FIFA Analysis

Joseph Mannarino

11/29/2020

What determines the overall success of a player in FIFA?

FIFA is a football simulation video game developed and released annually by Electronic Arts under the EA Sports label. As of 2011, the FIFA franchise has been released in 18 languages and available in 51 countries. It is listed in Guinness World Records as the best-selling sports video game franchise in the world. By 2019, the FIFA series had sold over 282.4 million copies. When the series began in late 1993, it was notable for being the first to have an official licence from FIFA, the world governing body of football. This means EA sports has many exclusively licensed leagues including leagues and teams from around the world. Allowing the use of real leagues, clubs and player names. All real world football players that are included in the game (roughly 18,000) are described by a set of attributes, which determine how good they are in-game. The Main indicator of the quality of a player is their overall score. The overall score is a net of all a player's statistics. There are a total of 34 attributes for each player. Five of which are used to describe goalkeeping abilities and 29 are used to describe abilities of an outfield player. All the players are described by all 34 stats, but goalkeeping abilities have no impact on outfield player's overall score and vice versa.

As of right now there are 10 million people who are enjoying the ability to create their own teams and play with their favorite players from around the world. A huge part of playing the game is building the best team possible with the best players possible. Gamer's will constantly be looking at player stats trying to decide if a player is good enough to put on their team. In a study by Grzegorz Chadysz, he attempts to divide players up into different categories each with their own advantages and disadvantages. In his study he concludes that there are Five different categories a player can put into, Offensive, technical, Defensive, Fast and dribbling. In turn, player belonging to a defensive group tends tends to have terrible attributes in every other part of the game (aside from passing). However, fast players are usually very physically weak. It would be advantageous to know what stats gamer's should be focused on when deciding a player to put on their team. In turn, it would be advantageous to know which players are underrated in terms of overall score and price. This would allow gamer's to get better players at a cheaper price. In this study we will use regression analysis to determine how a players overall score is achieved and the affect each attribute has on it. We will also dive into what affects the success of player's based on wage and overall score. In addition, we will use classification analysis to predict a players preffered foot and discuss several interesting topics found while exploring the data.

Data Exploration

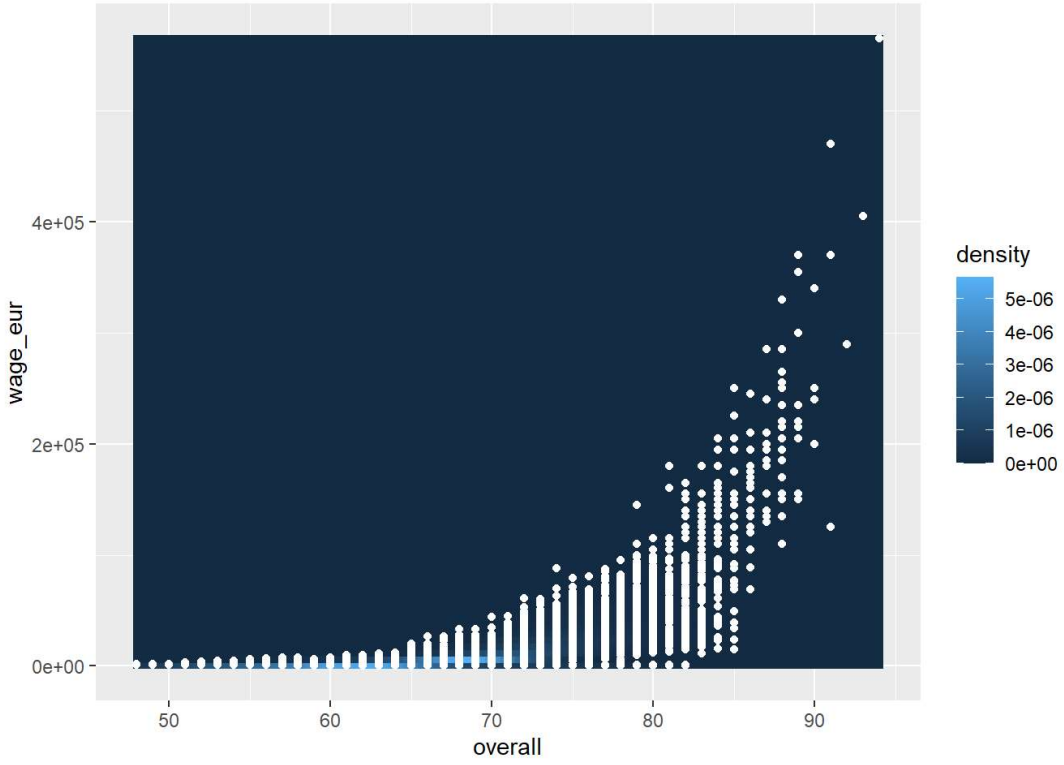
The data set was retrieved from Kaggle and consists of about 18,000 players from FIFA 20. There are 104 total columns, several of which will not be useful in this analysis. The seven main columns that will be significant are overall, pace, shooting, passing, dribbling, defending and physic. These are the seven attribute's shown on the front of every player card in FIFA. Overall is the combined score of all the attributes each player has. In turn, the average gamer will focus on these six attribute's when deciding if a player is good enough for their team. However, there are a multitude of other attribute's that make up a player's overall score. Below is a preview of the data set and the columns that come with it.

Preview of the main dataset

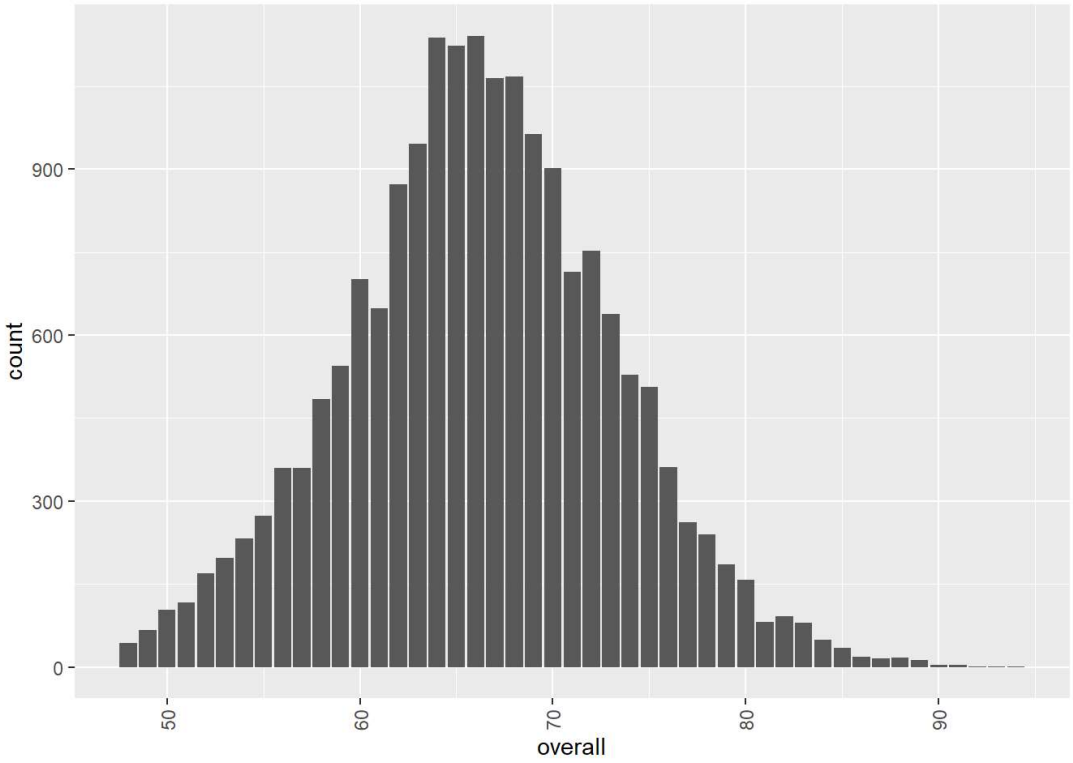
short_name	age	nationality	club	overall	wage_eur	player_positions	preferred_foot	weak_foot	skill_moves	wo
L. Messi	32	Argentina	FC Barcelona	94	565000	RW, CF, ST	Left	4	4	Me
Cristiano Ronaldo	34	Portugal	Juventus	93	405000	ST, LW	Right	4	5	Hig
Neymar Jr	27	Brazil	Paris Saint-Germain	92	290000	LW, CAM	Right	5	5	Hig
J. Oblak	26	Slovenia	Atlético Madrid	91	125000	GK	Right	3	1	Me
E. Hazard	28	Belgium	Real Madrid	91	470000	LW, CF	Right	4	4	Hig

Visualizations

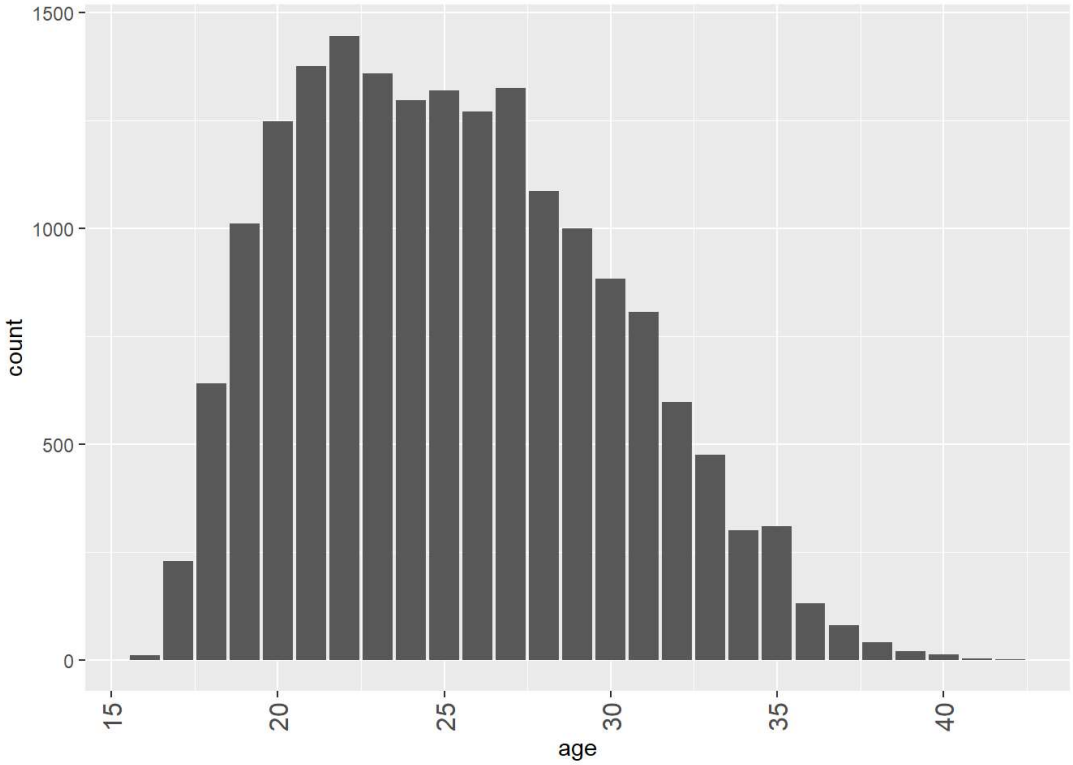
Now lets dive into some interesting relationships in the data and discuss the distribution of players in the data set.



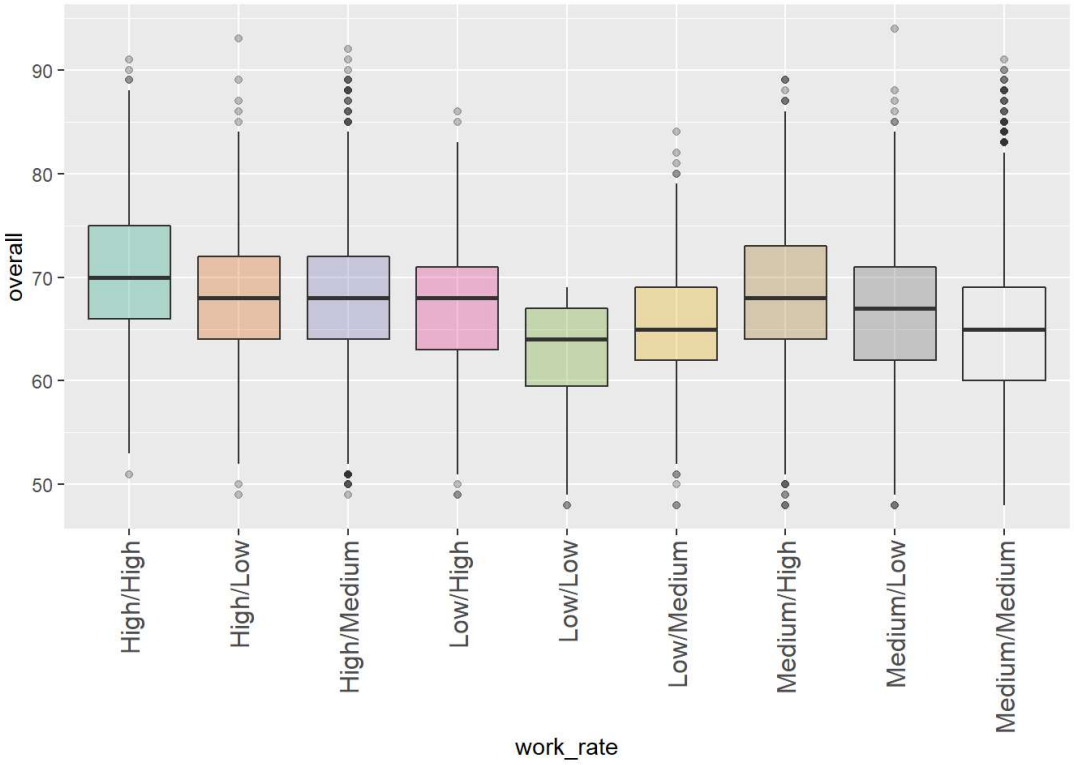
Here we can see the relationship between overall score and wage. The results show a positive relationship between overall score and wage. Thus, the higher the overall score, the higher the wage and vice versa. Since, these variables are measurements of success it would make sense that this is true.



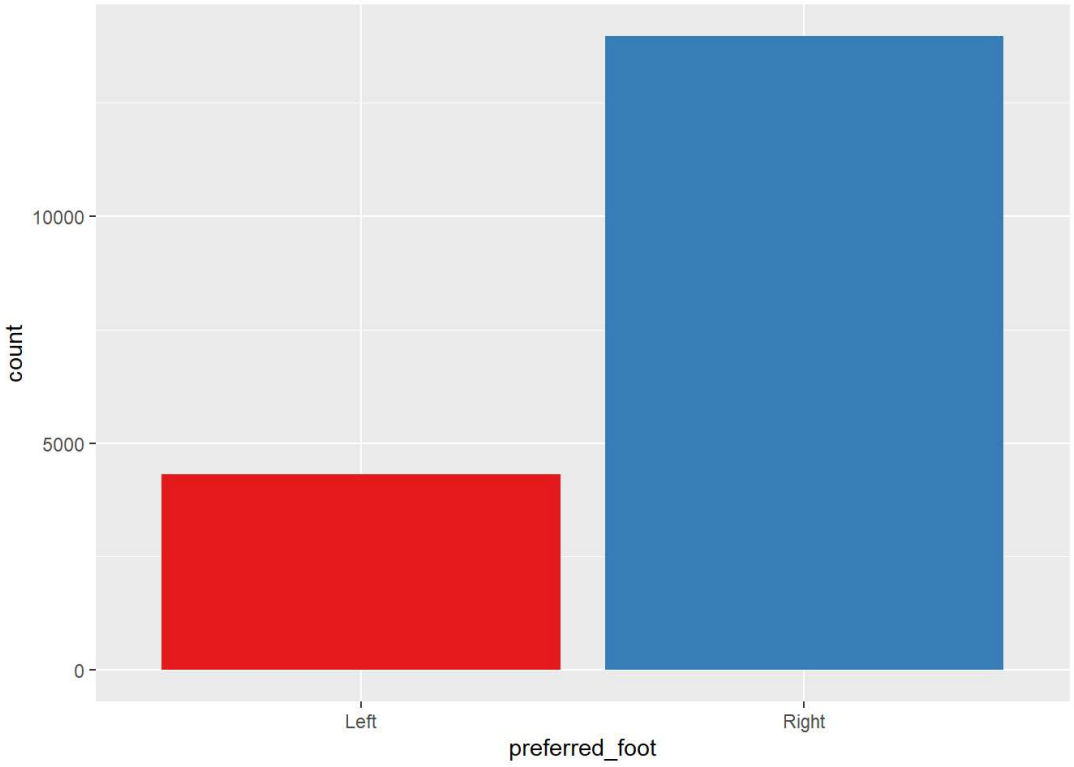
This graph shows the distribution of overall score in the data set. The results conclude that most of the players in the data set have an overall score in the range of 60-75. In turn, we can see that there are very few player's with an overall score over 85 and an overall score below 55.



Now let's discuss the distribution of player's age in the data set. The results of this graph tell us that most of the players in the data set are between the ages of 20-30. An interesting assumption is that most soccer players will start their professional career around the age of 17 and retire around the age of 35-40. Since age is such an important factor in professional sports, it is safe to assume that gamer's should be cautious choosing players that are above the age of 35 because as age increases work rate will decrease. Lets take a look at the relationship between work rate and overall score.



It is clear that player's with a high to high work rate yields on average the highest overall score and players with a low to low work rate yields on average the lowest overall score. This is important because it clearly shows that gamer's should be choosing players that have a high work rate. In turn, gamer's should definitely be avoiding players with a low work rate.



Lastly, in my experience there is an emphasis on player's who are left footed. This graph shows us the distribution of players who are left footed versus players who are right footed. From the graph we can see that there are about 3 times more right footed players in the data set than left footed players. This is interesting because it reveals not only that most soccer players are right footed, but also that people who play positions on the left side of the field will be harder to find.

Regression Analysis

As I stated before, there are seven main attribute's that are displayed on each player card in FIFA. These attribute's are overall, pace, shooting, passing, dribbling, defending and physic. In turn, it would be advantageous to know what stats gamer's should be focused on when deciding a player to put on their team. Our first regression model will be developed to answer this question based on the six main attribute's of a player. The regression model is:

$$\text{overall} = \beta_1 \text{pace} + \beta_2 \text{shooting} + \beta_3 \text{passing} + \beta_4 \text{dribbling} + \beta_5 \text{defending} + \beta_6 \text{physic}$$

Characteristic	Beta	95% CI[1]	p-value
pace	-0.01	-0.02, -0.01	<0.001
shooting	0.11	0.10, 0.12	<0.001
passing	0.07	0.06, 0.09	<0.001
dribbling	0.27	0.26, 0.29	<0.001
defending	0.11	0.11, 0.12	<0.001
physic	0.26	0.25, 0.26	<0.001

[1] CI = Confidence Interval

The regression model above is explaining about 70% of the variance in overall score and each independent variable is significant based on their p-values. We are not capturing all of the variance in overall score due to the fact that we left out the other attribute's in the data set and only included the six main attribute's. This is because we are trying to find which attribute's have the most influence on overall score when looking at the face of a player card. In other words, which attributes a gamer should be focused on when looking at the face of a player card. To answer this question we are going to discuss the interpretation of each independent variable and its coefficient. First off, the two attribute's with the highest coefficient's are dribbling and physique. For every 1 point that dribbling goes up, an extra .27 points will be added to the overall score. For every 1 point that physique goes up, an extra .26 points will be added to the overall score. Thus, a gamer's focus when looking at the face of a player card should be on the dribbling and physique attribute because they will have the biggest impact on overall score. The least influential attribute's are pace and passing with coefficients of -0.01 and 0.07.

```
##      pace  shooting   passing dribbling defending    physic
##  1.720788  4.533697  5.637610  6.699433  3.475353  1.787539
```

A test for multicolinearity shows us that passing and dribbling are highly correlated to other variables in the model. Thus, we can probably take one of the two out and not affect the variance being explained by overall score. Instead, we might want to create a model that looks more like this:

$$\text{overall} = \beta_1 \text{pace} + \beta_2 \text{shooting} + \beta_3 \text{passing} + \beta_4 \text{defending} + \beta_5 \text{physic}$$

Our next model is devised to find the most influential attributes including the ones not found on the face of a player card. In this model we will exclude attribute's that are specific to a player position and focus on the attributes that all player's have. The table below show's a summary of the results from the regression and do not include several attributes that were added to this model.

Characteristic	Beta	95% CI[1]	p-value
age	0.82	0.73, 0.91	<0.001
I(age^2)	-0.01	-0.02, -0.01	<0.001
right_footed	-0.18	-0.26, -0.09	<0.001

[1] CI = Confidence Interval

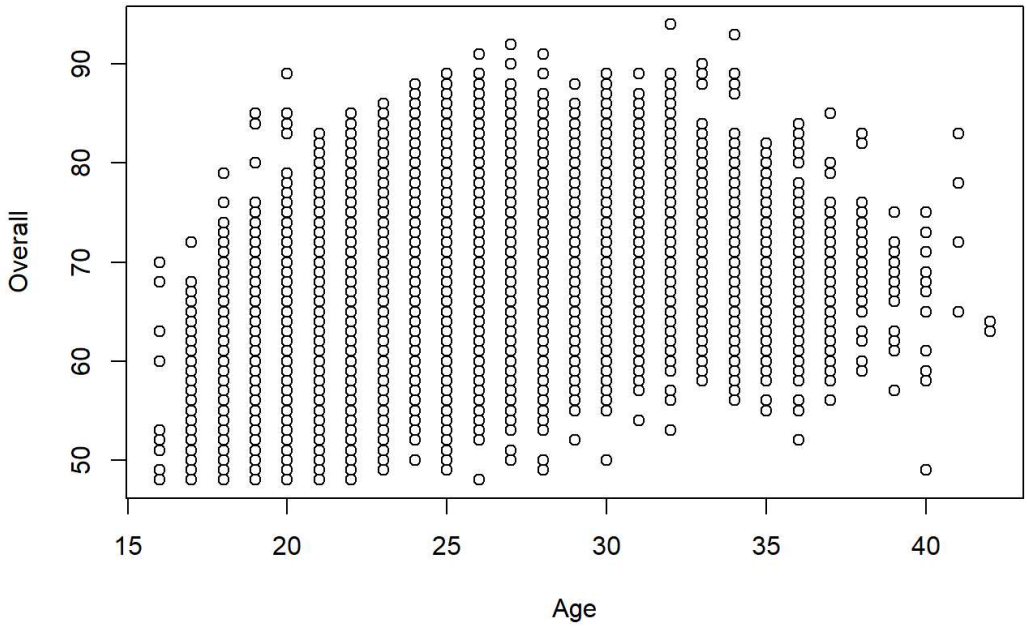
Characteristic	Beta	95% CI[1]	p-value
skill_moves	1.0	0.95, 1.1	<0.001
work_rate			
High/High	—	—	
High/Low	1.0	0.81, 1.3	<0.001
High/Medium	0.43	0.25, 0.60	<0.001
Low/High	1.6	1.3, 1.9	<0.001
Low/Low	-0.88	-1.7, -0.08	0.030
Low/Medium	1.3	1.0, 1.6	<0.001
Medium/High	0.66	0.47, 0.86	<0.001
Medium/Low	0.68	0.45, 0.91	<0.001
Medium/Medium	0.21	0.05, 0.38	0.012
shooting	0.11	0.10, 0.12	<0.001
passing	0.15	0.13, 0.17	<0.001

[1] CI = Confidence Interval

This regression model is explaining about 88% of the variation in overall score. An increase in R^2 is due to the fact that there are several other attributes included in this model that were left out in the first model. To answer our initial question we can take a look at the coefficients of each attribute and compare them to each other. The attributes with the most influential coefficients are skill moves and work rate.This means that when a gamer is deciding what player to put on their team, they should be focused on the players with a high work rate and high skill moves.

An interesting discovery in this regression is how players who are right footed inherently have a worse overall score than players who are left footed. This could be due to the fact that there are three times more right footed players in the data set. However, I believe that in general there are more right footed soccer players on this earth and because of that left footed players will have an advantage over defenders who are mostly defending right footed players. In turn, because there are so few left footed players, they are sought out for by teams and in some ways are rare.

In this model we include the interaction term age^2 in order to capture the law of diminishing returns for the attribute age. This means that as age goes up overall score will increase, but there is an inflection point where overall score will begin to decrease. Since, our data set only includes the ages 15-45 and there is a cap on overall score it will be difficult to find this inflection point. However, if we plot the relationship between overall score and age we can see the affect age has on overall score and we can see the law of diminishing returns in the graph.



In our last model we will be discussing which attribute's have the biggest affect on overall score for goal keepers. The model will include only attributes for goal keeping and exclude all other attribute's because they are not applicable for player's who's position is goal keeper. In turn, we will remodel the data to only include player's who are goal keepers because the other player's are not significant in this model. The regression model is:

overall = $\beta_1$movement_reactions + $\beta_2$gk_diving + $\beta_3$gk_positioning + $\beta_4$gk_kicking + $\beta_5$gk_handling + $\beta_6$gk_reflexes + $\beta_7$gk_speed

Characteristic	Beta	95% CI[1]	p-value
movement_reactions	0.11	0.11, 0.12	<0.001
gk_diving	0.21	0.20, 0.22	<0.001
gk_positioning	0.22	0.21, 0.23	<0.001
gk_kicking	0.05	0.05, 0.06	<0.001
gk_handling	0.21	0.20, 0.22	<0.001
gk_reflexes	0.21	0.20, 0.22	<0.001
gk_speed	0.00	0.00, 0.00	0.4

[1] CI = Confidence Interval

From the results of this regression we can see the affect each attribute has on overall score and determine which goal keeper attribute's will increase overall score the most. The coefficient's of each variable are very similar with goal keeper positioning being the highest at .22. This means that for every point goal keeper positioning goes up it will add an extra .22 points to overall score. This also means that if a gamer is choosing a goal keeper to put on his team they should focus on the positioning attribute. However, diving, handling and reflexes are not to far behind with a coefficient of .21. Ultimately, a gamer should be focused on all 4 of these attributes when choosing a goal keeper. On the other hand, the results tell us that the speed attribute has little to no affect on overall score. Thus, a gamer should not focus on speed and can pretty much ignore it because its coefficient is 0 and it p-value tells us that it is insignificant.

Advanced Modeling

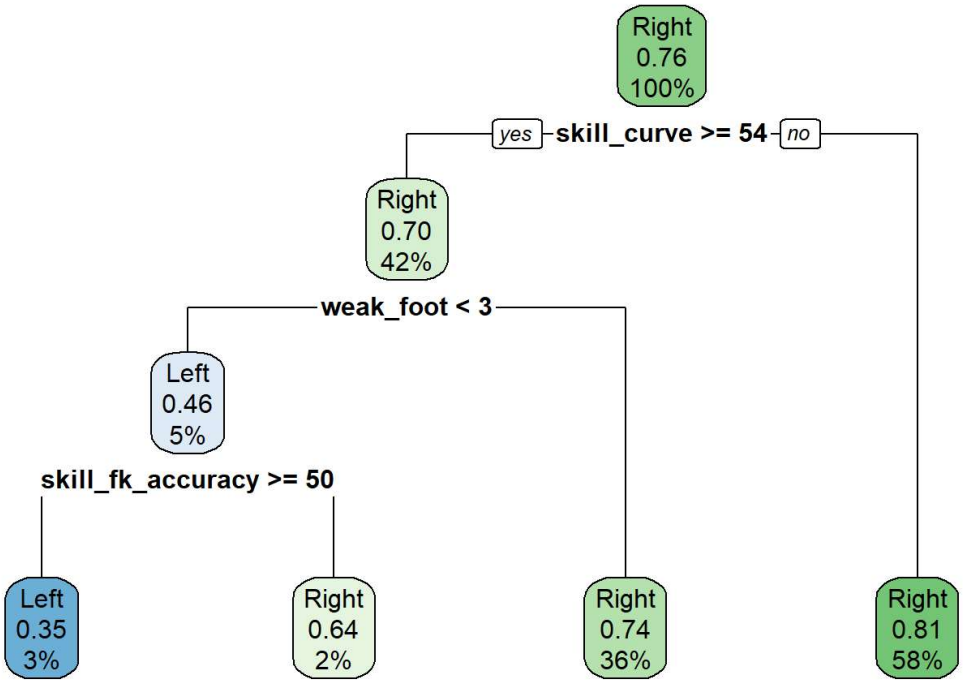
In this section of the analysis we will be using logistic regression and decision trees in order to predict the preferred foot of a player. We will discuss the advantages and disadvantages of both as well as their results.

The first step is to transform the preferred foot column into a 1 for Left footed player's and 0 for right footed player's. Then the data will be split and the regression model is trained on half of the data set. Ultimately, the model is tested on the other half of the data set. Below is a table with the predictions versus the actual players preferred foot.

	Left	Right
Left	262	1809
Right	139	6033

The model had a total accuracy of 67%. Overall, this is a terrible accuracy for a prediction model. Consequently, the logistic regression model will do a bad job of predicting the preferred foot of soccer players. This could be because of a number of things, however predicted someone's preferred foot based of off soccer attributes is not an easy task. It turns out that we are only explaining about 8% of the variance in preferred foot with this model. Which means our goodness of fit is terrible and the variables we are using are not significant to preferred foot at all. We may be over fitting the model and adding to many insignificant variables to the regression. By looking at the regression results we can see that the most significant values are weak foot, shooting, passing, skill_curve, skill_long_passing, mentality_vision. Although, there coefficients are small and will have at most a 0.0879 effect on the outcome of the prediction.

Let's try a decision tree approach and see if it will have a better accuracy when predicting the preferred foot of a player. Once again, we will split the data, train the regression model on half of the data and test it on the other half. Below is a plot of the categories and decisions that the tree consists of.



The plot is displaying the decisions it will make in order to predict whether or not a player is right or left footed. It starts off by looking at the skill curve attribute. If the attribute is greater than or equal to 54 it will make another decision. However, if the attribute is below 54, then it will be predicted as a right footed player. In order for the decision tree to predict a left footed player they need to have a skill curve above 54, a weak foot below 3 and a skill free kick accuracy above 50. Otherwise, the decision tree will call it a right footed player. Let's take a look at the accuracy of the decision tree and discuss how it compares to the regression model.

	Left	Right

	Left	Right
Left	215	1983
Right	110	6970

The accuracy of the decision tree is around 77% which is an increase from 67% in the logistic regression model. This can be explained by a decision tree's innate ability to capture the interaction's between variables. In addition, decision tree's do a better job conducting classification when the data is not linearly separable. Since, we are looking at soccer player's and their attributes it would make sense that the data values for each attribute are all over the place and do not produce a linear pattern.

Conclusion

In this analysis we wanted to discover what measures the success of a player in FIFA, uncover the affects of player attribute's on overall score and predict the preferred foot of player's using logistic regression and decision trees. In doing so, we discovered that when looking at the face of a player card, gamer's should focus on dribbling and physique because they will have the biggest affect on overall score. Overall, the most important attributes to look at when choosing a player to put on your team is skill moves and work rate. A player with high skill moves and high work rate will likely have a high overall score. These two attributes have the biggest effect on overall score when holding all other attributes not specific to a player position constant. In addition, we also looked at what attributes are the most important to focus on when choosing a goalie. It turned out that most of the goalie attributes are similarly significant. However, the main 4 attributes that a gamer should focus on are diving, handling, reflexes and most importantly positioning. Next, we discussed the advantage and disadvantages of logistic regression versus decision tree's, all the while predicting the preferred foot of player's using both approaches. We found out that a decision tree will ultimately have a %10 higher accuracy than a logistic regression model. This is due to a decision tree's innate ability to capture interaction terms and not be bogged down by data that does not represent a linear pattern. Finally, in the search for all of the answers to these questions we discovered some interesting facts about the data. For example, there are three times more right footed players than left footed players. Plus, if you are left footed you inherently have a higher overall score based on our model. Don't forget about the fact that most of the players in the data set are of age 20-30 or that we might assume that most professional player's will start their career around the age of 17 and retire around the age of 35-40.

To sum up, this may have been enlightening for many FIFA gamer's and could help them choose the right player's to put on their FIFA team. However, there are several things that could be done in order to improve upon this analysis. For example, more interaction terms on all of the linear and logistic regression models. Also, It would have been nice to have the ability to predict continuous variable's. This way instead of predicting the preferred foot of a player I can predict the overall score or wage or better yet the price! If I could have predicted the price of a player, this would be monumental for buying and finding underpriced players in FIFA. This could definitely benefit gamer's who buy and sell player's on FIFA for a profit in real life.