A Project on

**Football Players Rating, Wages and Value Prediction**

by

**Zakeer Ahmed Shaik**

**Akanksha Kottakota**

**Sukumar Vinnakota**

**Kushal Chavva**

**FACULTY NAME, Dr. Balaji Padmanabhan**

**University of South Florida**

**Muma College of Business**

**ISM 6251 Data Science Programming**

**16 NOV 2022**

Table of Contents

[1. INTRODUCTION 2](#_Toc119508395)

[1.1. Motivation 2](#_Toc119508396)

[1.2. Related work 2](#_Toc119508397)

[2. METHODOLOGY 3](#_Toc119508398)

[2.1 Techniques 3](#_Toc119508399)

[2.2 Evaluation Criteria 5](#_Toc119508400)

[2.3 Feature Engineering 5](#_Toc119508401)

[3. DATA 5](#_Toc119508402)

[4. MODELLING 8](#_Toc119508403)

[5. RESULTS 8](#_Toc119508404)

[6. CONCLUSION 9](#_Toc119508405)

# INTRODUCTION

## Motivation

Football is a team sport in which two teams of 11 players compete against each other on a field. Football is a popular sport played all over the world. Players belong to a football club, and these football clubs have a manager who decide who play in a certain game. Each football club has a rich history and culture and is a well-established institution. They are connected to their city, peculiar colors, fervent supporters, historical narratives, and, of course, the 11 players who compete each week on the field. They compete against clubs from their city and nation. They also compete for international titles with clubs from other countries. The players' salaries are the club's largest expense. Amortization costs are investments in transfers, and they are the second biggest expenses.

From the business point of view, it is important to know the value and wage of players. Earlier it was easy to do as the complexity of the issue was not too much. But from the past few years, when a particular player “Neymar” was sold for $270M, the market kept changing and became very unpredictable. The player who is supposed to cost $20M is now costing $60M. To understand this predictability, we would first like to predict a rating of the player after taking into consideration the player attributes such age, dribbling, defending, stamina, strength etc. After predicting the rating of a player, we decided to use this rating to predict the value of the player. Subsequently using the value of the player value, we want to predict the wage.

## Related work

* Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques

This article talks about the study and analysis of the factors that affect the market value of football players across several stages and the estimation of the value of players based on their relevant features.

In addition to the machine learning algorithms they used, we also tried to check if deep learning would be able to predict the rating and value of a player using multilayer perceptron model. This article supports the process we made use of in order to perform our analysis.

Link to the article: <https://ieeexplore.ieee.org/abstract/document/9721908>

* Computational Estimation of Football Player Wages

This study proposed an objective quantitative method to determine football players’ wages based on their skills. Machine learning algorithms were used to predict the weekly wage from variables that reflect skills and field performance of a football player.

Link to the article: <https://sciendo.com/abstract/journals/ijcss/16/1/article-p18.xml>

In addition to predicting the value and wage, we also predicted the player rating So in real time situations when the only information we have at hand are the player characteristics, our model will be able to predict the player ratings, player value using the predicted player ratings and player wage using the predicted player value.

# METHODOLOGY

The methodology that was employed in this analysis was supervised learning. Our models were allowed to train from a labeled set of data, and then these models were evaluated using data they had not seen before. The independent variables here were the player attributes and our dependent variables or our target variables are player rating, player wage and player value.

## 2.1 Techniques

**Step 1: Domain research to perform selection of required features**

The dataset that we are using has 84 player attributes. After careful study of the domain, the features that were would not have any impact on the target variables. These unnecessary attributes include name of the player, jersey number of the player, national team name, national team jersey number, etc.

**Step 2: Data Preprocessing**

The process of getting raw data ready for machine learning is known as "data preprocessing". It involves multiple steps such as data cleaning, data transformation, etc. Data cleaning is the process of eliminating or changing data that is inaccurate, incomplete, unnecessary, duplicated, or formatted incorrectly in order to prepare it for analysis.

In this process

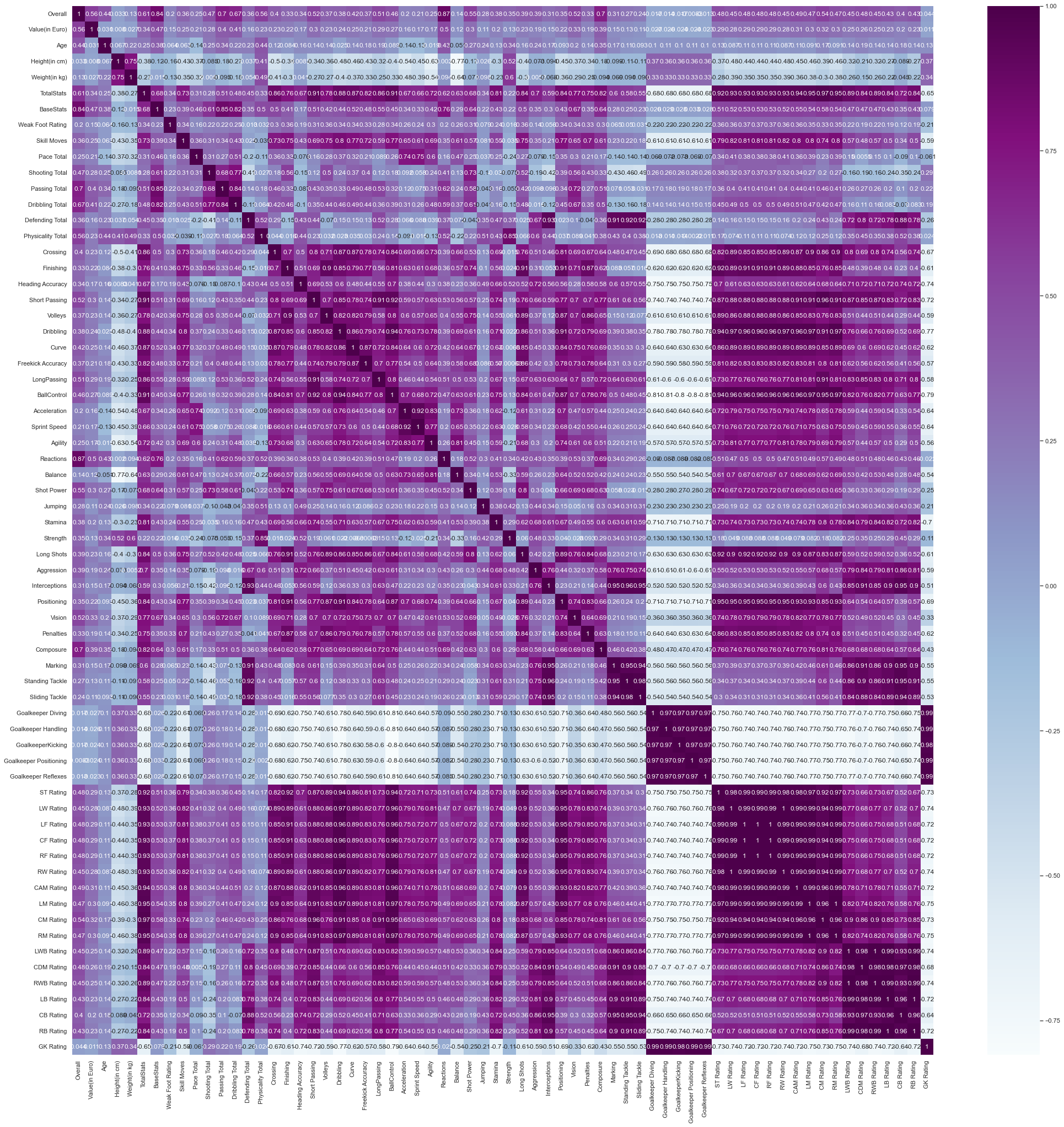
* We removed unnecessary factor columns’
* Converted categorical columns to numeric columns using one hot encoder
* Handled null values

**Step 3: Created Visualizations to understand the relationship between columns**

Data visualization allows us to view how the data is organized and what sort of relationships the qualities of the data hold. It is the quickest technique to determine whether the features match the output. We can interpret ML data with statistics by using Python recipes as a guide.

We made use of Heatmap to perform correlation analysis.

Through this we see that factors such as composure and reactions have high correlation with the overall rating of the player.



**Step 4: Data splitting**

Data splitting is done in order to avoid over-fitting. The data is split into two-part, train data and test data.

Train data is used by the model to learn, and test data is used to evaluate the model. We split the data in the 70:30 ratio of Training data to test data.

**Step 5: Training the machine learning and deep learning models**

Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so.

A set of labeled data with multiple layers or a neural network architecture are used to train deep learning models. These models must be used to train the training dataset.

**Step 6: Pipelining of models**

First, we use machine learning and deep learning models to predict the player rating. After understanding which models makes the best prediction for player rating, we predict the player rating using that model for the complete data including the test and train data. Now we include the predicted player rating as our independent variable to train our model to predict player value. Again, we use the model which predicts the player value the best to predict the player value for the whole dataset. Now we use the predicted player value to train models for player wage.

## 2.2 Evaluation Criteria

Mean Squared Error: Mean or Average of the square of the difference between actual and estimated values.

R squared: Measures how much variability in dependent variable can be explained by the model

We took into consideration the Mean Squared Error, as it is a very interpretable evaluation metric and gives a sense of the difference between the estimated value and the actual value.

## 2.3 Feature Engineering

Feature engineering uses data to generate new variables that aren't present in the training set. It improves the model accuracy as the model is trained better with useful features.

We created a new feature called contract length. This is because the wage of the player depends on the length of the contract. We had information of when the player joined and when the player’s contract is until. We used the difference of these attributes to calculate the contract length.

# DATA

Data Source: <https://www.kaggle.com/datasets/sanjeetsinghnaik/fifa-23-players-dataset>

Each row is data of individual player and his attributes.

Data Dictionary:

Known As: The player is called by

Full Name: The full name of the player

Overall: The overall rating of the player

Potential: The potential the player has

Value: The value of player in the market

Positions Played: The positions the players played at

Best Position: Players best position

Nationality: Country, the player belongs to

Age: Age of the player

Height: Height of the player

Weight: Weight of the player

TotalStats: Score of total statistics of a player

Club Name: The club the player belongs to

Wage: Wage, the player is paid

Release Clause: Amount a team supposed to pay to the former team to sign player

Club Position: Position the player plays at

Contract Until: year, in which his contract expires

Club Jersey Number: The jersey number of the player

Joined On: The year the player has joined on

On Loan: If the person is on Loan or not

Preferred Foot: If the player’s primary foot if left or right

Weak Foot Rating: rating of weak foot on scale of 1-5.

Skill Moves: Rating of skills on scale of 1-5.

International Reputation: International Reputation on scale of 1-5

National Team Position: Position played in country

National Team Jersey: Jersey number in country team

Attacking Work Rate: Categorized into High, medium, and low

Defensive Work Rate: Categorized into High, medium, and low

All the below columns are rated on a scale of 1 to 100

|  |  |  |
| --- | --- | --- |
| Pace Total | Volleys | Balance |
| Shooting Total | Dribbling | Shot Power |
| Passing Total | Curve | Jumping |
| Dribbling Total | Freekick Accuracy | Stamina |
| Defending Total | LongPassing | Strength |
| Physicality Total | BallControl | Long Shots |
| Crossing | Acceleration | Aggression |
| Finishing | Sprint Speed | Interceptions |
| Heading Accuracy | Agility | Positioning |
| Short Passing | Reactions | Vision |
| Penalties | ST Rating | LWB Rating |
| Composure | LW Rating | CDM Rating |
| Marking | LF Rating | RWB Rating |
| Standing Tackle | CF Rating | LB Rating |
| Sliding Tackle | RF Rating | CB Rating |
| Goalkeeper Diving | RW Rating | RB Rating |
| Goalkeeper Handling | CAM Rating | GK Rating |
| GoalkeeperKicking | LM Rating |
| Goalkeeper Positioning | CM Rating |
| Goalkeeper Reflexes | RM Rating |

# MODELLING

In order to train our model, different types of algorithms were used and tested. A mix of basic and advanced algorithms were tested. The algorithms used to train the model are as follows :

* Linear Regression
* MLP Regression
* XG Boost Regression
* Decision Tree Regression
* Random Forest Regression

# RESULTS

The model results for model with player rating are

A picture containing table

Description automatically generated

The model results for predicting player value is as follows:

Table

Description automatically generated

The model results for predicting player wage is as follows:

Table

Description automatically generated

According to the results, the random forest which is an ensemble method has the best results.

# CONCLUSION

We have used a data set that belongs to the calendar year 2022. We also have similar data sets from 2015 to 2021. We have ratings, wages, and values in the old data sets too. As we already said, there have been rapid changes in the wage and value of the player, as these get affected by external factors and change from time to time. So, we are planning to study the market in the future. For example, we can predict how the wage and value of a particular player are going to be in 2025, given the attributes of the player. For this, we can use time series analysis and forecast the values for the future.