Data analysis and player value prediction by FIFA 18 player dataset

Zhennong Chen
A99034238
zhc043@ucsd.edu

Yifan Zhou
A53238992
yfzhou@ucsd.edu

The dataset we plan to analyze is the complete FIFA 18 player dataset. The bundle consists three sets of data: (1) player personal data including club, nationality, wage etc, (2) player attribute data including aggression, ball control, dribbling, etc. shown as score from 0 to 100, and (3) player position data including player's performance in each position (e.g. forward, defense) shown as score from 0 to 100 as well. Obviously, the advantage of this dataset is that all the scores assessing players are using the one consistent scale from 0 to 100, avoiding the difficulty resulting from normalization.

During data preprocessing stage, the first general question we are interested to ask is the financial state of soccer players. Questions like "how much is the average of wage" , "who are top 10 wealthy players", "which 5 clubs burden the most in wage" or "how age affects wage" are worth asking. These topics are easy to be analyzed by counting, sorting and other simple math.

To dive deeper into the problem, we want to figure out the scoring system in this game. Each player is assigned to an overall score which may vary when they are playing different positions and an individual score in each attribute. The initial guess is that the overall score is heavily determined by around 10 attributes among all 40 which varies among different positions. More importantly, our (expected) quantitative analysis should coincide with human intuition, e.g. faster speed brings huge advantage, backfield usually have strong body, etc.

More specifically, a bunch of questions we want to solve are listed as follows. (1) figure out in general case and specific position case, respectively, how all attributes contribute to overall score and value, (3) find top ten determining attributes for each position, (2) given attribute of a player, which positions are most suitable.

Methodologically, we will implement some basic machine learning approaches. Predicting value and overall score can be cast into a regression problem, for example, using a linear combination as approximation. Pearson coefficient can be employed to sort most important attributes. Multi-class logistic regression is a good choice for position classification.

Overall, faced with this structured FIFA dataset, we will preprocess data using statistics and unsupervised learning. Then for prediction part, supervised learning will be our main method. We hope to gain a comprehensive insight and find hidden patterns behind the dataset through our analysis.

Reference:

[1] https://www.kaggle.com/thec03u5/fifa-18-demo-player-dataset