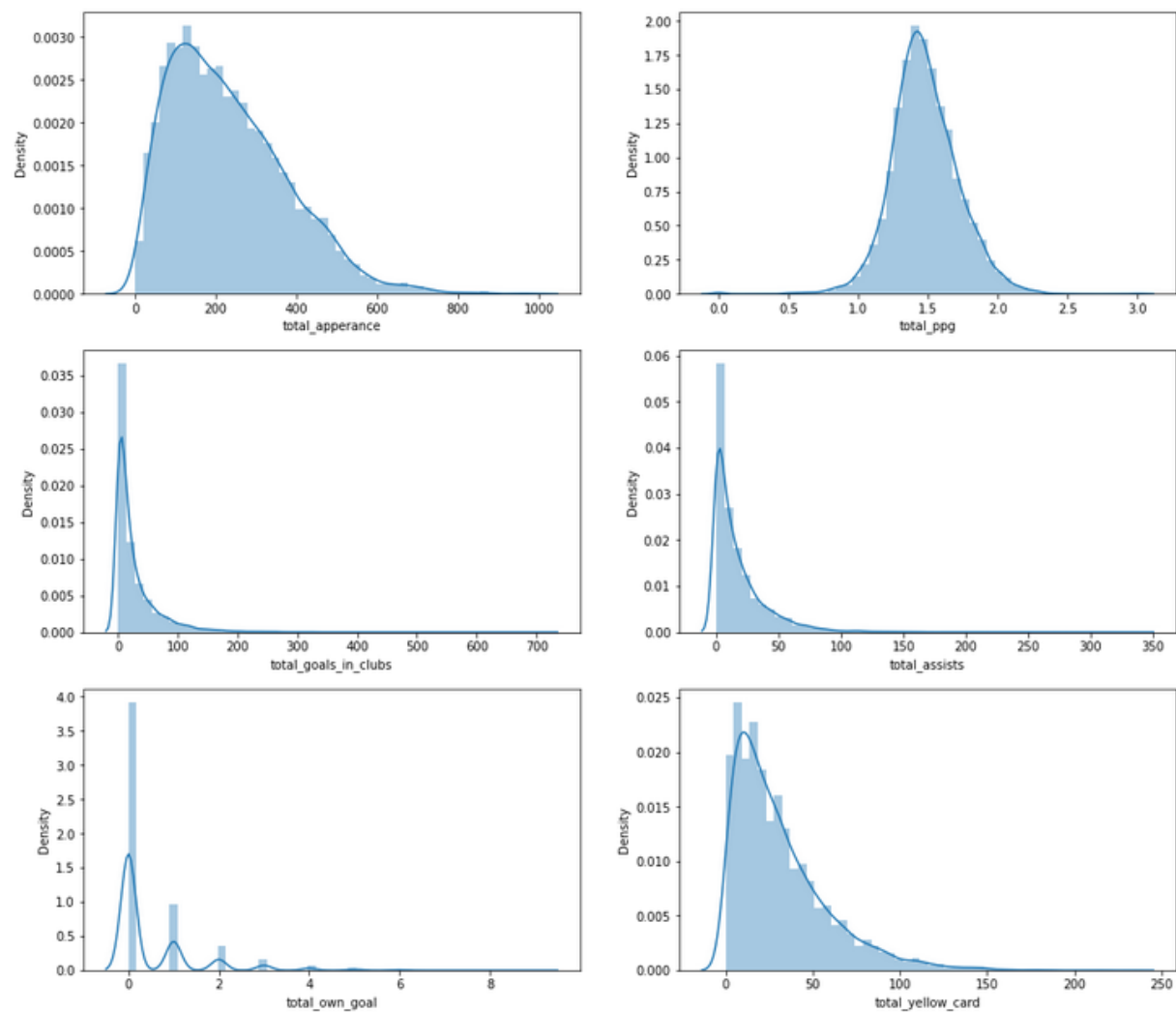


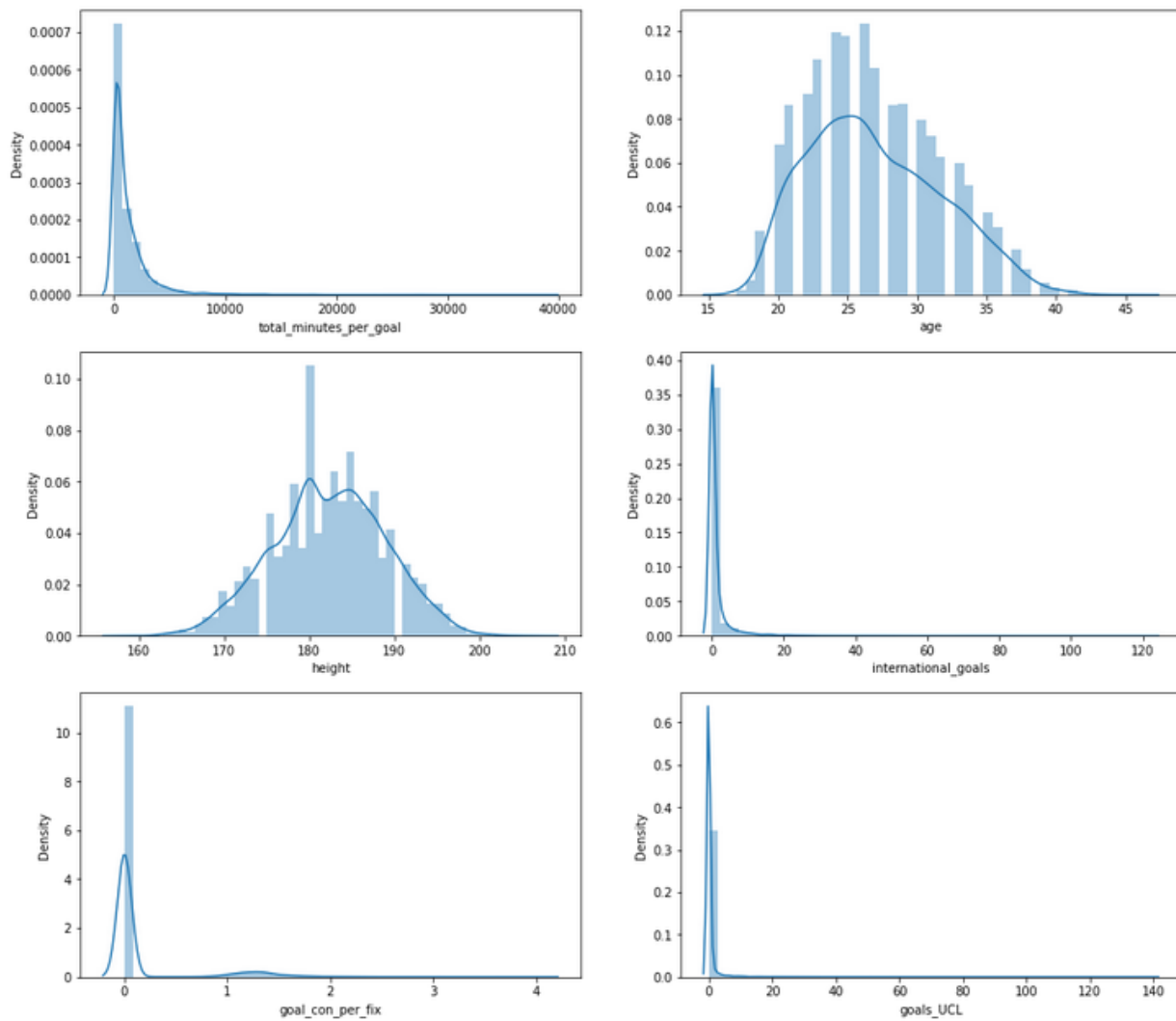
II. Box Plot





III. Distribution Plots





IV. Outlier Management

IQR:

```
need_outlier_management = ['total_appearance', 'age', 'height', 'total_goals_in_clubs', 'total_assists']

for item in need_outlier_management:
    q1 = df[item].quantile(0.25)
    q3 = df[item].quantile(0.75)
    iqr = q3 - q1

    lower_bound = q1 - 1.5*iqr
    upper_bound = q3 + 1.5*iqr

    df[item] = np.where((df[item] < lower_bound), lower_bound, df[item])
    df[item] = np.where((df[item] > upper_bound), upper_bound, df[item])
```

Z-SCORE:

```
upper_limit = df['total_ppg'].mean() + 3*df['total_ppg'].std()
lower_limit = df['total_ppg'].mean() - 3*df['total_ppg'].std()

print('Highest allowed', upper_limit)
print('Lowest allowed', lower_limit)

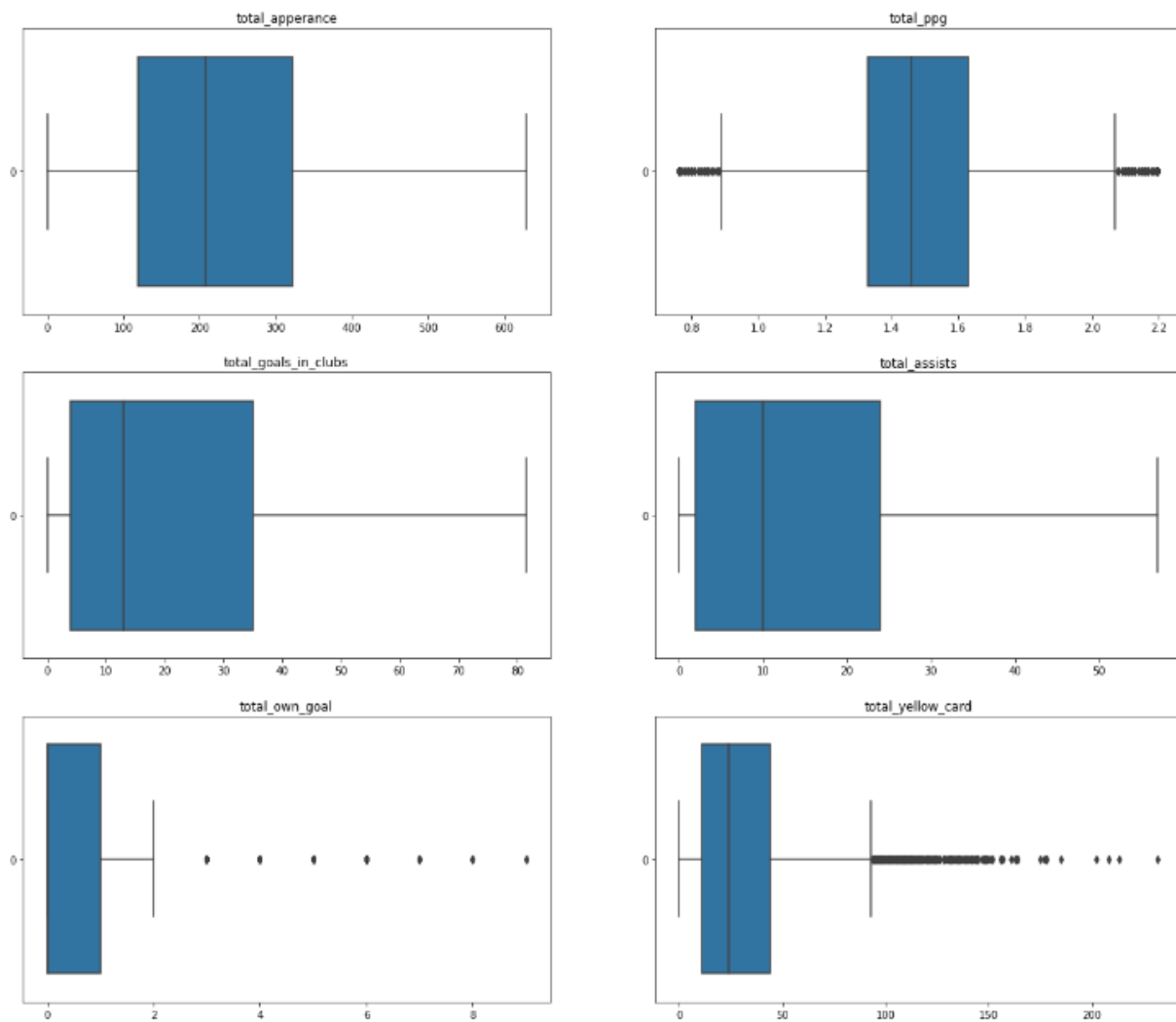
df['total_ppg'] = np.where(df['total_ppg'] > upper_limit, upper_limit, np.where(df['total_ppg'] < lower_limit, lower_limit, df['total_ppg']))
```

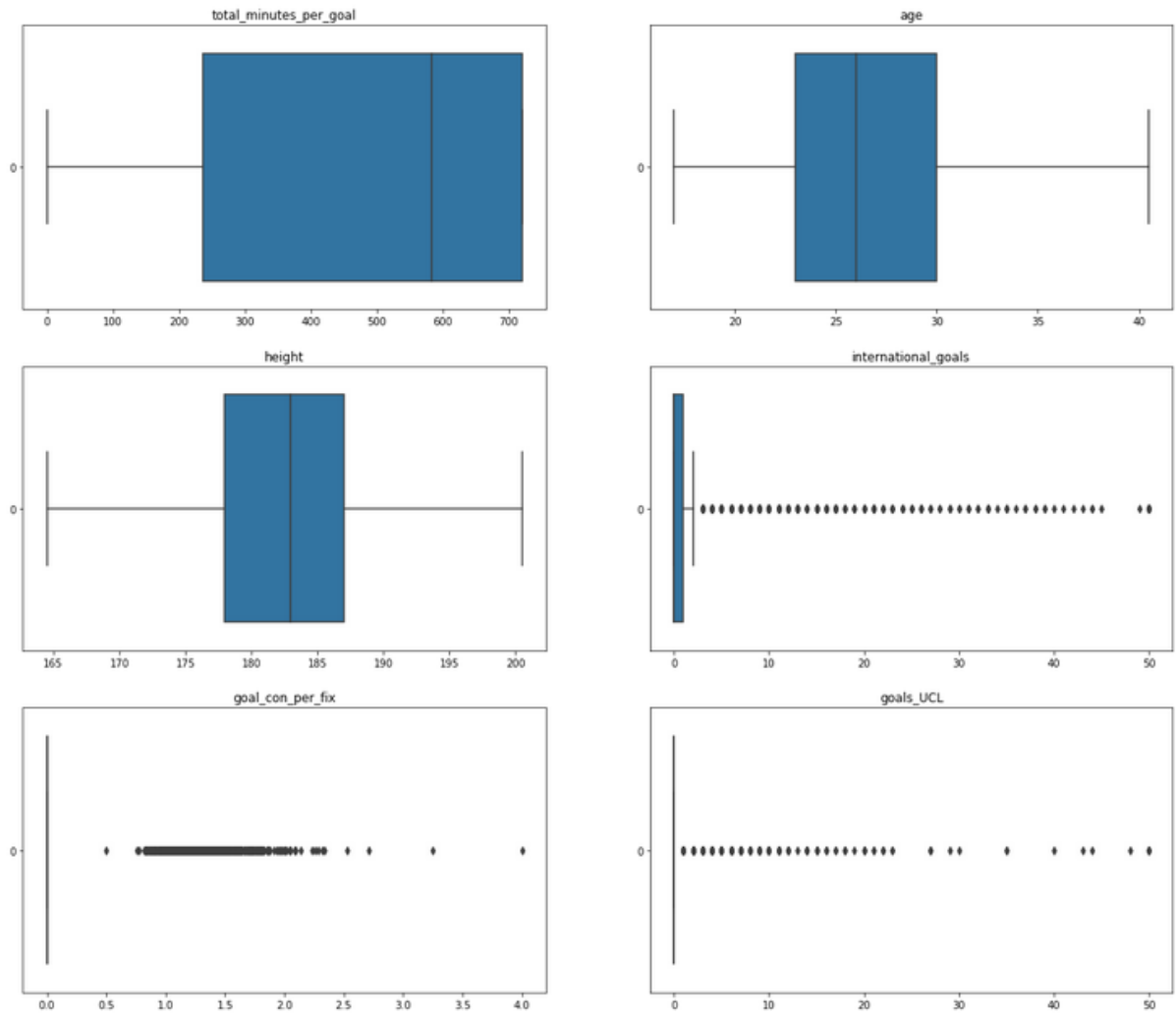
Manual Limits:

```
upper_bound = 50
df['goals_UCL'] = np.where((df['goals_UCL'] > upper_bound), upper_bound, df['goals_UCL'])

upper_bound = 120000000
df['Fee'] = np.where((df['Fee'] > upper_bound), upper_bound, df['Fee'])
```

V. Visualizing the Effect of Outlier Management Techniques on Data





Split the data

```
df = df.drop(columns='nationality')
```

```
y = df.loc[:, 'current_market_value']  
X = df.copy()  
X = X.drop(columns='current_market_value')
```

```
from sklearn.model_selection import train_test_split
```

```
X_train_val, X_test, y_train_val, y_test = train_test_split(X, y, test_size=0.1, shuffle=True)
```

```
X_train, X_val, y_train, y_val = train_test_split(X_train_val, y_train_val, test_size=0.1, shuffle=True)
```

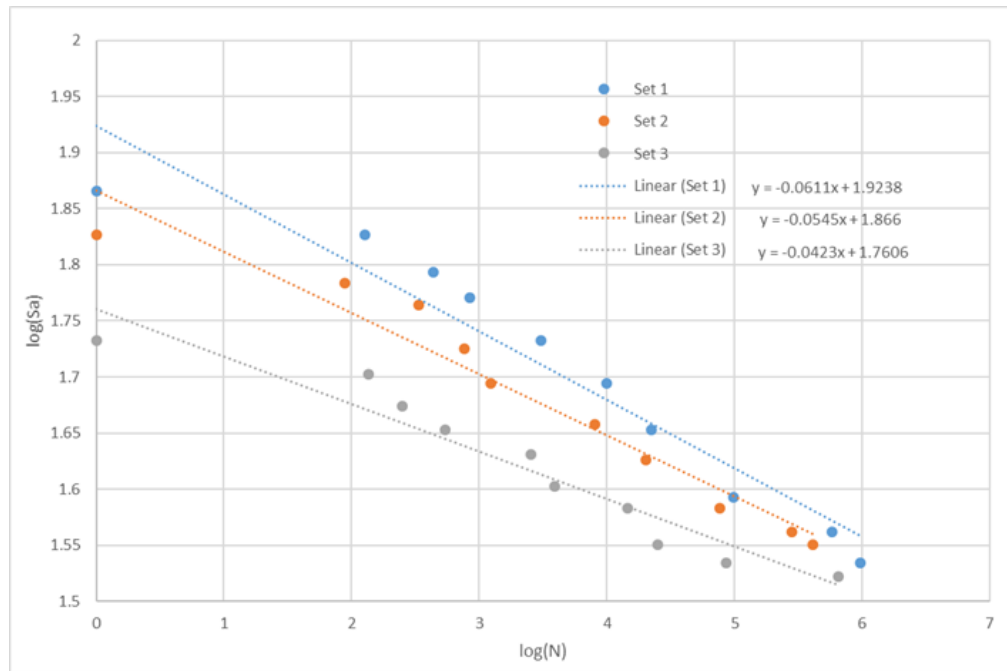
```
print(len(y_train))  
print(len(y_test))  
print(len(y_val))
```

6669

824

741

Select Model



Top 10 most expensive ever signings



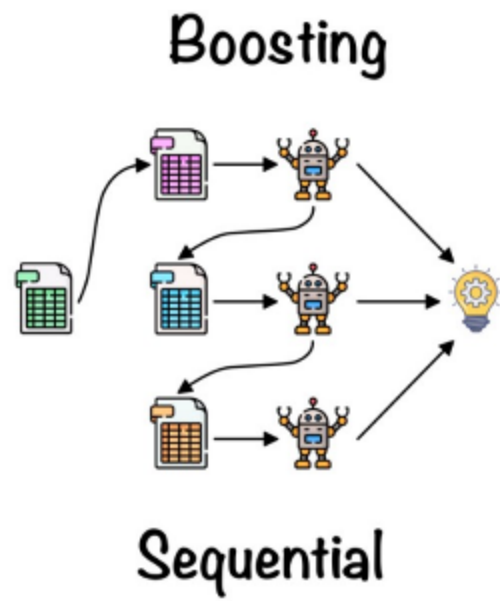
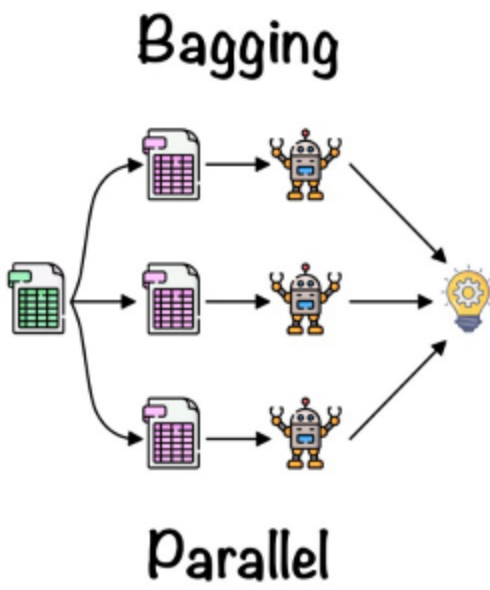
Player	Buying club	Selling club	Fee (£)	Year of transfer
Neymar	PSG	Barcelona	£198m	2017
Kylian Mbappe	PSG	Monaco	£163m	2018
Joao Felix	Atletico Madrid	Benfica	£113m	2019
Antoine Griezmann	Barcelona	Atletico Madrid	£107m	2019
Philippe Coutinho	Barcelona	Liverpool	£105m	2018
Jack Grealish	Manchester City	Aston Villa	£100m	2021
Romelu Lukaku	Chelsea	Inter Milan	£97.5m	2021
Ousmane Dembele	Barcelona	Dortmund	£97m	2017
Paul Pogba	Manchester United	Juventus	£89m	2016
Eden Hazard	Real Madrid	Chelsea	£89m	2019

TRIBUNA.COM

Most expensive deals for defenders

*Transfermarkt

1		Harry Maguire Leicester → Manchester United	€87M
2		Matthijs de Ligt Ajax → Juventus	€85.5M
3		Virgil van Dijk Southampton → Liverpool	€84.7M
4		Lucas Hernandez Atletico → Bayern	€80M
5		Aymeric Laporte Athletic → Manchester City	€65M
6		Benjamin Mendy Monaco → Manchester City	€57.5M
7		John Stones Everton → Manchester City	€55.6M
8		Aaron Wan-Bissaka Crystal Palace → Manchester United	€55M
9		Kyle Walker Tottenham → Manchester City	€52.7M
10		Éder Militão Porto → Real Madrid	€50M



Evaluating Performance of Models

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Linear regression : 0.48

Random Forest : 0.71

XGBoosting : 0.69

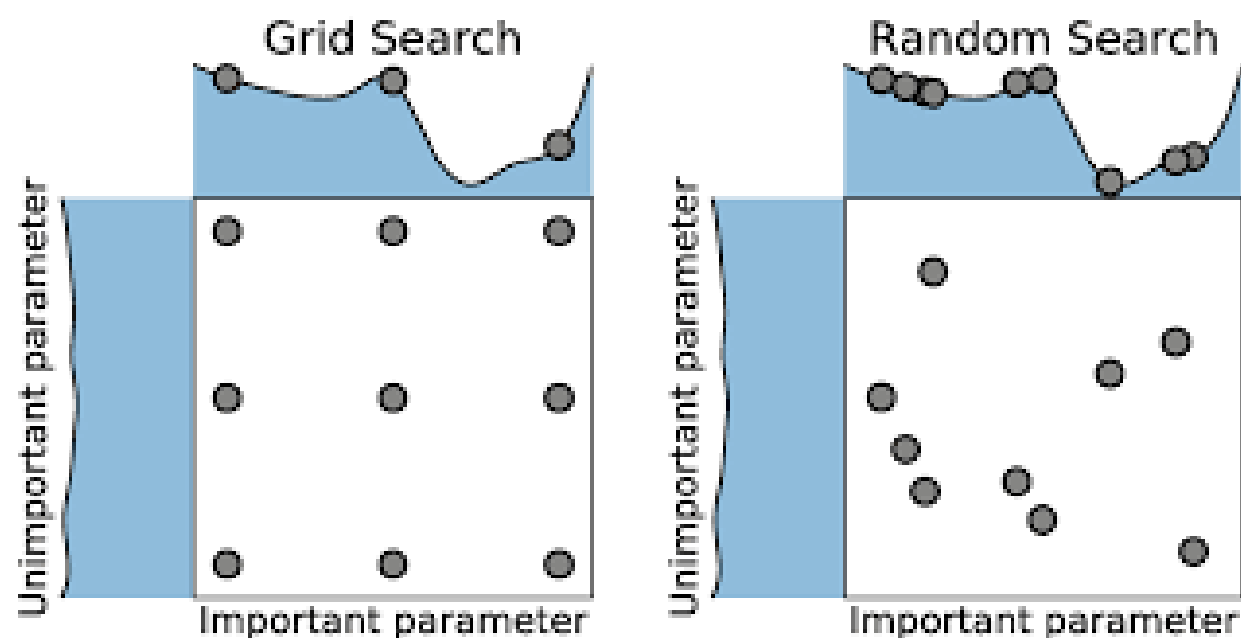
DROP : {'current_market_value', 'goals_UCL', 'assists_UCL', 'total_penalty'}

Linear regression : 0.48

Random Forest : 0.71

XGBoosting : 0.72

Tuning Hyperparameters



Gradient Boost :

Random Search

Test data:

Best Parameters: {'learning_rate': 0.09715939723698229, 'max_depth': 5,
R2 score on test data: 0.724

Val data:

R2 score on test data: 0.735

Table 8. Summary statistics of the predictions of our models on the out-of-sample test set.

Model	MAE	MAPE	R^2
xgbTree	3.60	67.47	0.77
xgbDART	3.64	68.90	0.76
glmer	4.11	69.05	0.74
glmnet	9.93	90.57	0.50
OLS	10.34	91.81	0.50