# Exploration of Soccer Players' Value Based on FIFA

## Multivariate Statistical Analysis Spring 2022

Group Members: Niu Shengjie, Jian Xinyao, Mao Wenhui

**Abstract**

In this paper, we build a comprehensive analysis and evaluation of FIFA players based on factor analysis and classification fitting. At first, thirty-four continuous variables of ability description are selected as the original evaluation indicators, after factor analysis, five common factors are confirmed. We use these five common factor variables and other descriptive variables(such as height, weight) to analyze the player's value. We think player of different team positions(forward, midfield, rearguard, goalkeeper) should be significant differences in the ability description variables. Therefore, we classify the data by team positions, then do the fitting analysis separately. The result shows that the classification fitting effect is really better, indicates that players in different team positions have different emphasis ability. Hence, we also conduct the clustering model to explore the differences in the styles of players from different team positions.

*Keywords: FIFA2016-2020, Analysis of Player's Value, Factor Analysis, Classification, Fitting, Feature Selection, Clustering*

# Contents

**1 Introduction** **3**

1.1 Background . . . 3

1.2 Research Objectives . . . 3

**2 Analysis of Players' Ability Evaluation Indicators** **3**

2.1 Introduction of FIFA Dataset . . . 3

2.2 Factor Analysis . . . 4

2.3 Exploratory Analysis . . . 7

**3 Analysis of Player's Value** **8**

3.1 Analysis Model of Player's Value . . . 8

3.1.1 Introduction to Model Building . . . 8

3.1.2 Data Split . . . 9

3.1.3 Classification Model . . . 9

3.1.4 Regression Model . . . 12

3.1.5 Model Results and Evaluation Effect . . . 14

3.1.6 Feature Importance . . . 15

3.2 Analysis Model of Player's Style . . . 16

**4 Conclusion** **17**

**References** **19**

**Appendix** **20**

4.1 Table-A . . . 20

4.2 Table-B . . . 23

# 1 Introduction

## 1.1 Background

As a famous football game, FIFA Soccer has been released for nearly 30 years. Under the authorization of FIFA, the player's data in the game comes from real-world player's data. It includes the data of more than 18,000 professional players around the world. The in-depth cooperation between EA and FIFA ensures that the copyrights of players and teams can be used, and more accurate data can be assessed on players' abilities. With this data, the game can simulate how the player will perform in reality.

## 1.2 Research Objectives

Numerous variables that describe player ability make it impossible for us to make an intuitive judgment on player ability. We hope to re-propose several player ability evaluation indicators to analyze the value of players.

In this paper, factor analysis is used to reduce the dimension of data variables, and according to the results, several effective player ability evaluation indicators are proposed. Using these player ability evaluation indicators, we evaluate and fit the players' comprehensive ability. For different team positions (forward, midfield, rearguard, goalkeeper), we believe that their comprehensive evaluation capabilities have different emphases on each description. Therefore, we classify them according to the different positions of the teams, perform fittings respectively, and judge whether the classifications are effective by the fitting effect. In addition, we use a clustering model to explore how the characteristics of players in different positions differ.

Through the processing and analysis of the data of the players, this paper infers the characteristics of the players' kicking and body, and further evaluates their positions on the field, or which tactical roles the players are suitable for. Have strong practical implications for the future development of the team, and the players self-planning. This is the goal of our research.

# 2 Analysis of Players' Ability Evaluation Indicators

## 2.1 Introduction of FIFA Dataset

We crawl data from the official FIFA football website, which is linked at https://fifauteam.com/fifa-20-attributes-guide/. It contains 3961 league data of Spanish male football players from 2016 to 2020, a total of 48 variables, including name, nationality, age, height, weight, club, position, preferred foot, weak foot, skill moves, international reputation, competition, team position, overall score, and thirty-four variables of ability description such as crossing and finishing. Among them, the weak foot, skill moves and international

reputation are all ranked variables of 1-5, the team positions are divided into five categories: forward, midfield, rearguard, goalkeeper and substitute, and the overall score is the corresponding next season ( 2017-2021) score, thirty-four variables of ability description are 0-100 continuous variables, which are described clearly in appendix Table-A.

## 2.2 Factor Analysis

**•Introduction of Factor Analysis**

Fundamental of factor analysis is to divide variables into groups with correlation in order to make same group's variables have higher correlation and different groups' variables have lower correlation. Each group variable represents a basic structure explained by common factors.

Purpose of factor analysis is to reduce dimensions. Reduction method of dimensionality is to use a few potential and unobservable random variables to describe covariance relationship of original variables.

Suppose p dimensional observable random vector $\boldsymbol{x} = (x_1, x_2, ..., x_p)^{'}$ and its mean and covariance matrix are $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_p)^{'}$ and $\boldsymbol{\Sigma} = (\sigma_{ij})$. The model of factor analysis is

$$\begin{cases} x_1 = \mu_1 + a_{11}f_1 + a_{12}f_2 ... + a_{1m}f_m + \varepsilon_1 \\ x_2 = \mu_2 + a_{21}f_1 + a_{22}f_2 ... + a_{2m}f_m + \varepsilon_2 \\ \qquad\qquad ... \\ x_p = \mu_p + a_{p1}f_1 + a_{p2}f_2 ... + a_{pm}f_m + \varepsilon_p \end{cases}$$

where unobservable random variables $f_1, f_2, ..., f_m$ are common factors. $\varepsilon_1, \varepsilon_2, ..., \varepsilon_p$ are special factors. Common factors are appeared in every original variable's expression. Namely, original variables have common factors. The above formula can be expressed in matrix

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{f} + \boldsymbol{\varepsilon}$$

where $x_i = \mu_i + a_{i1}f_1 + a_{i2}f_2 + ... + a_{im}f_m + \varepsilon_i$, $Cov(x_i, f_j) = \sum_{\alpha=1}^{m} a_{i\alpha}Cov(f_\alpha, f_j) + Cov(\varepsilon_i, f_j) = a_{ij}$, $a_{ij}$ is the covariance between $x_i$ and $f_j$.

If $\boldsymbol{x}$ is random vector with standardized component variables. Correlation coefficient of $x_i$ and $f_j$:

$$\rho\left(x_i, f_j\right) = \frac{Cov\left(x_i, f_j\right)}{\sqrt{V(x_i)V(f_j)}} = Cov(x_i, f_j) = a_{ij}$$

Now, $a_{ij}$ is the correlation coefficient between $x_i$ and $f_j$.

$$h_i^2 = \sum_{j=1}^{m} a_{ij}^2, i = 1, 2, .., p$$

$h_i^2$ reflects the influence of common factors to $x_i$, namely variance contribution of common factor $f_1, f_2, ..f_m$ to $x_i$, called **communality variance**. $\sigma_i^2$ is variance contribution of special factor $\varepsilon_i$ to $x_i$, called **specific variance**.

$$g_j^2 = \sum_{i=1}^{p} a_{ij}^2, j = 1, 2, .., m$$

$g_j^2$ reflects the influence of common factor $f_j$ to $x_1, x_2, ...x_p$, which is an important measure for $f_j$. $g_j^2$ can be explained as total variance contribution of $f_j$ to $x_1, x_2, ...x_p$.

$\sum_{j=1}^{m} g_j^2$, reflects the influence of common factors $f1, ....., f_m$ to $x_1, x_2, ..., x_p$, can be explained as accumulated contribution of $f_1, ...f_m$ to $x_1, ..., x_p$. Thus, only a few common factors are selected to make the cumulative contribution rate reach a relatively high level (e.g. 80%), so as to achieve the purpose of dimensionality reduction.

**•Selection and Analysis of Common Factors**

Suppose there are n player data, each player has p ability description values, and each ability description value of the i-th player is $x_{i1}, x_{i2}..., x_{ip}, i = 1, 2, ..., n$, then the observed data matrix, $\boldsymbol{X} = (x_{ij})_{p\times n}$, can be expressed as

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{21} & . & . & . & x_{n1} \\ x_{12} & x_{22} & . & . & . & x_{n2} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ x_{1p} & x_{2p} & . & . & . & x_{np} \end{bmatrix}$$

Let $\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, j = 1, ..., p$ be the sample mean of the j-th index of these n funds, $s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2$ be the sample variance. In order to ensure the unity of data magnitude, we standardize the original data by $z_{ij} = (x_{ij} - \overline{x}_j)/\sqrt{s_j}$.It is obvious that the covariance matrix of the standardized sample is exactly the sample correlation matrix of original data:

$$\boldsymbol{R} = \begin{bmatrix} r_{11} & r_{12} & . & . & . & r_{1p} \\ r_{21} & r_{22} & . & . & . & r_{2p} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ r_{p1} & r_{p2} & . & . & . & r_{pp} \end{bmatrix}$$

where $r_{kl} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ik} - \overline{x}_k)(x_{il} - \overline{x}_l)/s_k s_l (k = 1, ..., p, l = 1, ..., p)$ represents the correlation between the k-th and l-th samples with $r_{kk} = 1$.

Starting from the correlation matrix $\mathbf{R}$, we obtain the eigenvectors by solving the characteristic equation $|\lambda \boldsymbol{I} - \boldsymbol{R}|$, denoted as $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq ... \geq \widehat{\lambda}_1$. When $\frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{p} \lambda_j} > 80\%$, we can conclude that these m common factors can explain original p variables.

We perform dimensionality reduction on thirty-four 0-100 continuous variables from "crossing" to "reflexes".

The gravel chart can give polylines of the eigenvalues from large to small, and the recommended common factor size. We have drawn the gravel plot (both) of principal component analysis and factor analysis, as follows :

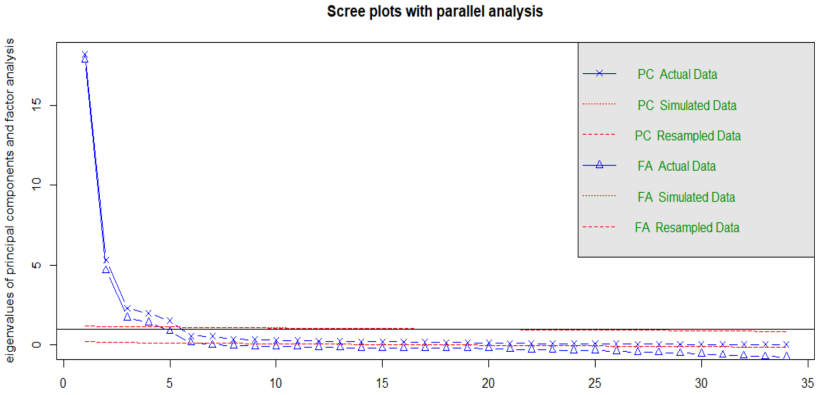



Figure 1: Scree Plots with Parallel Analysis

From the results of the gravel plot, it can be seen that taking 5 common factors can explain most of the variance.

From the orthogonal unrotated model, its factor analysis diagram(figure 2) show the interpretation of the five common factors is not good. Then we use the principal axis iteration method to construct the oblique factor model (we think that the common factors are related), and rotate the factor loading matrix (rotation=varimix). Then we obtain the corresponding factor analysis diagram(figure 3). According to this result, we can get five common factors with good interpretation.

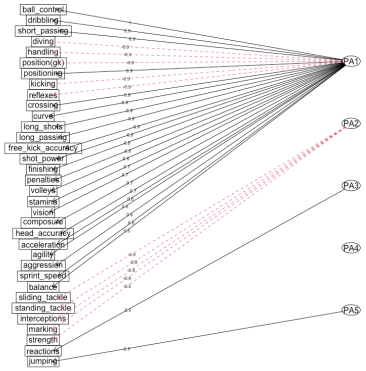



Figure 2: Factor Analysis of Orthogonal Unrotated Model

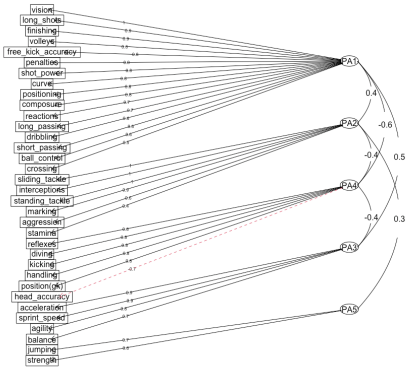



Figure 3: Factor Analysis of Oblique Rotated Model

Accordingly, we redefine the five common factors as:

- **Shooting_Passing** : it describes the shooting and passing skills.
- **Defend** : this focuses on the player's defensive ability.
- **Goalkeeping** : it contains some goalkeeping's special attributes.
- **Pace** : describes the speed of a player.
- **Physical** : non-skillful, the player's own qualities such as strength, jumping.

The descriptive variables of player ability reflected by the five common factors are listed in appendix Table-B.

## 2.3 Exploratory Analysis

**•Physical Fitness Indicators**

Three physical fitness variables of age, height and weight of each player are given in the dataset. The range of age is 16-40. It can be seen from the histogram (figure 4) that the age of players is mostly 20-30 years old; the range of height is 162-200 (cm)(figure 5), it can be inferred that the height of the player roughly follows a normal distribution; the weight range is 120-215. From the histogram (figure 6), it can be inferred that the weight of the player roughly obeys a normal distribution.

We believe that the player's age, height and weight reflect the player's physical quality and will affect the player's overall score, so we add these to the indicators of the player's value analysis.

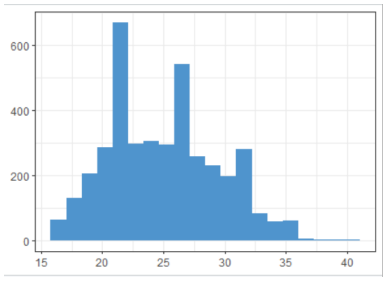



Figure 4: Histogram of Age

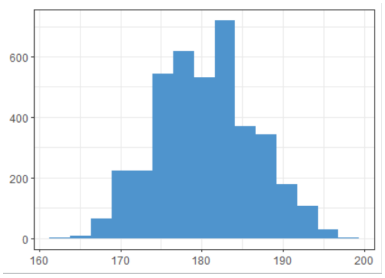



Figure 5: Histogram of Height

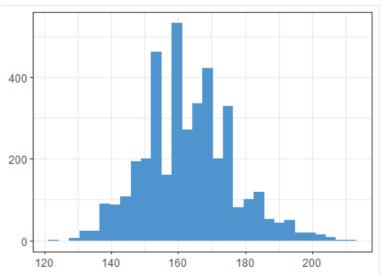



Figure 6: Histogram of Weight

**•Other Indicators**

The dataset also includes three rating variables ranging from 1 to 5, namely weak foot, skill moves and international reputation. Through one-way ANOVA, we believe that these three variables have significant impact on the players' overall score.

The data set contains both position and team position. The classification of team position covers the refined classification of position. We are more concerned about the difference between players in different team positions and do not require detailed positions, so we consider the team position. The classification of team_position is as follows (Table 1). It has been verified (one-way ANOVA) that team position has a significant impact on the players' overall score.

| backcourt | frontcourt | goalkeeper | substitute | midfield |
|---|---|---|---|---|
| 636 | 241 | 164 | 2181 | 739 |

Table 1: Table of team position

# 3 Analysis of Player's Value

## 3.1 Analysis Model of Player's Value

### 3.1.1 Introduction to Model Building

**•Purpose: predict player ability and build player evaluation model**

**•Idea: classification + prediction**

Our goal is to establish a player evaluation model, that is, to evaluate the comprehensive ability of players through their performance in various aspects. However, in a football team, for players in different team positions, in order to make the greatest contribution, their comprehensive abilities must have their own emphasis. For example, a good goalkeeper will inevitably deviate from the ability required by a good striker.

Therefore, establishing the same regression model for all players may not be able to grasp the deviation in this regard. We believe that establishing separate regression models for players in different positions will obtain a more accurate prediction effect. The team positions can be roughly divided into four categories: forward, midfield, rearguard and goalkeeper, and can be subdivided into 25 categories such as left forward, center forward, and right forward. Therefore, we divide the evaluatation model into two parts. The first part is the classification model, which judges the player as one of the four team positions of goalkeeper, frontcourt, midfield, and backcourt; the second part is the regression model, make prediction for players separately and evaluate their overall score.

### 3.1.2 Data Split

**•Dataset Division Methods & Indicators Description (y and x)**

We divide the original dataset into two parts: training set and test set. The data of the training set consists of all non-substitute players, with a total of 1780 samples, accounting for 44.93% of the total data; the data of the test set consists of all substituted players, with a total of 2181 samples, accounting for 55.06% of the total.

Both classification and regression models are trained on the training set. The corresponding variable of the classification model is "team position", and the response variable of the regression model is "overall score". For the two models, we select three categorical variables of players' basic physical information (age, height and weight), weak foot, skill moves, international reputation and five common factor variables obtained by dimensionality reduction as explanatory variables.

### 3.1.3 Classification Model

Based on the non-substitute players train dataset (1780 observations), we train the support vector machine (SVM), random forest (RF), recursive segmentation tree (RPART), k-nearest neighbor (KNN), Adaboost, and artificial neural network (ANN) six classification models.

**•Support Vector Machine (SVM)**

SVM is a kind of generalized linear classification that performs binary classification on data according to supervised learning, and its decision boundary is the maximum margin hyperplane that solves the learning samples. Given a set of training instances, each marked belonging to one or the other class, the SVM training algorithm builds a model that assigns new instances to one of the two classes, making it a non-probabilistic two Binary Linear Classifier. The SVM model is to represent instances as points in space such that the mapping makes instances of different classes separated by as wide a noticeable interval as possible. Then,

map the new instances to the space and predict the class they belong to based on which side of the interval they fall on.

**•Decision Tree (DT)**

A decision tree is a type of supervised machine learning used to categorize or make prediction based on how the previous set of questions were answered.

The classification tree predicts the classification result based on one or more input variables and combined with the division conditions. The splitting process starts from the root node of the classification tree: at each node, the algorithm will check whether the input variable needs to be divided intermittently to the left cotyledon and the right cotyledon according to the division conditions. Stop splitting when a child node (end point) of any classification tree is reached.

**•K-Nearest Neighbor (KNN)**

The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. For k-NN classification, an input is classified by a majority vote of its neighbors. That is, the algorithm obtains the class membership of its k neighbors and outputs the class that represents a majority of the k neighbors.

**•Random Forest (RF)**

In machine learning, a random forest is a classifier that consists of multiple decision trees, and its output category is determined by the mode of the categories output by the individual trees.

Lin and Jeon in 2002 pointed out the relationship between the random forest algorithm and the K-nearest neighbor algorithm (k-NN). It turns out that both algorithms can be viewed as so-called "weighted neighbor schemes". These are in the dataset $\{(x_i, y_i)\}_{i=1}^{n}$ computes a prediction $\hat{y}$ for a new point x by looking at the point's neighbors, and these neighbors are weighted using the weight function W.

**•Adaptive Boosting(Adaboost)**

AdaBoost, short for Adaptive Boosting, is a statistical classification meta-algorithm formulated by Yoav Freund and Robert Schapire, who won the 2003 Gödel Prize for their work. It can be used in conjunction with many other types of learning algorithms to improve performance. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier. AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers. In some problems it can be less susceptible to the overfitting problem than other learning algorithms. The individual learners can be weak, but as long as the performance of each one is slightly better than random guessing, the final model can be proven to converge to a strong learner.

AdaBoost is a very popular boosting technique that aims at combining multiple weak classifiers to build one strong classifier.

**•Artificial Neural Network (ANN)**

ANN refers to a complex network structure formed by interconnecting a large number of processing units (neurons), and is a kind of abstraction, simplification and simulation of the structure and operation mechanism of the human brain. Artificial Neural Network (ANN), which simulates neuron activity with a mathematical model, is an information processing system established based on imitating the structure and function of the neural network of the brain.

The artificial neural network is divided into multi-layer and single-layer, each layer contains several neurons, and the neurons are connected by directed arcs with variable weights. Change the weight of neuron connection to achieve the purpose of processing information and simulating the relationship between input and output. It does not need to know the exact relationship between input and output, and does not need a large number of parameters, but only needs to know the non-constant factors that cause output changes, that is, non-quantitative parameters. Therefore, compared with traditional data processing methods, neural network technology has obvious advantages in processing fuzzy data, random data, and nonlinear data, and is especially suitable for large-scale, complex and unclear information systems.

**Classifiers' effect comparison:**

It can be seen from the Table 2 below that KNN and ANN may have over-fitting, while SVM and RF have good accuracy on the test set although the training set is not particularly effective. Adaboost not only achieves an astonishing 1 accuracy on the training set, but also has a classification accuracy that is not inferior to SVM and RF on the test set.

| | Accuracy(test) | ROC | Accuracy(train) |
|---|---|---|---|
| DT | 0.7734082 | 0.7557 | 0.8113965 |
| KNN | 0.8670412 | 0.8327 | 0.9686998 |
| SVM | 0.9044944 | 0.8788 | 0.9012841 |
| RF | 0.8932584 | 0.8709 | 0.8796148 |
| ANN | 0.8501873 | NA | 0.9558587 |
| Adaboost | 0.8988764 | NA | 1 |

Table 2: Classifiers' effect comparisone

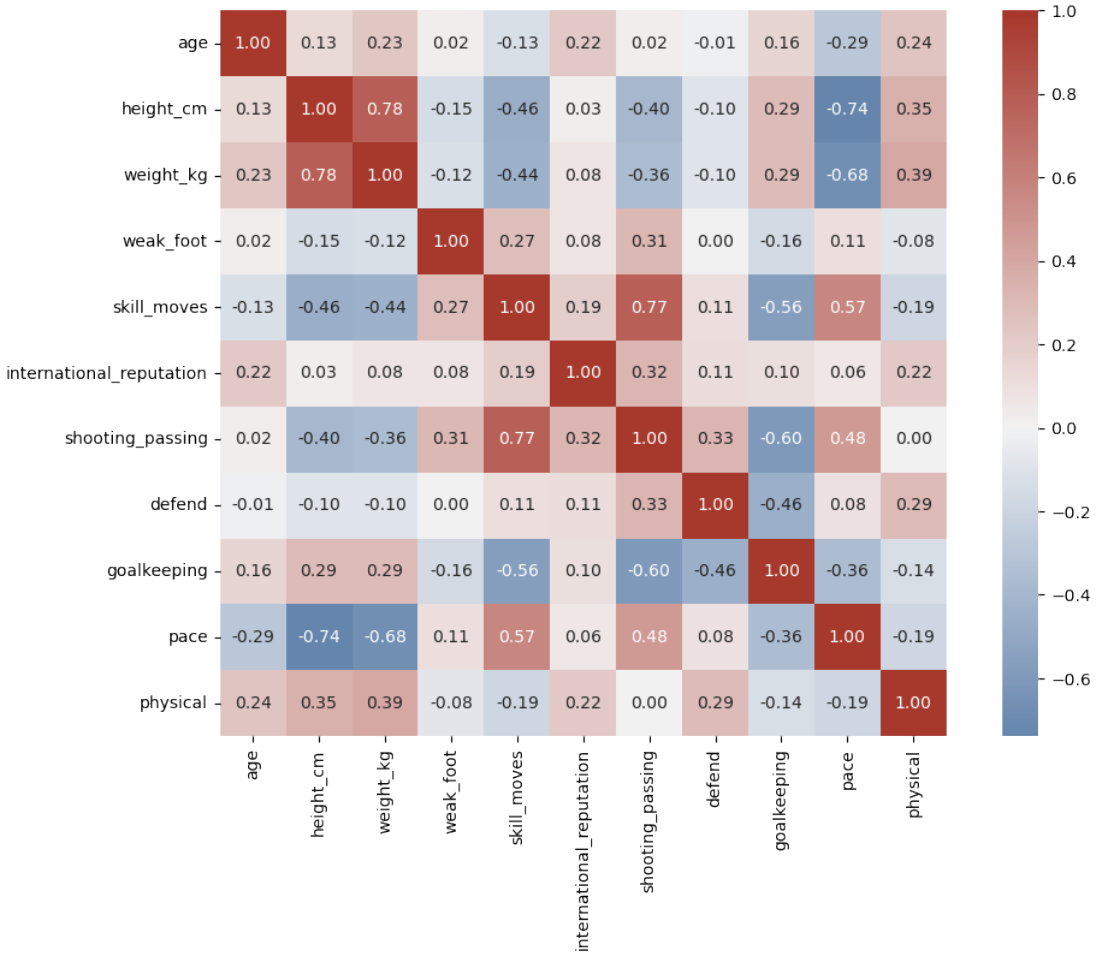



Figure 7: heatmap of variables

### 3.1.4 Regression Model

**•Linear Regression**

In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

**• Ridge Regression**

Multicollinearity in regression refers to the case when one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy.

From the figure 7, we can find that there exist the significant correlations between pairs of independent

variables in the dataset, So we have good reason to check for multicollinearity.

We can check the tolerance or the variance inflation factor (tolerance $< 0.1$ or VIF $> 10$) to detect the multicollinearity.

| | VIF | Tolerance | | VIF | Tolerance |
|---|---|---|---|---|---|
| age | 51.4602 | 0.0194 | shooting passing | 3.8901 | 0.2571 |
| height | 531.4928 | 0.0019 | defend | 1.5826 | 0.6319 |
| weight | 435.4220 | 0.0023 | goalkeeping | 2.6258 | 0.3808 |
| weak foot | 32.1762 | 0.0311 | pace | 1.9446 | 0.5142 |
| skill moves | 39.1301 | 0.0256 | physical | 1.6912 | 0.5913 |
| international reputation | 7.4249 | 0.1347 | | | |

And we can see that multicollinearity exists for “age”, “height”, “weight”, “weak_foot” and “skill_moves” . Existence of multicollinearity will lead some terrible consequences:

(1) Difficult to test individual regression coefficients due to inflated standard errors.

(2) Unstable coefficient estimates, sensitive to small change in the model.

So next we try to use ridge regression to alleviate this problem. Ridge Regression is an adaptation of the popular and widely used linear regression algorithm. It enhances regular linear regression by slightly changing its cost function, which results in less overfit models.

Ridge regressionis a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization.

• **Desicion Tree(GBDT,XGBoost)**

**GBDT (Gradient Boosting Decision Tree)** works well in data analysis and prediction. It is an ensemble algorithm based on decision tree. Among them, Gradient Boosting is an algorithm in the integrated method boosting, which iterates the new learner through gradient descent. We use CART (regression tree with gain with gini index) as base classifier. It can handle various types of data flexibly, and can use some robust loss functions, which is more robust to outliers.

**XGBoost** can be regarded as an implementation of an upgraded version of the GBDT algorithm (second-order Taylor expansion is performed, and second-order gradient information is used).

XGBoost adds a regular term to the cost function to control the complexity of the model. The regular term includes the number of leaf nodes of the tree, etc. The regular term reduces the variance of the model,

making the learned model simpler and preventing overfitting, which is also a feature of XGBoost superior to traditional GBDT. In addition, XGBoost can also carry out parallel learning.

#### 3.1.5 Model Results and Evaluation Effect

In order to compare the quality of different models, we use MSE to compare the quality of different combinations of classification models and regression models. At the same time, as a comparison, we also used the same process to directly train the regression model on the no classifier. The calculation formula and results are as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{4}\sum_{j=1}^{n_i}(y_{ij} - \widehat{y}_{ij})^2$$

| | SVM | Decision Tree | Random Forest | KNN | Adaboost | ANN | No Classifier |
|---|---|---|---|---|---|---|---|
| Linear Regression | 9.94 | 16.42 | 9.85 | 9.64 | 9.78 | 12.88 | 25.83 |
| Ridge Regression | 10.20 | 16.74 | 10.11 | 9.91 | 10.03 | 12.89 | 25.84 |
| GBDT | 11.45 | 12.73 | 11.34 | 11.25 | 11.35 | 10.77 | 17.14 |
| XGBoost | 9.34 | 9.67 | 8.74 | 9.25 | 8.69 | 8.60 | 12.54 |

Table 3: MSE of different models

Comparison of the classification models shows that the classification effect of decision trees is less suitable compared with other models, Adaboost can achieve the lowest MSE after classification.

A comparison of the regression models shows that XGBoost obtains a relatively low MSE after each classifier.

In general, we can see that the best result can be achieved by combining the ANN classification model with the XGBoost regression model.

For the unclassified model XGBoost achieves a low MSE, but most of the models after classification achieve this level. This indicates that our proposed model does have a very large improvement in prediction accuracy and validates the reasonableness of the model.

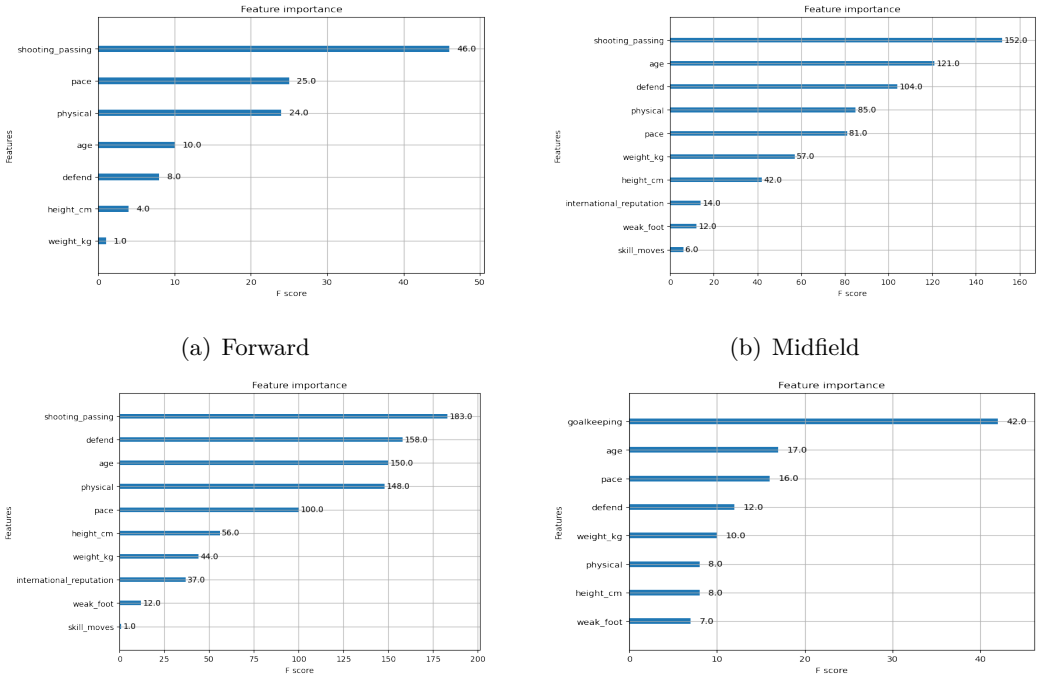
Then in each row of the table 4, the ratio of the combined model to the direct regression model is calculated using the regression results without classification as the benchmark. At this point when the ratio is more than 1, the model effect is stronger than the effect of the direct regression model.

| | SVM | Decision Tree | Random Forest | KNN | Adaboost | ANN |
|---|---|---|---|---|---|---|
| Linear Regression | 0.38 | 0.64 | 0.38 | 0.37 | 0.38 | 0.50 |
| Ridge Regression | 0.39 | 0.65 | 0.39 | 0.38 | 0.39 | 0.50 |
| GBDT | 0.67 | 0.74 | 0.66 | 0.66 | 0.66 | 0.63 |
| XGBoost | 0.74 | 0.77 | 0.70 | 0.74 | 0.69 | 0.69 |

Table 4: ratio of MSE compared to No Classifier

### 3.1.6 Feature Importance

In addition to our best regression model, XGBoost and ANN classifiers, which allow us to predict a player's overall score, we further explore which indicators need to be taken into account when selecting players for different positions.




(a) Forward

(b) Midfield

(c) Rearguard

(d) Goalkeeper

Figure 8: Feature Importance

- **Forward**:The most important thing for players in the Forward is the ability to shoot and pass the ball. Speed and physicality are also very important.
- **Midfield**: For midfielders, the ability to shoot and pass is as important as the ability to defend. Age,

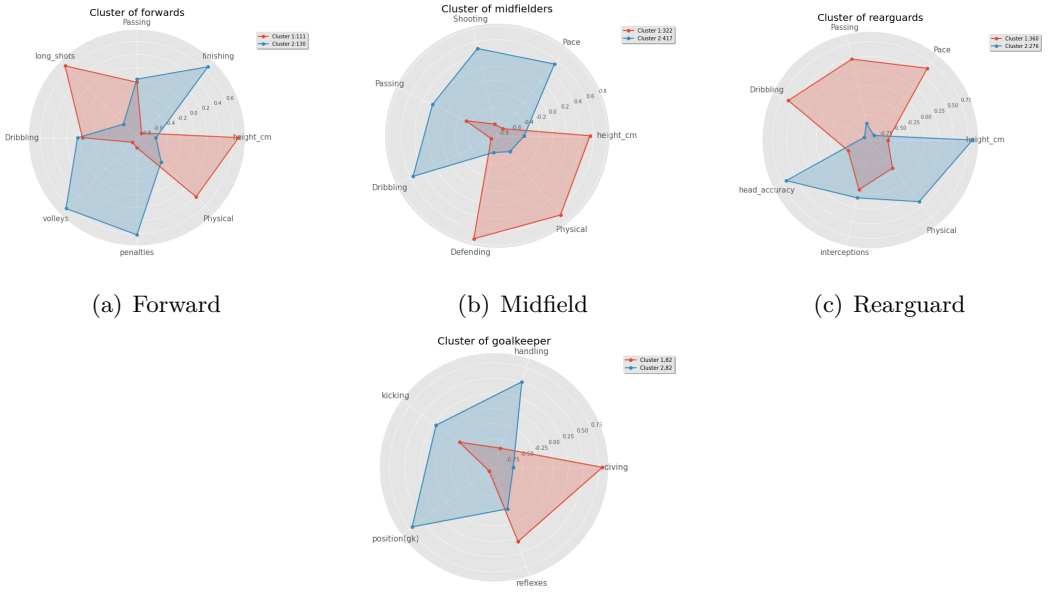
speed and physicality are also important indicators, just like in the front.

- **Rearguard** :For the Rearguard players, the most important abilities are passing, defending and speed, while age and fitness are also very important factors. Height, weight, international reputation and the ability to play against the foot are not as important.
- **Goalkeeper**: The most important thing for a goalkeeper is naturally the ability to keep the ball. Other aspects of ability are relatively average.

Overall, the ability to pass and shoot is important at all positions except goalkeeper. Physicality and age are more important at each position. The importance to offensive and defensive ability varies from position to position.

### 3.2 Analysis Model of Player's Style

Even though players in the same position have some similarity, they can have different style characteristics. Therefore we clustered for four positions: forwards, midfielder, rearguard and goalkeeper. For each position, we select features that are highly relevant to that position based on our experience, and use the **k-means** model to cluster into 2 classes respectively. The clustering results were obtained as shown in the radar chart below, with blue and red representing the performance of the two clustered classes on different indicators.




(a) Forward (b) Midfield (c) Rearguard

(d) Goalkeeper

Figure 9: Clustering Result

- **Forward**:(divided into center and winger)
  - **BLUE** These players are stronger in volleys, shot power, and penalties, and are better suited as wingers, playing as one-touch players for the team.
  - **RED** This group of players is better at long range shooting, taller and more physical, more suitable for the position of winger.
- **Midfield**: (divided into offense and defense) Midfield is an area where both offense and defense are important, so the players in this position will be divided into defensive and offensive types of players accordingly.
  - **BLUE** These players leans towards the offensive side of the game, they are fast, shoot well and are also good at dribbling and passing.
  - **RED** This category of players is more adept at defending, they are tall, physical, have a good sense of pressure and strong defensive ability.
- **Rearguard**: (divided into sides and backs)
  - **BLUE** These players are taller, have a strong physical presence, are accurate with their headers, and are better suited in the middle of the backfield.
  - **RED** These players are shorter, but faster, stronger on the ball and in the passing game, and better suited at the wing-back position.
- **Goalkeeper**: The goalkeeper's position is rather specific, so only the five attributes related to the goalkeeper are used in the clustering. relatively average.
  - **BLUE** This type of goalkeeper has solid basic skills, plays steadily and has a good sense of position and hand shape. They also have certain striking attributes and are good at driving the ball to assist their teammates in attack
  - **RED** This type of goalkeeper is more agile and better at pouncing on the ball. They may be able to save the day in some important games, but their play may not be consistent.

# 4 Conclusion

**How should the ability of soccer players be predicted and evaluated?**

In order to predict and evaluate players' ability, we firstly obtained five player ability evaluation indexes ("shooting_passing", 'defence", "goalkeeping", "space", "physical") based on the factor analysis.

Secondly, we propose a "classification + regression" value analysis model, which first assigns a player to a specific position in the team (forward, midfield, rearguard and goalkeeper) by classification, and then uses the regression model corresponding to that position to predict his ability. The MSE fitted by this model in the test set showed a substantial decrease compared to that of the direct regression without classification. The best combination model accounted for 60-70% of the original MSE, while linear regression and ridge regression were able to achieve 30%-40% of the MSE of the original model after classification. We try six classification models (svm, randomForest, recursive split tree, KNN, ANN, Adaboost) and four regression models (linear regression, ridge regression, GBDT, XGBoost), where the best prediction was the combination of ANN classifier and XGBoost regression.

**What type of indicators should be focused on when selecting a certain type of player?**

Based on the optimal model, we further explore the characteristics that are most important for each position. For the forwards, the most important is the ability to shoot and pass, the midfielder is both offensive and defensive, and the rearguard is more defensive, cooperative and passing oriented. The goalkeeper is naturally the ability to keep the ball. And for each position, physicality and age are of some importance.

**How to choose the right player?**

In addition to the prediction and evaluation of players' abilities, we also used the k-means clustering algorithm to cluster the players in each position separately. According to the results of clustering, both forward and rearguard can be divided into two types: middle (good physical quality) and sides (fast), midfield into offense and defense, while goalkeeper can be divided into two types: stable technique and flexible reaction. Accordingly, by judging the types of players according to their characteristics and placing them in the proper positions, they can best bring out their abilities and thus improve the overall level of the team.

## References


[1] Christopher M. Bishop. (2006). Pattern Recognition and Machine Learning(PRML).

[2] Robert I. Kabacoff (2011). R in Action, Data Analysis and Graphics with R.

[3] Guo-Liang Tian, Xue-Jun Jiang (2020). Mathematical Statistics.

[4] Olah, C. (2015). Understanding lstm networks.

[5] Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 270(2), 654-669.

[6] Paul Wilmott(2019). Machine Learning An Applied Mathematics Introduction.

[7] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2013). An Introduction to Statistical Learning.

[8] Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(8):1979–1993, 2018.

(2018). Single event sensitivity analysis of bandgap reference. Diànzǐ Jìshù Yīngyòng, 44(12)

[9] Chih-Chung, Chang, Chih-Jen, Lin. (2011). LIBSVM: A library for support vector machines.10.1145/1961189.196

[10] Ren, M., Zeng, W., Yang, B., and Urtasun, R. Learning to reweight examples for robust deep learning. In Proceedings of the 35th International Conference on Machine Learning, pp. 4331–4340, 2018.

[11] https://fifauteam.com/fifa-20-attributes-guide/

[12] https://www.bilibili.com/read/cv14271986

# Appendix

## 4.1 Table-A

| Name | Description |
|---|---|
| Crossing | This attribute measures how accurately the player crosses the ball during both normal running and free kick set pieces. |
| finishing | Finishing is the accuracy of shots using foot, inside the penalty area. |
| head_accuracy | This stats measures the heading accuracy of the player for either a pass or a shot. |
| short_passing | This attribute ranks how well a player performs a short / ground pass to his teammate. In other words, it determines a player' s accuracy and speed of passing over a short distance. |
| volleys | This attribute measures the accuracy and power of volleys at goal. It affects the technique and accuracy of shots taken while the ball is in the air. This tends to be coupled with the balance trait if he is not fully facing the goal. |
| dribbling | Dribbling is the player' s ability to carry the ball and past an opponent. A higher value means the player will be able to keep better possession of the ball whilst dribbling because he will keep the ball closer, making it harder for the opponent to win it off of him. |
| curve | Curve is used to measures the player' s ability to curve the ball when passing and shooting. The higher the value the more curve/curl the player is capable of putting on the ball. |
| free_kick_accuracy | Free Kick is a FIFA attribute used to measures the player' s accuracy for taking Free Kicks. The higher the value the better the accuracy of a direct free kick on goal. |
| long_passing | This stat is used to classify how well a player performs a long pass in the air to his teammate. It doesn' t affect long ground passes. |
| ball_control | Ball control is the ability of a player to control the ball as he receives it. The higher the value, the less likely the ball is to bounce away from the player after controlling it. |
| sprint_speed | Sprint speed measures how fast the player runs while at top speed. |

| | |
|---|---|
| acceleration | Acceleration is the increment of a player' s running speed. The higher the value, the shorter the time needed to reach maximum speed, no matter what that is. |
| agility | Agility measures how agile the player is while moving or turning. In other words, how fast and graceful a player is able to control the ball. |
| reactions | Reactions measures how quickly a player responds to a situation happening around him. |
| balance | Balance attribute is the ability to maintain balance after a physical challenge. |
| shot_power | Shot Power evaluates how hard the player hits the ball when taking a shot at goal. It is the amount of power a player can put into a shot while still keeping it accurate. |
| jumping | Jumping is the player' s ability and quality for jumping from the surface for headers. The higher the value is, the higher the player can jump. |
| stamina | Stamina determines the rate at which a player will tire during a game. It evaluates how tired your player gets as the match approaches half time or full time. |
| strength | Strength is about the quality or state of being physically strong. The higher the value, the more likely the player will win a physical challenge. |
| long_shots | This attribute measures the accuracy of shots from outside the penalty area. |
| aggression | The aggression level of a player measures the frequency and the aggression of jostling, tackling and slide tackling. It is the attribute which determines the player' s power of will or commitment to a match. |
| interceptions | Interception determines the ability to read the game and intercept passes. |
| positioning | Positioning is the player' s ability to take up good positions on the field during a game. The higher this stat, the more likely a player is to make enough space to receive the ball in dangerous areas. |
| vision | Vision ranks the player' s awareness of the position of his teammates & opponents around him. It is the attribute that increases (or reduces) the possibilities of a successful long pass. |
| penalties | This attribute measures the accuracy of shots from inside the penalty area. |

| | |
|---|---|
| composure | This attribute determines at what distance the player with the ball starts feeling the pressure from the opponent. This then affects the chances of the player making an error when he shoots, passes and crosses. The higher the value, the better the player performs when under pressure from an opponent. |
| marking | Marking is the ability to track and defend an opposing player. In other words, it is your player' s ability to stay close to an opposing attacker and stop him getting to a cross/pass from a teammate. Also contributes to tracking runs. |
| standing_tackle | This stats measures the ability of the player to time sliding tackles so that they win the ball rather than give away a foul. |
| sliding_tackle | This stats measures the ability of the player to time sliding tackles so that they win the ball rather than give away a foul. |
| handling | Handling is an exclusive goalkeeper attribute used to measures how cleanly he catches the ball and does he hold on to it. |
| diving | Diving is GK' s ability to make a save whilst diving through the air. It is directly affected by the player' s height. |
| kicking | Kicking it' s another attribute that only goalkeepers have, used to measures the length and accuracy of goal kicks, from out of the hands or on the ground. |
| position(gk) | Positioning for goalkeepers is slightly different. It is the GK' s ability to position himself correctly when saving shots. It also affects the way how a goalkeeper reacts to crosses. |
| reflexes | Reflexes stat is the agility of the goalkeeper when making a save. |

### 4.2 Table-B

| Factor | Variables | Factor | Variables |
|---|---|---|---|
| Shooting_ Passing | vision | Defend | sliding_tackle |
| | long_shots | | interceptions |
| | finishing | | standing_tackle |
| | volleys | | marking |
| | free_kick_accuracy | | aggression |
| | penalties | | stamina |
| | shot_power | goalkeeping | reflexes |
| | curve | | diving |
| | positioning | | handling |
| | composure | | kicking |
| | reactions | | position(gk) |
| | long_passing | | head_accuracy |
| | dribbling | Pace | acceleration |
| | short_passing | | sprint_speed |
| | ball_control | | agility |
| | crossing | | balance |
| Physical | jumping | | |
| | strength | | |