Player Overall Scores Analysis and Prediction by FIFA 18 player dataset

Zhennong Chen A99034238

  Yifan Zhou A53238992

1. **Introduction**

FIFA 19 is the latest version of a famous series of football video games, FIFA soccer. In order to reproduce the reality in the virtual world, the game developers track the performance of every player in the real matches and assess their abilities by assigning ratings from 0 to 100 for their attributes such as finishing or dribbling. Each player is also assigned an overall score from 0 to 100, one number summarizing all abilities. We were very interested with the rating calculation of this one-number overall score from all player’s attributes.

In this project, we utilized the FIFA 18 player dataset from Kaggle (<https://www.kaggle.com/thec03u5/fifa-18-demo-player-dataset>) to unravel the hidden rating calculation of overall scores. We wanted to find the most determining attributes in the calculation by implementing the thesis of correlation and the technique of machine learning. We were also curious about whether the set of determining attributes would vary among different player’s positions (forward, midfield, defender and goalkeeper). Our result included:

1. The distribution of overall scores in the player dataset
2. The correlation between overall scores and individual attribute.
3. The search of determining attributes by machine learning using all players in all positions
4. The search of determining attributes by machine learning using players in a specific position

**2. Data processing**

The raw dataset was a complete FIFA18 player dataset from Kaggle, 17981 players in total. It had 75 columns of information for each player including player personal attributes (Nationality, Club, Age, etc.), player performance attributes (Overall, Finishing, Dribbling, etc.) and player preferred subdivided positions (CF, CM, CB, etc.).

The code doing the data processing can be found in appendix part 1 “Data processing”, which includes:

1. Extraction of player performance attributes and preferred subdivided positions from all information, which we considered as the most relevant to our project. There were 29 individual attributes and 1 overall score in the player performance attributes categories.
2. Conversion of float numbers to integers easier for further processing.
3. Filling of null space by 0.
4. Assignment of one preferred position (only four categories: forward, midfield, defender and goalkeeper) for each player based on preferred subdivided positions.

**3. Result**

**3.1 Distribution of overall scores**

We first investigated the distribution of overall scores. Simple statistics showed that there were 17981 players with mean 66.25 and standard deviation 6.99 in overall scores.

We plotted overall scores for all players in Figure 1, which inspired our hypothesis that the distribution could be approximated as a normal distribution. Thus, we fitted a gaussian distribution N ~ (66.25, 49) onto the overall score distribution, where 66.25 and 49 were the

mean and variance we directly got from simple statistics. The perfect fitting confirmed our hypothesis. This normal distribution also made sense in reality since most of the players are just mid-class and only few players can reach the top-class.

The code for this part could be found in appendix part 2 “Distribution of overall scores”.

|  |
| --- |
|  |
| Figure 1. The distribution of overall scores of all players. Total 17981 players were examined with the mean 66.25 and standard deviation 6.99 in overall score. We fitted a Gaussian distribution N ~ (66.25, 49) to the overall score distribution. The fitted result confirmed the hypothesis that the overall score rating in FIFA18 can be approximated as a normal distribution. |

**3.2 Correlation between overall score and attributes**

There are 29 attributes of interest in total in the dataset including finishing, dribbling, reaction and etc. We calculated the correlation between overall score and each of 29 attributes. The correlation is a function as:

where x represents overall score and y represents individual attribute. Larger correlation indicated that individual attribute is more determining in the rating calculation of overall scores. Figure 2 plotted all the correlations in descending order, in which we realized top 10 determining ones were Reactions (0.84), Composure (0.63), Short passing (0.49), Vision (0.48), Long passing (0.47), Ball control (0.45), Shot power (0.43), Curve (0.41), Long shots (0.41) and Aggression (0.40). It was interesting that reactions had such an importance.

In next section we were going to show the top 10 determining attributes examined by machine learning technique and compared with those examined by correlation. The code for correlation could be found in appendix part 3, “Correlation between overall score and attributes”.