Final Project

Kyle Morris

May 29, 2019

## Purpose

We are investigating here and running statistical analysis on soccer data. As a longtime sports fan, statistics are now the name of the game and I was interested in seeing how electronic players held up to their real world counterparts, and whether there is easily detected bias to the ratings systems which ultimately are objective.

We’ll be using a number of packages in our analysis, including:

* ggplot2, a graphing package.
* pastecs, statistical analysis.
* dplyr, for data cleaning.

## Our Data

Our data comes to us from Kaggle. More specifically, we have three data sets we are looking at:

* Kaggle – FIFA 19 Ultimate Team  <https://www.kaggle.com/stefanoleone992/fifa-19-fifa-ultimate-team>
* Kaggle – FIFA 19 Player Database  <https://www.kaggle.com/karangadiya/fifa19/>
* Kaggle – World Cup Players  <https://www.kaggle.com/djamshed/fifa-world-cup-2018-players>

All of these are related to soccer (or Football, as those across the pond would insist on it being called.) The first two datasets are from FIFA 19 by Electronic Arts, the juggernaut that has seen over 20 million units sold to date alone. While it does face competition from Pro Evolution Soccer, FIFA is still the undisputed champion of the simulated soccer world.

Our last dataset concerns the 2018 World Cup. We were interested in how simulated soccer royalty compared to real world soccer royalty.

To begin with, let’s import our data.

fifaUltimate <- read.csv("fifa19ultimate.csv", header = TRUE)  
fifaGame <- read.csv("fifagamedata.csv", header = TRUE)  
worldCup <- read.csv("wc2018-players.csv", header = TRUE)  
worldCup <- worldCup[-c(5)]

## Cleaning the data.

Some preliminary analysis of our data and the form it comes to us.

fifaUltimateNA <- fifaUltimate[complete.cases(fifaUltimate), ]  
fifaGameNA <- fifaGame[complete.cases(fifaGame), ]  
worldCupNA <- worldCup[complete.cases(worldCup), ]  
  
badFifaUltimate <- nrow(fifaUltimate) - nrow(fifaUltimateNA)  
badFifaGame <- nrow(fifaGame) - nrow(fifaGameNA)  
badWorldCup <- nrow(worldCup) - nrow(worldCupNA)  
  
badFifaUltimate \* 100 / nrow(fifaUltimate)

## [1] 100

badFifaGame \* 100 / nrow(fifaGame)

## [1] 0.3295436

badWorldCup \* 100 / nrow(worldCup)

## [1] 0

For the ultimate data:

* There is not a single complete entry in the entire 18831 rows of data.
* That being said, there are 95 variables and the vast majority have almost every column.
* Missing fields are NA.

For the FIFA data:

* Only 0.33% of the rows are missing any data. There are 18207 rows!
* There are 89 variables tracked.
* Missing results are marked NA. There are only 60 total incomplete rows.

For the World Cup data:

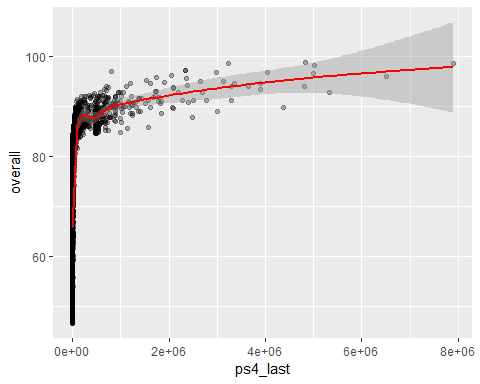
* Confirmed list of player names and birth date by country.
* Dataset has been curated based on FIFA records. This is the only complete data set we have – there is no missing data.
* Data is from official rosters provided by FIFA.
* No information on how it was originally collected. Small enough data set it could be done by hand.
* There are 10 variables tracked and 736 observations.

## Preliminary Analysis

We will now take a look at scatterplots of our data.

ggplot(subset(fifaUltimate, ps4\_last > 0), aes(x = ps4\_last, y = overall)) +  
 geom\_point(position = "jitter", alpha = 0.3) +  
 geom\_smooth(col = "red")

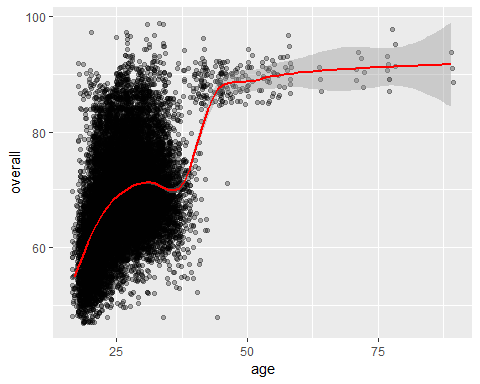
## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'



This graph charts the last sale price of the card on the PS4 versus the overall rating.

ggplot(fifaUltimate, aes(x = age, y = overall)) +  
 geom\_point(position = "jitter", alpha = 0.3) +  
 geom\_smooth(col = "red")

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'



This graph shows the Ultimate rating versus age. Some of the players are “legacy” in that they are famous players from the past and it shows the overall age.

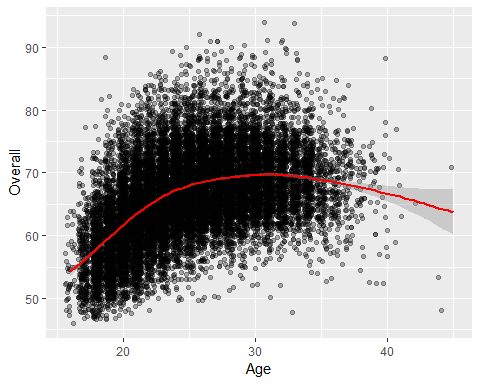
round(stat.desc(fifaUltimate$overall, basic = FALSE), digits = 3)

## median mean SE.mean CI.mean.0.95 var   
## 67.000 68.201 0.059 0.116 66.311   
## std.dev coef.var   
## 8.143 0.119

A look at our distribution of overall ratings.

ggplot(fifaGame, aes(x = Age, y = Overall)) +  
 geom\_point(position = "jitter", alpha = 0.3) +  
 geom\_smooth(col = "red")

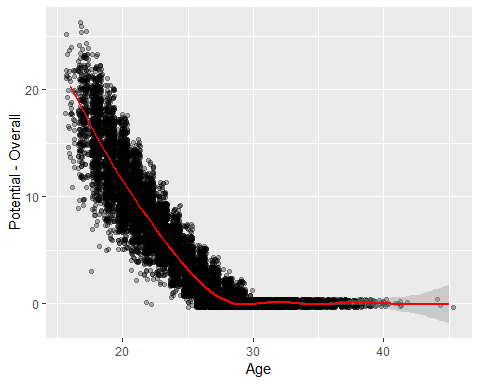
## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'



Same age versus overall plot, but this time for the FIFA stats.

ggplot(fifaGame, aes(x = Age, y = Potential - Overall)) +  
 geom\_point(position = "jitter", alpha = 0.3) +  
 geom\_smooth(col = "red")

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

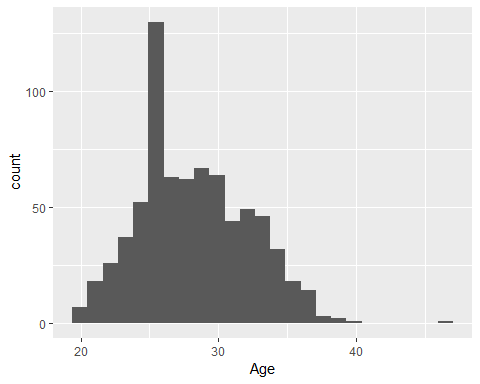


This graph is a measure of age versus room for growth i.e. $ potential - overall$.

round(stat.desc(fifaGame$Overall, basic = FALSE), digits = 3)

## median mean SE.mean CI.mean.0.95 var   
## 66.000 66.239 0.051 0.100 47.733   
## std.dev coef.var   
## 6.909 0.104

bw <- 2 \* IQR(worldCup$Age) / length(worldCup$Age)^(1/3)  
ggplot(worldCup, aes(x = Age)) +  
 geom\_histogram(binwidth = bw)



A histogram of the ages of players in the most recent World Cup.

round(stat.desc(worldCup$Age, basic = FALSE, norm = TRUE), digits = 3)

## median mean SE.mean CI.mean.0.95 var   
## 28.000 28.236 0.146 0.286 15.582   
## std.dev coef.var skewness skew.2SE kurtosis   
## 3.947 0.140 0.262 1.453 -0.106   
## kurt.2SE normtest.W normtest.p   
## -0.295 0.984 0.000

cleanFifaGame <- subset(fifaGame, select = c("Name", "Age", "Overall", "Potential", "Club","Position", "Height", "Weight", "Crossing", "Finishing", "HeadingAccuracy" , "ShortPassing" , "Volleys" , "Dribbling" , "Curve", "FKAccuracy", "LongPassing" , "BallControl" , "Acceleration" , "SprintSpeed" , "Agility" , "Reactions", "Balance" , "ShotPower" , "Jumping" , "Stamina" , "Strength" , "LongShots", "Aggression" , "Interceptions" , "Positioning", "Vision" , "Penalties" , "Composure", "Marking" , "StandingTackle" , "SlidingTackle" , "GKDiving" , "GKHandling", "GKKicking", "GKPositioning", "GKReflexes"))  
cleanFifaGameNA <- cleanFifaGame[complete.cases(cleanFifaGame), ]  
  
head(cleanFifaGameNA)

## Name Age Overall Potential Club Position  
## 1 L. Messi 31 94 94 FC Barcelona RF  
## 2 Cristiano Ronaldo 33 94 94 Juventus ST  
## 3 Neymar Jr 26 92 93 Paris Saint-Germain LW  
## 4 De Gea 27 91 93 Manchester United GK  
## 5 K. De Bruyne 27 91 92 Manchester City RCM  
## 6 E. Hazard 27 91 91 Chelsea LF  
## Height Weight Crossing Finishing HeadingAccuracy ShortPassing Volleys  
## 1 5'7 159lbs 84 95 70 90 86  
## 2 6'2 183lbs 84 94 89 81 87  
## 3 5'9 150lbs 79 87 62 84 84  
## 4 6'4 168lbs 17 13 21 50 13  
## 5 5'11 154lbs 93 82 55 92 82  
## 6 5'8 163lbs 81 84 61 89 80  
## Dribbling Curve FKAccuracy LongPassing BallControl Acceleration  
## 1 97 93 94 87 96 91  
## 2 88 81 76 77 94 89  
## 3 96 88 87 78 95 94  
## 4 18 21 19 51 42 57  
## 5 86 85 83 91 91 78  
## 6 95 83 79 83 94 94  
## SprintSpeed Agility Reactions Balance ShotPower Jumping Stamina Strength  
## 1 86 91 95 95 85 68 72 59  
## 2 91 87 96 70 95 95 88 79  
## 3 90 96 94 84 80 61 81 49  
## 4 58 60 90 43 31 67 43 64  
## 5 76 79 91 77 91 63 90 75  
## 6 88 95 90 94 82 56 83 66  
## LongShots Aggression Interceptions Positioning Vision Penalties  
## 1 94 48 22 94 94 75  
## 2 93 63 29 95 82 85  
## 3 82 56 36 89 87 81  
## 4 12 38 30 12 68 40  
## 5 91 76 61 87 94 79  
## 6 80 54 41 87 89 86  
## Composure Marking StandingTackle SlidingTackle GKDiving GKHandling  
## 1 96 33 28 26 6 11  
## 2 95 28 31 23 7 11  
## 3 94 27 24 33 9 9  
## 4 68 15 21 13 90 85  
## 5 88 68 58 51 15 13  
## 6 91 34 27 22 11 12  
## GKKicking GKPositioning GKReflexes  
## 1 15 14 8  
## 2 15 14 11  
## 3 15 15 11  
## 4 87 88 94  
## 5 5 10 13  
## 6 6 8 8

We’ve now cleaned up the FIFA stats as those will be the most useful. We kept just the most useful fields for our analysis. Now, we can begin!

## What are some of the questions we might be looking to answer?

Electronic Arts has done their best to score over 19000 football players around the world. On their platform, all skills should in theory be ranked accordingly. The Overall score is an easy catch-all explanation of player skill, but how is that calculated? Is it the average of all scores? Is there a league that has, in general, a stronger overall player base than any other? If I was utilizing their Career mode, I would want to know the beginning set up that gave my future superstar the highest overall potentional!

In order to address that,, we should look to address the following: which league has the highest overall skill rating, how exactly does FIFA weight the various skill measurements to determine overall rating, and how we can maximize the potential of our new recruits.

## Our Model

Let’s create our model. I’m going to split the data at this point as well, because I suspect that the model for goalkeepers is going to look significantly different than the model for all players, or just other position players.

cleanGK <- subset(cleanFifaGameNA, cleanFifaGameNA$Position == "GK")  
cleanOthers <- anti\_join(cleanFifaGameNA, cleanGK)

## Joining, by = c("Name", "Age", "Overall", "Potential", "Club", "Position", "Height", "Weight", "Crossing", "Finishing", "HeadingAccuracy", "ShortPassing", "Volleys", "Dribbling", "Curve", "FKAccuracy", "LongPassing", "BallControl", "Acceleration", "SprintSpeed", "Agility", "Reactions", "Balance", "ShotPower", "Jumping", "Stamina", "Strength", "LongShots", "Aggression", "Interceptions", "Positioning", "Vision", "Penalties", "Composure", "Marking", "StandingTackle", "SlidingTackle", "GKDiving", "GKHandling", "GKKicking", "GKPositioning", "GKReflexes")

gkTest <- cleanGK %>% sample\_frac(.2)  
gkTrain <- anti\_join(cleanGK, gkTest)

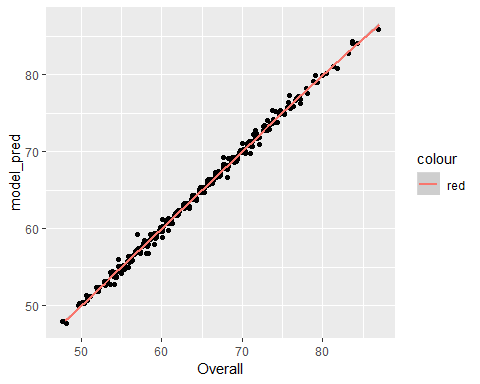
## Joining, by = c("Name", "Age", "Overall", "Potential", "Club", "Position", "Height", "Weight", "Crossing", "Finishing", "HeadingAccuracy", "ShortPassing", "Volleys", "Dribbling", "Curve", "FKAccuracy", "LongPassing", "BallControl", "Acceleration", "SprintSpeed", "Agility", "Reactions", "Balance", "ShotPower", "Jumping", "Stamina", "Strength", "LongShots", "Aggression", "Interceptions", "Positioning", "Vision", "Penalties", "Composure", "Marking", "StandingTackle", "SlidingTackle", "GKDiving", "GKHandling", "GKKicking", "GKPositioning", "GKReflexes")

offenseTest <- cleanOthers %>% sample\_frac(.2)  
offenseTrain <- anti\_join(cleanOthers, offenseTest)

## Joining, by = c("Name", "Age", "Overall", "Potential", "Club", "Position", "Height", "Weight", "Crossing", "Finishing", "HeadingAccuracy", "ShortPassing", "Volleys", "Dribbling", "Curve", "FKAccuracy", "LongPassing", "BallControl", "Acceleration", "SprintSpeed", "Agility", "Reactions", "Balance", "ShotPower", "Jumping", "Stamina", "Strength", "LongShots", "Aggression", "Interceptions", "Positioning", "Vision", "Penalties", "Composure", "Marking", "StandingTackle", "SlidingTackle", "GKDiving", "GKHandling", "GKKicking", "GKPositioning", "GKReflexes")

overallModel <- glm(Overall ~ Age + Crossing + Finishing + HeadingAccuracy + ShortPassing + Volleys + Dribbling + Curve + FKAccuracy + LongPassing + BallControl + Acceleration + SprintSpeed + Agility + Reactions + Balance + ShotPower + Jumping + Stamina + Strength + LongShots + Aggression + Interceptions + Positioning + Vision + Penalties + Composure + Marking + StandingTackle + SlidingTackle + GKDiving + GKHandling + GKKicking + GKPositioning + GKReflexes, data = cleanFifaGameNA)  
  
gkModel <-glm(Overall ~ Age + Crossing + Finishing + HeadingAccuracy + ShortPassing + Volleys + Dribbling + Curve + FKAccuracy + LongPassing + BallControl + Acceleration + SprintSpeed + Agility + Reactions + Balance + ShotPower + Jumping + Stamina + Strength + LongShots + Aggression + Interceptions + Positioning + Vision + Penalties + Composure + Marking + StandingTackle + SlidingTackle + GKDiving + GKHandling + GKKicking + GKPositioning + GKReflexes, data = gkTrain)  
  
offenseModel <- glm(Overall ~ Age + Crossing + Finishing + HeadingAccuracy + ShortPassing + Volleys + Dribbling + Curve + FKAccuracy + LongPassing + BallControl + Acceleration + SprintSpeed + Agility + Reactions + Balance + ShotPower + Jumping + Stamina + Strength + LongShots + Aggression + Interceptions + Positioning + Vision + Penalties + Composure + Marking + StandingTackle + SlidingTackle + GKDiving + GKHandling + GKKicking + GKPositioning + GKReflexes, data = offenseTrain)  
  
gkTest$model\_pred <- round(predict(gkModel, gkTest, type = "response"))  
  
offenseTest$model\_pred <- round(predict(offenseModel, offenseTest, type = "response"))  
  
ggplot(gkTest, aes(x = Overall, y = model\_pred)) +  
 geom\_jitter() +  
 geom\_smooth(aes(color = "red"))

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'

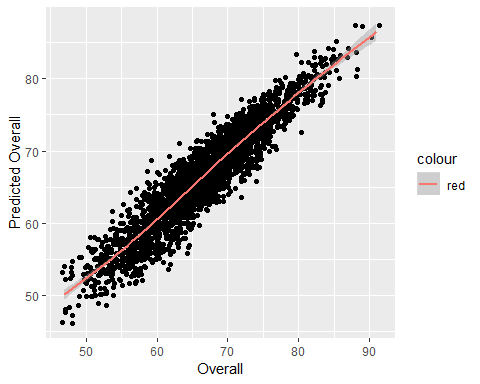


Wow, our model for goalkeepers looks really, really nice. There are a few outliers but for the most part the model looks very good at predicting the overall grade. It looks like there might not be as much nudge factor as expected.

Here’s the same graph but for the test set of regular players:

ggplot(offenseTest, aes(x = Overall, y = model\_pred)) +  
 geom\_jitter() +  
 geom\_smooth(aes(color = "red"))+  
 ylab("Predicted Overall")

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'



Overall a pleasing graph. I added some jitter due to the number of datapoints but generally there is an overall trend.

What variables are most significant? We have all of our data on the same scale so this analysis is slightly easier.

summary(gkModel)

##   
## Call:  
## glm(formula = Overall ~ Age + Crossing + Finishing + HeadingAccuracy +   
## ShortPassing + Volleys + Dribbling + Curve + FKAccuracy +   
## LongPassing + BallControl + Acceleration + SprintSpeed +   
## Agility + Reactions + Balance + ShotPower + Jumping + Stamina +   
## Strength + LongShots + Aggression + Interceptions + Positioning +   
## Vision + Penalties + Composure + Marking + StandingTackle +   
## SlidingTackle + GKDiving + GKHandling + GKKicking + GKPositioning +   
## GKReflexes, data = gkTrain)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.54487 -0.26869 -0.01792 0.26840 1.50250   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.346e-01 9.813e-02 8.505 < 2e-16 \*\*\*  
## Age -2.701e-03 2.062e-03 -1.310 0.1904   
## Crossing -3.229e-03 2.489e-03 -1.297 0.1948   
## Finishing -3.947e-03 3.140e-03 -1.257 0.2090   
## HeadingAccuracy 1.656e-04 2.327e-03 0.071 0.9433   
## ShortPassing 3.927e-04 1.461e-03 0.269 0.7881   
## Volleys 1.463e-03 2.852e-03 0.513 0.6081   
## Dribbling 3.040e-04 2.382e-03 0.128 0.8985   
## Curve -8.983e-04 2.282e-03 -0.394 0.6939   
## FKAccuracy -2.725e-03 2.058e-03 -1.324 0.1856   
## LongPassing -3.849e-04 1.409e-03 -0.273 0.7847   
## BallControl -9.169e-04 1.714e-03 -0.535 0.5927   
## Acceleration 1.278e-03 1.455e-03 0.878 0.3798   
## SprintSpeed 5.763e-04 1.425e-03 0.404 0.6859   
## Agility -5.519e-04 9.248e-04 -0.597 0.5508   
## Reactions 1.101e-01 1.466e-03 75.076 < 2e-16 \*\*\*  
## Balance 3.704e-04 9.411e-04 0.394 0.6940   
## ShotPower 5.322e-03 1.322e-03 4.026 5.93e-05 \*\*\*  
## Jumping -1.506e-03 9.457e-04 -1.592 0.1115   
## Stamina -1.117e-03 1.277e-03 -0.874 0.3821   
## Strength -8.127e-04 8.204e-04 -0.991 0.3220   
## LongShots -2.509e-05 2.872e-03 -0.009 0.9930   
## Aggression 8.417e-04 1.197e-03 0.703 0.4821   
## Interceptions 3.054e-03 1.945e-03 1.570 0.1166   
## Positioning -5.149e-03 2.948e-03 -1.746 0.0809 .   
## Vision 8.195e-04 7.554e-04 1.085 0.2781   
## Penalties 3.000e-04 1.413e-03 0.212 0.8319   
## Composure 1.603e-03 8.548e-04 1.875 0.0609 .   
## Marking -4.823e-04 1.498e-03 -0.322 0.7475   
## StandingTackle 8.399e-04 3.070e-03 0.274 0.7844   
## SlidingTackle 4.154e-04 2.987e-03 0.139 0.8894   
## GKDiving 2.117e-01 2.733e-03 77.475 < 2e-16 \*\*\*  
## GKHandling 2.151e-01 2.345e-03 91.734 < 2e-16 \*\*\*  
## GKKicking 4.960e-02 1.662e-03 29.848 < 2e-16 \*\*\*  
## GKPositioning 2.075e-01 2.266e-03 91.567 < 2e-16 \*\*\*  
## GKReflexes 2.097e-01 2.616e-03 80.137 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.1051547)  
##   
## Null deviance: 93236.26 on 1619 degrees of freedom  
## Residual deviance: 166.57 on 1584 degrees of freedom  
## AIC: 986.19  
##   
## Number of Fisher Scoring iterations: 2

summary(offenseModel)

##   
## Call:  
## glm(formula = Overall ~ Age + Crossing + Finishing + HeadingAccuracy +   
## ShortPassing + Volleys + Dribbling + Curve + FKAccuracy +   
## LongPassing + BallControl + Acceleration + SprintSpeed +   
## Agility + Reactions + Balance + ShotPower + Jumping + Stamina +   
## Strength + LongShots + Aggression + Interceptions + Positioning +   
## Vision + Penalties + Composure + Marking + StandingTackle +   
## SlidingTackle + GKDiving + GKHandling + GKKicking + GKPositioning +   
## GKReflexes, data = offenseTrain)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -11.0625 -1.6495 -0.0432 1.6076 10.8006   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.5567530 0.3458387 30.525 < 2e-16 \*\*\*  
## Age 0.0583794 0.0065084 8.970 < 2e-16 \*\*\*  
## Crossing 0.0067719 0.0029048 2.331 0.019752 \*   
## Finishing 0.0303307 0.0035554 8.531 < 2e-16 \*\*\*  
## HeadingAccuracy 0.0875119 0.0031353 27.912 < 2e-16 \*\*\*  
## ShortPassing 0.1073836 0.0056594 18.974 < 2e-16 \*\*\*  
## Volleys -0.0046531 0.0030880 -1.507 0.131875   
## Dribbling 0.0079945 0.0046647 1.714 0.086588 .   
## Curve 0.0058707 0.0030584 1.920 0.054935 .   
## FKAccuracy 0.0086442 0.0027070 3.193 0.001410 \*\*   
## LongPassing -0.0149297 0.0039644 -3.766 0.000167 \*\*\*  
## BallControl 0.1625651 0.0059796 27.187 < 2e-16 \*\*\*  
## Acceleration 0.0477304 0.0045726 10.438 < 2e-16 \*\*\*  
## SprintSpeed 0.0340296 0.0041919 8.118 5.17e-16 \*\*\*  
## Agility -0.0008424 0.0034821 -0.242 0.808846   
## Reactions 0.2685227 0.0045647 58.826 < 2e-16 \*\*\*  
## Balance -0.0163480 0.0030838 -5.301 1.17e-07 \*\*\*  
## ShotPower 0.0209864 0.0033419 6.280 3.50e-10 \*\*\*  
## Jumping 0.0024485 0.0022967 1.066 0.286395   
## Stamina 0.0103999 0.0026428 3.935 8.36e-05 \*\*\*  
## Strength 0.0398028 0.0029535 13.477 < 2e-16 \*\*\*  
## LongShots -0.0218054 0.0034420 -6.335 2.45e-10 \*\*\*  
## Aggression -0.0032853 0.0024290 -1.353 0.176229   
## Interceptions -0.0034589 0.0034978 -0.989 0.322746   
## Positioning -0.0490919 0.0033449 -14.677 < 2e-16 \*\*\*  
## Vision -0.0399565 0.0036763 -10.869 < 2e-16 \*\*\*  
## Penalties 0.0078018 0.0030352 2.570 0.010169 \*   
## Composure 0.1388401 0.0038834 35.753 < 2e-16 \*\*\*  
## Marking 0.0345941 0.0027791 12.448 < 2e-16 \*\*\*  
## StandingTackle 0.0280537 0.0051862 5.409 6.44e-08 \*\*\*  
## SlidingTackle -0.0215724 0.0047697 -4.523 6.16e-06 \*\*\*  
## GKDiving 0.0061678 0.0071661 0.861 0.389425   
## GKHandling 0.0046242 0.0072611 0.637 0.524239   
## GKKicking -0.0027113 0.0070230 -0.386 0.699457   
## GKPositioning -0.0154661 0.0072236 -2.141 0.032289 \*   
## GKReflexes 0.0105602 0.0071758 1.472 0.141146   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 6.249683)  
##   
## Null deviance: 591428 on 12906 degrees of freedom  
## Residual deviance: 80440 on 12871 degrees of freedom  
## AIC: 60319  
##   
## Number of Fisher Scoring iterations: 2

For goalkeepers, there are 5 variables that have quite an effect on the overall score:

1. GKHandling
2. GKDiving
3. GKReflexies
4. GKPosition
5. Reactions

For regular players, it is slightly different:

1. Reactions
2. Ball Control
3. Composure
4. Short Passing
5. Heading Accuracy

Reactions is important for both sets of players, but that is it.

gkTest <- mutate(gkTest, differential = (model\_pred - Overall))  
offenseTest <- mutate(offenseTest, differential = (model\_pred - Overall))  
  
overachievingOffense <- subset(offenseTest, differential < 0)  
overachievingGK <- subset(gkTest, differential < 0)  
  
underachievingOffense <- subset(offenseTest, differential > 2)  
underachievingGK <- subset(gkTest, differential > 2)

There are 15 Goalkeepers who had ranked slightly higher than the model predicts and 1327 regular players. this corresponds to 3.704% of the goalkeepers and 41.122% of the regular players. Yikes! However, if we look at it a different way:

overachievingOffense.2 <- subset(overachievingOffense, abs(differential) >= 2)  
overachievingGK.2 <- subset(overachievingGK, abs(differential) >= 2)  
  
overachievingOffense.5 <- subset(overachievingOffense, abs(differential) >= 5)

There are now 0 goalkeepers that are 2 or more points away from their expected, and 860 regular players. So it seems goalkeepers are spot on but players are a little trickier!

However, 122 players are 5 or more points higher than the model predicts.

So who are the players we would find most desirable? Those would be the players that have an overall rating higher than our model predicts. They are more than the sum of their parts, basically.

We’ll now look at the top 6 players per position, based on them being better than the model predicts.

overachievingOffense <- overachievingOffense[order(overachievingOffense$differential),]  
  
overachievingGK <- overachievingGK[order(overachievingGK$differential),]  
  
print("Central Attacking Midfielder:")

## [1] "Central Attacking Midfielder:"

head(select(subset(overachievingOffense,overachievingOffense$Position == "CAM"), Name, Overall, differential))

## Name Overall differential  
## 300 Z. Jovanovic 64 -8  
## 2402 N. Silva 74 -8  
## 1178 Carlitos 64 -7  
## 167 E. LÃ³pez 72 -6  
## 723 L. BÃ©nes 69 -6  
## 775 RÃ©gis 74 -6

print("Center Back:")

## [1] "Center Back:"

head(select(subset(overachievingOffense,overachievingOffense$Position == "CB"), Name, Overall, differential))

## Name Overall differential  
## 2290 R. Schlegel 62 -8  
## 79 T. Davies 62 -6  
## 119 J. Trtovac 61 -6  
## 405 S. Ngezana 64 -6  
## 518 G. Hanley 72 -6  
## 935 T. Baack 58 -6

print("Central Defensive Midfielder:")

## [1] "Central Defensive Midfielder:"

head(select(subset(overachievingOffense,overachievingOffense$Position == "CDM"), Name, Overall, differential))

## Name Overall differential  
## 1615 S. AscacÃ­bar 78 -5  
## 189 L. Fejsa 82 -4  
## 2395 Markel Bergara 77 -4  
## 24 J. McCarthy 77 -3  
## 2088 S. Marreh 67 -3  
## 2397 N. Pelaitay 62 -3

print("Center Forward:")

## [1] "Center Forward:"

head(select(subset(overachievingOffense,overachievingOffense$Position == "CF"), Name, Overall, differential))

## Name Overall differential  
## 596 Luis Alberto 82 -4  
## 1027 B. Kuwas 74 -4  
## 637 Toni Villa 73 -3  
## 160 G. Caprari 73 -2  
## 2710 J. Vargas 67 -2

print("Center Midfield:")

## [1] "Center Midfield:"

head(select(subset(overachievingOffense,overachievingOffense$Position == "CM"), Name, Overall, differential))

## Name Overall differential  
## 900 N. KeÃ¯ta 83 -5  
## 786 Unai LÃ³pez 75 -4  
## 2603 B. Halimi 69 -4  
## 203 P. Galdames 70 -3  
## 226 J. Fuentes 68 -3  
## 325 A. GrgiÄ‡ 68 -3

print("Left Attacking Midfield:")

## [1] "Left Attacking Midfield:"

head(select(subset(overachievingOffense,overachievingOffense$Position == "LAM"), Name, Overall, differential))

## Name Overall differential  
## 1771 D. Buitrago 72 -8  
## 782 Leordinho Paes 72 -3  
## 1791 H. Abe 68 -2  
## 2910 Paulolettinho 71 -1

print("Left Back:")

## [1] "Left Back:"

head(select(subset(overachievingOffense,overachievingOffense$Position == "LB"), Name, Overall, differential))

## Name Overall differential  
## 1454 Elbis 62 -5  
## 431 Bai Jiajun 68 -4  
## 777 Jordi Alba 87 -4  
## 1464 Y. Armougom 59 -4  
## 1498 J. Pendant 63 -4  
## 1632 RaÃºl Llorente 69 -4

print("Left Center Back:")

## [1] "Left Center Back:"

head(select(subset(overachievingOffense,overachievingOffense$Position == "LCB"), Name, Overall, differential))

## Name Overall differential  
## 88 O. Ba 65 -6  
## 480 F. Fontanini 72 -6  
## 337 G. Nauber 68 -5  
## 1591 A. Sedlar 67 -5  
## 2942 G. Valsvik 69 -5  
## 2968 A. El-Abd 67 -5

print("Left Center Midfielder:")

## [1] "Left Center Midfielder:"

head(select(subset(overachievingOffense,overachievingOffense$Position == "LCM"), Name, Overall, differential))

## Name Overall differential  
## 247 K. Kampl 83 -4  
## 1933 Fran Villalba 69 -4  
## 2064 E. FernÃ¡ndez 68 -4  
## 2861 C. Pinares 71 -4  
## 1576 F. VÃ¡zquez 82 -3  
## 2171 L. Torreira 82 -3

print("Left Forward:")

## [1] "Left Forward:"

head(select(subset(overachievingOffense,overachievingOffense$Position == "LF"), Name, Overall, differential))

## Name Overall differential  
## 202 Jonathan Viera 82 -3

print("Left Midfielder:")

## [1] "Left Midfielder:"

head(select(subset(overachievingOffense,overachievingOffense$Position == "LM"), Name, Overall, differential))

## Name Overall differential  
## 661 S. Boufal 77 -7  
## 82 Felipe Anderson 83 -6  
## 816 J. Aquino 77 -6  
## 1664 E. Frear 63 -6  
## 2452 D. Gray 76 -6  
## 2797 T. Usami 74 -6

print("Left Striker:")

## [1] "Left Striker:"

head(select(subset(overachievingOffense,overachievingOffense$Position == "LS"), Name, Overall, differential))

## Name Overall differential  
## 2678 M. Balotelli 83 -6  
## 166 Douglas 72 -3  
## 266 Marc Gual 70 -2  
## 380 Santi Mina 80 -2  
## 454 Mata 76 -2  
## 1284 L. Pratto 78 -2

print("Left Wing:")

## [1] "Left Wing:"

head(select(subset(overachievingOffense,overachievingOffense$Position == "LW"), Name, Overall, differential))

## Name Overall differential  
## 1452 Isco 88 -8  
## 2867 L. Insigne 88 -7  
## 41 Y. Konoplyanka 79 -4  
## 521 H. St Clair 56 -4  
## 956 A. Al Qahtani 64 -4  
## 1407 J. Ngando 60 -4

print("Left Wing Back:")

## [1] "Left Wing Back:"

head(select(subset(overachievingOffense,overachievingOffense$Position == "LWB"), Name, Overall, differential))

## Name Overall differential  
## 2673 B. PittÃ³n 68 -3  
## 1742 M. Pedersen 68 -2  
## 2951 D. Lafferty 67 -2  
## 1988 Mossa 71 -1  
## 2150 R. Tait 67 -1  
## 3218 L. Carole 72 -1

print("Right Attacking Midfield:")

## [1] "Right Attacking Midfield:"

head(select(subset(overachievingOffense,overachievingOffense$Position == "RAM"), Name, Overall, differential))

## Name Overall differential  
## 1665 Y. Cabrera 69 -6  
## 2562 J. Cuadrado 84 -3  
## 2747 Emerson Avintes 71 -2

print("Right Back:")

## [1] "Right Back:"

head(select(subset(overachievingOffense,overachievingOffense$Position == "RB"), Name, Overall, differential))

## Name Overall differential  
## 3174 J. Risdon 72 -5  
## 157 T. Rieder 63 -4  
## 758 A. Wan-Bissaka 74 -4  
## 1580 Liu Boyang 54 -4  
## 393 V. Salazar 71 -3  
## 968 IvÃ¡n RodrÃ­guez 68 -3

print("Right Center Back:")

## [1] "Right Center Back:"

head(select(subset(overachievingOffense,overachievingOffense$Position == "RCB"), Name, Overall, differential))

## Name Overall differential  
## 1319 B. N'Gala 62 -8  
## 1543 M. Rahn 65 -8  
## 627 Manuel da Costa 71 -6  
## 664 G. Margreitter 74 -6  
## 939 O. Gonzalez 73 -6  
## 1512 S. Takahashi 62 -6

print("Right Center Midfielder:")

## [1] "Right Center Midfielder:"

head(select(subset(overachievingOffense,overachievingOffense$Position == "RCM"), Name, Overall, differential))

## Name Overall differential  
## 1969 Canales 80 -7  
## 988 Pablo Sarabia 82 -4  
## 1737 Sergi Darder 79 -4  
## 2331 J. Clasie 76 -3  
## 2586 F. Gino 66 -3  
## 3085 Javi Lara 70 -3

print("Right Forward:")

## [1] "Right Forward:"

head(select(subset(overachievingOffense,overachievingOffense$Position == "RF"), Name, Overall, differential))

## Name Overall differential  
## 766 Zhang Xizhe 72 -4  
## 2840 D. Moberg Karlsson 69 -4  
## 3148 G. Notsuda 66 -1

print("Right Midfielder:")

## [1] "Right Midfielder:"

head(select(subset(overachievingOffense,overachievingOffense$Position == "RM"), Name, Overall, differential))

## Name Overall differential  
## 2070 Lucas Moura 83 -7  
## 1944 Ã\201lvaro JimÃ©nez 74 -6  
## 2793 S. Kaneko 70 -6  
## 2937 B. AlÄ±cÄ± 73 -6  
## 206 J. Ibe 74 -5  
## 533 O. Romero 76 -5

print("Right Striker:")

## [1] "Right Striker:"

head(select(subset(overachievingOffense,overachievingOffense$Position == "RS"), Name, Overall, differential))

## Name Overall differential  
## 2527 M. BolaÃ±os 73 -6  
## 1120 K. Schindler 72 -4  
## 2159 Deulofeu 80 -4  
## 503 K. Billiat 75 -3  
## 2642 N. Citro 68 -3  
## 179 Z. IbrahimoviÄ‡ 85 -2

print("Right Wing:")

## [1] "Right Wing:"

head(select(subset(overachievingOffense,overachievingOffense$Position == "RW"), Name, Overall, differential))

## Name Overall differential  
## 1354 F. El Mellali 66 -6  
## 2980 R. Sterling 86 -6  
## 1425 R. Nelson 70 -5  
## 2237 C. Musonda 75 -5  
## 78 Kuki Zalazar 64 -4  
## 125 Lucas VÃ¡zquez 83 -4

print("Right Wing Back:")

## [1] "Right Wing Back:"

head(select(subset(overachievingOffense,overachievingOffense$Position == "RWB"), Name, Overall, differential))

## Name Overall differential  
## 2993 Pablo Maffeo 78 -4  
## 3076 Johannesson 70 -3  
## 51 M. Doherty 75 -2  
## 925 A. Amade 58 -2  
## 279 R. Laursen 65 -1  
## 1508 D. Sundgren 69 -1

print("Striker:")

## [1] "Striker:"

head(select(subset(overachievingOffense,overachievingOffense$Position == "ST"), Name, Overall, differential))

## Name Overall differential  
## 2511 K. Sierhuis 68 -5  
## 346 S. Demhasaj 64 -4  
## 1378 B. Gschweidl 64 -4  
## 1802 C. Huanca 58 -4  
## 2116 Tiago Marques 64 -4  
## 3224 L. FernÃ¡ndez 74 -4

head(select(overachievingGK, Name, Overall, differential))

## Name Overall differential  
## 20 K. Navas 87 -1  
## 110 J. Weaver 54 -1  
## 151 M. Hassen 71 -1  
## 159 Pablo CacharrÃ³n 58 -1  
## 165 T. Masuda 61 -1  
## 169 W. Yarbrough 68 -1

## Analysis

Overall, how accurate was our model? Here we will define inaccuracy as +- 2 points from predicted.

For regular players:

accuracyOffense <- (1 - ((nrow(underachievingOffense) + nrow(overachievingOffense.2)) / nrow(offenseTest))) \* 100

For goalkeepers:

accuracyGK <- (1 - ((nrow(underachievingGK) + nrow(overachievingGK)) / nrow(gkTest))) \* 100

The model is accurate for 96.3% of Goalkeepers in our test set, and 58.6% of players in our regular set. Still, when one considers the implication none of our players are more than 9 points off from prediction so it is actually a pretty decent model despite being off by 2 points for 40% of players.

## Conclusions

Based on our analysis, the overall skill is a fairly linear relationship between the different skills measured in our players. While there are some outliers, given a set of skills we can reasonably predict within a few points where that player will lie. While the initial analysis of players may have its own set of biases, once a player is graded their skills are weighted equally across the board. Goalkeepers in particular adhere fairly strictly to this.

We have now identified the top 5 skills for both sets of players in order to maximize your overall gains.These are the skills you should focus on to have the greatest overall effect.

This analysis is solely limited to FIFA 19. We may be able to improve the accuracy of the model by controlling for player position, something that we did not do other than by splitting out the goalkeepers.

Overall, I was impressed with how much of a relationship existed that was quantifiable. I was worried initially that there would be some liberties taken with scoring. I was pleasantly surprised to find that was not the case.