A Model to Predict a Young Player's Potential Using the FIFA Dataset

Final project for Data Mining course

Group number - 14

Email address for correspondence:
adamd2@gmail.com

Abstract

In the world of sports, whether it's live on a pitch or virtual through video games or even fantasy leagues, statistical data and numeric values of players' abilities are prevalent everywhere. Teams use gathered data to try to improve their players where the metrics were lacking and try to optimize their conditioning, skill, diet and schedule. In our experiment we use a large database of players and their virtual attributes to predict young players' potential by creating a new weighted average of their individual attributes and applying a linear regression model of FIFA's "Overall" and "Potential" of young players (under 23) data to predict our recommended player potential. Our regression model is accurate for up to 3.79 RMSE and an r=0.79 coefficient correlation. With this model, FIFA players can better choose a long-term team using young players that will have high ratings in the future.

1 Introduction

1.1 Background

FIFA is the world's leading football video game, allowing people from around the world to play football matches against each other.

The game comprises of real players from real teams, and their quality is simulated by numerous attributes that affect their play in each position on the field. With each edition that is released yearly, the game tries to paint a more in-depth picture of the players real-world skills and abilities.

We believe that the ratings shown for players in FIFA can be biased (towards popularity, for example) and are not showing a true overall rating of a player. In many instances, a player will receive an extremely high rating while having a very low score in one of his attributes. For this reason, we will build our own weighted average to determine a player's overall.

We will apply this weighted overall to young players under the age of 23 and choose the best ones in each position to create the "future super team" and show their potentials using a linear regression model.

1.2 Literature Review

Due to a lack of cohesive, centralized real world football data regarding players, teams, games and leagues there have been cases of other studies using the FIFA dataset trying to analyze certain patterns or behaviors that could be extrapolated to live football cases. In this article, Using FIFA Soccer video game data for soccer analytics[1] the authors used the FIFA dataset to analyze two cases: the contrast between the Brazilian and German National teams in 2014 and FC Barcelona's distinguished style in the 2012/13 season. In their study they used a linear regression model to analyze player qualities

over time for the two national squads and for the second analysis a PCA model and K-means clustering model were used to determine the causes of Barcelona's style in that season.

Another example of a study using the FIFA dataset would be Empirical Comparisons for Combining Balancing and Feature Selection Strategies for Characterizing Football Players Using FIFA Video Game System[2] where the authors used multiple machine learning and data mining models (PCA, random forests) to model individual player performance.

2 Material and Methods

2.1 Materials

The dataset we used was taken from Kaggle[3]. This dataset gives a large amount of player measurements in the 2021-2022 season. This includes physical data (height, weight, strength, etc.), performance data (shooting, passing, defending, etc.) and general attributes (age, nationality, current club, etc.).
The dataset contains 19,239 players and 110 different attributes, including numeric data and strings.

2.2 Methods

2.2.1 Preprocessing

The first step of our preprocessing was cleaning the data and removing all irrelevant columns. That included columns like club information, national team information, financial data and in game data (such as facial features, tags, ID numbers, etc.) and duplicates (such as age/DOB and positions).
We noticed that there were several empty values, mostly in attribute columns where the values are irrelevant to the player, such as field attributes for goalkeepers. To prevent potential errors, we replaced the null values with 0.
We also noticed that players were listed in multiple positions, with their main position shown first and other positions the player can play in afterwards. We have chosen to show players only in their main positions, as we will group multiple positions into one to build a team in a 4-3-3 formation in a later stage. We did that by splitting each cell in the positions column and only keeping the first element in the array.

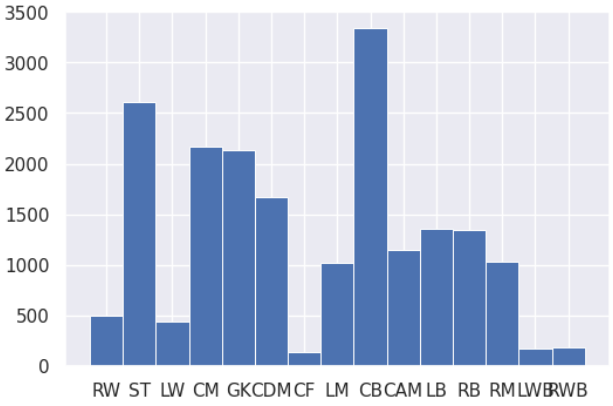

Fig 1: Shows the number of players per position

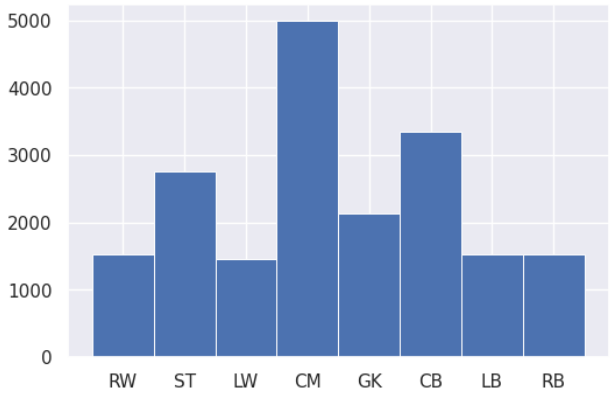

Fig 2: Players per position after combining positions

Our next step was filtering our data by age, removing all players over the age of 23. During the age restriction step 59% of the data was removed.

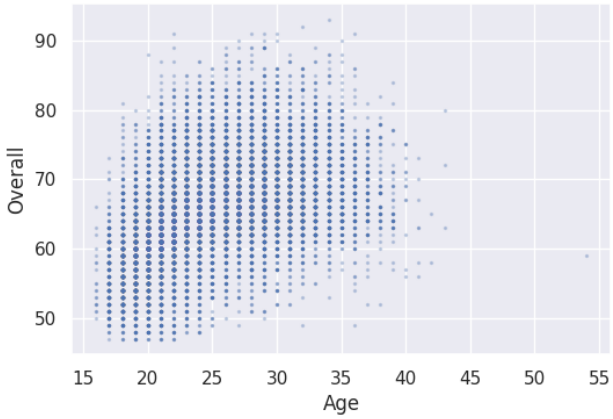

Fig 3: Shows the 'overall' density distribution by ages

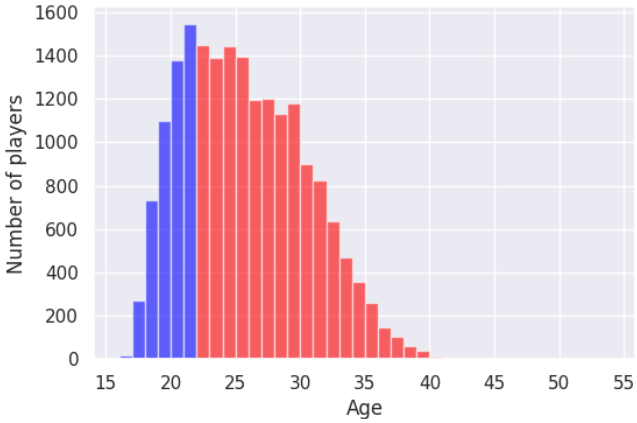

Fig 4: Shows the age distribution of under and over 23 players

	short_name	overall	potential	pace	shooting	passing	dribbling	defending	physic
6	K. Mbappé	91	95	97.0	88.0	80.0	92.0	36.0	77.0
21	G. Donnarumma	89	93	0.0	0.0	0.0	0.0	0.0	0.0
29	E. Haaland	88	93	89.0	91.0	65.0	80.0	45.0	88.0
44	T. Alexander-Arnold	87	92	79.0	68.0	88.0	80.0	80.0	72.0
45	J. Sancho	87	91	81.0	76.0	82.0	91.0	36.0	65.0

Fig 5: The table after the preprocessing

2.2.2 Linear Regression

In order to predict the players' potential, we first need to create a new "overall" metric. We created a weighted average formula, giving each position a different weight set to calculate the new overall outcome. We created a new table and removed every line that doesn't have the desired position name and applied the weighted average on the remaining rows and created a "new overall" column that will be used to predict the potential.

We used a linear regression model to check the correlation between the given overall and potential columns, and used the resulting coefficient and intercept to calculate a new potential score. The model was trained with 80% of the data and the rest of the dataset was used to test the results. Our regression model is accurate for up to 3.79 RMSE and an r=0.79 coefficient correlation.

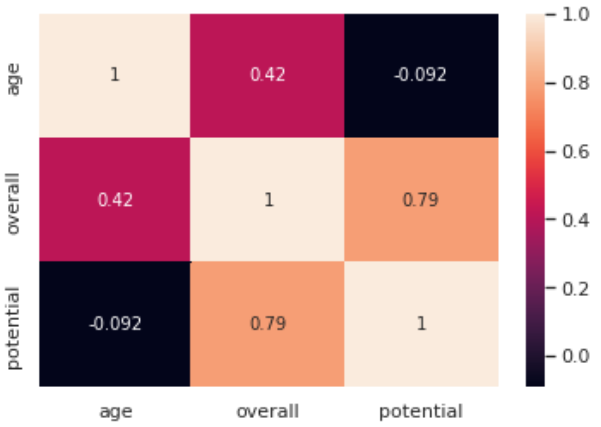

Fig 6: Correlation matrix

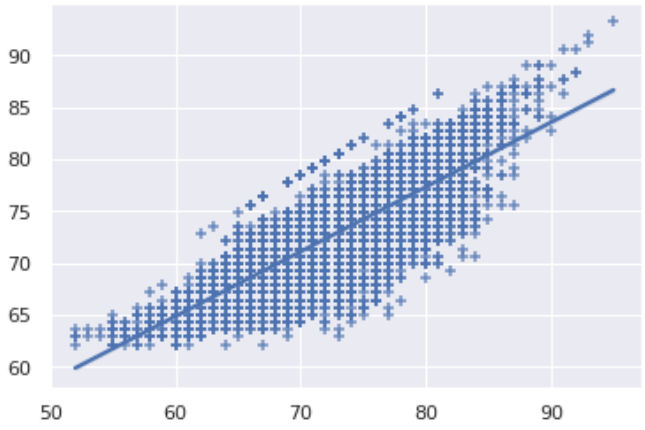

Fig 7: Linear regression plot

3 Numerical Analysis and Results

3.1 Results

The results that we received, both from the weighted average calculation and the linear regression based potential result were different from FIFA's data. Our weighted overall was about 5% lower than FIFA's overall, this due to our hypothesis that the calculations are not the same for each player and each position. There might be a popularity bias that we cannot gauge in this study, however we can see for example that the main factor that lowered an attacking player's weighted overall was the weight of the defensive attribute, that while it held a very small weight in our calculation, it appears to hold no weight at all in some of FIFA's scores.

We can also see in our results that 6 of the 11 players selected with FIFA's overall and potential attributes are different in our results, affirming our assertion that the possibility of bias in FIFA's ratings exists.

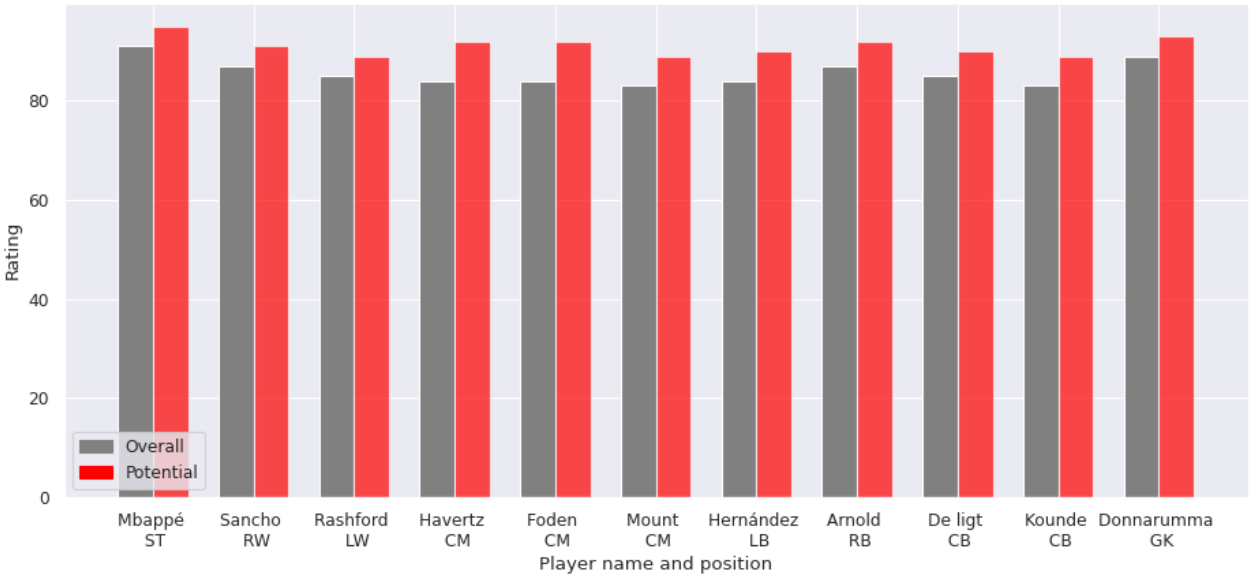

Fig 8: The 11 player squad chosen using FIFA's raw overall and potential ratings

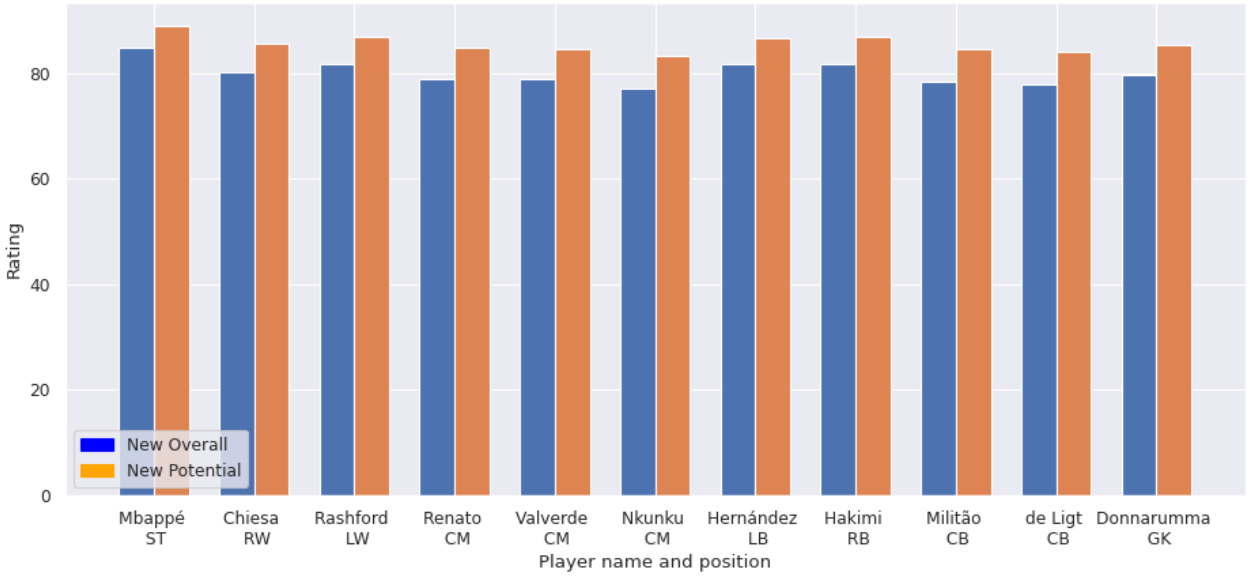

Fig 9: The 11 players that were chosen to our young stars squad, with their new overall and potential attributes

3.2 Discussion

There are many ways to gauge a player's quality. We chose an analytic approach using attributes from a video game that tries to portray a player's real-world quality in its features. There are other ways, such as looking at datasets from previous years and gauging a player's growth to predict his value and attributes in the next season. We chose to use only the current year's dataset as we believed that it would have the best portrayal of young players, due to the high volatility in attributes in these ages over the seasons.

4 Conclusions

In this paper we show an alternative way of gauging a player's attributes through FIFA's video game data set for 2022. We believed that there was bias in the initial calculations, based on popularity and financial motives.
Our model has shown that players could be considered to be overvalued with their attributes when more parameters and different weights are applied to the calculation.

In future work, we recommend using the results of this study as a first iteration and track the young players in this study every year when a new dataset is released. As the years go by the accuracy of the model should improve as the samples can be compared to new data. New models and algorithms should be introduced to this model once the data is looked at over time.

5 Appendix - References

[1] Cotta, L., de Melo, P. O. V., Benevenuto, F., & Loureiro, A. A. (2016). Using fifa soccer video game data for soccer analytics. In *Workshop on large scale sports analytics*.

[2] Al-Asadi, M. A., & Tasdemír, S. (2021). Empirical Comparisons for Combining Balancing and Feature Selection Strategies for Characterizing Football Players Using FIFA Video Game System. *IEEE Access*, *9*, 149266-149286.

[3] https://www.kaggle.com/stefanoleone992/fifa-22-complete-player-dataset?select=players_22.csv