Statistical Analysis on How to Play FIFA Strategically

Xiyu Chen, Dongru Jia, Ning Hu, Ruizhe Li

1. Introduction

This project report aims to answer the question of "How to play FIFA 19, a soccer simulation video game, strategically?", by taking the approaches of statistical learning models. But before diving right into any analysis and model, what the game, FIFA 19, is? FIFA is a soccer simulation video game that uses various attributes and an overall score to mimic tens of thousands of real soccer players from the world's top leagues. As a player, you can recruit and even train your own team that consists of real player's avatars. And then use your team to compete with your friends or other players online.

Back to the topic, in order to answer this abstract question,our team divided it into three modules. First, we will try to explore different styles among players, for example will there be any group pattern that separates players. Then, we will predict and interpret the overall score for each player, and try to understand what factors would lead to a high overall score, for example one's real market value or wage. Finally, we will also identify players with great potential, the hidden gems, the players whose overall score can be further enhanced.

Each module has a specific goal to achieve and uses multiple supervised or unsupervised learning models. In total, all of the results will be interpreted and used to base our overall conclusion, answer the "big" question and give strategies for any FIFA fan.

2. Methodology

2.1 Data Overview

The raw data is collected from Kaggle. For the data cleaning part, we converted the wage column from string to numerical values. We also converted the height values from string to numerical values. Then we removed the unwanted columns Photo, Flag and Club Logo which are in png forms.

The data after cleaning contains 53 columns, which are shown in the graph below. The columns after “Weight” are all the columns about the player's attributes, each containing values from 0 to 100. And these attributes will be used in linear regression for overall rating calculation.

```{r}
fifa <- read.csv("fifa_cleaned.csv")
df <- read.csv("data.csv")
colnames(fifa)
```

```
 [1] "ID"                        "Name"                  "Age"                "Overall"            "Potential"
 [6] "Club"                      "Value"                 "Wage"               "Special"            "Preferred.Foot"
[11] "International.Reputation" "Weak.Foot"             "Skill.Moves"        "Work.Rate"          "Body.Type"
[16] "Position"                 "Height"                "Weight"             "Crossing"           "Finishing"
[21] "HeadingAccuracy"          "ShortPassing"          "Volleys"            "Dribbling"          "Curve"
[26] "FKAccuracy"               "LongPassing"           "BallControl"        "Acceleration"       "SprintSpeed"
[31] "Agility"                  "Reactions"             "Balance"            "ShotPower"          "Jumping"
[36] "Stamina"                  "Strength"              "LongShots"          "Aggression"         "Interceptions"
[41] "Positioning"              "Vision"                "Penalties"          "Composure"          "Marking"
[46] "StandingTackle"           "SlidingTackle"         "GKDiving"           "GKHandling"         "GKKicking"
[51] "GKPositioning"            "GKReflexes"            "Release.Clause"
```

Fig. All Columns in the Dataset

The figure below shows the distribution of ages in the dataset. We can see that the Age data is skewed to the right. Most of the players in the dataset are between 20 to 30 years old.

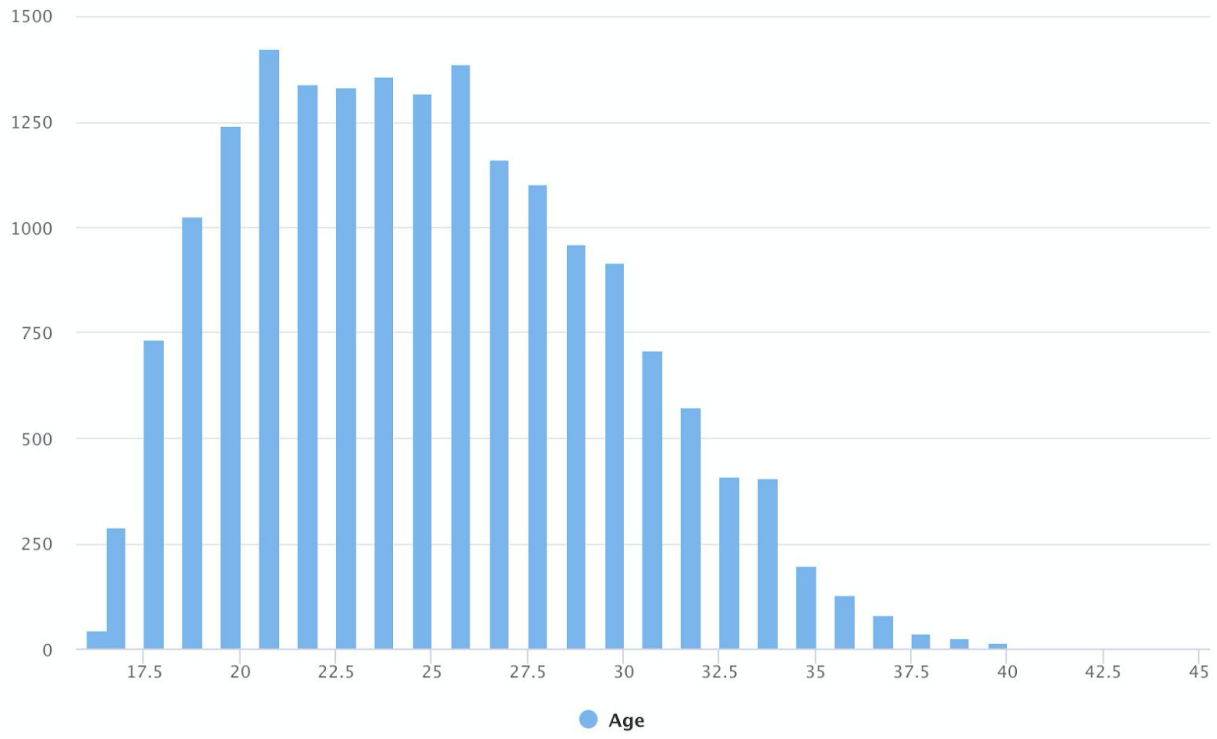


Fig. Distribution of Age

The graph below shows the distribution of Overall Rating in the dataset. It looks normally distributed. This shows the balance in the game, with most players having average ratings.

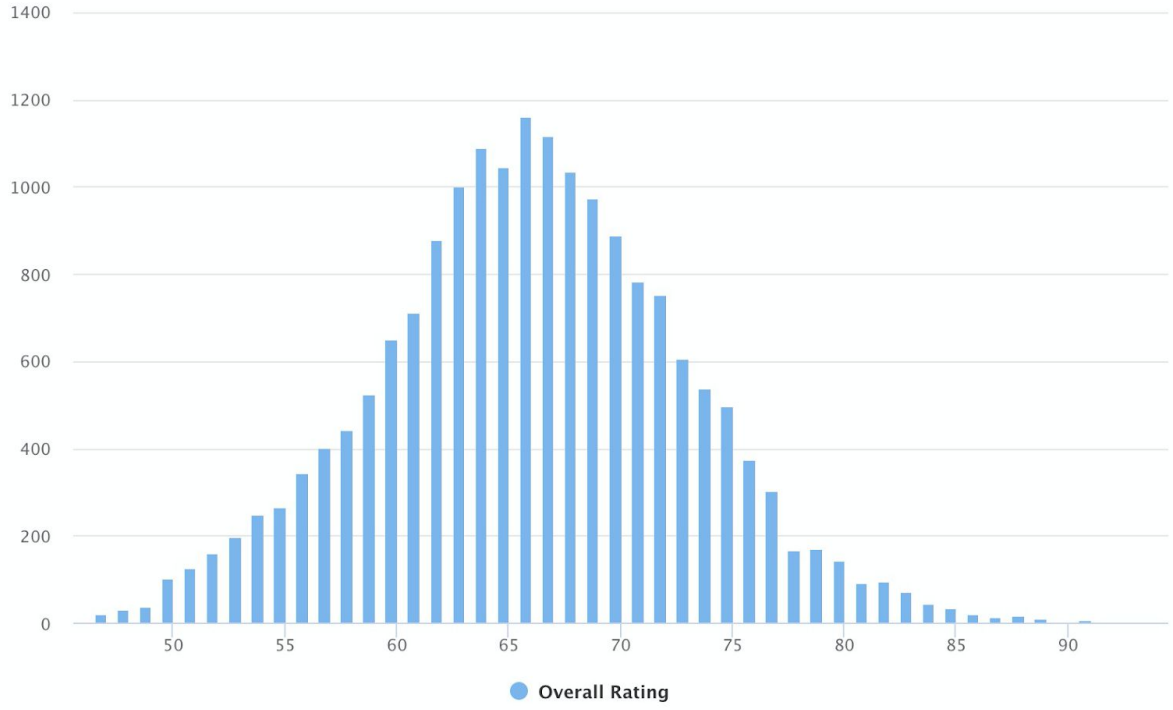


Fig. Distribution of Overall Rating

At last, the bar chart below shows the distribution of positions of players in FIFA19. Due to the modern styles of football, some positions are abandoned by the coaches. Therefore we can see that there are not so many players playing under positions such as “LF” and “RF”. The top three most popular positions are “ST” (Striker), “GK” (GoalKeeper) and “”CB” (Center Back).

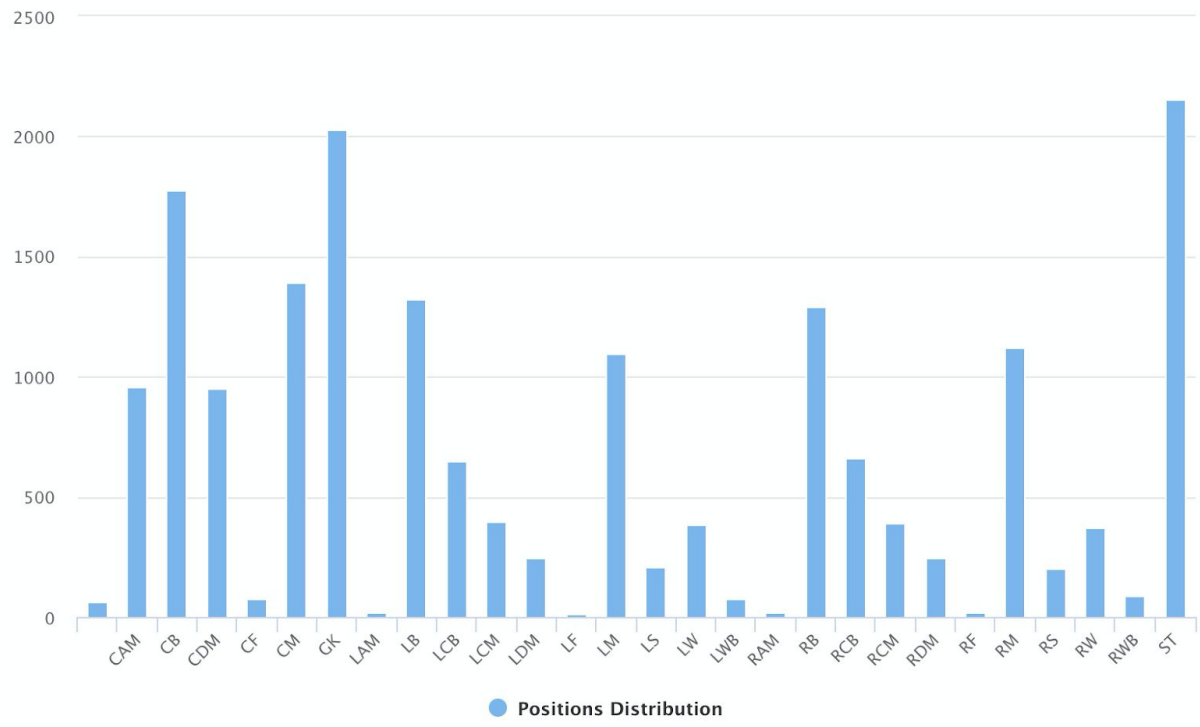


Fig. Distribution of Players' positions

2.2 Statistic Methods

2.2.1 K-Means Clustering

There are some game features defined by the game in the FIFA19 determined players' style and skill in the game. For example, there are some features helping to determine if the player is good at shooting and defending. So we decide to conduct k-means clustering on those features. We will look into our results to see if there are any interesting findings on the clusters. We will evaluate the results by silhouette and by the real meaning of the data. The results are interpreted in the Results section 3.1.

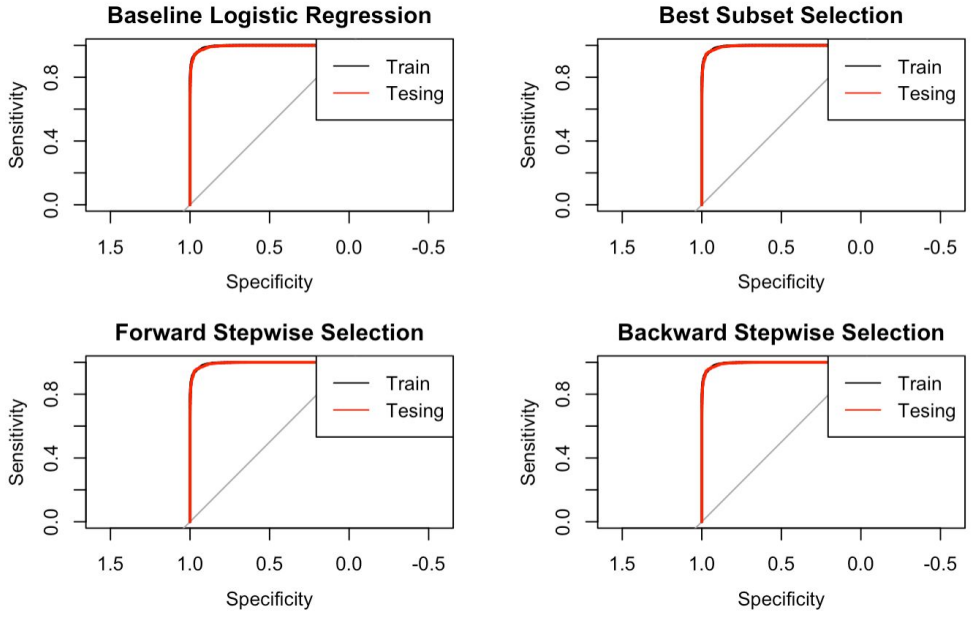
2.2.2 Linear Regression

We wanted to use linear regression to analyze how overall rating is being calculated in Fifa19. In Fifa19, there are 36 attributes that a player contains. There are also 27 positions that players can play on court. We found that while calculating overall rating of a player, Fifa19 only uses several attributes in the formula based on the player's position. We used Lasso regression to select features that are calculated in the formula

of overall rating. Then we used linear regression to check how accurate the models are. The results are interpreted in the Results section 3.2.

2.2.3 Logistic Regression

To identify players with great potential, we trained logistic regression, which will be elaborated here, and tree-based models to make predictions. And to define what having great potential means, there is simply a column in the dataset that indicates potential. So, we can just predict that using classifications. For logistic regression, we fitted multiple models by applying best subset, forward and backward stepwise feature selections. All of the top-picked models from these feature selection methods, using BIC as an evaluation metric, gave excellent results in terms of test accuracy, 95.64%, as shown in the following ROC plots. Furthermore, the output models are also consistent with each other, picking the model with 6 features, including Wage, Value, Position, Overall, Height, and Age. The results are interpreted in the Results section 3.3.



2.2.4 Tree Methods

To further explore which features can mostly affect the result, a classification tree was applied to the data. The same column as we did in the logistic regression was used as target variable. Generalized Boosted Regression Model was applied to further study the relative influence of each variable. Random Forest was also used to train the data and the relative importance was generated.

The results are interpreted in Results section 3.4.

3. Results

3.1 K-Means Clustering Results

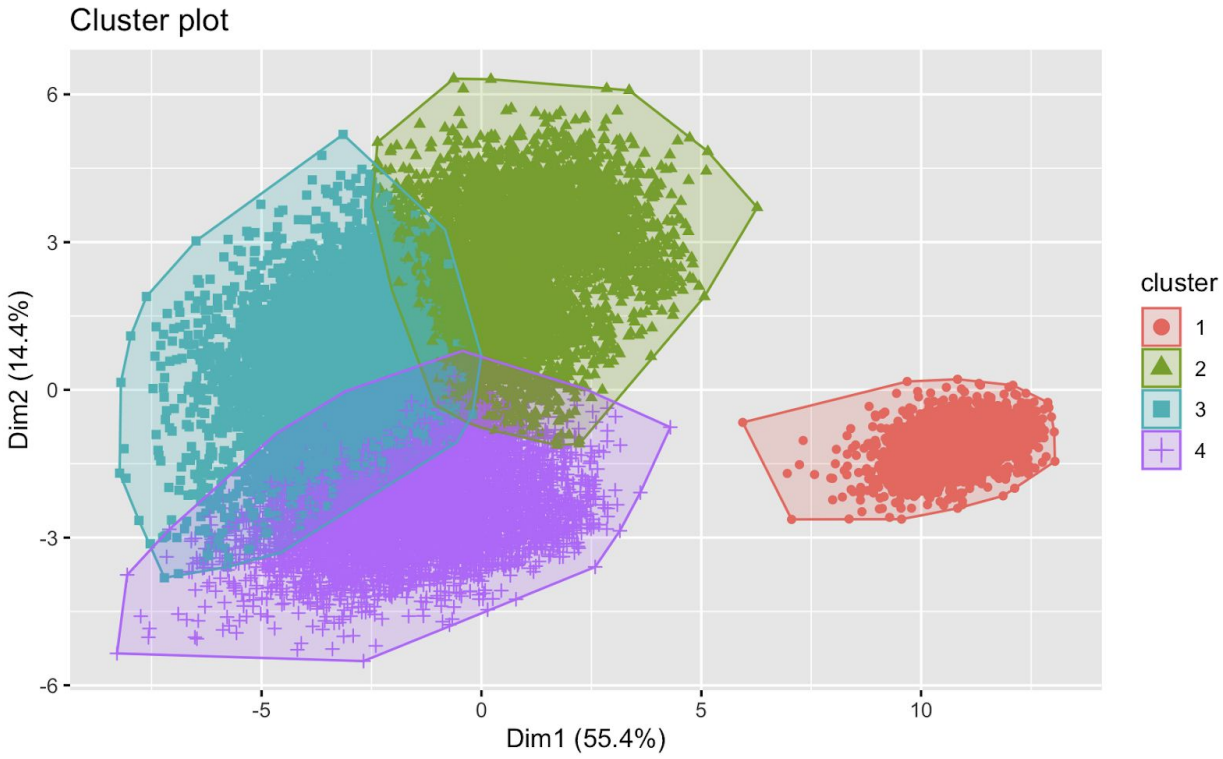


Fig. Clustering Results

The graph above shows the clustering results of k means with k=4 in two dimensions. We can easily see that there is one group away from the other three groups. The right group as group 1 contains all the goalkeepers. Goalkeeper is a very special position on the soccer field. The three bar graphs below show top 5 positions in the other three groups.

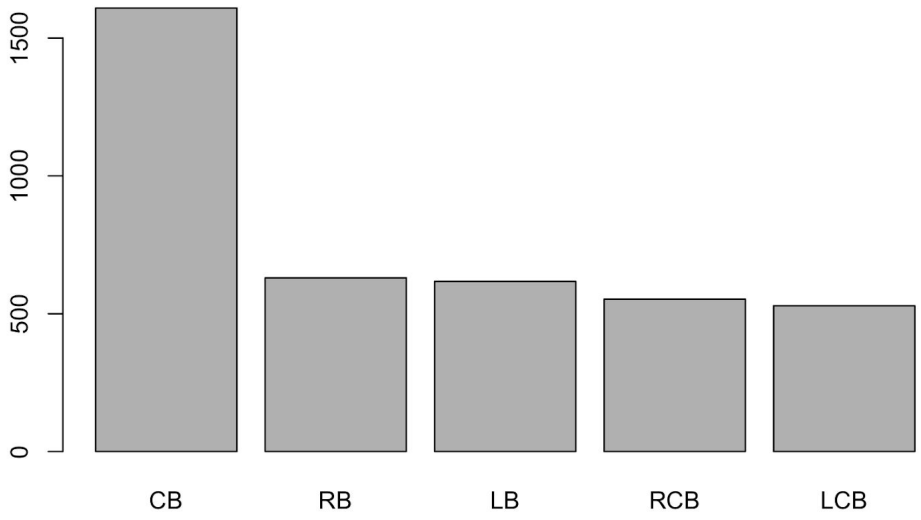


Fig. Position components group 2

That is the bar plot for group 2. It is easy to see that those are all backs. So it implies that this group contains most defensive players in soccer.

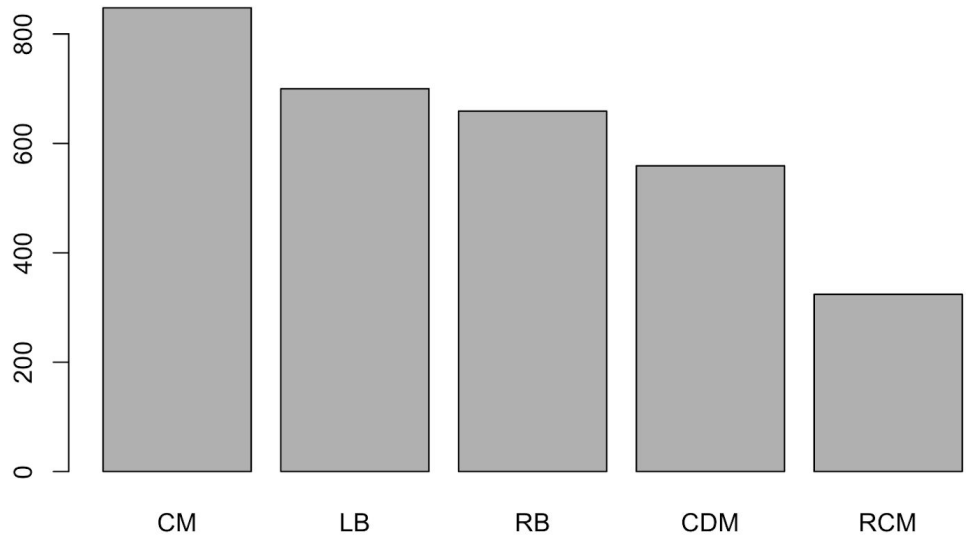


Fig. Position components group 3

That is the bar plot for group 3. It is easy to see that those are all middlefield. So it implies that this group contains most middlefield players in soccer.

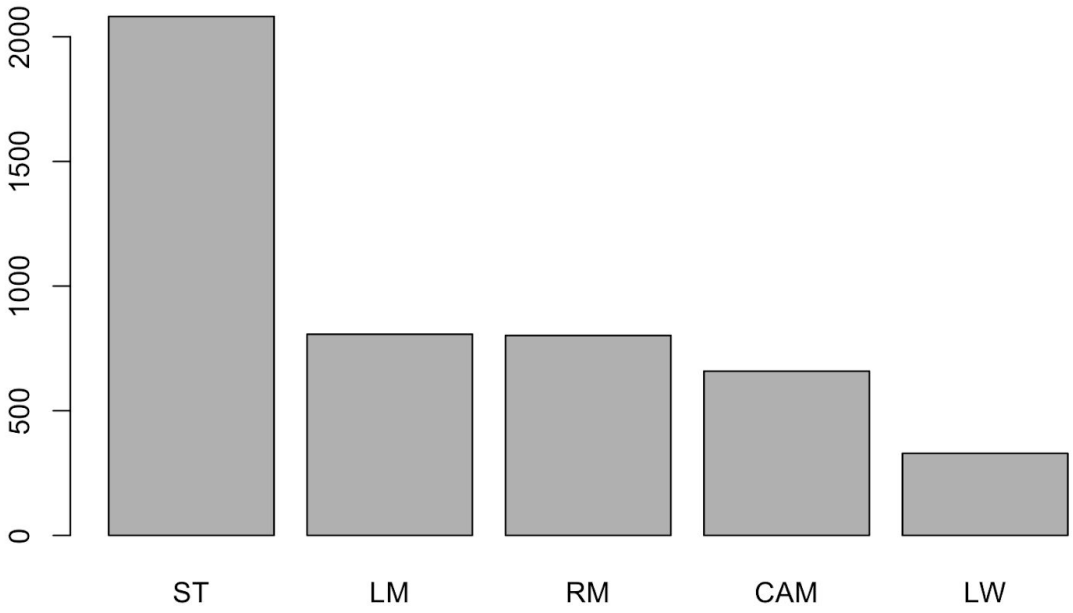


Fig. Position components group 3

That is the bar plot for group 4. It is easy to see that those are all attacking midfield and striker. So it implies that this group contains most middlefield players in soccer.

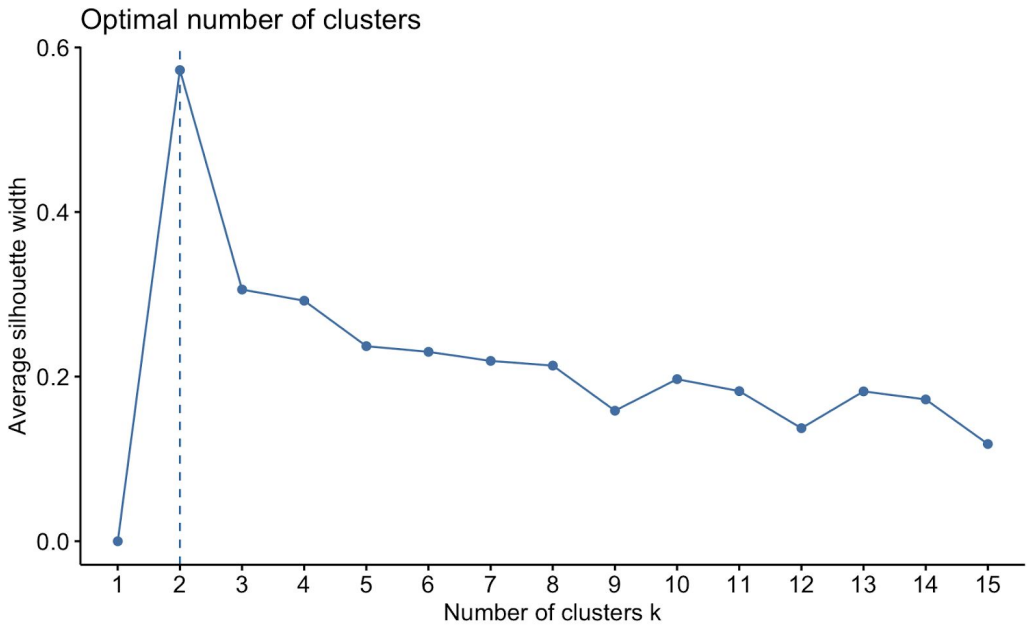


Fig. silhouette vs. k

The above graph shows the relationship between silhouette and k. Although it implies that a lower k has a better score, we still choose to set k equals 4. Because it makes more sense to the real world soccer game. Cluster the players as Goalkeeper, forward, midfield and backs.

3.2 Linear Regression Results

We take our analysis on goal keeper's overall rating as an example. After doing Lasso analysis with all the attributes from goalkeepers in the data, the model tells us that there are only several attributes that relate to goal keeper left to make up the formula. So we came up with a theory that based on the position of a player, the calculation of overall rating only considers several attributes that relate to the position.

```
```{r}
rownames(lasso.coef)[which(lasso.coef!=0)]
```
```

```
 [1] "(Intercept)"    "Reactions"      "GKDiving"       "GKHandling"     "GKKicking"      "GKPositioning"  "GKReflexes"
 [8] "Release.Clause"
```

Fig. Features Selected on "GK" by Lasso

Then we try to fit these attributes into linear regression, the result is shown below.

```
Call:
lm(formula = Overall ~ GKDiving + GKHandling + GKPositioning +
    GKKicking + GKReflexes + Release.Clause, data = train, na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6840 -0.4320  0.0339  0.4714  2.6148

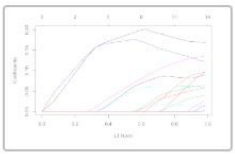
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.033e-02  2.030e-01   0.248    0.804
GKDiving       2.223e-01  6.233e-03  35.670  < 2e-16 ***
GKHandling     2.311e-01  5.550e-03  41.642  < 2e-16 ***
GKPositioning  2.421e-01  5.039e-03  48.044  < 2e-16 ***
GKKicking      5.374e-02  3.960e-03  13.572  < 2e-16 ***
GKReflexes     2.555e-01  5.854e-03  43.645  < 2e-16 ***
Release.Clause 1.344e-08  2.581e-09   5.207  2.2e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

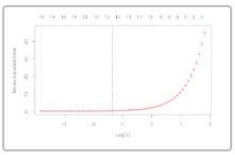
Residual standard error: 0.7172 on 1410 degrees of freedom
Multiple R-squared:  0.9913,    Adjusted R-squared:  0.9913
F-statistic: 2.683e+04 on 6 and 1410 DF,  p-value: < 2.2e-16
```

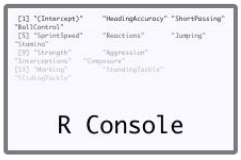
Fig. Results about Linear Regression on “GK”

The R-squared value is 0.99, which shows that the model fits extremely well for the calculation of goalkeeper’s overall rating. We then test the theory using Lasso analysis on other positions. We choose the position “CB” (center back) as another example. The graph below shows the attributes selected by Lasso analysis for the position “CB”.

```
rownames(lasso.coef)[which(lasso.coef!=0)]
```








```
 [1] "(Intercept)"     "HeadingAccuracy" "ShortPassing"    "BallControl"
 [5] "SprintSpeed"     "Reactions"       "Jumping"         "Stamina"
 [9] "Strength"        "Aggression"      "Interceptions"   "Composure"
[13] "Marking"         "StandingTackle"  "SlidingTackle"
```

Fig. Features selected by Lasso on “CB”

And after fitting these attributes into linear regression, we get the following result.

```
Call:
lm(formula = Overall ~ HeadingAccuracy + ShortPassing + BallControl +
    SprintSpeed + Reactions + Jumping + Strength + Interceptions +
    Marking + StandingTackle + SlidingTackle + Aggression + Reactions +
    Stamina + Composure, data = fifa.back)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7931 -0.2706 -0.0275  0.2530 13.1219

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.443084   0.160595   8.986  < 2e-16 ***
HeadingAccuracy 0.095377   0.002360  40.414  < 2e-16 ***
ShortPassing    0.051953   0.001971  26.355  < 2e-16 ***
BallControl     0.041912   0.001991  21.046  < 2e-16 ***
SprintSpeed     0.017389   0.001202  14.472  < 2e-16 ***
Reactions       0.054771   0.002641  20.739  < 2e-16 ***
Jumping         0.028997   0.001176  24.667  < 2e-16 ***
Strength        0.097810   0.001727  56.640  < 2e-16 ***
Interceptions   0.126186   0.003111  40.558  < 2e-16 ***
Marking         0.138696   0.002515  55.158  < 2e-16 ***
StandingTackle  0.172272   0.004483  38.427  < 2e-16 ***
SlidingTackle   0.096675   0.003915  24.691  < 2e-16 ***
Aggression      0.065015   0.001562  41.635  < 2e-16 ***
Stamina         0.001293   0.001347   0.960  0.33739
Composure       0.006468   0.002010   3.219  0.00131 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4879 on 1763 degrees of freedom
Multiple R-squared:  0.9945,    Adjusted R-squared:  0.9945
F-statistic: 2.284e+04 on 14 and 1763 DF,  p-value: < 2.2e-16
```

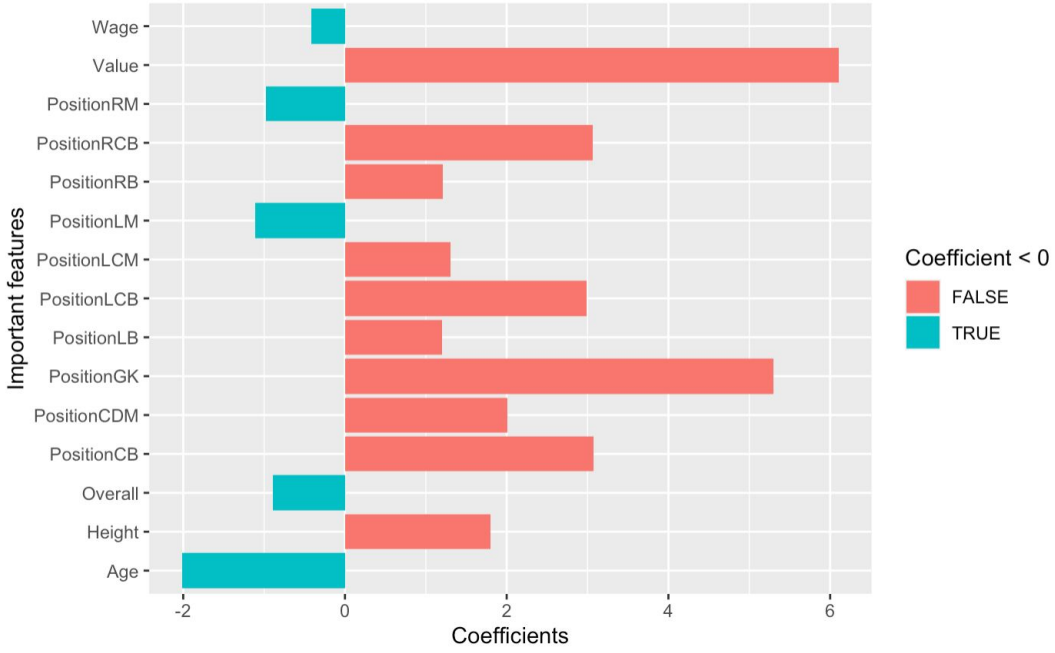
Fig. Results about Linear Regression on “CB”

We see that except Stamina and Composure are less significant than other features, all other attributes selected by Lasso Analysis look significant in the linear model. And the R-squared value is 0.9945, which shows that this model fits really well too. We

concluded that the Lasso analysis can help us choose the right attributes that will be calculated into the overall rating based on the position of the player.

3.3 Logistic Regression Results

In addition to being able to predict players with great potential, we also want to understand what are the decisive factors for being potential. Given the features in the model picked by Best Subset, Forward Stepwise and Backward Stepwise Selection, as partially shown in the barplot, lots of the positive and negative influences of features on potential are intuitive and not hard to understand.



For example, a player's actual market value, the higher it is, the more possible he or she has great potential and could reach a higher overall score. It is also intuitive to understand the negative influence of age. Obviously, young players have longer careers and hence have more room to improve. However, we also spotted an interesting pattern for positions. To be specific, most of the positively related features to potential are defensive positions, for example, CB, CDM, and GK. In other words, players who play defense are generally more likely to reach a higher overall score. Combined with our knowledge and observations in real life soccer, we speculate that the reason is that defensive players normally have longer careers than offensive players. In other words, they have more time to grow compared to players in other positions. Another reason is that defensive players rely more heavily on their experience and positioning, instead of acceleration and pace which are crucial for offensive players. Think of any strikers or wings, for example, Neymar or Messi, they are all fast runners. So, as a defense

becomes older, he or she can only become more experienced and sophisticated, and hence have greater potential. However, for strikers and players in other offensive positions, when they get older, they would only gradually become slower, which results in little room for potential.

After understanding the statistically significant influence brought by certain features, FIFA players then can take these inferences into account when they recruit their team in the game. Find good players with the lowest cost and avoid players with poor potential.

3.4 Tree Methods Results

In the classification tree we generated, only two features, the Age and Position were selected as predicting variables, which is consistent with the logistic regression result. The two figures below in 3.4.1 and 3.4.2 show the confusion matrix of test results. The one on the left is the result of the classification tree and the test accuracy is 94.7%. The other one is the test result of the random forest. It slightly improved the test accuracy to 95.5%. The result shows that both models have really good fit.

	test.Improved	
tree.pred	FALSE	TRUE
FALSE	1076	71
TRUE	126	2309

Fig. 3.4.1

	test.Improved	
rf1.test	FALSE	TRUE
FALSE	1110	69
TRUE	92	2311

Fig. 3.4.2

Figure 3.4.3 and 3.4.4 is the feature importance result generated by fitting a Generalized Boosted Regression Model to the classification tree, which shows the relative influence of each variable. Age plays a significant part in classifying the potential of the player in tree classification as it has 92% relative influence. Position has some influence on predicting the potential of the player. Players in some specific back field positions at an older age will have the potential to improve. Other features such as wage and height have relative influence lower than 1% for each of the variables.

var <fctr>	rel.inf <dbl>
Age	92.1990135
Position	6.3402169
Value	0.7591368
Wage	0.2519933
Height	0.1681037
Contract.Duration	0.1621021
Weight	0.1194336
Preferred.Foot	0.0000000

Fig. 3.4.3

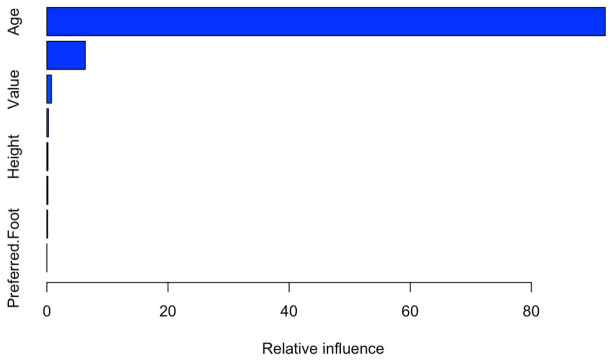


Fig. 3.4.4

In the figure 3.4.5 below is the feature importance plot generated from the random forest. Random forest avoids considering the strongest predictors in part of its splits. Therefore, the variables other than 'Age' have relatively more importance than they were in the classification tree, but Age still has a significant influence in the model.

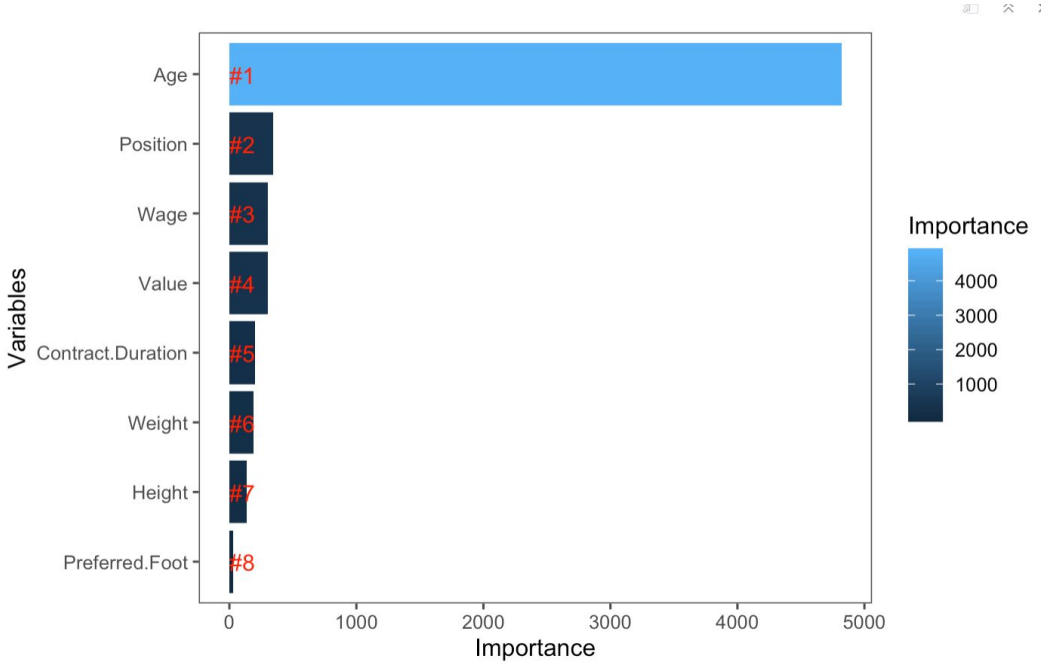


Fig. 3.4.5

From the results of the classification tree and the random forest, we find that players at younger age would have potential to improve. Younger players have more strength than older players and as they gain more experience in the game, they would easily have a boost in their value.

Older players at some defensive positions such as Central Back and Goal Keeper would have the potential to improve. This is because defensive players will need more time to

become matured. Experience is more important for defensive players. They don’t need the explosive power like the offensive players

4. Conclusion

With all the analysis, we can conclude that we should have a balance team first. If we want to replace some players on the market. We need to see if the players are close to each other or whether they are the same type. Fifa is fairly doing a great job on classifying similar type players.

Second, when we are playing my career mode, we should think about what kinds of training we should take. Because the overall score of the player is strongly related to different features for different positions. Choosing the right training can faster the player's growth. So we can spend a short time creating a strong character and have a longer career.

Third, if we are playing Fifa like managing a soccer club for decades. We should always pay attention to those players who have potential and with a low price. Then we should combine that with our clustering to find potential and cheap young players. In that way, we can have a glorious team for a very long time.

In conclusion, FIFA is a very great soccer simulation game. It helps people still enjoy soccer when they cannot go outside. Future work can be more on the real opponents and games such as predicting the results of games or even the ranking of a season.

Appendix

KMeans clustering: clustering.pdf

Exploratory Data Analysis: EDA.html

Linear Regression: LinearRegression.pdf

Logistic Regression: LogisticRegression.pdf

Tree-based models: tree.pdf