**Predicting player ratings for the FIFA video game**

**Matt Watson**

---------------------------------------------------------------------------------------------------------------------

**Abstract:**

*Objective*: The goal of this paper is to use real life statistics from soccer matches to predict player ratings in EA Sports’ FIFA 20 video game. FIFA is one of the most popular video game franchises of all time, but how EA comes up with player ratings is not known outside of the company.

*Methods*: Random Forest, Gradient Boosting, ADA Boosting, Neural Network and Support Vector Machine models will be used with cross validation for both Regression and Classification. For Regression, FIFA 20 rating will be predicted and for Classification, change in player rating will be predicted. Various feature selection methods, such as Stepwise Recursive Backwards, Wrapper and Univariate feature selection will be used for each model. All models besides Support Vector Machine will be run both with and without feature normalization, while SVM will be ran only with feature normalization.

*Results*: The Regression models gave RMSEs ranging from 1.5 to 2.2 (which is good given than the target feature has a standard deviation of around 5.3) and Explained Variances ranging from 0.8 to 0.92. The Classification models gave Accuracies ranging from 0.62 to 0.82 and AUCs ranging from 0.71 to 0.90. The features selected from the various feature selected methods used across all models related to 4 of the 6 main player attributes in FIFA: Shooting, Passing, Dribbling and Defense (no features fit into Pace or Physical).

*Conclusion*: Both the Regression and Classification models had quite good results. The findings of this paper suggest that EA likely uses similar data when creating player ratings for the FIFA video game franchise, but other features, such as Age or 40-yard dash time, are likely used as well.

---------------------------------------------------------------------------------------------------------------------

**1. Introduction:**

Soccer, or Football as it is known outside of America, is arguably the most popular sport in the world. In 2009, the sporting industry made $64 billion, 43% of which came from soccer [1]. With tournaments like the FIFA World Cup and EUFA Euro Cup every 4 years for international teams, and the Champions League every year for European Club teams, it should come as no surprise that soccer is vastly popular. The sport has gone through its shares of ups and downs in America, however. Before Major League Soccer (MLS) debuted in 1996, there had been numerous attempts to create a professional soccer league in America. All of the previous leagues had failed to stay around, due to negative coverage in the media as well as a general lack of interest from the US population. This might be due to people thinking of soccer as “foreign,” and therefore inferior to the “Big 3” American Sports: American Football, Baseball and Basketball [2]. With the release of the first FIFA video game in late 1993 and the US hosting the 1994 World Cup, interest in the sport definitely increased in America.

Since its first release in 1993, FIFA has become one of the best-selling video game franchises of all time, with more than 150 million copies sold as of 2017 [3]. Not only is it a popular game amongst fans, but with players as well. Famous players such as John Terry would often host tournaments amongst his teammates, while Zlatan Ibrahimović would play multiple hours a day. Younger players, such as Alex Iwobi, develop their skills by mimicking moves seen/used in the video game. This relationship between playing sports video games and playing real life sports is something that had not been researched until recently [4, 5]. Using data from video games to analyze real life sports is something that has become more common as well. However, doing the opposite is something that has not been done to my knowledge, at least when it comes to FIFA. The goal of this research is to attempt to predict the FIFA player ratings for the newest game, FIFA 20 (2019-20 season) by using data from the 2018-2019 season for English Premier League (EPL), the top division in England, as well as the FUT 19 player ratings. These ratings are a combination of various attributes, such as Pace and Shooting, which are further broken down into a total of 40+ characteristics, such as Pace being the average of Sprint Speed and Acceleration. The way Electronic Arts (EA), who makes FIFA, creates these ratings is unknown, but it is likely a combination of game statistics as well as scouting reports from employees around the world. Using real life data, I will attempt to predict FIFA 20 rating (Regression) as well as predict whether the players’ ratings decreased or increased from FIFA 19 to FIFA 20 (Classification).

**2. Literature Review:**

*2.1: Impact of Video Games*

Up until the past decade or so, most research on video games has focused on the negative outcomes instead of the positive. Many studies have been done on the effect of violent video games, often focusing on the Violent Content Hypothesis, which states that repeatedly exposing children to media violence increases later aggression, even into adulthood. Anderson et all wanted to compare this hypothesis against the Competition hypothesis, which states that competitive situations stimulate aggressiveness [6]. They did this by creating a study which compared various questionnaire results about aggressiveness for people who played a nonviolent sports game versus people who played a violent sports game. Across three different experiments, the results supported the violent-content hypothesis. This was the first study done comparing the two different hypotheses.

Other studies, however, have shown that there are positive outcomes related to playing sports video games. Alivari et all found that computer games can positively affect the stress system and the perceptual-cognitive system. Even though this impact was not significant in most cases, the changes in cognitive and hormonal test and also in brain waves were visible [7]. This study involved analyzing brain waves through EKGs while playing FIFA, analyzing stress levels through saliva collected before/after playing, as well various surveys. Sports games, in general, are much less violent than other types of video games, such as First-Person Shooters like Call of Duty.

Adachi and Willoughby performed two different studies with the goal of looking for a relationship between sports video games and real-life sports play [4, 5]. One study focused on High School Students, with surveys taken once a year over the course of 4 years [5], while the other focused on college students/emerging adults who also took a survey once a year for 4 years [4]. Both studies showed that playing sports video games has a positive long-term bidirectional relationship with participating real-life sports, with increased self-esteem as a result of playing video games as an important factor in predicting real life sports play. These studies were the first to find this relationship. This is greatly important, as getting youth to participate more in sports could help combat the high obesity rates present in the US.

*2.2: Studies on Sports Video Game data*

Using data from sports video games to analyze real-life sports is something that has started to become popular in recent years. Especially for soccer, a sport where it is difficult to gather data outside of traditional statistics such as passing, shooting and tackles, using large databases of video game data can provide interesting insights. Cotta et all used player attribute data from FIFA to analyze two recent events in soccer: Germany’s domination of Brazil in the 2014 World Cup Semi Final, and the dominance of FC Barcelona from 2011 to 2013 through their use of “Tika-Taka” [1]. Rather famously and unpredictably, Germany beat Brazil 7-1 in the 2014 World Cup. To try to get a better idea of how this happened, the authors studied the changes in the top 20 forwards, midfields and defenders for each team over the previous 5 years. They found that overall, the German players improved more than the Brazilians in the metrics that mattered, especially when it came to midfielders. FC Barcelona dominated La Liga (the top division in Spain) during the 2011 to 2013 seasons through their use of “Tike-Taka,” essentially just one touch passes which enabled them to dominate possession. The authors again used player attributes, just for midfielders as they usually possess the ball the most, for all teams in La Liga. Using PCA to create a component for each team, and then Clustering to see if there were differences amongst the teams, the authors found that Barcelona was in its own cluster. This indicates that the play style cannot really duplicated by other teams, as it was the players on Barcelona that made the system work so well. This finding makes sense, as the coach of Barcelona at the time, Pep Guardiola, could not make the system work well with the next team he managed, Dortmund. Although the findings for Barcelona are valid, there are some questions about the Germany vs Brazil game. First, the study ignored the fact that Brazil’s two best players, Neymar and Thiago Silva, were both injured and did not play against Germany. Second, the study used the top 20 players for each nation for each of the 3 positions, while only 23 players were on each team. So even though German players developed more before the World Cup, not all of them were on the roster for it. Despite the potential lack of validity, this study is interesting and shows the potential for using data from video games to analyze real-life sports.

Soto-Valero used similar data to perform a Gaussian mixture clustering model with the goal of characterizing football players [8]. He first performed PCA, then the clustering model and lastly a regression tree model to look at feature importance. Interestingly, the results of the clustering model gave 3 unique clusters, each corresponding to a position: Forward, Midfielder and Defender. The regression tree showed that the most important feature for distinguished between the clusters was dribbling. This makes sense, as Midfielders usually have the best dribbling skills, followed by Forwards and then Defenders, who are not really know for doing fancy moves. Both of these studies used the 40+ attributes available for each player in EA Sports’ FIFA. My study will be using real life data that corresponds to some of these attributes

*2.3: Studies using real life data to analyze sports video games*

To my knowledge, there has been no other research done using statistics from real life sports to analyze sports video games. The only related work to my project that I have found is an article on “Towards Data Science” [9]. The author of this article only used things such as salary and potential rating from FIFA to predict rating through linear regression. Although he obtained good results, I do not believe that the results of his study are valid. Obviously, players who make more money are going to have higher ratings, as the best players make the most money. Potential is also a confusing feature, as for the good players, they have already reached their max potential, so their rating is often the same as potential. How EA comes up with this feature is not known, so it is likely not a good to use it in analysis. A common saying, “Correlation does not imply causation” definitely relates to this study. Although features such as salary and potential might be correlated with Rating, it is very unlikely that they are used by EA when creating ratings for players in FIFA.

**3. Data:**

The data I will be using comes from the website Football Reference and contains in game statistics from players in the English Premier League during the 2018-2019 season. The data is stored in 4 different tables: Standard Stats, Shooting, Passing and Miscellaneous Stats. There is a total of 508 players and 73 features. I removed all of the “Expected” stats before downloading the data, as they do not say how these were generated. Next, I downloaded data from WEFUT for both the 2019 and 2020 FIFA player ratings.

For preprocessing, some features were removed from each table due to missing values. For example, in the Passing table, “CMP%\_S” was missing due to “CMP\_S” and “ATT\_S” both being 0. I also removed all players who are Goalies, as there were only 38 total players, and this would result in a lot of missing values if they were kept. There were two other tables, Goalkeeping and Advanced Goalkeeping I could’ve downloaded if I decided to keep the Goalies. When attempting to add the player ratings from FIFA, I ran into quite a few issues as well. The biggest issue was that the player names, the condition on which I used to merge data frames, were different. For example, there is player on the Tottenham Hotspurs who’s name in the Football Reference data was listed as “Lucas Moura,” however in FIFA he is listed as “Lucas.” Since the bottom 3 teams in the EPL are relegated to the English Football League (EFL) Championship, the 2nd division in English Football, after each season, I also needed to download the data for these teams. Many players transferred teams as well; for these players I had to manually add their ratings. There were some players that needed to be removed, as they either retired or transferred to teams in leagues that are not in FIFA, such as to the QNB Stars League, the top division in Qatar soccer.

There were some features that were clearly linear combinations of other features, so I removed them. Some features that were removed include: “BORN,” “G/SH” and “90s.” “BORN” is directly related to “AGE,” “G/SH” is goals per shot, and is a linear combination of Goals and Shots and “90s” is Minutes Played divided by 90 (as there are 90 minutes per match). For the Classification models, I created a new, binary feature called “Dif,” which is the difference between ratings in FIFA 19 and FIFA 20. If the player’s rating decreased, “Dif” will have a value of 0, while if the rating stayed the same or increased, it will have value of 1. Figure 1 below shows the counts for “Dif.” After all of the preprocessing, the final data set was reduced to 447 players and 53 features. Six of the final 53 features were not used for analysis, as they included categorial things such as Name, Team and Position.

A screenshot of a cell phone

Description automatically generated

Figure 1: Class distribution for “Dif”

**4. Methodology:**

*4.1: Models*

Both Regression and Classification models will be used to attempt to predict player ratings in FIFA 20. For Regression, all 47 features will be used as possible predictors for rating in FIFA 20. For Classification, all of the features will be used to attempt to predict whether player ratings either decreased or increased/stay the same from FIFA 19 to FIFA 20. Various Machine Learning models will be used, included Random Forest (RF), Gradient Boosting (GB), ADA Boosting (AB) and Neural Networks (NN), both with and without feature normalization. Support Vector Machine (SVM) will also be used, but only with normalized features. The next paragraph will go into brief detail about all of these models.

RF is an ensemble method that uses a number of Decision Trees (DT), with a subset of features used at each split in the tree. Weaker DTs are created than a single DT model, with the idea of collective intelligence in mind. Even though each DT is weak, when they are put together, the model performs well. Both GB and AB are similar to RF, in that a number of DT are created, but the way they are created differs. In GB, each tree is made one at a time and the model learns from its mistakes by comparing the residual errors. In AB, each tree is also made one a time, but it learns from its mistakes by increasing the weights of misclassified points. NN attempts to mimic the way the human brain works by creating “neurons” (nodes) that are linked together and have weights and activation functions. I will be using a Multilayer Perceptron model, which has at least 3 layers of nodes: input, hidden and output. Information flows across the neurons as decisions are made. SVM uses a kernel to separate data in a high dimensional plane. This is especially useful in this research, as there are large number of features. Each of the various models will be run with 5-fold cross validation. In k-fold cross validation, a different 1/kth of the data is held out to be used for testing, resulting in k different runs.

*4.2: Feature Selection methods*

In addition to all of these ML models, I will be using various feature selection (FS) methods. I will not be focusing as much on hyperparameter tuning, as there a large number of features in the data relative to the number of observations. I will be using Stepwise Recursive Backwards (SRB) FS, Wrapper FS and Univariate FS with Mutual Info. SRB fits the model by starting with all available features and then removes the weakest feature until the desired number of features remain, which in this study is either 5 or 10. Wrapper FS is a greedy search algorithm that checks possible subsets of features and selects the one that performs the best. For both SRB and Wrapper, a RF model was used. Univariate FS selects the best N features based on some statistical measure, in this case Mutual Info. I decided to use just the top 10 features for this method.

**5. Results:**

In total, there were 45 different combinations of models for both the Regression and Classification problem. RF, GB, AB and NN were each run with 5 different feature selection methods, both with and without feature normalization. SVM was run with feature normalization and 5 feature selection methods. Table 1 shows the top models from the Regression problem and Table 2 shows the top models from the Classification problem.

As it can be seen in both Table 1 and Table 2, feature normalization did not impact model performance besides for the NN models. Interestingly, the only model that performed better with a different feature selection method after features were normalized was RF in the Regression problem. Most of the models had similar performance; even though the values varied slightly, the ranges from cross validation overlapped, indicating the differences are not significant. In both the Regression and Classification problems, the NN model without feature normalization performed the worst.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model Type | Feature Normalization | Feature Selection | RMSE | Explained Variance | Runtime (seconds) |
| RF | No | NONE | 1.94 +/- 0.32 | 0.87 +/- 0.02 | 1.01253 |
| RF | Yes | SRB w/ 5 | 1.92 +/- 0.29 | 0.87 +/- 0.02 | 0.67516 |
| GB | No | NONE | 1.71 +/- 0.26 | 0.90 +/- 0.02 | 0.89420 |
| GB | **Yes** | **NONE** | **1.71 +/- 0.27** | **0.90 +/- 0.02** | **0.87820** |
| AB | No | SRB w/ 10 | 1.81 +/- 0.20 | 0.88 +/- 0.03 | 0.65315 |
| AB | Yes | SRB w/ 10 | 1.81 +/- 0.16 | 0.87 +/- 0.02 | 0.67516 |
| NN | No | SRB w/ 5 | 2.79 +/- 2.20 | 0.69 +/- 0.46 | 1.76140 |
| NN | Yes | SRB w/5 | 1.97 +/- 0.43 | 0.86 +/- 0.06 | 1.58534 |
| SVM | Yes | NONE | 1.77 +/- 0.11 | 0.89 +/- 0.04 | 0.31907 |

Table 1: Best results for each model with and without feature normalization for Regression.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model Type | Feature Normalization | Feature Selection | Accuracy | AUC | Runtime (seconds) |
| RF | No | SRB w/ 5 | 0.71 +/- 0.05 | 0.78 +/- 0.05 | 0.77618 |
| RF | Yes | SRB w/ 5 | 0.71 +/- 0.07 | 0.77 +/- 0.05 | 0.79819 |
| GB | No | SRB w/ 10 | 0.72 +/- 0.11 | 0.78 +/- 0.07 | 0.47011 |
| GB | Yes | SRB w/ 10 | 0.72 +/- 0.10 | 0.78 +/- 0.07 | 0.48611 |
| AB | No | SRB w/ 10 | 0.74 +/- 0.07 | 0.81 +/- 0.07 | 0.73517 |
| AB | Yes | SRB w/ 10 | 0.74 +/- 0.07 | 0.81 +/- 0.07 | 0.74517 |
| NN | No | NONE | 0.63 +/- 0.15 | 0.62 +/- 0.16 | 2.61959 |
| NN | Yes | NONE | 0.72 +/- 0.10 | 0.80 +/- 0.09 | 3.37927 |
| SVM | **Yes** | **Univariate w/ Mutual Info** | **0.72 +/- 0.09** | **0.81 +/- 0.09** | **0.07602** |

Table 2: Best results for each model with and without feature normalization for Classification

For Regression models, a Gradient Boost model with feature normalization and no feature selection gave the best results. The model had a RMSE of 1.71 +/- 0.27 and an Explained Variance of 0.90 +/- 0.02. Getting an Explained Variance of 90% is quite high and would be considered a good model in most fields. For this model, 100 estimators (trees) were used, the loss function used was “ls” (least squares regression), the learning rate was 0.1, max depth was 3 and the minimum number of samples to split at was 3.

For Classification models, a Support Vector Machine model with feature normalization and Univariate feature selection using Mutual Info performed the best. The model had an accuracy of 0.72 +/- 0.09 and an AUC of 0.81 +/- 0.09. It also had the fastest run time out of all of the models in the tables above, at 0.07602 seconds. The accuracy and AUC scores were better than what seen in most of the homeworks for this course. For this model, a linear kernel was used, the kernel coefficient “gamma” was set to ‘scale’ and the regularization parameter “C” was 1.

Tables 3 and 4 below show the features that were selected from the various FS methods used for the best models from Tables 1 and 2. Interestingly, the features selected by each method did not change with feature normalization. Since some features were on completely different scales, for example ‘Sh’ ranged from 0 to 135 and ‘Att\_M’ ranged from 0 to 2,474, it would make sense if normalizing the features resulted in different features selected, but this was not the case.

|  |  |
| --- | --- |
| Feature Selection Method | Features |
| SRB w/ 5 features | 'Sh', 'Cmp\_M', 'Att\_M', 'Left', 'Rating19' |
| SRB w/ 10 features | 'Age', 'MP', 'Min', 'Sh', 'Cmp\_M', 'Att\_M', 'Att\_L', 'Left', 'Right', 'Rating19' |

Table 3: Features selected for best Regression models

|  |  |
| --- | --- |
| Feature Selection Method | Features |
| SRB w/ 5 features | 'Age', 'Min', 'Right', 'Dr\_Att', 'Rating19' |
| SRB w/ 10 features | 'Age', 'Min', 'Sh', 'Att\_S', 'Left', 'Right', 'Pass13', 'Dr\_Att', 'Tkl', 'Rating19' |
| Univariate w/ Mutual Info | 'Age', 'MP', 'Starts', 'CrsPA', 'Fls', 'TklW', 'Dr\_Att', 'Megs', 'Tkl', 'Past\_D' |

Table 4: Features selected for best Classification models

**6. Discussion:**

*6.1: Summarization of findings*

The goal of this paper was to see if I could accurately predict player ratings in FIFA 20 using real life data from the previous season, as well as the previous rating in FIFA 19. The results indicated that EA Sports likely use similar data to what was used in this paper when creating ratings. However, even though the RMSE and Explained Variance from the Regression models and the Accuracies and AUCs from the Classification model were all quite good, I believe that EA likely uses other data as well when creating ratings.

Although I do not believe the findings of Ganiyu to be valid, some of the features used in his study might be important as well [9]. For example, potential in FIFA 19 might be useful in helping predict rating in FIFA 20, especially for younger players, as their potential is often much higher than their rating. Table 5 below shows how the features selected relate to the various player attributes in FIFA. Most of the features that were selected fit into 4 of the 6 main attributes: Shooting, Passing, Dribbling and Defense. Data that includes features related to the Pace and Physical attributes could potentially give better results, but these are difficult things to accurately measure in real life play.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Pace | Shooting | Passing | Dribbling | Defense | Physical |
|  | SH | Att\_S | DR\_ATT | Tkl |  |
|  |  | Att\_M | Megs | TklW |  |
|  |  | Cmp\_M | Past\_D | Fls |  |
|  |  | Att\_L |  |  |  |
|  |  | Right |  |  |  |
|  |  | Left |  |  |  |
|  |  | Pass13 |  |  |  |
|  |  | CrsPA |  |  |  |

Table 5: Feature selected related to FIFA attributes. Not in table: Age, Min, MP, Starts, Rating 19

*6.2: Limitations/Future Work*

The main limitations in this research were that only league games were included in the data, club tournaments and international games were not included, and that only 1 league was used. For example, Roberto Firmino, who plays for Liverpool FC, played in 34 EPL games during the 2018-19 season, but played a total of 48 games for the club during the season. 4 of his 16 goals came from non-league games, so not having this data could potentially influence the results. Almost every country in the world has at least 1 professional soccer league, with most having multiple, especially in Europe.

For future work, I would like to include all games played during the season, as well as include other leagues. Adding data for every game played would give a better understanding of each individual players performance throughout the season. Domestic Cups, such as the FA Cup in England, and international club tournaments, such as the Champions League and Europa League, are often where players play their best, as there is more on the line. Unless I am able to find a different source for the data, this would require a lot of web-scraping. By including more observations (players from other leagues), I would be able to increase the validity of my findings, as the models I created might only work well for the EPL, although I doubt this is the case. Using other features outside of in game statistics might be useful as well, such as top speed and weight/height to try to cover more of the attributes from FIFA.

**7. Conclusions:**

EA Sports’ algorithm for creating player ratings in FIFA is likely something that they will never share publicly. The findings of this research indicate that EA likely uses some combination of real life in game statistics and other features that are related to the players, such as Age, Weight/Height and things that would one would see tested at something like the NFL Combine: 40 yard dash, reps at a bench press or vertical jump, when creating player ratings. As most of the models for both Regression and Classification did not vary in performance, one can see how there is not necessarily a correct Machine Learning model to be used in most cases. This shows the importance of additional techniques such as feature selection and hyperparameter tuning, as making small changes to the model can result in large differences in the results. Although there are not any real-world implications from this research, I have been a big FIFA/Soccer fan for most of my life, and this project is something that I have thought about doing for a while.

**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] Delgado, Fernando (08/1997). "MAJOR LEAGUE SOCCER The Return of the Foreign

Sport". Journal of sport and social issues (0193-7235), 21 (3), p. 285.

[3] Smith, R. (2016, Oct 14). To know your enemy, just press start.*New York Times (1923*

*Current File)* Retrieved from <http://ezproxy.depaul.edu/login?url=https://search-proquest->com.ezproxy.depaul.edu/docview/2310093378?accountid=10477

[4] Adachi, P. J., & Willoughby, T. (2016). Does playing sports video games predict increased

involvement in real-life sports over several years among older adolescents and emerging

adults?. Journal of youth and adolescence, 45(2), 391-401.

[5] Adachi, P. J., & Willoughby, T. (2015). From the couch to the sports field: The longitudinal

associations between sports video game play, self-esteem, and involvement in sports.

Psychology of Popular Media Culture, 4(4), 329.

[6] Anderson, C. A., & Carnagey, N. L. (2009). Causal effects of violent sports video games on

aggression: Is it competitiveness or violent content?. Journal of Experimental Social Psychology, 45(4), 731-739.

[7] Aliyari, H., Kazemi, M., Tekieh, E., Salehi, M., Sahraei, H., Daliri, M. R., ... & Hadipour, M.

M. (2015). The effects of fifa 2015 computer games on changes in cognitive, Hormonal

and brain waves functions of young men volunteers. Basic and clinical neuroscience,

6(3), 193.

[8] Soto-Valero, C. (2017). A Gaussian mixture clustering model for characterizing football

players using the EA Sports' FIFA video game system.[Modelo basado en agrupamiento

de mixturas Gaussianas para caracterizar futbolistas utilizando el sistema de videojuegos FIFA de EA Sports]. RICYDE. Revista Internacional de Ciencias del Deporte. doi: 10.5232/ricyde, 13(49), 244-259.

[9] Ganiyu, Mubarek. (2019). How To Predict Your Best Footballers’ FIFA 20 Ratings.

<https://towardsdatascience.com/how-to-predict-your-best-footballers-fifa-20-ratings-9d1147ced401>

*Links to data:*

<https://fbref.com/en/comps/9/1889/stats/2018-2019-Premier-League-Stats>

<https://wefut.com/player-database/20>