

The Effect of Player Attributes on Overall Player Rating in FIFA 19

I. Arvelo, S. Ryu, W. Jeffries, H. Walsh

May 26, 2021

Abstract: This paper attempts to replicate the algorithm to determine the overall rating of players in the FIFA 19 video game to understand which player attributes are most important for each position. The analysis specifically focuses on models that predict the overall rating of defenders. A variety of statistical methods were employed to make sense of the data including Multiple Linear Regression, Lasso Regression, and a Neural Network. Our findings indicate the different skill sets are necessary to be highly rated in each position and standing tackle, interceptions, heading accuracy, sliding tackle, and marking are specifically important in the defensive end of the field. The results of this study apply specifically to the FIFA 19 game since the professional soccer and iterations of the FIFA video game are constantly changing.

Keywords: Multiple Linear Regression, Statistical Interaction, Model Comparison

Introduction

The motivation of this study was to discover which attributes and skills were the most important in determining a player's overall rating. People should care about the findings of this study if they are a fan of the FIFA 19 video game, as this will give them some insight into how the game works and why certain players are rated higher than others. Our response variable of interest is overall rating, which is an integer from 0 to 99 on a player's card which is a weighted average of attributes and international reputation that essentially describes how good a player is at soccer.

Hypotheses:

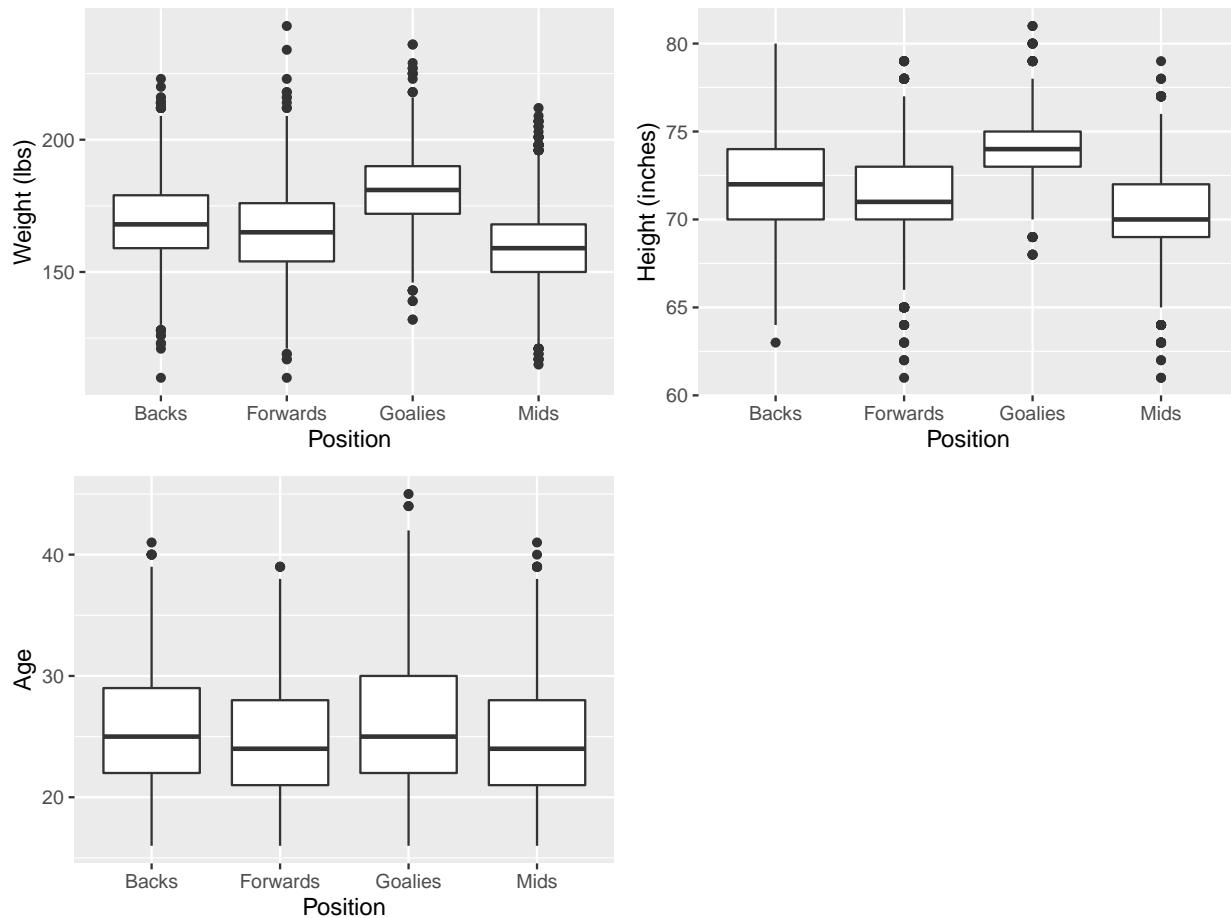
H_0 : All skills and abilities are equally important in determining a player's rating for all positions. H_A : All skills and abilities are not equally important in determining a player's rating for all positions.

Data Collection and Preparation

Data Collection: The “FIFA 19 Complete Data Set” was obtained on Kaggle. It is an 8.72MB data set with 18,207 unique players and 89 variables for each. There were no issues downloading the data set, as it was simply a matter of downloading the .csv file.

Data Preparation: In terms of preparing the data, we first had to clean up the variables and decided to do preliminary data cleaning in JMP. We had to convert Wages and Value from character type in Euros to a continuous numerical variable in dollar units. We also had to remove lbs off the end of Weight and convert it to a continuous numeric variable. There were also variables we had to eliminate because they were not as relevant to our analysis. Photo, flag, logo columns, variables related to the video game itself (i.e. RealFace and Body Types), ID, Jersey number, Release Clause, and Special were the variables we deleted from our data set. At this point, we imported the cleaned data into R and changed height from ft'in character format to numerical variables (in inches) and removed players who did not have data populated for position. A lot of the missing values in our original data set were removed through this process.

Data Collection and Preparation



Goalies tend to be the tallest and subsequently heaviest position group. There are not many unusually old, young, tall or short players either. Overall, physical characteristics are fairly similar across positions

Name	Overall
L. Messi	94
Cristiano Ronaldo	94
Neymar Jr	92
De Gea	91
K. De Bruyne	91
E. Hazard	91
L. Modrić	91
L. Suárez	91
Sergio Ramos	91
J. Oblak	90

Many of the players with the highest overall ratings are household names such as Lionel Messi and Cristiano Ronaldo. Although overall rating is out of 99, the highest ranked players do not surpass a rating of 94. Almost all 10 of these players play in La Liga or the Premier League.

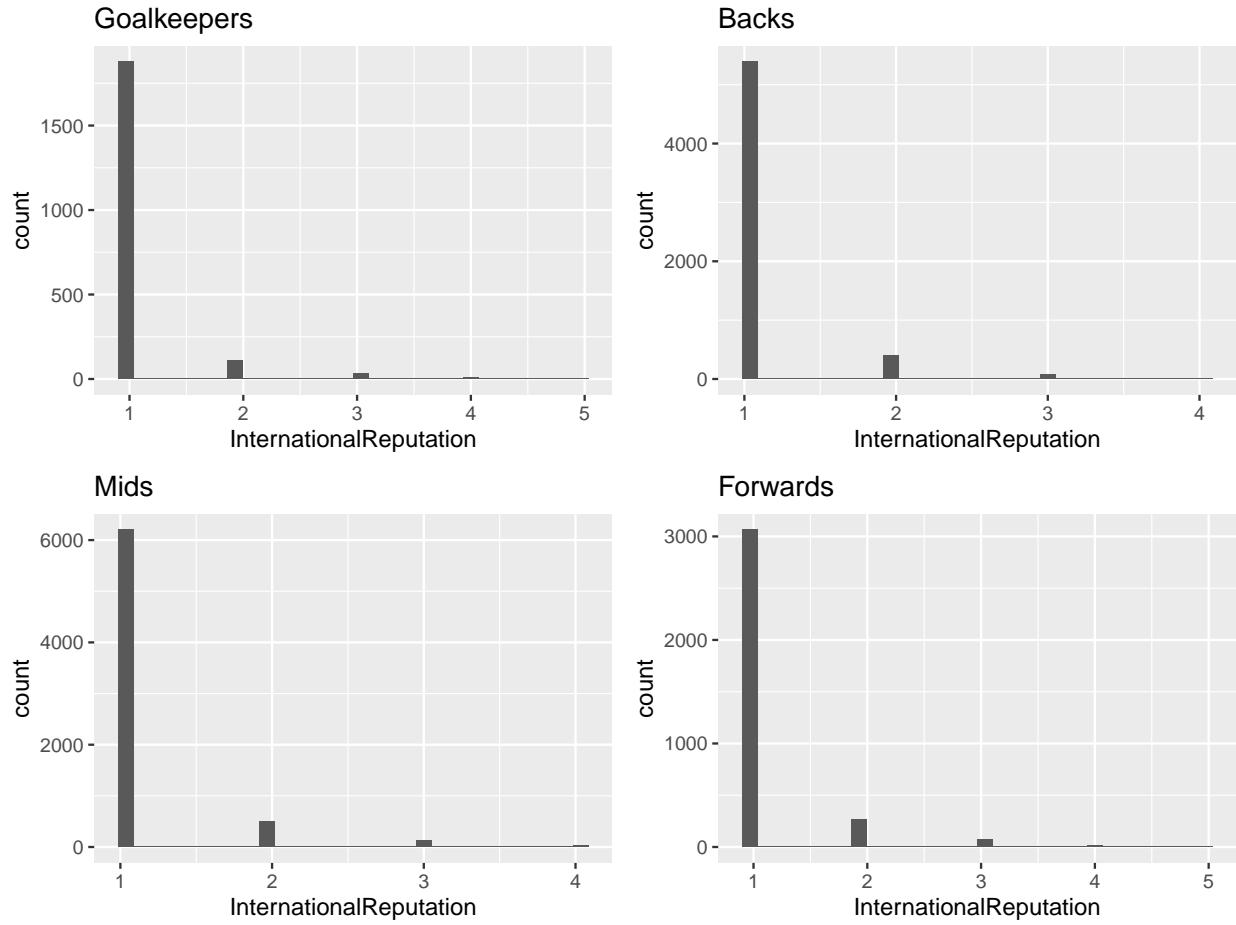
Club	medianRanking
Juventus	84
FC Bayern München	83
FC Barcelona	82
Atlético Madrid	81
Inter	81
Manchester City	81
Napoli	81
Liverpool	80
Real Madrid	80
Tottenham Hotspur	80

It is interesting to note that there are some big time names missing in the top ten teams with the highest median player rankings in Fifa 19 such as Milan, Paris Saint Germain, and Roma. This is most likely due to the fact that these teams, while high powered, have lacking benches according to Fifa ratings and are therefore deemed as lower overall by median rating while they would be higher ranked by a mean ranking.

Nationality	medianRanking
United Arab Emirates	77
Dominican Republic	72
Oman	72
São Tomé & Príncipe	72
Israel	71.5
Mozambique	71.5
Brazil	71
Cape Verde	71
Fiji	71
New Caledonia	71

The countries that have the highest median player rating are mostly in the Eastern Hemisphere, many falling between the Middle East and Africa.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



As we can see from the above, it is interesting that the international reputation figures matter so little for so many players. International reputation is a statistic that basically takes into account the teams overall popularity in relation to the player and then boosts their stats accordingly. The international reputation graphs for all position groups are heavily right skewed and it is interesting to note that there are only a few goalkeepers and forwards that have international reputations of a five.

We then segmented the data by the four major position groups: Defenders, Midfielders, Forwards, and Goalie. We imported a data file with information from each position group into JMP and created regression trees to determine the 10 to 12 most important skills for determining Overall Rating based on which variables the data split on for each position.

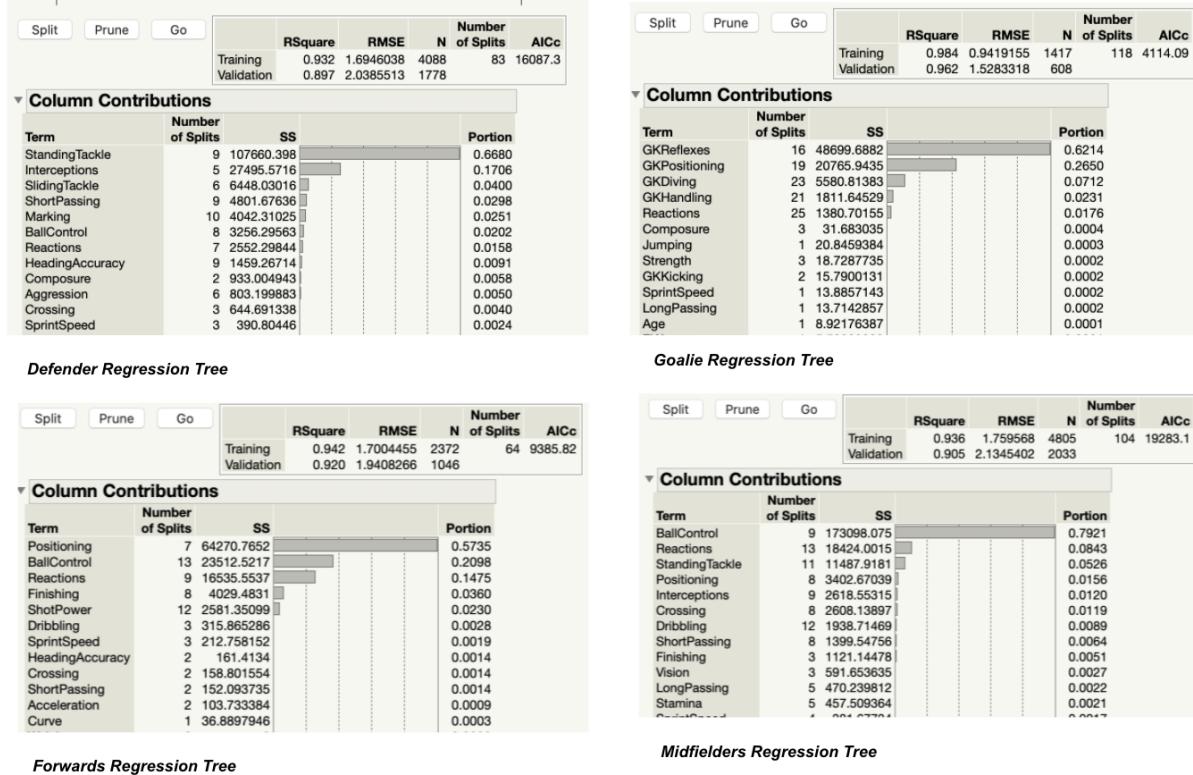


Figure 1: JMP Regression Trees

From this point on in the analysis, we will focus on the variables that split the data the most in the regression trees. We will also explore and include age, weight, and height to control for physical characteristics between players. The last variable we are choosing to include is international reputation since, using domain knowledge, we know that it goes into the calculation for overall rating so we want to control for it when considering which player attributes are important.

Forwards

The highest correlation we observed for this subset was between Long Shots and Volleys (0.72), while the lowest correlation was between Standing Tackle and Jumping (-0.0015). Out of all the variables, Volley, Strength, and Agility all seemed to have fairly high correlations with other variables. Height and Standing Tackle generally had the least correlation with the other variables. At the same time, we observe a few negative correlations involving either Agility or Standing Tackle.

Based on the histograms, each of the variables were unimodal and normal or close to it. However, we noticed that Standing Tackle is skewed to the right and Strength has a slight left skew. Except for Standing Tackle, most players are in the “good” range for all variables.

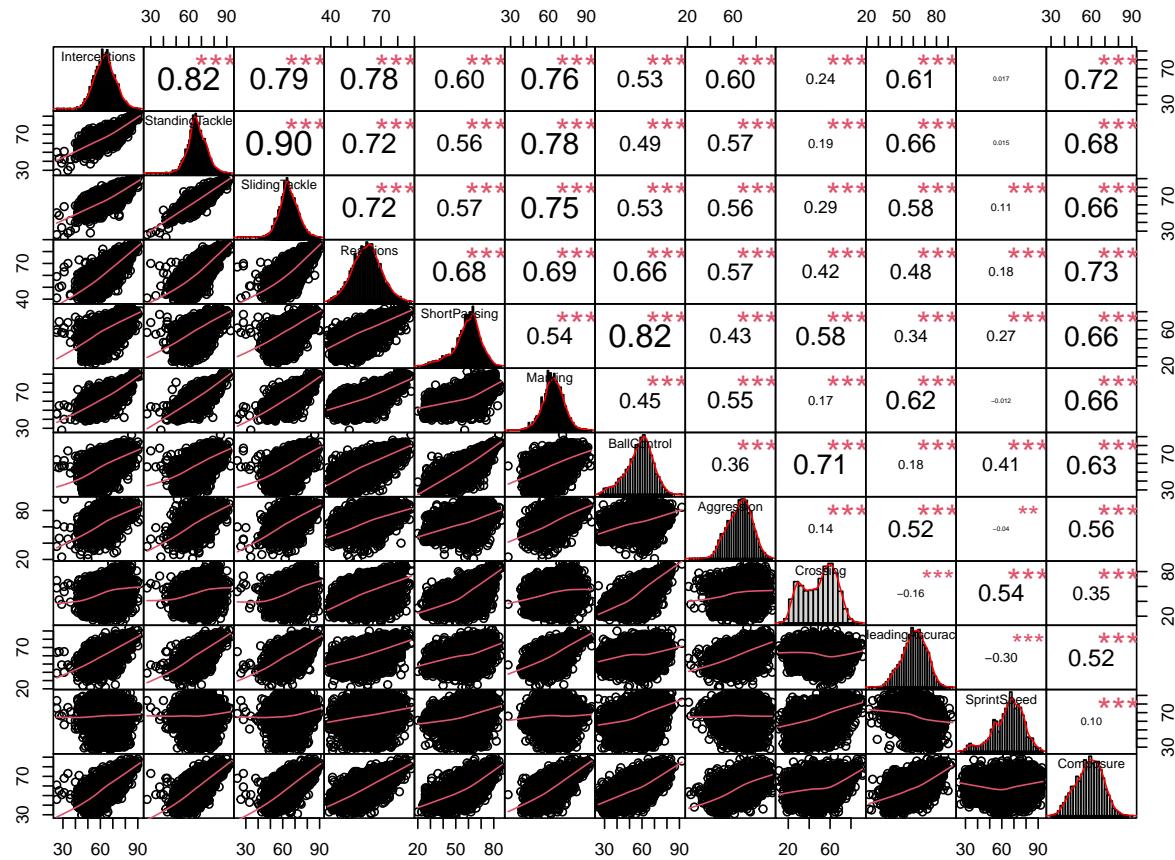
Strength has the highest spread, with a standard deviation of just over 13 and a mean of 66.26 All variables have a mean in the high 50’s to low 70’s with a standard deviation around 10, once again except for Standing Tackle.

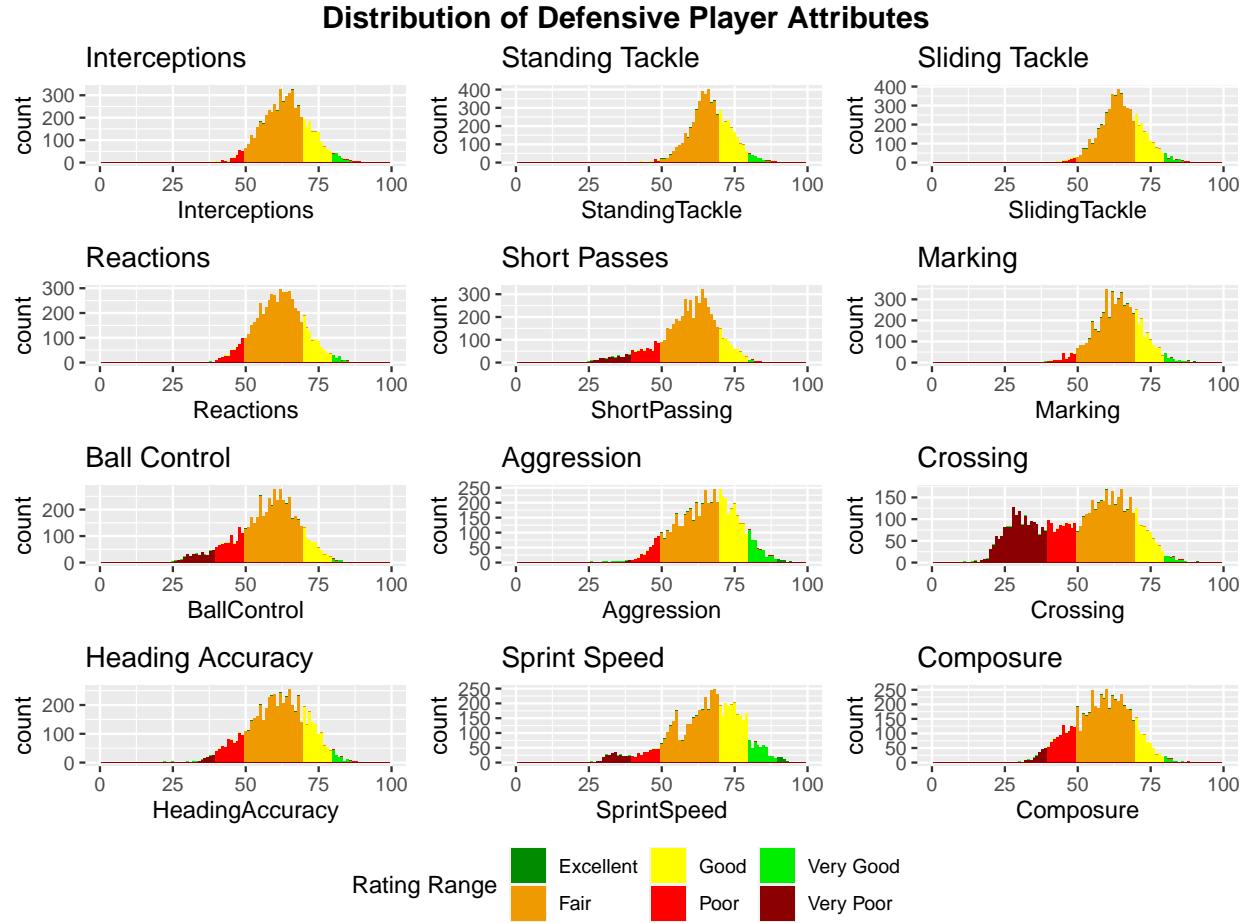
Midfielders

The highest positive correlation is between Standing Tackle and Interceptions (.88), while the lowest correlation is between Standing Tackle and Crossing (-0.033). We also noticed that this subset includes negative correlations between variables like Standing Tackle and Sprint Speed (-0.41). This makes some sense because midfielders are more versatile players so they may need a wider variety of skill sets to successfully play the position. Overall, stamina has low correlation with other variables.

Most of the player attributes are unimodal, and nearly normal with centers between 49 and 69. Interceptions and Standing Tackle resemble a more uniform distribution. Most of the players are in the fair and good range for each attribute. Standing Tackle has the largest spread (Range: 80, SD: 16.39). Interceptions and Standing Tackle have the greatest variability compared to other attributes.

Defenders





Many of the mental and physical attributes of players are positively correlated. This makes sense because we would want a group of well rounded professional defenders to not only make it into the Premier League by being good at slide tackles and crossing. The highest correlation is between Slide and standing tackles (0.90) and the weakest correlation is between Marking and Sprint Speed (-0.012).

Most of the player attributes are unimodal, and nearly normal with centers between 54 and 68. Some interesting points are that Crossing has the largest spread (Range:80, SD: 16.2) with two separate peaks around 28 and 57, most likely a group of defenders that are not very good at accurately crossing the ball. Ball Control, Sprint Speed, Composure have the greatest variability among the remaining attributes. Unlike Midfielders, Standing Tackle has the smallest spread.

Goalkeepers

The highest correlation is between Goalkeeper Diving and Goalkeeper reflexes (0.90), which makes sense because you need good reflexes in order to dive for the ball at the right time. The lowest correlation is between Goalkeepers' Long Shot rating and Age (0.26) and Long Shot has very low correlation, overall, with all other attributes.

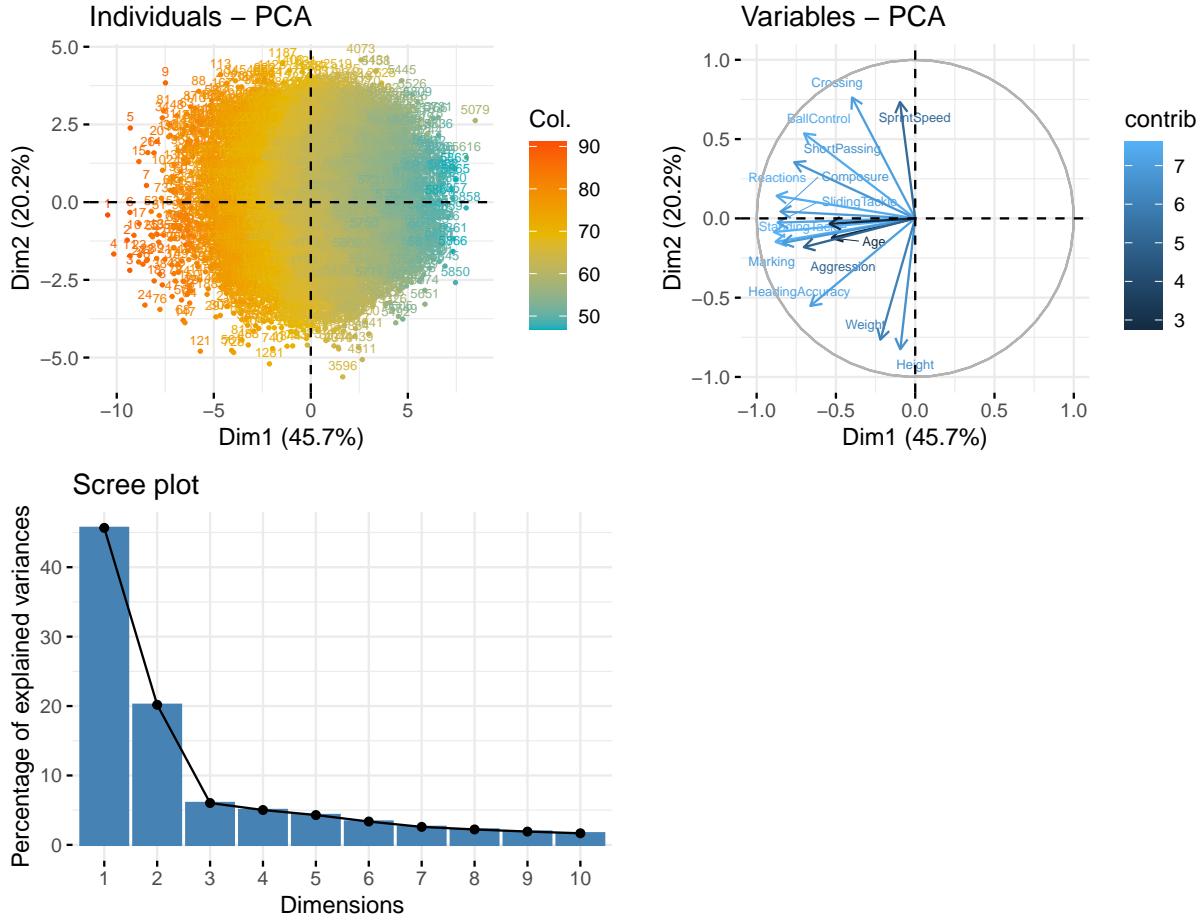
All goalkeeper attributes are fairly unimodal and centered between 58 and 65. We notice that most goalkeepers are between the ages of 20 to 25, and its plot is slightly right skewed. It is interesting that goal keeper's reaction rating has the biggest spread. The Reflexes plot has the highest centered distribution (66.1), since goalkeepers need to have high reflexes.

Results

Analysis: We learned that several variables are significant in predicting a defender's overall rating, including interceptions, standing tackle, and marking. This meant that all skills were not valued equally for defenders. It also became extremely clear early on in our exploratory data analysis that the different soccer positions valued different abilities, which makes quite a bit of sense. Different positions have different roles on the field that require unique skill sets. Thus, we could reject the portion of our null hypothesis that stated that skills required do not vary across positions. Thus, we did not have to generate new hypotheses, but could clearly reject our original null hypothesis on both points that it made.

Principal Components:

When we took a look at principal components, we found that PC1 was more or less an average of each of the players stats, while PC2 had appeared to have a positive correlation with skills associated with defender specific abilities such as marking and interceptions but had a negative correlation with more general soccer skills, such as ball control, passing, and reactions.



However, when we plotted Overall rating versus PC2, we could not discern an obvious or meaningful relationship. We decided not to go with principal components for the same reason principal components are not extremely common: they are exceedingly difficult to interpret. While PC1 made a fair deal of sense, PC2 would be nearly impossible to explain to a FIFA fan who did not have knowledge of statistics. Also, with our two linear models we got R² values around 0.95, and with such a high percentage of data explained by a more easily explainable model, there is no reason to use a more convoluted model that only explains roughly 70% of the variation in the data.

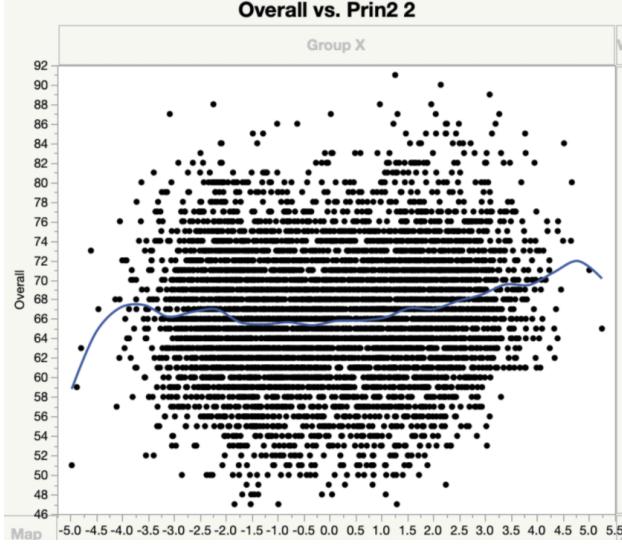


Figure 2: Overall vs PC2

Models and Summaries: We constructed four different models to capture the relationship between player attributes and overall rating for defenders in FIFA19. We first fit a minimum BIC Backwards Stepwise Regression from all player attributes of interest and then expanded this into a more complex model, which minimized BIC with all of the player attributes of interest and their interaction terms. Next, we took the same variables of interest and fit a lasso regression model with 20 parameters and finally fit a double layer neural network.

Across all these models the “variables of interest” are the top 12 player skills determined to be important by the regression trees along with physical characteristics such as age, weight, and height and international reputation since we know from domain knowledge that this is a part of the overall rating algorithm.

Multiple Linear Regression:

First, we built a model with all of these variables and ran a backwards stepwise regression using minimum AIC and BIC. Interesting parts of the output of the model that minimizes BIC are shown below.

```
##
## Call:
## lm(formula = Overall ~ Age + Weight + Interceptions + StandingTackle +
##     SlidingTackle + Reactions + ShortPassing + Marking + BallControl +
##     Aggression + Crossing + HeadingAccuracy + SprintSpeed + Composure +
##     InternationalReputation, data = BacksMainAttr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -6.2409 -0.8044  0.0064  0.8250 17.5928 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 3.982328  0.324551 12.270 < 2e-16 ***
## Age         0.037091  0.005166  7.180 7.83e-13 ***
## Weight      0.019535  0.001562 12.503 < 2e-16 ***
```

```

## Interceptions      0.103131  0.004735 21.783 < 2e-16 ***
## StandingTackle    0.159474  0.006979 22.849 < 2e-16 ***
## SlidingTackle     0.085810  0.006041 14.204 < 2e-16 ***
## Reactions          0.095873  0.004266 22.471 < 2e-16 ***
## ShortPassing       0.076969  0.003384 22.743 < 2e-16 ***
## Marking            0.128552  0.004025 31.938 < 2e-16 ***
## BallControl         0.055975  0.003620 15.461 < 2e-16 ***
## Aggression         0.050409  0.002364 21.325 < 2e-16 ***
## Crossing           -0.011435 0.001929 -5.927 3.27e-09 ***
## HeadingAccuracy    0.070212  0.002979 23.567 < 2e-16 ***
## SprintSpeed        0.058500  0.001997 29.298 < 2e-16 ***
## Composure          0.038807  0.003189 12.170 < 2e-16 ***
## InternationalReputation 0.348956  0.057831  6.034 1.70e-09 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.381 on 5850 degrees of freedom
## Multiple R-squared:  0.9544, Adjusted R-squared:  0.9543
## F-statistic:  8163 on 15 and 5850 DF,  p-value: < 2.2e-16

```

The minimum AIC and minimum BIC model were identical and include all the variables we started with, except height.

In order to decide on a model, we decided to perform cross validation on these three preliminary models (all predictors, min AIC, min BIC) to see which which one tends to have the highest R^2 and lowest RMSE on the test set.

Cross validation indicates that the minimum BIC model tends to perform best on test sets because it has on average, it fit the model with the lowest SEE and highest R^2 on 100 differnt randomly selected test sets.

JMP has some interesting analysis features in the Fit Model dialog so went into JMP and fit a mimumum BIC model with the same variables to look at effect summaries and other interesting visualizations. In our more simple model, the model without the interaction terms, we learned that each 1 point increase in standing tackle, marking, and interceptions had the greatest influence in determining a player's overall rating.

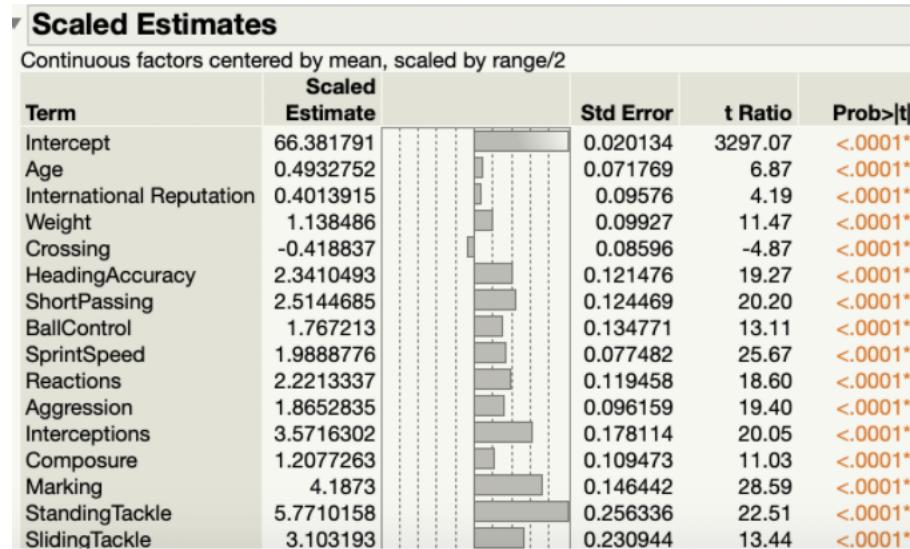
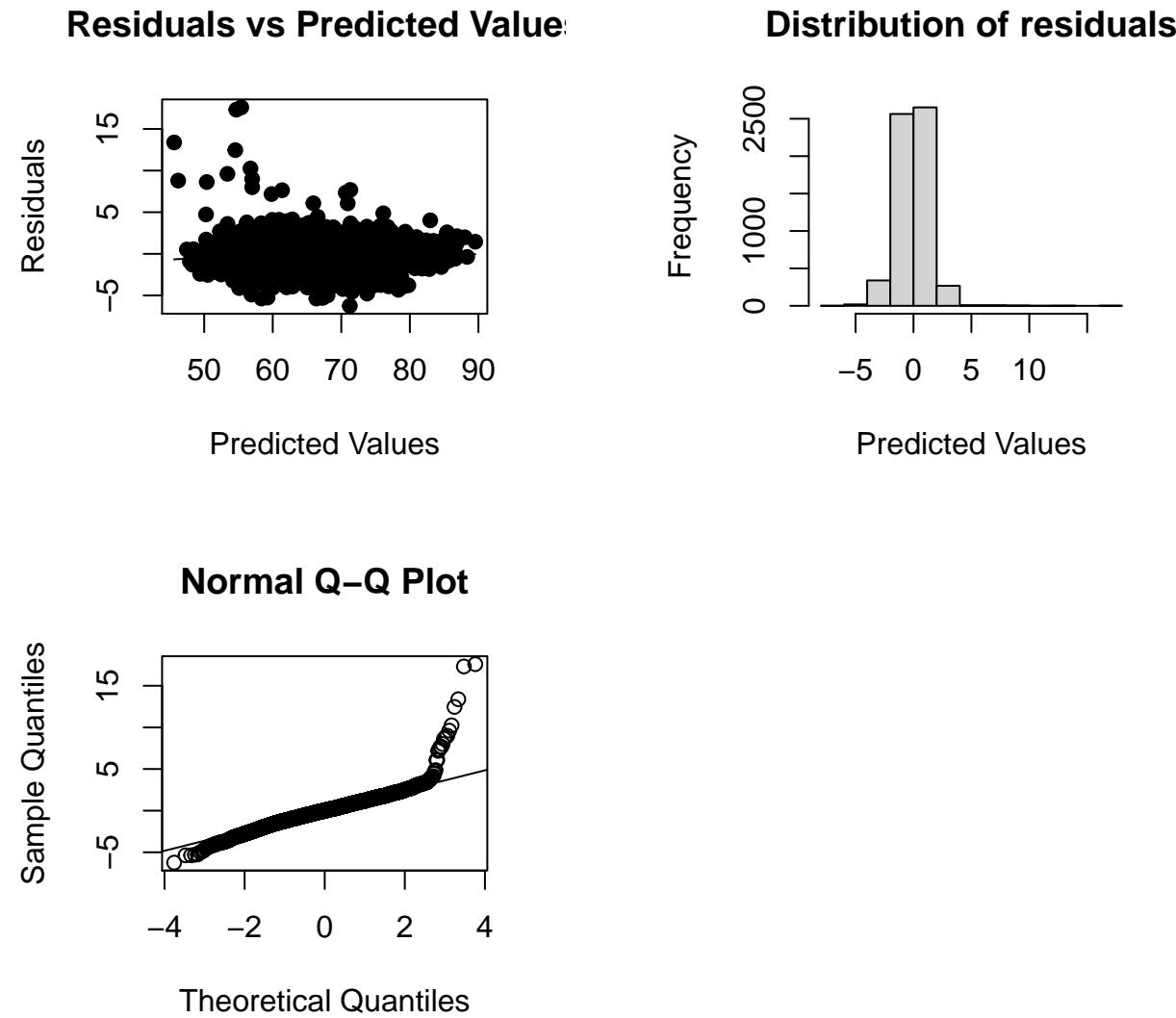


Figure 3: Scaled Estimates

The parameter estimates are highly dependent on the scale of their different corresponding factors (comparing

height in inches and rating (1 to 99) is not very helpful) so scaled estimates allow us to provide a scale invariant means to examine parameter estimates. A large number of other variables, there being 15 in total, were significant, with none, including the ones previously mentioned, having estimated slope coefficients above 0.4.

The R^2 and adjusted R^2 are both round to 0.95, which is very good, but makes sense considering the game literally used our predictors to determine our response variable. The root mean squared error is 1.38, which is fairly low considering our response variable can be any integer under 99. However, it is important to note that overall rating must be an integer, so small changes in RMSE are effectively worthless.



The assumptions and conditions for a linear model do not appear to be fully satisfied. The residuals are fairly heteroskedastic so nonconstant variance is satisfied. The horizontal smoother across the predicted vs residuals indicates decent linearity. Our main concern here is that the errors are skewed to the right which means we are making a lot of predictions are too high with this model. We also see that the qq plot seems to be somewhat linear, although imperfect towards the right end.

The next step was to make the model more robust and see whether the effect of certain player attributes depend on the level of other skills. We created a model with all of the variables of interest and their interaction terms and did a minimum BIC stepwise regression to build a more complex model.

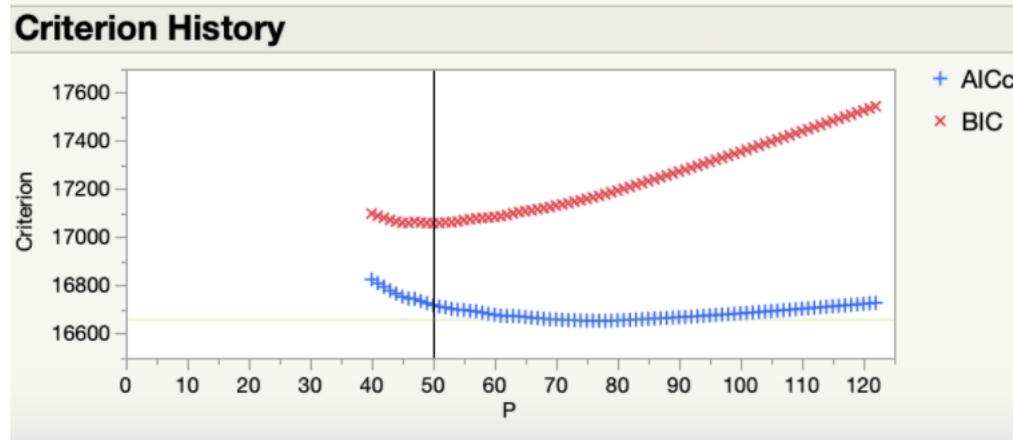


Figure 4: Criterion History

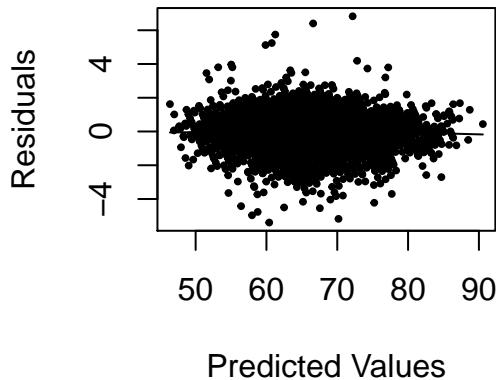
Summary of Fit	
RSquare	0.97766
RSquare Adj	0.97739
Root Mean Square Error	0.975118
Mean of Response	66.4301
Observations (or Sum Wgts)	4106

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Model	49	168777.77	3444.44	3622.467	
Error	4056	3856.67	0.95		<.0001*
C. Total	4105	172634.44			

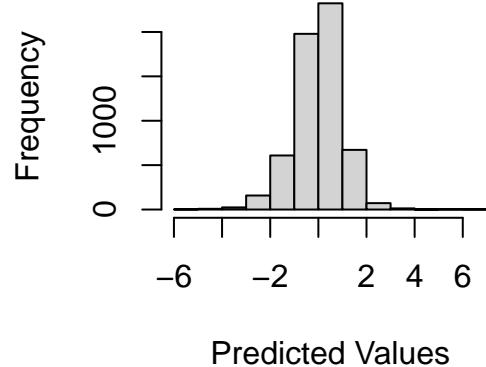
Figure 5: Big Summary

The criterion history plot clearly indicates that BIC “imposes” a more severe penalty for adding more predictors, as it hits a minimum way before the AIC selection criteria. Our more complex model had more than three times the number of predictors of our more simple model, so where our simple model had a few predictors that clearly stood out, that was not the case for our complex model. For example, an increase of one point for interceptions is associated with, on average, an increase of 0.125 in predicted overall rank. We found that a huge number of significant variables were interaction terms, which meant that how much any skill mattered in determining one’s overall rank varied greatly with the other skills that player possessed. With 49 significant predictor variables, no variable or interact term particularly dominated in the final calculation for predicted overall rating. While the RMSE did decrease from 1.38 to just over 1.0 going from our simpler to more complex model, as stated before, this change in RMSE means very little because our response variable is discrete.

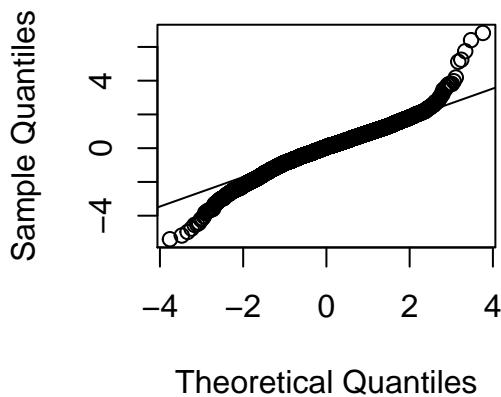
Residuals vs Predicted Value



Distribution of residuals



Normal Q-Q Plot



The assumptions for our larger model were roughly met. As one can see, our residuals vs predicted plot shows no pattern meaning that there is a linear relationship between our predictors and our response. The histogram of our residuals is roughly normally distributed (could definitely be better, but will count it as passing), so our residuals are normal. Lastly, the qq plot shows a roughly linear pattern that means that the residuals are independent and so all in all I would say the assumptions of our residuals are met. This is an improvement from the diagnostics of our smaller model.

Lasso Regression:

Our next step was to try to find a model that balanced accuracy and interpretability. Using Lasso Regression, we use biased regression estimates to reduce the variance introduced by multicollinearity. By minimizing $PRSS = \min(y - Xb)(y - Xb) + \lambda \sum |bi|$, we found that Lasso Regression reduces the model with all of the variables and interactions down to 57 parameters with a λ penalty of shrinkage parameter of 6.28.

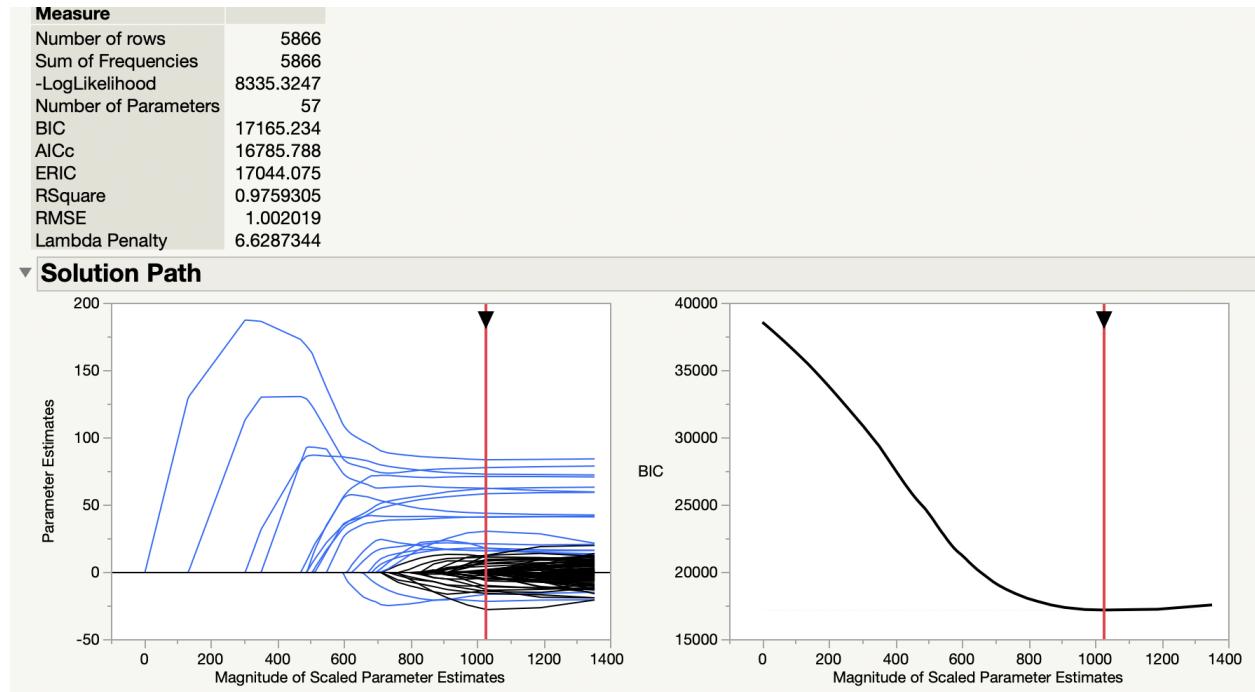


Figure 6: Original Lasso

However, we were interested in exploring how we could use lasso regression to come up with a model that does not necessarily have the highest R², but explains a larger portion of the variability in overall rating than the smaller stepwise regression model with less parameters than the complex interaction model with 49 predictor variables.

We found that a lambda penalty of 165.71 resulted in a model with only 20 parameters with an R² of 0.965 and RMSE of 1.2055. This model balances robustness and interpretability because it has a higher R² and lower RMSE than our previous smaller model, but it has less than half of the coefficients in the larger interaction model making it a more meaningful model to interpret.

The most important variables in the model, which are Standing Tackle, Marking, Heading Accuracy, and Interceptions.

We also fit a lasso with only the linear predictors and found that the lambda penalty was less than 1 ($\lambda = 0.01116$ minimized our estimate and $\lambda = 0.10072$ was the heuristic choice to produce a simpler model).

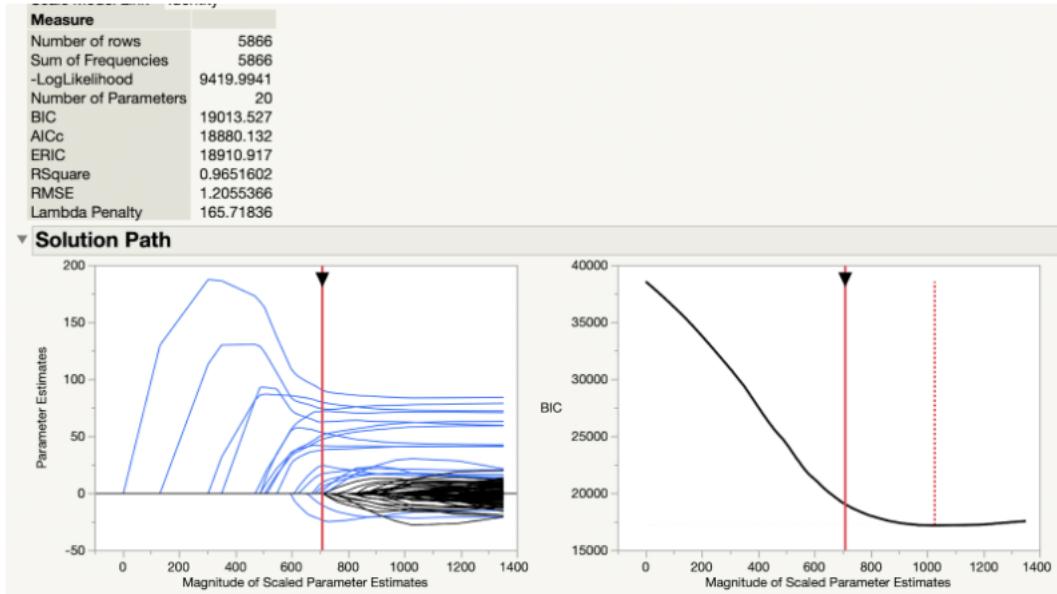


Figure 7: Adjusted Lasso

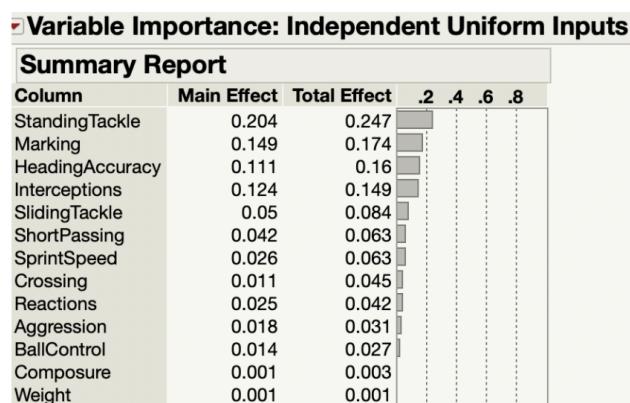
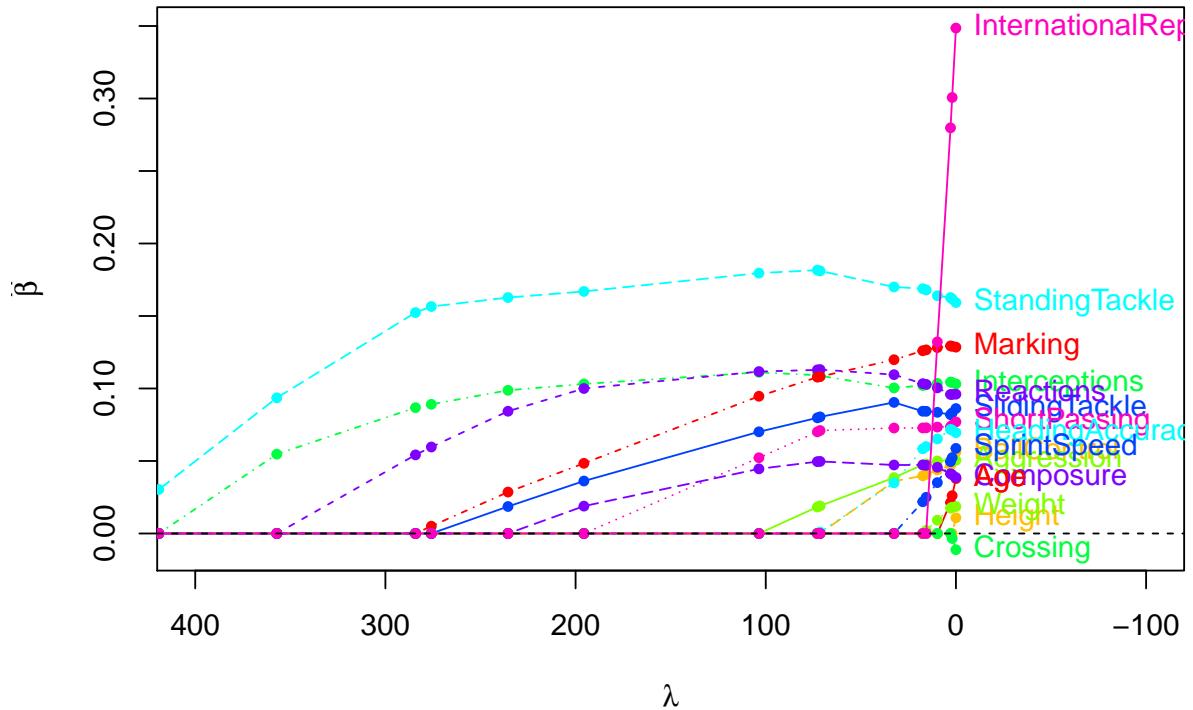
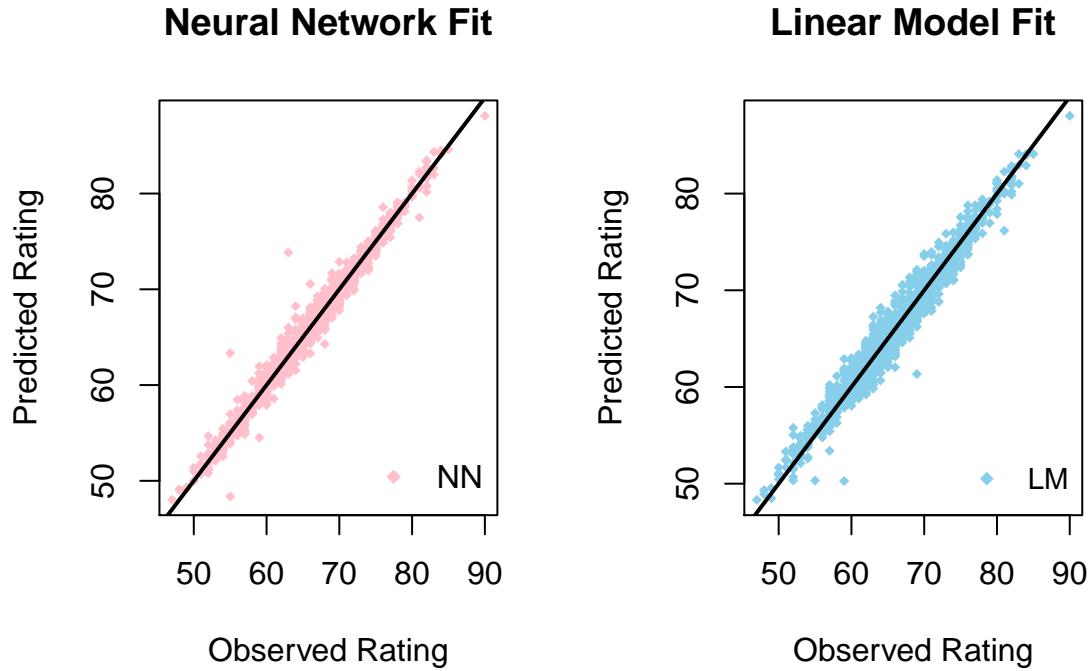


Figure 8: Lasso Variable Importance



Neural Network

The last model we attempted to fit was a neural network. In order to do so, we fit a two layer model with 4 nodes in the first layer and 8 nodes in the second layer (closest to the X's). In the creation of this model we boosted with 3 models and performed 5 tours to come to a model. By minimizing SSE and the squares of the weights, we created a model that captures almost all of the variability in overall rating, as seen in the high R² in the training and validation sets, but it is a black box and does not tell us a lot about the data.

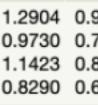
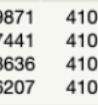


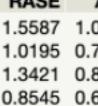
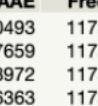
The actual vs predicted plots shown above show the overall effectiveness of ours model. The neural network has extremely low variance and is quite accurate at predicting the overall rating of a player in the fifa video game. There are very few outliers and leverage points and the model clearly has a distinct positive linear pattern.

Comparing Models: To compare our models, we were interested in looking at which variables each model treated as the most important as well as their cross validation performance. In order to conduct such an analysis, we used Model Comparison on JMP to compare all four models on randomly selected training (Test = 0), test (Test = 1), and validation (Test = 2) sets all at once.

Unsurprisingly, the neural network had the best performance on all three data sets, but as mentioned previously, this model does not provide a lot of insights on the relationship between different player attributes and overall rating. All of the models had R2 values all within 0.2 for the training set but when we fit these models on the test and validation sets we started to see differences in how well they fit the data. The larger stepwise model with interaction performed the second best and the simple stepwise model had the lowest R2 and highest RASE for both sets. It is interesting to note that we use Root Average Squared Error(RASE) instead of RMSE in these analyses because there is no known unbiased estimator for neural networks. In the test set, the smallest model only accounted for 94.2% of the variability in overall rating, which in other contexts is an incredibly high R2, but in our analyses well below the performance of other models. With that being said, all four models had R2 values within 0.0436 of one another and RASE within 0.7042 so none of the models are clearly better or worse, it simply depends on how much the user of the model places on precision versus interpretability.

Model	Simple BIC	Complex BIC	Lasso	Neural Network
Top 5 Variables in order of importance	Standing Tackle Marking Interceptions Sliding Tackle Reactions	Sliding Tackle Standing Tackle Interceptions Ball Control Reactions	Standing Tackle Interceptions Heading Accuracy Marking Sprint Speed Speed	Standing Tackle Short Passing Heading Accuracy Sprint Speed Crossing

Model Comparison Test=0						
Predictors						
Measures of Fit for Overall						
Predictor	Creator	.2.4.6.8	RSquare	RASE	AAE	Freq
SimpleBIC	Fit Least Squares		0.9604	1.2904	0.9871	4106
ComplexBIC	Fit Least Squares		0.9775	0.9730	0.7441	4106
Lasso20	Fit Generalized Adaptive Lasso		0.9690	1.1423	0.8636	4106
NN	Neural		0.9837	0.8290	0.6207	4106

Model Comparison Test=1						
Predictors						
Measures of Fit for Overall						
Predictor	Creator	.2.4.6.8	RSquare	RASE	AAE	Freq
SimpleBIC	Fit Least Squares		0.9420	1.5587	1.0493	1173
ComplexBIC	Fit Least Squares		0.9752	1.0195	0.7659	1173
Lasso20	Fit Generalized Adaptive Lasso		0.9570	1.3421	0.8972	1173
NN	Neural		0.9826	0.8545	0.6363	1173

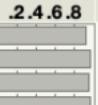
Model Comparison Test=2						
Predictors						
Measures of Fit for Overall						
Predictor	Creator	.2.4.6.8	RSquare	RASE	AAE	Freq
SimpleBIC	Fit Least Squares		0.9358	1.5827	1.0903	587
ComplexBIC	Fit Least Squares		0.9694	1.0939	0.7814	587
Lasso20	Fit Generalized Adaptive Lasso		0.9540	1.3402	0.9396	587
NN	Neural		0.9794	0.8978	0.6444	587

Figure 9: Model Comparison

Standing Tackle was the only player attribute that was calculated to be the most important variable in four out of the five models and was the only skill that appeared in top 5 most variables for every single model. Other variables that were within the 3 most important variables for determining overall rating across three of the four models were sliding tackle, and interceptions. Overall, the five most important variables across all four models were standing tackle, interceptions, heading accuracy, sliding tackle, and marking which are all key skills to effectively defend in soccer. Interestingly, age and height were of the least importance across all of the models.

Inference

	Ssewankambo	Median	Saldaña
Marking	62	64	65
Ball Control	59	59	58
Aggression	66	66	61
Crossing	61	54	66
Heading Accuracy	58	62	48
Composure	62	59	53
Sprint Speed	89	67	83
Interceptions	61	64	64
Standing Tackle	63	66	60
Sliding Tackle	60	65	63
Reactions	58	62	71
Short Passing	62	60	58
Prediction (small)	65.99	66.607	65.273

	Ssewankambo	Median	Saldaña
Prediction (large)	66.038	65.946	66.622

Above is an example of our actual model in use. We picked two backs that had the same rating (66) yet different characteristics and tried to figure out if we could predict the overall of the player. The left side is all the ratings of different attributes for Ssewankambo, the right is all the ratings of different attributes for Saldaña, and the middle is all the median ratings of the same different attributes. One can see clearly that the two stack up differently in terms of which attributes that have ranked higher compared to the median and which they have ranked lower compared to the median. The bottom two lines are the point forecast predictions for both players and the median ratings based on both our small and large model. One can clearly see that Ssewankambo's point forecast for the small model and the large model were right on target (65.99, 66.038). Saldaña's point forecasts for the small model missed the mark by an eighth of a point (65.273) and the large model overestimated by a smidge (66.622). The median value point forecasts were slightly more accurate than Saldaña, but not more accurate than Ssewankambo as they had point forecasts of 66.607 for the small model and 65.946 for the large model. Interesting to note that the small model gave a lower point forecast than the large model for both players, while it actually gave a higher point forecast based on the median values. Lastly, the 95% confidence intervals and prediction intervals for both models contained the correct rating (66) for both players showing the fact that overall our models are quite accurate.

Conclusions

We set out to find what skills were the most important in how FIFA ranked its players in FIFA 2019. Specifically, we focused on modeling this relationship for defenders. We did this by looking at data for 5,866 backs in the FIFA 2019 data set and fitting the overall rating against all other predictors and narrowing our model down with the goal of finding a reasonable model that was both accurate and not overly complicated. We ran a regression tree to narrow our data down from the original 89 variables in the data set to a much smaller number, then ran a stepwise regression to minimize BIC, giving us our simple model. For our more complex model, we took the variables from the regression tree and included the interaction terms between all of those variables and then ran a stepwise regression to minimize BIC, leaving us with many more predictors than the simpler model. These two models gave us slightly different pictures of the algorithm FIFA used to rank its players, with the smaller model having a few variables that had a somewhat greater impact on overall rank but with the larger model having no variables or interaction terms that were particularly larger than others. We also fit lasso regression and neural network models that had similar performance in terms of fitting test data and predicting overall rating, but these were a little harder to interpret if one wants to understand the influence of each player attribute. When it comes to defenders in soccer, certain skills absolutely matter more than others, and the importance of each of them varies with the other variables involved especially standing tackle, interceptions, heading accuracy, sliding tackle, and marking.

These models only truly works for defenders in this game, as the other positions largely depend on entirely different variables. By the time a new game comes out they may change what skills they value or how much they value. It might be better to make a model with data from several games so that we might predict how good a player would be in a new game. It could also be interesting to try to make a model that could predict the overall rank for a player of any position, so long as we could keep track of all of the dummy variables necessary for such a model.