# Data Analysis of EA Sports' FIFA: Using Machine Learning Techniques to Predict Player Ratings, Skills, and Positions

Danny Kim

## I. Introduction

EA Sports' FIFA is one of the most popular video games in the world: players from across the world enjoy this hyper-realistic soccer game. In fact, this remarkable similarity between FIFA and the real world, both with respect to its gameplay and its representation of real professionals, is part of what makes it so popular to both fans and sport researchers alike. In our collective capacity as both soccer fans and data scientists, we aimed to use our knowledge of machine learning techniques to better understand a range of elements, including player attributes and professional teams' coaching capabilities.

## II. The Data

We scraped our data from sofifa.com, a website containing game data on the players in several of the most recent editions of FIFA. This scraping process proved to be quite arduous, but we were ultimately able to create a robust and polished dataframe.

The general structure of the website is such that its home page includes a list of up to 60 players (which is customizable by search) as well as a few select attributes (which is also customizable in table form) for any given edition of the game (for example, FIFA 20). Most importantly, each player in the list has a hyperlink to a personalized page where visitors can see the complete list of that individual's attributes and stats.

Thus, we were able to leverage this structure by looping through each of the pages for the 6 years of interest (FIFA 15-20). We essentially sorted the players in a descending order by overall rating and looped through each of the pages in order to get data for every player in a given edition of the game. Using BeautifulSoup, we scraped each of the players' names along with their personalized links, ID numbers, and "Big 6" attributes (Pace, Shooting, Passing, Dribbling, Defending, Physical), which we were unable to take from the individual pages, as we will discuss further below.

Once we had this information, we then needed to make sure that we were only dealing with individuals who appeared in every year of interest in the game. We made this decision since we knew we wanted time-series data where we could use past years to predict the current, so players needed to be present across all editions of interest. It also decreased the time-complexity of the scraping. Thus, we compared the IDs we had for each year, and dropped the individuals for whom we knew we would not have data for every year. Then, we looped through the links in each six editions of the dataframes described, and pulled attribute data for each player.

This part proved to be one of the most challenging aspects of the data collection process. The data storage both within and across pages was fairly inconsistent, so we had to deal with many edge cases in order to make sure we had all of the data. For example, some of the attributes, such as nationality, were stored via labeled images, while others such as skill ratings were stored through list elements, and real positional ratings were stored in a custom format with custom tags. Additionally, the “Big 6” attributes were stored via an unlabeled image of a hexagon, so it was impossible to pull this data from these pages, which is why we had to go back and get them from the home page. Across players, we saw that some players were missing data others had, such as a national team or a club team, and others had data stored in different orders (club before national or vice-versa). Additionally, across years, we saw that there were different attributes that came and went, such as the release clause, which wasn’t there in FIFA 15.

Ultimately, this non-uniformness in the data meant we had to consider a host of edge cases, some of which we didn’t even notice until after we had spent several hours letting our scraping code run, as they only appeared on a small handful of pages. This resulted in our scraping code being quite complex, but ultimately successful, in large part due to a number of conditionals enabling it to work on every single player in every single year.

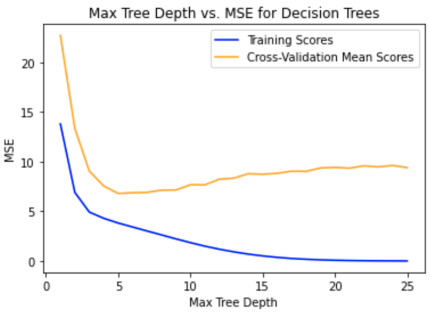
Our final dataset had 5,173 rows and 581 columns. As discussed, this included all of the players who appeared in all 6 editions of the game, and features across the 6 years. There were a handful of “missing values,” but these were not much of an issue as these all provided information: for example, a player not having a national team position or number is useful information, as it means that player is not one of the top players of their nationality.

# III. Modeling

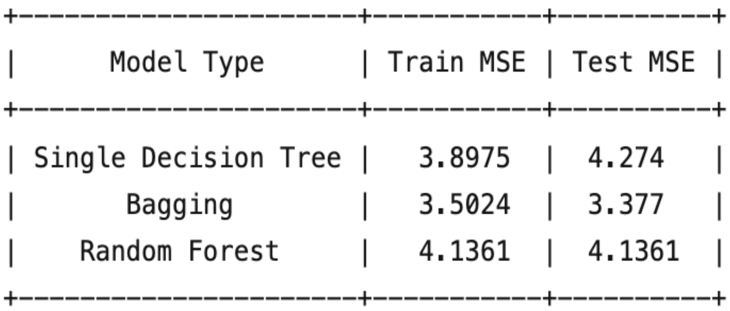
## Part A) Ranking Players by Predicting Overall Ratings

The first exercise we undertook was to try to predict players’ overall ratings using other attributes from the previous year. We trained using the attributes from FIFA18 as the predictors and the overall from FIFA19 as the response variable for all individuals in our dataset. Our first thought was to use a linear regression, in order to try and maintain some model interpretability and enable us to make statements about what we should expect to happen to the overall rating when a single attribute increases by a point by using our model coefficients. However, upon further thought, we recognized that since we had so many predictors with strong correlations between them, we would likely get small coefficients that are confounded by each other and overfit to the data, which would not serve much use. Thus, we decided to proceed with a decision tree regression model in order to maintain some interpretability but also make sure that we were weighting the most important predictors most heavily.

We began by tuning our max_depth hyper-parameter to determine the best possible depth for a decision tree on this data. This exercise is incredibly important, because too-shallow trees can underfit while too-deep trees can overfit, so this step can help us optimize



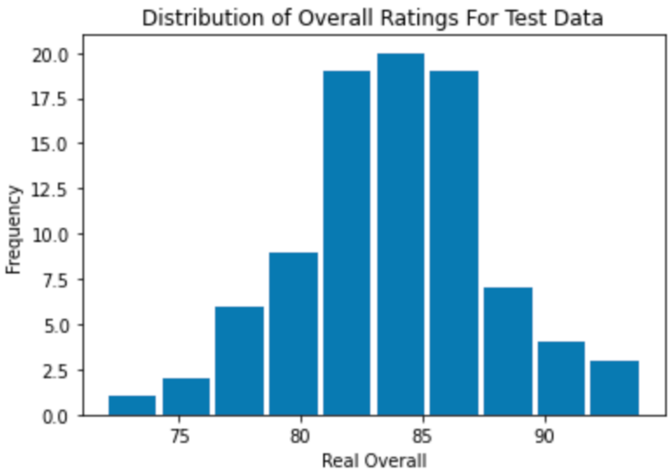
our model. We did this tuning via 10-fold cross validation, considering depths between 1 and 25, and found that the optimal tree-depth is 5, as can be seen in the figure here.



```
+----------------------+-----------+----------+
|      Model Type      | Train MSE | Test MSE |
+----------------------+-----------+----------+
| Single Decision Tree |   3.8975  |  4.274   |
|       Bagging        |   3.5024  |  3.377   |
|    Random Forest     |   4.1361  |  4.1361  |
+----------------------+-----------+----------+
```

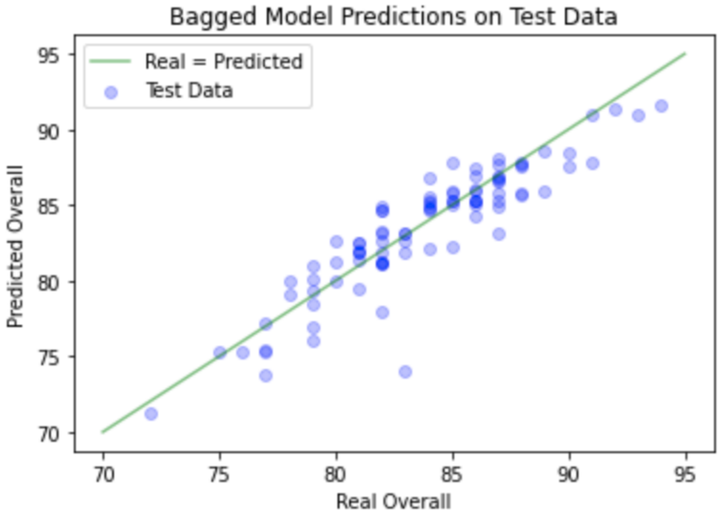
We then went on to take steps to try to enhance our model, namely by Bagging and Random Forest. Using our tuned max_depth hyper-parameter of 5, we fit both of the above to the train data, and then computed train and test MSEs for all three models, as seen in the figure at left.

Bagging significantly outperforms both the single decision tree and the Random Forest with respect to MSE, and, intuitively, this makes sense. We saw when looking at the feature importance of our tree that by far the most important feature was the market value, given in EUR, of the player. In Random Forest, when we limit the amount of predictors we can use at each split, we limit our ability to use what could be our very best predictor, which makes our model worse overall than the Bagging model. This hypothesis is corroborated by the fact that our Random Forest model performs even worse than the single tree on the train data.



Looking at the test data to assess our Bagging model's performance, it is helpful to first take a look at the distribution of our response variable, which we can see in the histogram above and at right. Notably, the overall ratings are quite high. This intuitively makes sense, given that we are using a set of players who play for the top clubs in the world.

Going on to actually assess our performance graphically, we can observe the scatterplot at left. This graph compares the individuals' real overall ratings with our predictions of their overall ratings. The green line is the y = x line, and points that lie on this line are points that we have predicted perfectly. Overall, we see that our

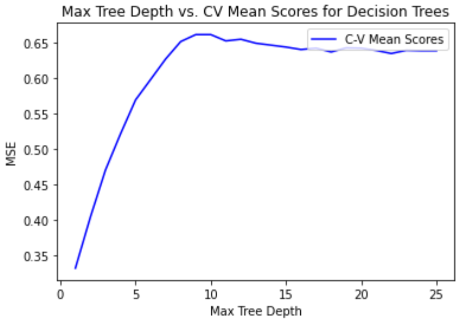


model did quite well on the test data, as many of the points lie on the line.

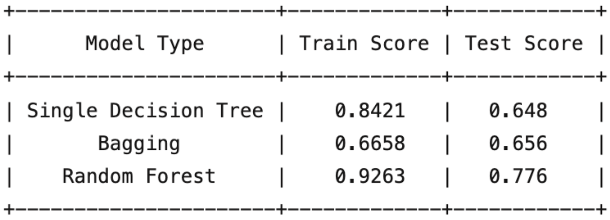
Therefore, if we were to store our results and recombine them with their player names, we could use our model to determine a predicted ranking of players for a future year's game (for example, our model could use FIFA 22 attributes to predict a ranking of FIFA 23 players).

## Part B) Classifying Player Position

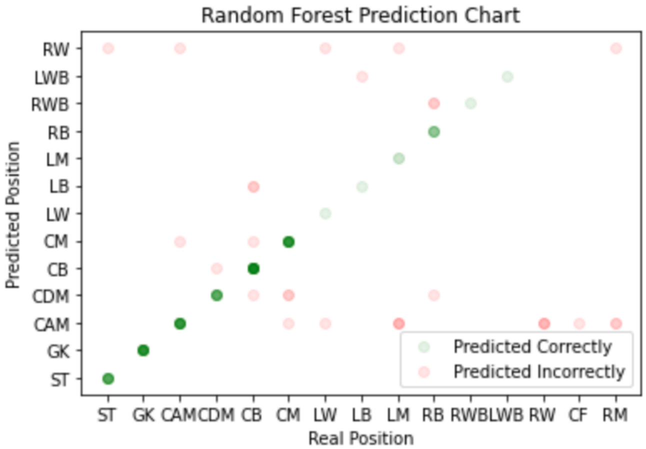
Every soccer player has their strengths and weaknesses, which is portrayed by their skill ratings. For example, some players may be better than others at shooting, while others may be better at skills like passing. From this knowledge, it is plausible to think that certain traits and skills that the players possess may determine what kind of position they play. A defensive player generally has higher defending and physical stats, while attackers are faster and having better shooting stats. We were able to do an exploratory data analysis of the data at hand. We looked at a few of the most important stats of a soccer player, such as pace, passing, shooting, defending, etc. It was evident that there were patterns that displayed how players in different positions had certain skill sets that were common. As a result, with the numerous predictors that display their skills and traits, we ran multiple classification models to attempt to accurately predict a player's position.



The first model that we ran was a decision tree. To figure out which tree depth was the most optimal, we ran a 10-fold cross validation. From this analysis, we concluded that a max_depth of 10. After fitting the decision tree and running it on our test data, we ended with a test accuracy of .656. To see if improvements could be made, we ran a random forest model as well to predict player positions. This model had a test accuracy of .776, which was significantly higher than our decision tree. Lastly, we ran a bagging model with the hope that it would perform better than our random forest model, but it had a lower test accuracy of .656. These test accuracies were surprisingly high, as it would be difficult to see the difference between similar positions, such as left wings/right wings and left backs/right backs. In the plot to the right, we were able to see which positions we were able to predict correctly and which we were not. The green points indicate



```
+----------------------+-------------+------------+
|      Model Type      | Train Score | Test Score |
+----------------------+-------------+------------+
| Single Decision Tree |    0.8421   |   0.648    |
|       Bagging        |    0.6658   |   0.656    |
|    Random Forest     |    0.9263   |   0.776    |
+----------------------+-------------+------------+
```



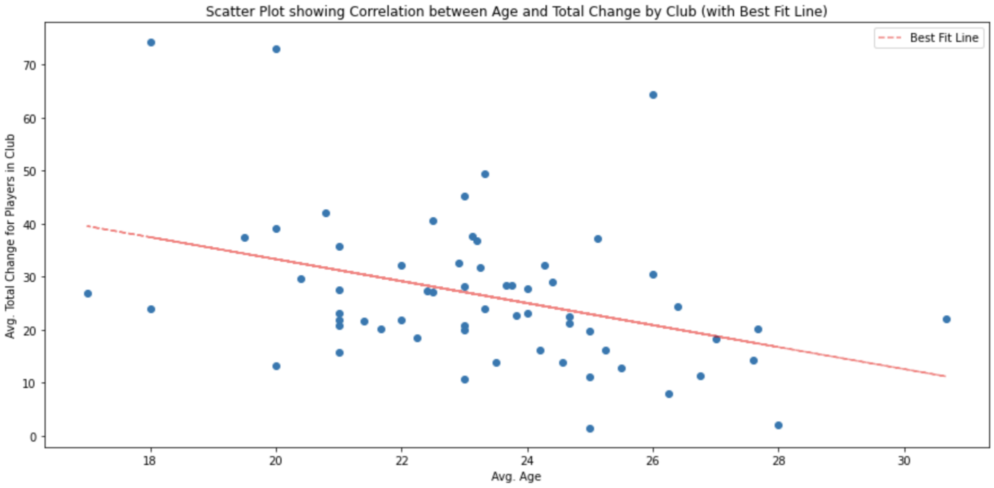
the correctly predicted players, and the red points indicate the incorrectly predicted ones. Most of our points are green, but we see that some of the positions on the left and the right side were confused, which is understandable. However, overall, we were able to end up with a relatively good prediction random forest model to classify the player's position based on their skill ratings and traits.

## Part C) Which Club Has The Best Staff?

Determining which club has the best staff is not a purely objective task, but rather requires decision making as to which variables reflect staff performance. Additionally, there is no metric for staff rating on sofifa.com so we had no values to which we could compare our model. So, for this task, we chose to examine differences in a player’s overall rating, his big six attributes (pace, shooting, passing, dribbling, defending, physical), and his value between 2015 and 2020. We chose these variables out of the range of available predictors since we believed increases in these variables best reflected the success of staff in developing players. Improved overall, big six attributes, and value indicate a better player in 2020 than 2015 which in turn indicates at least some degree of staff success in developing these better players (especially averaged across players in a club).

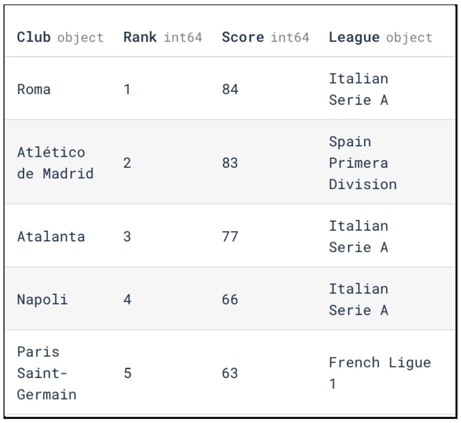
To properly weight these different predictors, we chose to sum changes in the big six and overall scores into an “attribute change” value with each of the big six weighted at ⅙ that of the overall (thinking an increase in overall represents a bigger improvement in player than any one statistic). We scaled this “attribute change” and the “value change” with MinMaxScaler to obtain equal scales, and took a simple average of these two to determine what we called “total_change” which reflected staff performance on a 0-100 scale. By averaging, we assumed that changes in player value and player attributes were equally significant in determining club staff success, which seems to be a reasonable assumption.

However, one concern we had was that this “total_change” variable, representing a staff’s success in improving players, may be correlated with age and thus misrepresented a club staff’s true ability to improve players. Clubs with younger players on average may have higher score improvement without necessarily being better at development since younger players tend to improve with time while older players tend to get worse. Regression analysis confirmed this as it showed the correlation between age and “total_change.”



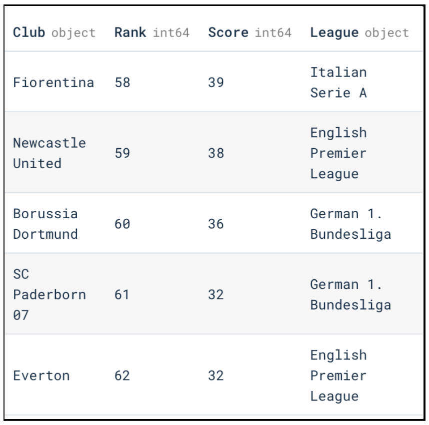
To solve this issue, we used the residuals from our regression for our score. Residuals represent the relative performance of the staff relative to what we'd predict given the age of their players, a more accurate statistic. Using this "relative performance" metric, we standardized and scaled to create a score variable for the club staff with mean 50 and standard deviation of 10, and ranked clubs accordingly. From this we determined the Top 5 and Bottom 5 Clubs:

**Top 5**



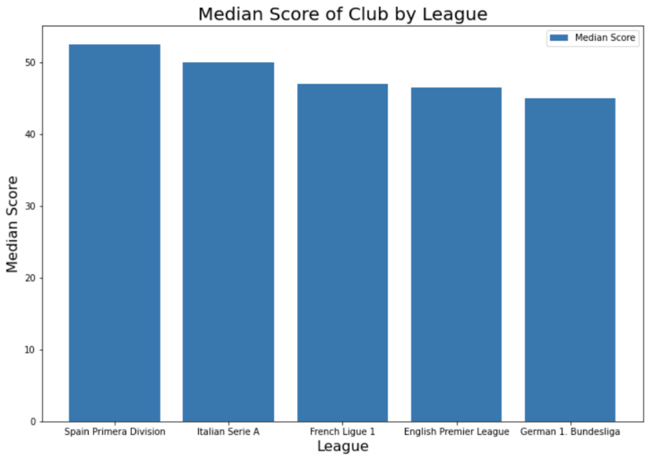
| Club object | Rank int64 | Score int64 | League object |
|---|---|---|---|
| Roma | 1 | 84 | Italian Serie A |
| Atlético de Madrid | 2 | 83 | Spain Primera Division |
| Atalanta | 3 | 77 | Italian Serie A |
| Napoli | 4 | 66 | Italian Serie A |
| Paris Saint-Germain | 5 | 63 | French Ligue 1 |

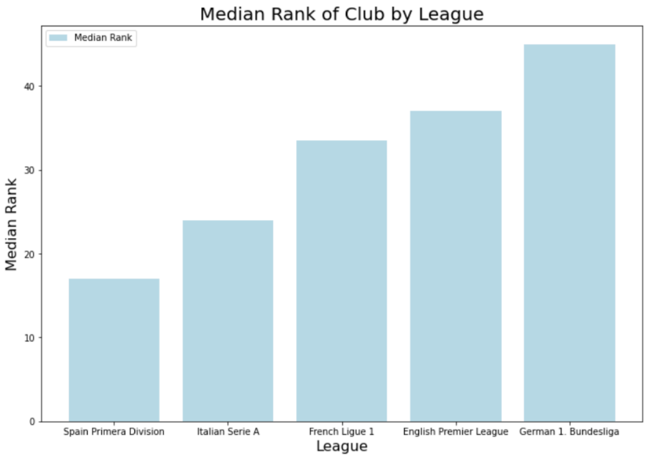
**Bottom 5**



| Club object | Rank int64 | Score int64 | League object |
|---|---|---|---|
| Fiorentina | 58 | 39 | Italian Serie A |
| Newcastle United | 59 | 38 | English Premier League |
| Borussia Dortmund | 60 | 36 | German 1. Bundesliga |
| SC Paderborn 07 | 61 | 32 | German 1. Bundesliga |
| Everton | 62 | 32 | English Premier League |

We were also able to determine the success of the club staff across the different leagues, determining that La Liga and Serie A had the highest median scores of the Big 5 leagues and Bundesliga had the lowest out of the five. One note of interest is that there are a few outlier data points on the upper end of score since a few clubs (e.g. Roma) had only one or two players across the 5 years and him/they improved greatly. Generally, these outliers are averaged away within a club with many members, however with few players this doesn't happen and thus our scores are right-skewed.

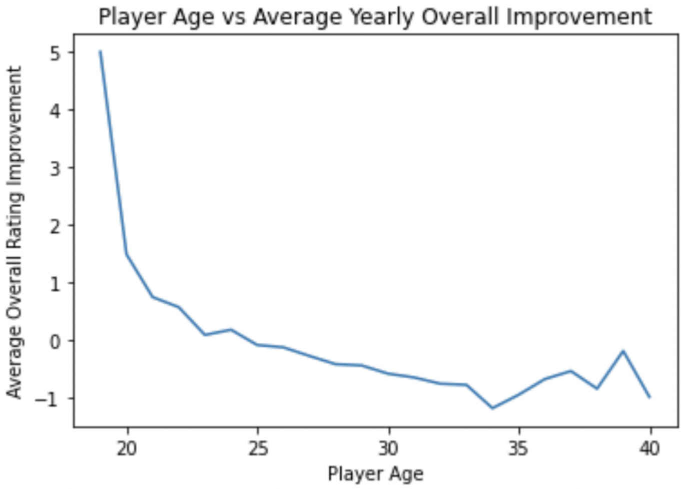




## Part D) How Things Will Be In 2021

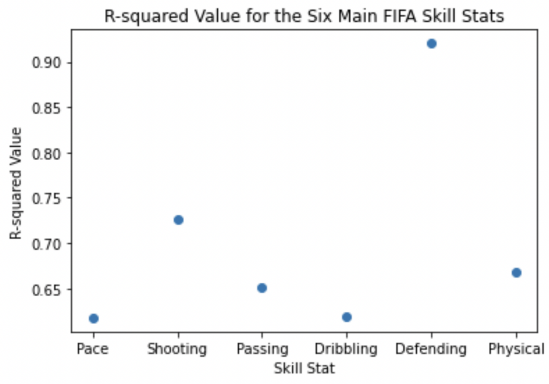
In Part D, the task at hand was to predict each player's six main skill statistics (pace, shooting, passing, dribbling, defending, and physical) using skill statistics from prior years. Many predictor variables were used since each player has 29 skill statistics that feed into their main statistics. For example, sprint speed, acceleration, and agility all contribute to a player's general "pace" rating.

Before beginning the data cleaning and modeling aspects, we first performed some exploratory data analysis (EDA). The most finding during our EDA was the relationship between a player's age and the improvement (or decline) in their skill statistics. As shown in the accompanying graph, the typical player's overall rating improves by 3-5 points year-over-year when they are teenagers. This shows that they improve a lot when they are young since they have not yet maximized their potential. Once a player reaches age 25, their overall rating tends to start to decrease. This finding is relevant since it shows us that age will be a relevant predictor of a player's skill rating the following year: A young player will likely see an improvement in their skill statistics while an older player will likely see a decline.



After EDA we cleaned and stored the data. We quickly realized that this part requires running six separate models, one for each of the main skill statistics listed above. Rather than having one x_train, y_train, x_test, and y_test, we had one for each major skill statistic (e.g. an x_train_pace, y_train_pace, x_test_pace, and y_test_pace). This is because for each player we are building a model to make six separate predictions, one for each skill statistic.

To create each of these train and test data sets, we first looked at our dataset and manually categorized our predictors based on what skill stat they may reasonably predict (for example, ‘shot power’ is a reasonable predictor of a player’s ‘shooting’ rating but not of their ‘dribbling’ rating). We then created an x_train dataframe for each of the six skill stats. Each one of these DataFrames contained data on over 5,000 players for each of the data points deemed to be a reasonable predictor. Our y_train was the resulting main skill statistic. Our testing sets were structured similarly, except they contain data from the following year.



We decided to apply a linear regression model to the data. For each skill statistic, we calculated the train and test mean squared error (MSE) as well as the r-squared value. The accompanying graph shows the resulting r-squared values for each of the six main skill statistics. We found these results interesting since the predictive power of the model varied greatly across skill areas. For example, our model explains nearly all of the variation for players’ “defending” statistics, but only 62 percent of the variation for players’ “dribbling” statistics.

# IV. Discussion

*Conclusions:* With our machine learning techniques, we were able to create relatively accurate models to predict player attributes and professional teams’ coaching capabilities. Now, we have a better understanding of what attributes constitute a good player, and what kind of teams have good coaching staffs. It is amazing to see how data from a video game allows people to see the important factors in soccer players and teams, and there are so many more new findings we could strive to uncover through FIFA.

*Limitations:* One key limitation in our project is the time inconsistency of the data. Specifically, many players do not exist in versions 15 through 20 of the FIFA game. We had to decide whether to include all players that made a single appearance in any version or only those who exist across all years. Ultimately, we chose to limit the set to only those existing across versions to obtain more robust data, albeit on fewer players. This was a tradeoff we had to accept and a limitation of the data itself.

Another data limitation is that we are restricted to only data from the FIFA game. While trying to predict future overall or future skill attribute values, it would be helpful to have real-world data for players. This real-world data directly influences the future FIFA ratings and would be helpful in a prediction. For instance, if a player has a breakout season, this would be hard to predict using just previous years' FIFA data but easy using current real-world data.

Additionally, our model from Part C relies on large assumptions about what factors reflect success of club staff in developing players. Without a dataset of club staff ratings to compare to, our model relies heavily on assumptions and is really what we make it out to be. Thus, while we attempted to be thoughtful and meticulous in creating that model, we recognize there is inherent bias in which predictors we use and which we do not.

*Future Directions:* Though our analysis is comprehensive, we are left with several exciting areas that we could dive deeper into in the future. For example, it is possible that FIFA data could be used to predict the outcomes of real-world soccer games. In September of each year, FIFA publishes its ratings for the following year (for example, FIFA 22 ratings are published in September of 2021). Since the video game is designed to be as realistic as possible, it is reasonable to expect that team ratings may predict which teams have successful seasons the following year. Similarly, player ratings may predict which players have award-winning seasons the next year.

The methods used in this study could also apply to soccer scouting (i.e. identifying players to recruit to a professional team). Although we never created such a model for this project, we scraped the necessary data to be able to predict a player's wage and transfer market value based on their skill statistics and overall rating. Building this model could allow us to identify players that are undervalued based on FIFA's assessment of their abilities. Within the game, this model could allow players to find the most cost-effective players to buy for their team. In real life, such a model could help professional soccer scouts find undervalued players in the transfer market. Of course, this model would be more effective if it were built using real-world data (i.e. real player wages, valuations, and in-game performance statistics). This idea of using a data-driven approach to find undervalued players was popularized in the book *Moneyball* by Michael Lewis, which later became a successful film.

---

Data provided by: sofifa.com