FINAL CAPSTONE PROJECT

ALY6015

INTERMEDIATE ANALYTICS

TRUNG PHAM LE MINH, TAIZHOU LI, XINGKUN CHEN

MATTHEW GOODWIN

NORTHEASTERN UNIVERSITY – SILLICON VALLEY

FIFA19: HOW DOES ONE CALCULATE THE POTENTIAL OF A PLAYER?

Problem Introduction

EA FIFA or FIFA game, also known as FIFA Football or FIFA Soccer is a series of association football video games released annually by EA under EA Sports label. It is the most anticipated sports game for football fans around the world. What makes it so special is in the game's ability to replicate a real-life football game, from the dimensions of a coach to the facial expressions of the players, FIFA has it all. One of the best aspects about the game is the idea of "Overall" stats of a specific player. Each player has an overall rating from 1 – 100 depending on their skill levels and attributes. Overall ratings are usually correlated to many of the attributes of the players, such as shooting, dribbling and passing. All of these attributes are also rated from 1 – 100 and the Overall rating is just taking the average of all the attributes. However, in FIFA, a player's overall rating doesn't stop at just that specific rating, players can improve and that would improve their Overall ratings. A very interesting statistic in FIFA called "Potential" is essentially the ceiling of a player's overall rating, meaning a player with a low overall right now can have a really high potential rating in the future, if used correctly in the game. Potential ratings are different from Overall ratings because it predicts the future outcome of something, and in this case it's the future potential rating of a player. Therefore, our group decides to answer the question: **how can one predict the potential ratings of a player and what factors contribute to a player's potential in FIFA19?**

Dataset

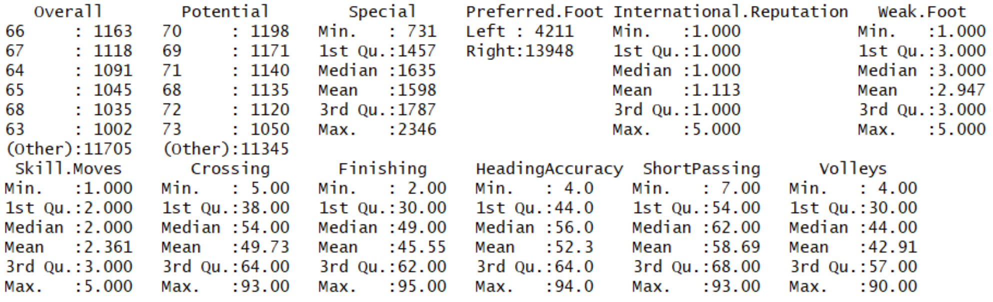
The dataset we have chosen is taken from the public website at waggle.com, where the initial dataset contains 18,207 observations with 53 different variables. The dataset contains a detailed attributes for every player registered in the latest edition of FIFA19 database. Apart from a player's name and ID which is a unique identifier for each player, there are 52 variables/attributes that contributes to a player's Overall Rating or Potential Ratings. A link to the dataset will be attached to this project for more details on the overall attributes of the players. It is noted that a potential rating of a player can only take integers, ranging from 48 to 95.

Preparing the Data

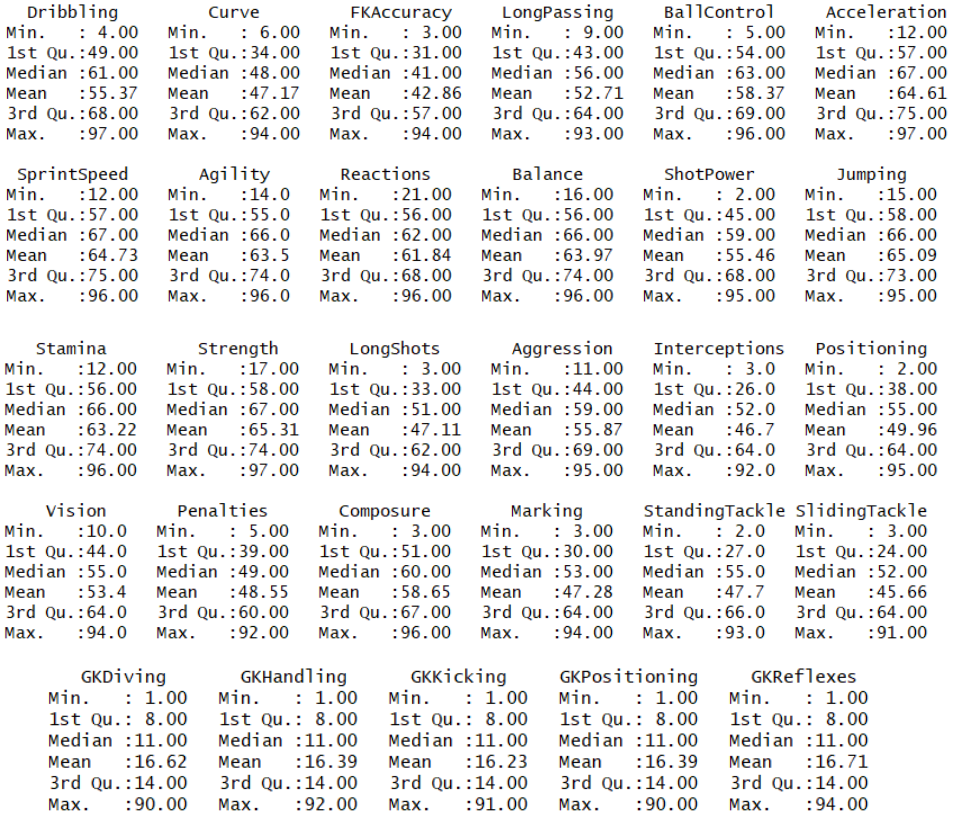
There were 48 NA's, or missing values and therefore, our group decided to only use rows with complete cases, hence 48 NA's were removed from the data. We have also changed some of the variable's class as either factor or character depending on its variables (details on R code document). We also left out 9 variables that we believe does not affect a player's Potential rating: Name, Nationality, Club, Contract.Valid. Until, Value, Wage, Position, Height and Weight. We will talk more about these variables later on. After cleaning out the data, we are left with 44 variables with 17,918 observations.

Descriptive Analysis

To get started, we look at the different descriptive statistics of our concerned dataset:



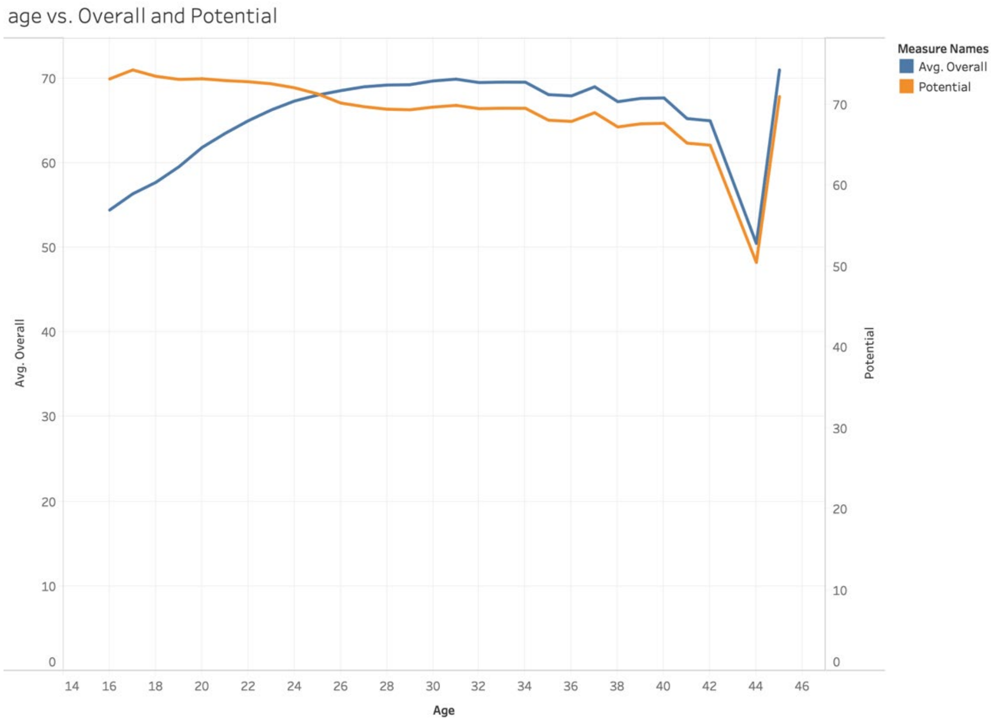
```
    Overall        Potential         Special      Preferred.Foot International.Reputation   Weak.Foot
 66      : 1163   70      : 1198   Min.   : 731   Left : 4211    Min.   :1.000            Min.   :1.000
 67      : 1118   69      : 1171   1st Qu.:1457   Right:13948    1st Qu.:1.000            1st Qu.:3.000
 64      : 1091   71      : 1140   Median :1635                  Median :1.000            Median :3.000
 65      : 1045   68      : 1135   Mean   :1598                  Mean   :1.113            Mean   :2.947
 68      : 1035   72      : 1120   3rd Qu.:1787                  3rd Qu.:1.000            3rd Qu.:3.000
 63      : 1002   73      : 1050   Max.   :2346                  Max.   :5.000            Max.   :5.000
 (Other):11705    (Other):11345
  Skill.Moves       Crossing         Finishing      HeadingAccuracy  ShortPassing      Volleys
 Min.   :1.000   Min.   : 5.00   Min.   : 2.00   Min.   : 4.0    Min.   : 7.00   Min.   : 4.00
 1st Qu.:2.000   1st Qu.:38.00   1st Qu.:30.00   1st Qu.:44.0    1st Qu.:54.00   1st Qu.:30.00
 Median :2.000   Median :54.00   Median :49.00   Median :56.0    Median :62.00   Median :44.00
 Mean   :2.361   Mean   :49.73   Mean   :45.55   Mean   :52.3    Mean   :58.69   Mean   :42.91
 3rd Qu.:3.000   3rd Qu.:64.00   3rd Qu.:62.00   3rd Qu.:64.0    3rd Qu.:68.00   3rd Qu.:57.00
 Max.   :5.000   Max.   :93.00   Max.   :95.00   Max.   :94.0    Max.   :93.00   Max.   :90.00
```



```
   Dribbling         Curve         FKAccuracy      LongPassing      BallControl      Acceleration 
 Min.   : 4.00   Min.   : 6.00   Min.   : 3.00   Min.   : 9.00   Min.   : 5.00   Min.   :12.00  
 1st Qu.:49.00   1st Qu.:34.00   1st Qu.:31.00   1st Qu.:43.00   1st Qu.:54.00   1st Qu.:57.00  
 Median :61.00   Median :48.00   Median :41.00   Median :56.00   Median :63.00   Median :67.00  
 Mean   :55.37   Mean   :47.17   Mean   :42.86   Mean   :52.71   Mean   :58.37   Mean   :64.61  
 3rd Qu.:68.00   3rd Qu.:62.00   3rd Qu.:57.00   3rd Qu.:64.00   3rd Qu.:69.00   3rd Qu.:75.00  
 Max.   :97.00   Max.   :94.00   Max.   :94.00   Max.   :93.00   Max.   :96.00   Max.   :97.00  

  SprintSpeed       Agility        Reactions        Balance        ShotPower        Jumping     
 Min.   :12.00   Min.   :14.0   Min.   :21.00   Min.   :16.00   Min.   : 2.00   Min.   :15.00  
 1st Qu.:57.00   1st Qu.:55.0   1st Qu.:56.00   1st Qu.:56.00   1st Qu.:45.00   1st Qu.:58.00  
 Median :67.00   Median :66.0   Median :62.00   Median :66.00   Median :59.00   Median :66.00  
 Mean   :64.73   Mean   :63.5   Mean   :61.84   Mean   :63.97   Mean   :55.46   Mean   :65.09  
 3rd Qu.:75.00   3rd Qu.:74.0   3rd Qu.:68.00   3rd Qu.:74.00   3rd Qu.:68.00   3rd Qu.:73.00  
 Max.   :96.00   Max.   :96.0   Max.   :96.00   Max.   :96.00   Max.   :95.00   Max.   :95.00  

    Stamina         Strength       LongShots       Aggression    Interceptions   Positioning   
 Min.   :12.00   Min.   :17.00   Min.   : 3.00   Min.   :11.00   Min.   : 3.0   Min.   : 2.00  
 1st Qu.:56.00   1st Qu.:58.00   1st Qu.:33.00   1st Qu.:44.00   1st Qu.:26.0   1st Qu.:38.00  
 Median :66.00   Median :67.00   Median :51.00   Median :59.00   Median :52.0   Median :55.00  
 Mean   :63.22   Mean   :65.31   Mean   :47.11   Mean   :55.87   Mean   :46.7   Mean   :49.96  
 3rd Qu.:74.00   3rd Qu.:74.00   3rd Qu.:62.00   3rd Qu.:69.00   3rd Qu.:64.0   3rd Qu.:64.00  
 Max.   :96.00   Max.   :97.00   Max.   :94.00   Max.   :95.00   Max.   :92.0   Max.   :95.00  

     Vision       Penalties       Composure        Marking      StandingTackle SlidingTackle  
 Min.   :10.0   Min.   : 5.00   Min.   : 3.00   Min.   : 3.00   Min.   : 2.0   Min.   : 3.00  
 1st Qu.:44.0   1st Qu.:39.00   1st Qu.:51.00   1st Qu.:30.00   1st Qu.:27.0   1st Qu.:24.00  
 Median :55.0   Median :49.00   Median :60.00   Median :53.00   Median :55.0   Median :52.00  
 Mean   :53.4   Mean   :48.55   Mean   :58.65   Mean   :47.28   Mean   :47.7   Mean   :45.66  
 3rd Qu.:64.0   3rd Qu.:60.00   3rd Qu.:67.00   3rd Qu.:64.00   3rd Qu.:66.0   3rd Qu.:64.00  
 Max.   :94.0   Max.   :92.00   Max.   :96.00   Max.   :94.00   Max.   :93.0   Max.   :91.00  

    GKDiving       GKHandling       GKKicking      GKPositioning     GKReflexes   
 Min.   : 1.00   Min.   : 1.00   Min.   : 1.00   Min.   : 1.00   Min.   : 1.00  
 1st Qu.: 8.00   1st Qu.: 8.00   1st Qu.: 8.00   1st Qu.: 8.00   1st Qu.: 8.00  
 Median :11.00   Median :11.00   Median :11.00   Median :11.00   Median :11.00  
 Mean   :16.62   Mean   :16.39   Mean   :16.23   Mean   :16.39   Mean   :16.71  
 3rd Qu.:14.00   3rd Qu.:14.00   3rd Qu.:14.00   3rd Qu.:14.00   3rd Qu.:14.00  
 Max.   :90.00   Max.   :92.00   Max.   :91.00   Max.   :90.00   Max.   :94.00  
```

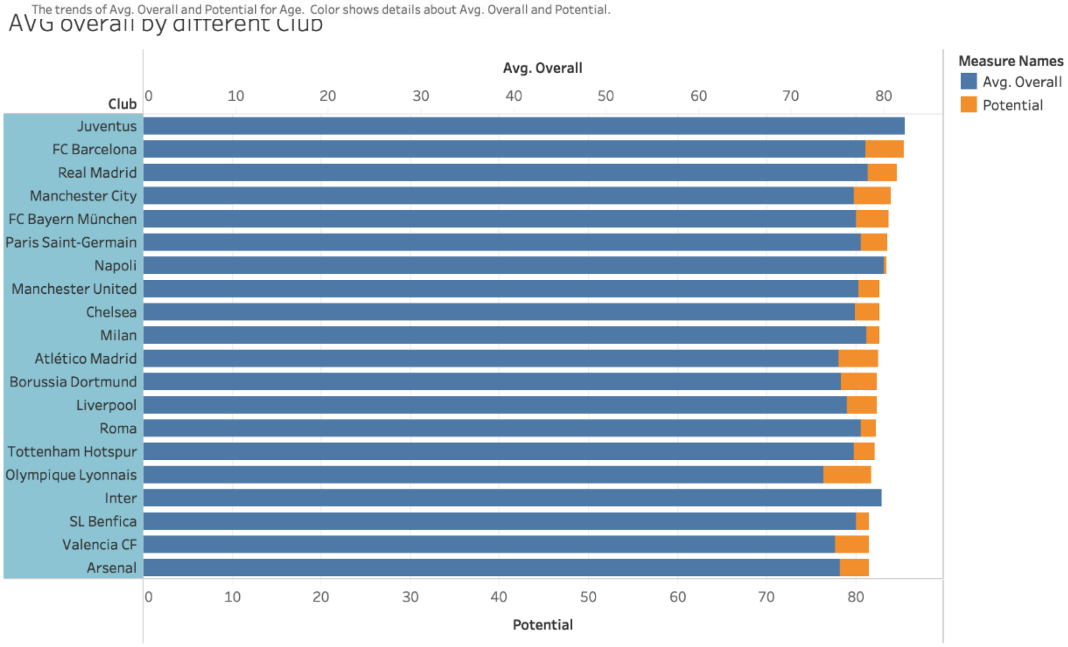
As we can see, there are a lot of instances of players with a rating of 70, and there are more right footed players than left footed. From the descriptive analytics, most of the attributes range from 1 – 100, but there are no players that have an attribute rating of 100. The closest is a player with an attribute rating for acceleration of 97. Four variables that are in a different scale are Special (continuous variable), International Reputation (1 to 5), Weak Foot ability (1 to 5) and Skill Moves ability (1 to 5). In order to do further analysis, we have to put these variables in the same scale as the other variables.

Data Visualization



The trends of Avg. Overall and Potential for Age. Color shows details about Avg. Overall and Potential.

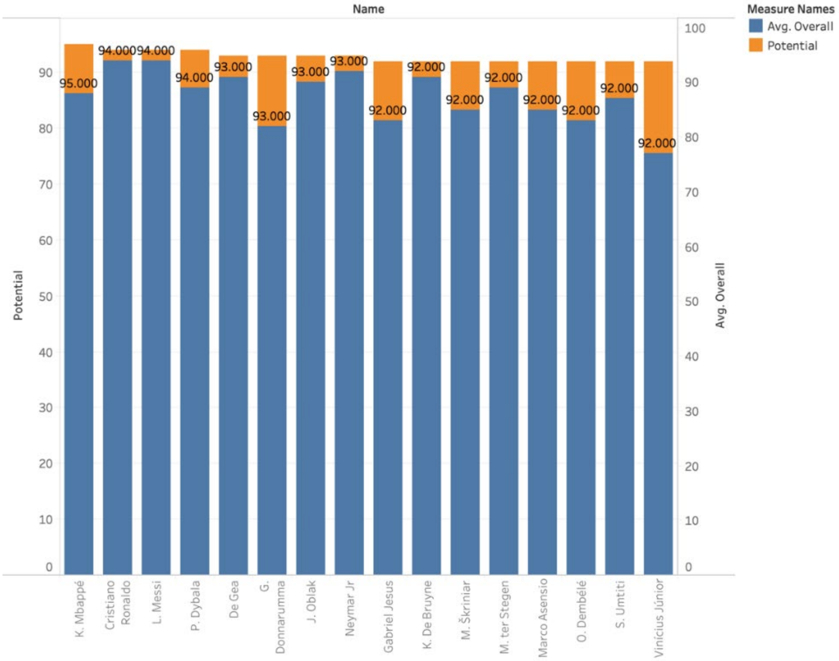
The First graph illustrates the average overall score and potential score within different ages. It is obvious that younger players have a higher potential score whilst the overall is not high. But after age 25, the players' overall score exceeds their potential score and continue descending. The potential score will stay in a same pattern as overall score after players' age 28.



Potential and Avg. Overall for each Club. Color shows details about Potential and Avg. Overall. The view is filtered on Club, which keeps 20 of 651 members.

The second graph shows the top 20 club with highest average overall score in FIFA 19. We can still see some of them still have some potential.

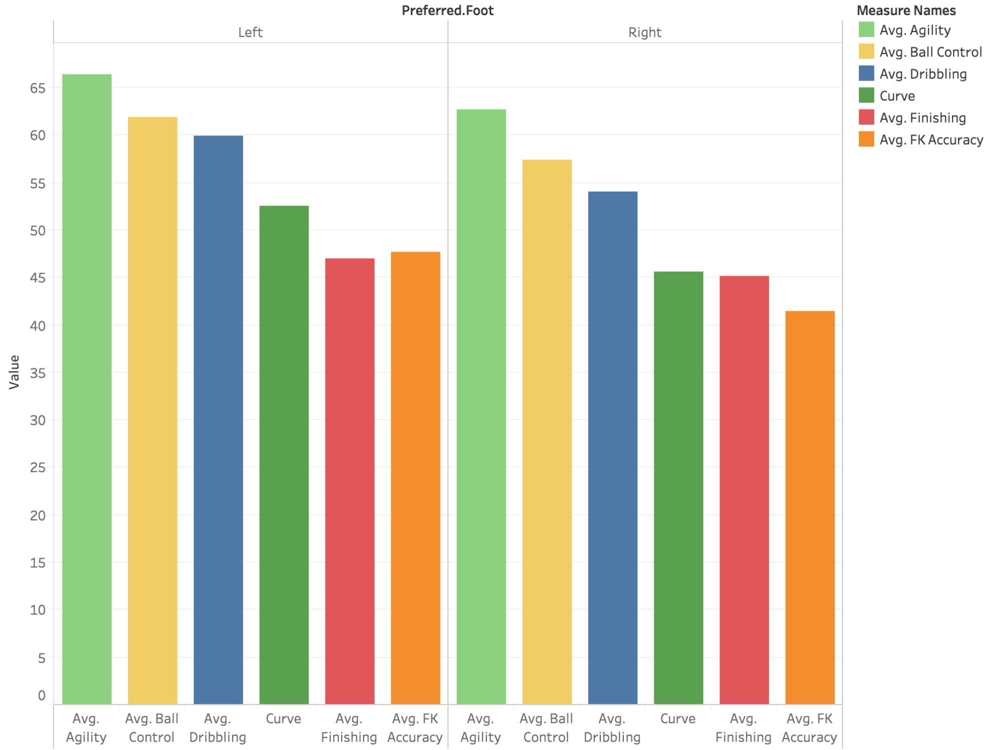
Top 16 in Potential score



Potential and Avg. Overall for each Name. Color shows details about Potential and Avg. Overall. For pane Average of Overall: The marks are labeled by Potential. The view is filtered on average of Potential, which ranges from 92.000 to 95.000.

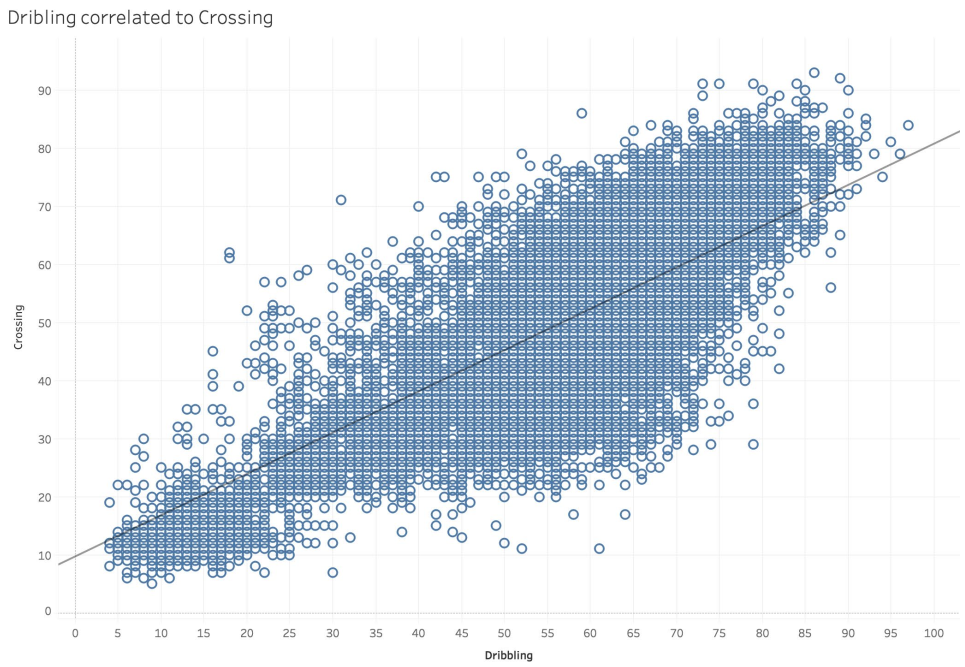
The third graph shows us the top 16 players with greatest potential in this graph, the overall score is given in it too. We can know there are some players who are now in relatively low overall level have a very good potential to become better in the future. The football managers should keep their eyes on them.

Prefered foot with their abilities



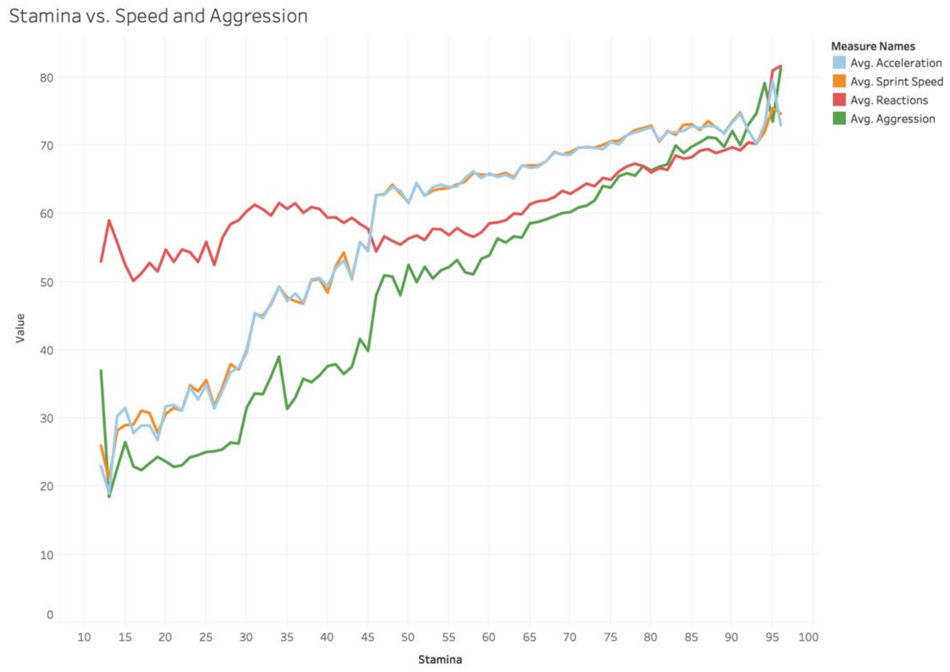
Avg. Dribbling, Avg. FK Accuracy, Avg. Finishing, Avg. Agility, Curve and Avg. Ball Control for each Preferred.Foot. Color shows details about Avg. Dribbling, Avg. FK Accuracy, Avg. Finishing, Avg. Agility, Curve and Avg. Ball Control.

There are more players right footed than left footed. From the bar graph, players who prefer using their left foot perform better in testing their agility, ball control skill, dribbling skill, curve skill, fishing skills and FK Accuracy.



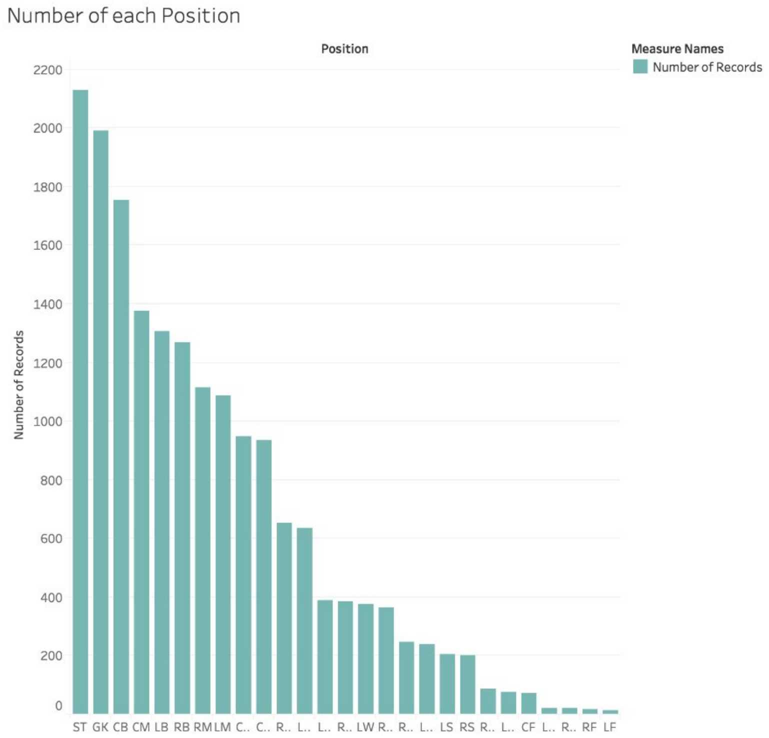
In this graph, it is clear the dribbling skill and crossing skill are positive related, players with better dribbling skill can perform better in crossing others.

Dribling vs. Crossing.



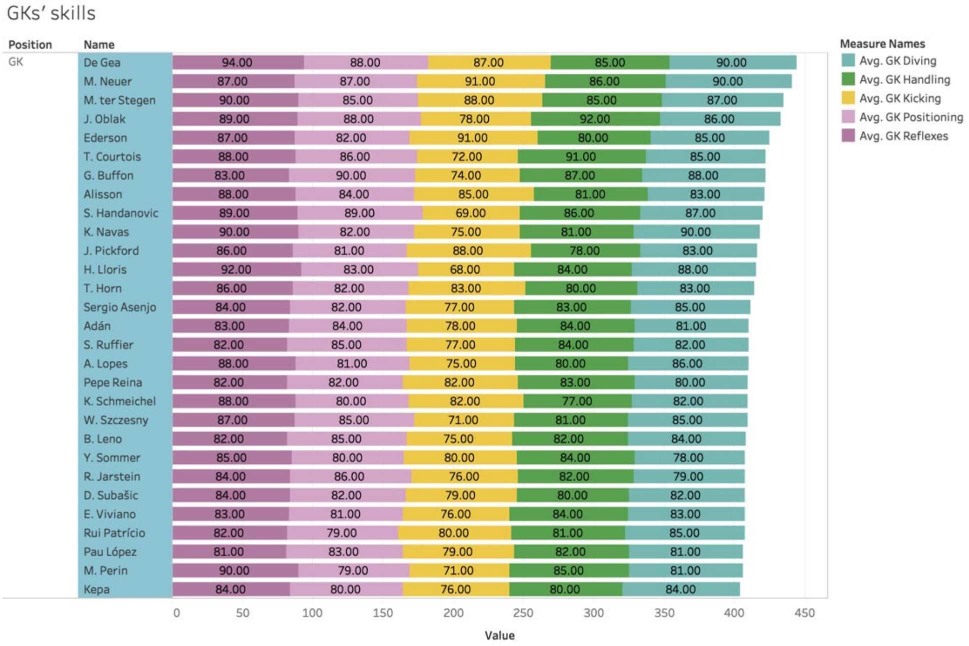
Players in a very strong physical conditions are also more agility, which is said their accelerating skill, sprint speed and reactions are also positive related with stamina.

The trends of Avg. Acceleration, Avg. Sprint Speed, Avg. Reactions and Avg. Aggression for Stamina. Color shows details about Avg. Acceleration, Avg. Sprint Speed, Avg. Reactions and Avg. Aggression.



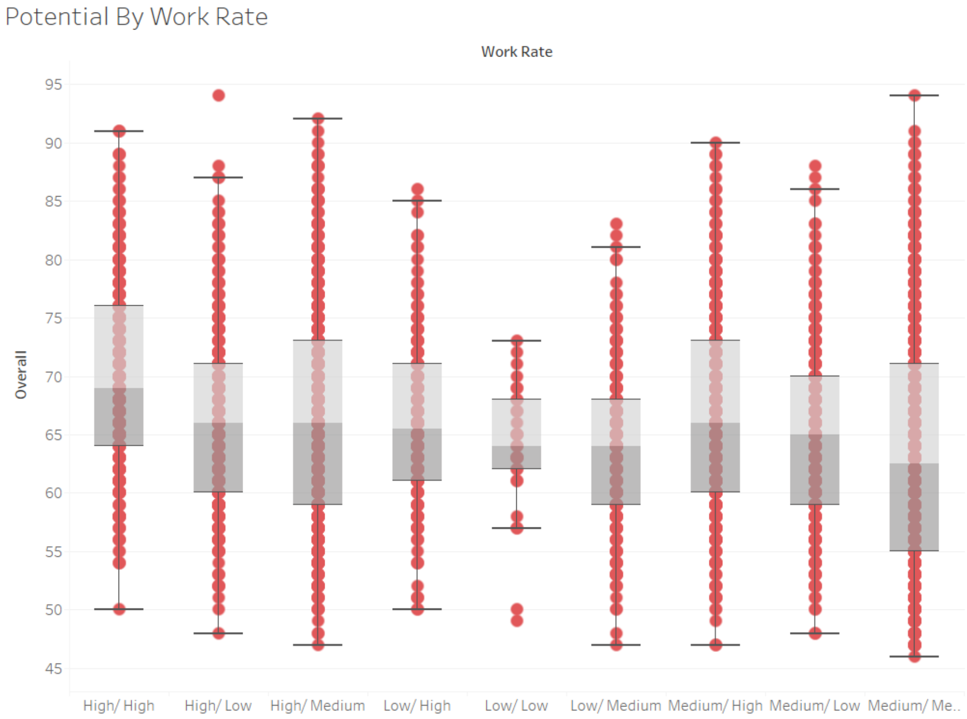
Number of Records for each Position. Color shows details about Number of Records.

Here is the number of records in each different position. There are more than 2100 players position in ST (striker) and the 2nd most position in the dataset is GK (goalkeeper).

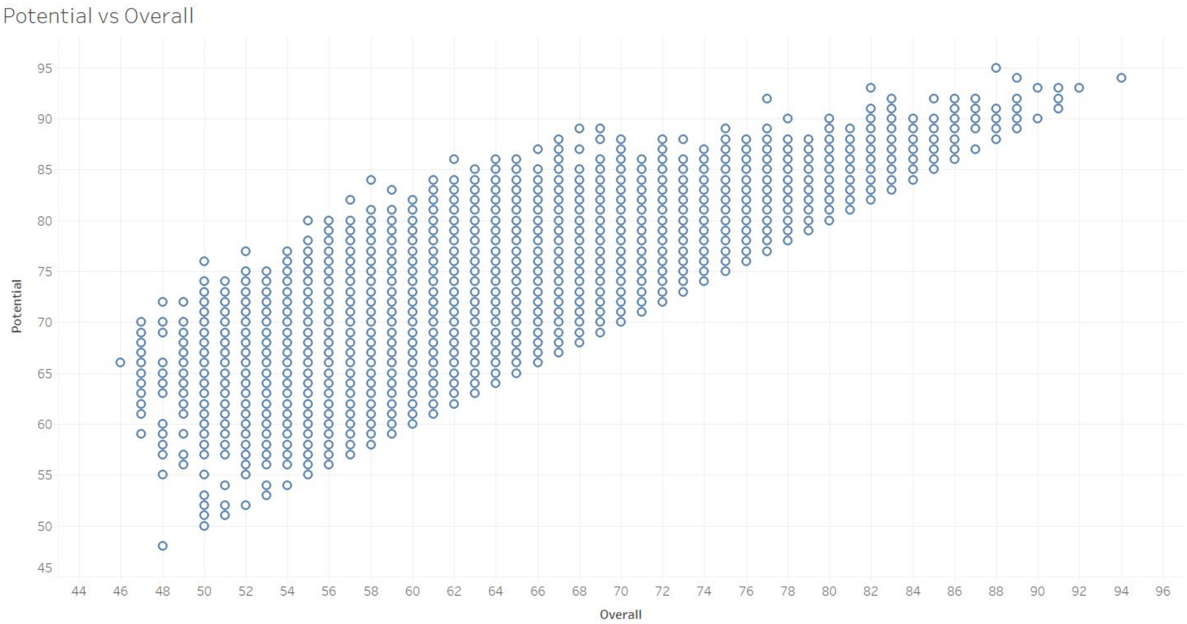


Avg. GK Diving, Avg. GK Handling, Avg. GK Kicking, Avg. GK Positioning and Avg. GK Reflexes for each Name broken down by Position. Color shows details about Avg. GK Diving, Avg. GK Handling, Avg. GK Kicking, Avg. GK Positioning and Avg. GK Reflexes. The marks are labeled by Avg. GK Diving, Avg. GK Handling, Avg. GK Kicking, Avg. GK Positioning and Avg. GK Reflexes. The view is filtered on Position and Name. The Position filter keeps GK. The Name filter keeps 29 of 16,923 members.

In this graph, we can see some of the best goal keepers in FIFA game. I calculate the average value of each skills include diving, handing, kicking, positioning and reflexes, and add them together to see who the top goal keepers in this game are.



There are many types of players with different work rate, and we can see that players that have high work rate will often have a better overall. For example, the median for Low/Low work rate players are lower than High/High work rate players.



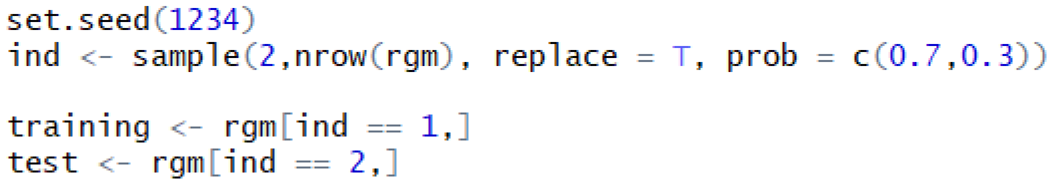
One of the most important aspect of our study is to see if collinearity exist in Overall and Potential of a player. From the graph, we can say that there does lie a linear relationship but not a collinearity, and hence we can use this variable in our analysis.

Data Mining Practices

After checking for collinearity and outliers, we will be using machine learning techniques to predict our outcome variable: Potential, with 43 other dependent variables. Since our outcome variable ranges from 0 – 100, we are considering a numerical integer variable, hence linear regression model can be used. We would also use randomForest model and compare the different results each model contains.

Data Partition

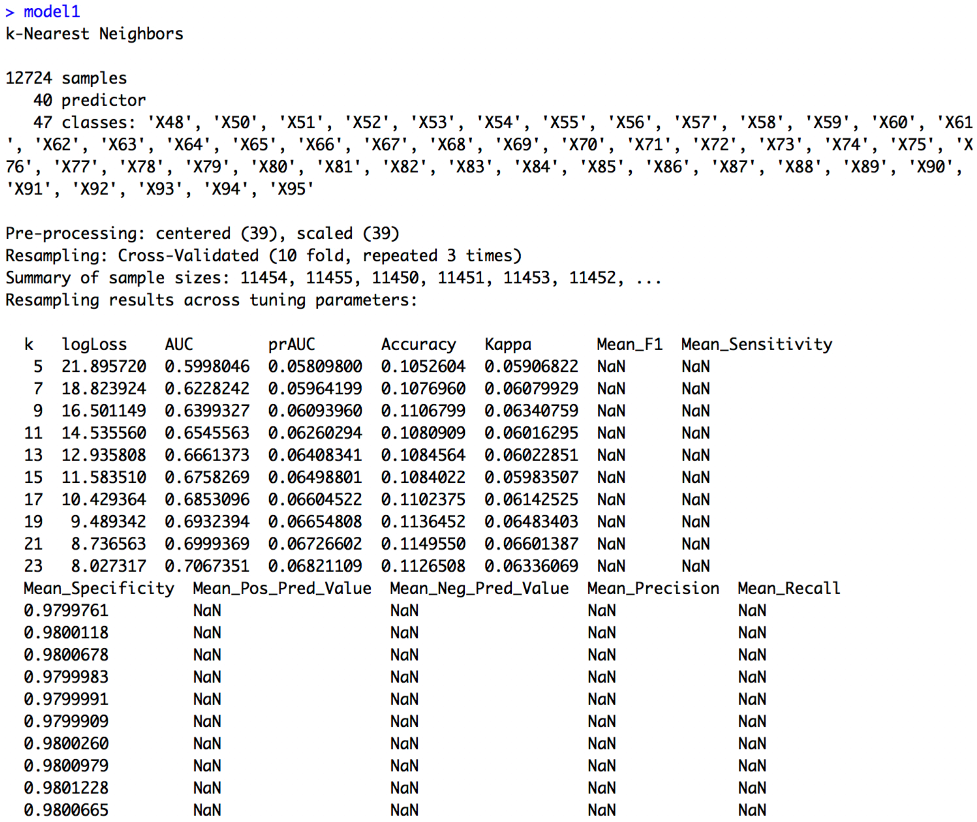
Our first step is to split our data into two set: train and test. The ratios we will be using to split our data will be 70/30. The piece of code below helps us by doing that:



```
set.seed(1234)
ind <- sample(2,nrow(rgm), replace = T, prob = c(0.7,0.3))

training <- rgm[ind == 1,]
test <- rgm[ind == 2,]
```

Our training dataset now has 12,527 observations while our testing dataset now has 5391 observations with each having 44 variables including the predictor variable in Potential. Now we are ready to use both supervised and unsupervised learning techniques in order to see which model predicts best for Potential ratings of a player in FIFA19.

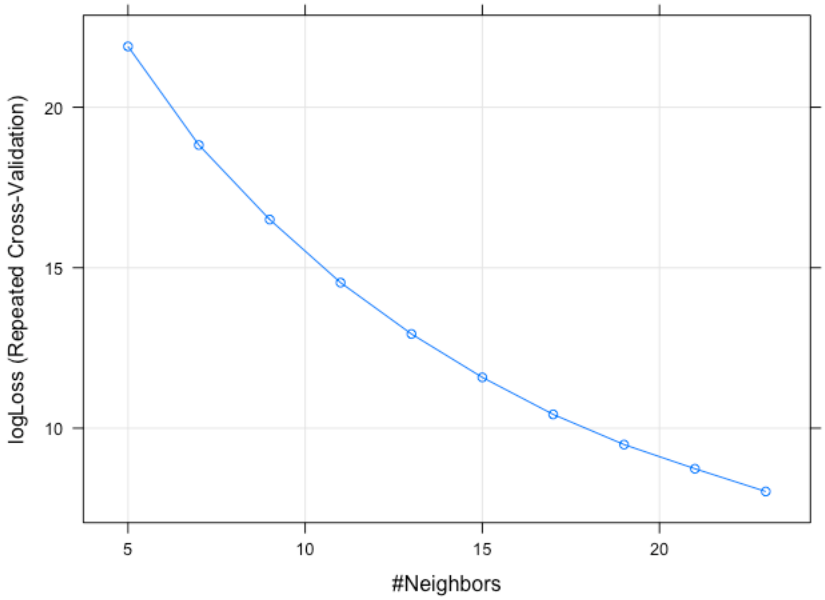
Supervised Learning: K-NN model



```
> model1
k-Nearest Neighbors

12724 samples
   40 predictor
   47 classes: 'X48', 'X50', 'X51', 'X52', 'X53', 'X54', 'X55', 'X56', 'X57', 'X58', 'X59', 'X60', 'X61
', 'X62', 'X63', 'X64', 'X65', 'X66', 'X67', 'X68', 'X69', 'X70', 'X71', 'X72', 'X73', 'X74', 'X75', 'X
76', 'X77', 'X78', 'X79', 'X80', 'X81', 'X82', 'X83', 'X84', 'X85', 'X86', 'X87', 'X88', 'X89', 'X90',
'X91', 'X92', 'X93', 'X94', 'X95'

Pre-processing: centered (39), scaled (39)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 11454, 11455, 11450, 11451, 11453, 11452, ...
Resampling results across tuning parameters:

  k   logLoss    AUC        prAUC       Accuracy   Kappa       Mean_F1  Mean_Sensitivity
   5  21.895720  0.5998046  0.05809800  0.1052604  0.05906822  NaN      NaN
   7  18.823924  0.6228242  0.05964199  0.1076960  0.06079929  NaN      NaN
   9  16.501149  0.6399327  0.06093960  0.1106799  0.06340759  NaN      NaN
  11  14.535560  0.6545563  0.06260294  0.1080909  0.06016295  NaN      NaN
  13  12.935808  0.6661373  0.06408341  0.1084564  0.06022851  NaN      NaN
  15  11.583510  0.6758269  0.06498801  0.1084022  0.05983507  NaN      NaN
  17  10.429364  0.6853096  0.06604522  0.1102375  0.06142525  NaN      NaN
  19   9.489342  0.6932394  0.06654808  0.1136452  0.06483403  NaN      NaN
  21   8.736563  0.6999369  0.06726602  0.1149550  0.06601387  NaN      NaN
  23   8.027317  0.7067351  0.06821109  0.1126508  0.06336069  NaN      NaN
  Mean_Specificity  Mean_Pos_Pred_Value  Mean_Neg_Pred_Value  Mean_Precision  Mean_Recall
  0.9799761         NaN                  NaN                  NaN             NaN
  0.9800118         NaN                  NaN                  NaN             NaN
  0.9800678         NaN                  NaN                  NaN             NaN
  0.9799983         NaN                  NaN                  NaN             NaN
  0.9799991         NaN                  NaN                  NaN             NaN
  0.9799909         NaN                  NaN                  NaN             NaN
  0.9800260         NaN                  NaN                  NaN             NaN
  0.9800979         NaN                  NaN                  NaN             NaN
  0.9801228         NaN                  NaN                  NaN             NaN
  0.9800665         NaN                  NaN                  NaN             NaN
```

According to KNN model, our training model is choosing k = 23 as its optimal k value using the smallest logloss value. We can also see that KNN model does not really predict well the Potential Ratings of a player since the most optimal K value yields only an accuracy of 0.112 or 11.2%.

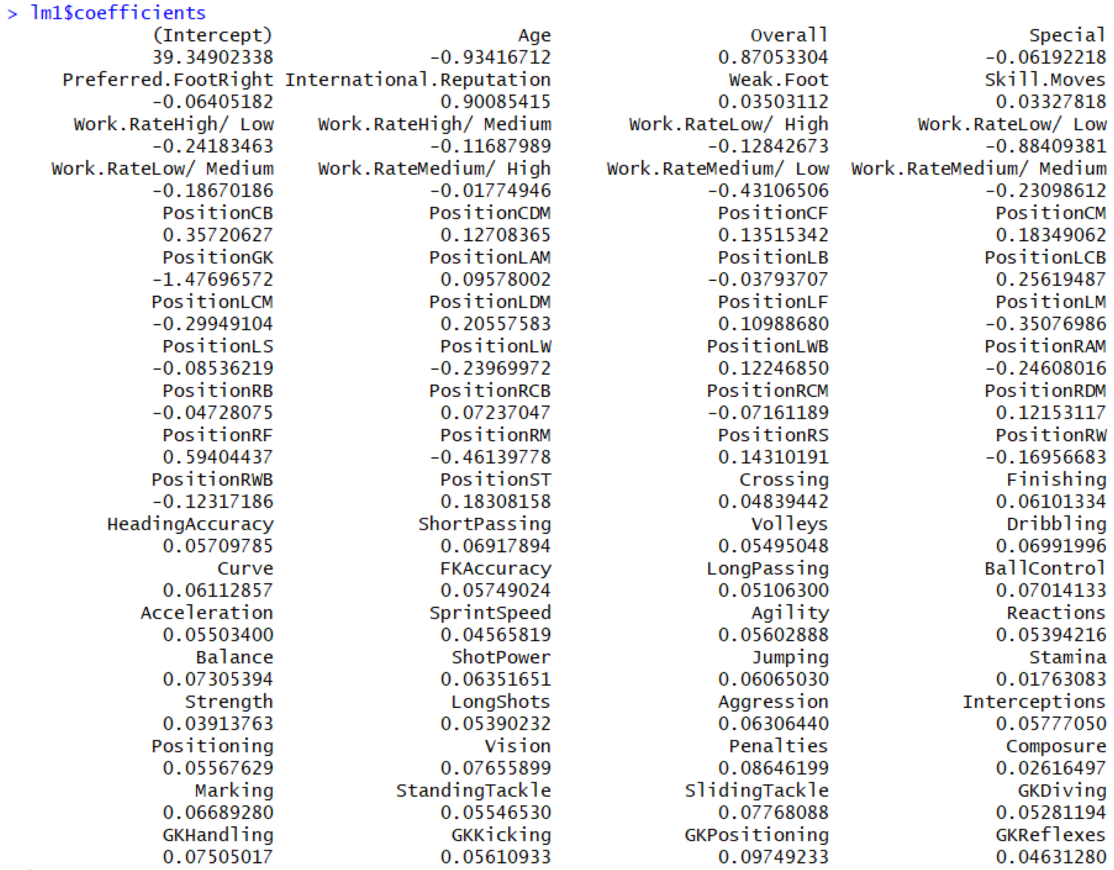
We can see the variation in logloss with k value in the graph:



Since the model performed very bad against its training dataset, our group decided to abandon K-NN model method and to use supervised methods in order to predict the Potential ratings of players in FIFA19.

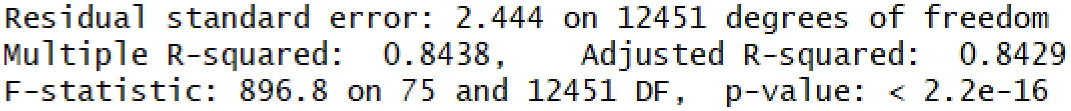
Unsupervised Learning Model: Multiple Linear Regression Model

Now, we are ready to build a linear regression model. After running the linear regression model of our predictor variable (Potential) against 43 other variables, we have the following coefficients:



```
> lm1$coefficients
             (Intercept)                      Age                  Overall                  Special
             39.34902338              -0.93416712               0.87053304              -0.06192218
     Preferred.FootRight International.Reputation                Weak.Foot              Skill.Moves
             -0.06405182               0.90085415               0.03503112               0.03327818
      Work.RateHigh/ Low    Work.RateHigh/ Medium       Work.RateLow/ High        Work.RateLow/ Low
             -0.24183463              -0.11687989              -0.12842673              -0.88409381
    Work.RateLow/ Medium    Work.RateMedium/ High     Work.RateMedium/ Low  Work.RateMedium/ Medium
             -0.18670186              -0.01774946              -0.43106506              -0.23098612
              PositionCB              PositionCDM               PositionCF               PositionCM
              0.35720627               0.12708365               0.13515342               0.18349062
              PositionGK              PositionLAM               PositionLB              PositionLCB
             -1.47696572               0.09578002              -0.03793707               0.25619487
             PositionLCM              PositionLDM               PositionLF               PositionLM
             -0.29949104               0.20557583               0.10988680              -0.35076986
              PositionLS               PositionLW              PositionLWB              PositionRAM
             -0.08536219              -0.23969972               0.12246850              -0.24608016
              PositionRB              PositionRCB              PositionRCM              PositionRDM
             -0.04728075               0.07237047              -0.07161189               0.12153117
              PositionRF               PositionRM               PositionRS               PositionRW
              0.59404437              -0.46139778               0.14310191              -0.16956683
             PositionRWB               PositionST                 Crossing                Finishing
             -0.12317186               0.18308158               0.04839442               0.06101334
         HeadingAccuracy             ShortPassing                  Volleys                Dribbling
              0.05709785               0.06917894               0.05495048               0.06991996
                   Curve               FKAccuracy              LongPassing              BallControl
              0.06112857               0.05749024               0.05106300               0.07014133
            Acceleration              SprintSpeed                  Agility                Reactions
              0.05503400               0.04565819               0.05602888               0.05394216
                 Balance                ShotPower                  Jumping                  Stamina
              0.07305394               0.06351651               0.06065030               0.01763083
                Strength                LongShots               Aggression            Interceptions
              0.03913763               0.05390232               0.06306440               0.05777050
             Positioning                   Vision                Penalties                Composure
              0.05567629               0.07655899               0.08646199               0.02616497
                 Marking           StandingTackle            SlidingTackle                 GKDiving
              0.06689280               0.05546530               0.07768088               0.05281194
              GKHandling                GKKicking            GKPositioning               GKReflexes
              0.07505017               0.05610933               0.09749233               0.04631280
```

The most important part in analyzing a regression model is in its error terms, in this case we are concerned with R-squared value and the significant level of our model in p-value:

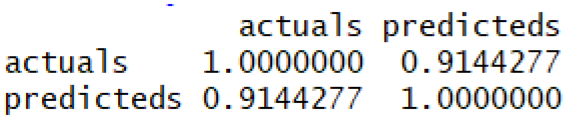


```
Residual standard error: 2.444 on 12451 degrees of freedom
Multiple R-squared:  0.8438,    Adjusted R-squared:  0.8429
F-statistic: 896.8 on 75 and 12451 DF,  p-value: < 2.2e-16
```

The R-squared value tells us the proportion of variation in the dependent (response) variable that has been explained in this model. Since we are looking at a nested model, we should use the adjusted R-squared value which is 0.8429 or 84.29%. This is a relatively high adjusted R-squared value, which tells us that the datapoints are relatively very close to the regression line.

Next, we check for the p-value, which is associated with the null and alternative hypothesis. In linear regression, the null hypothesis is that the coefficients associated with the variables are equal to zero, meaning that there does not exist a relationship between the independent variable and the dependent variables. The alternative hypothesis however, is the opposite of the null, meaning that there does exist a relationship between the independent in question and the dependent variables. We could see that our p-value is really low, and lower than our significance level of 0.05, we can reject our null hypothesis and say that there is sufficient evidence in concluding that there is a relationship between Potential ratings of a player and his/her attributes. (Note: Again, we could look at the F-statistic value and would yield the same result, since 896.8 is greater than 1.96)

Once we have determined that the model is statistically significant, we will now use our test data in order to judge how good our model is performing. Data frame "actual_preds" is created in R which entails the predicted Potential ratings from our linear model and the actual Potential ratings of the tested players.

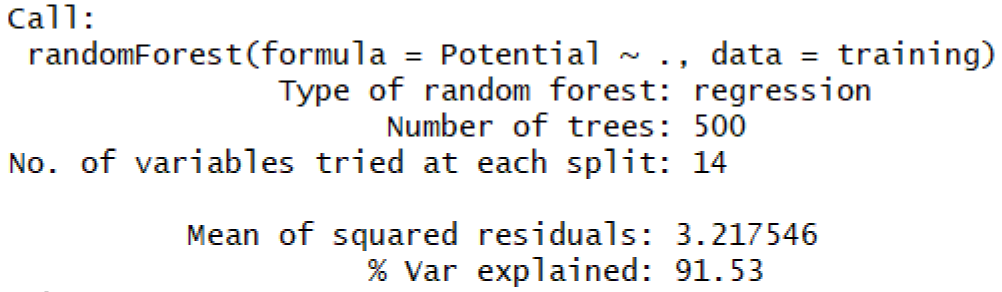


```
             actuals predicteds
actuals    1.0000000  0.9144277
predicteds 0.9144277  1.0000000
```

Running the correlation of this table gives us the accuracy of our model, which is what we want. As it stands, the accuracy of model is currently at 91.44%, which is quite high. To put it in perspective, we will be 9/10 times correct with our model in predicting the Potential of the FIFA19 soccer players, if there were new soccer players being added to our dataset.

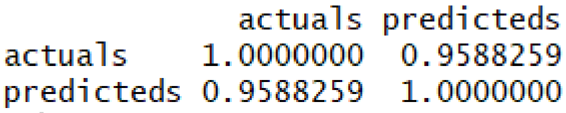
<u>Unsupervised Learning Model: Random Forest Model</u>

Now that we have analyzed our first model, as data analysts, we should always look for better performing models in order to yield the most accurate results. Hence, our goal is to beat the accuracy of our previous models. Therefore, the next machine learning technique we will be using is called a randomForest model.



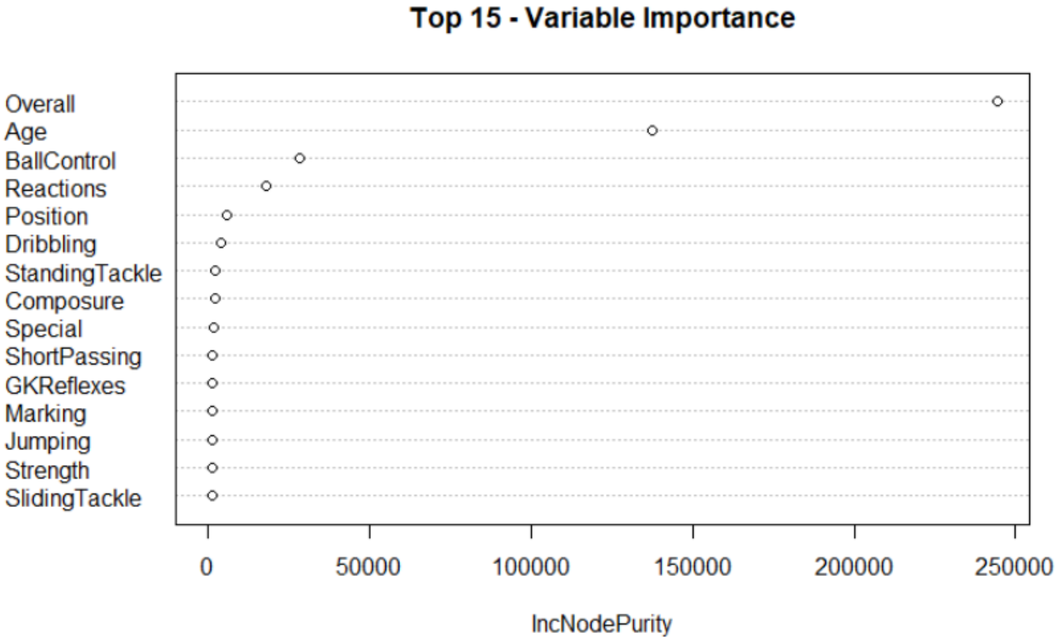
```
Call:
 randomForest(formula = Potential ~ ., data = training)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 14

          Mean of squared residuals: 3.217546
                    % Var explained: 91.53
```

When running a randomForest model by default without any restrictions, R will test on a total of 500 trees. As we look at our output above, we can see that the model chooses 14 variables at each split, with 91.48% of the variance explained (which is essentially the R-squared value). This shows that our randomForest model works better on the training dataset against our initial linear regression model since it has a higher R-squared value. However, the better training model might not mean that it is better at actually predicting the test data. To look at this model further, we can again use the same test data and see the accuracy of this randomForest model:



```
             actuals predicteds
actuals    1.0000000  0.9588259
predicteds 0.9588259  1.0000000
```

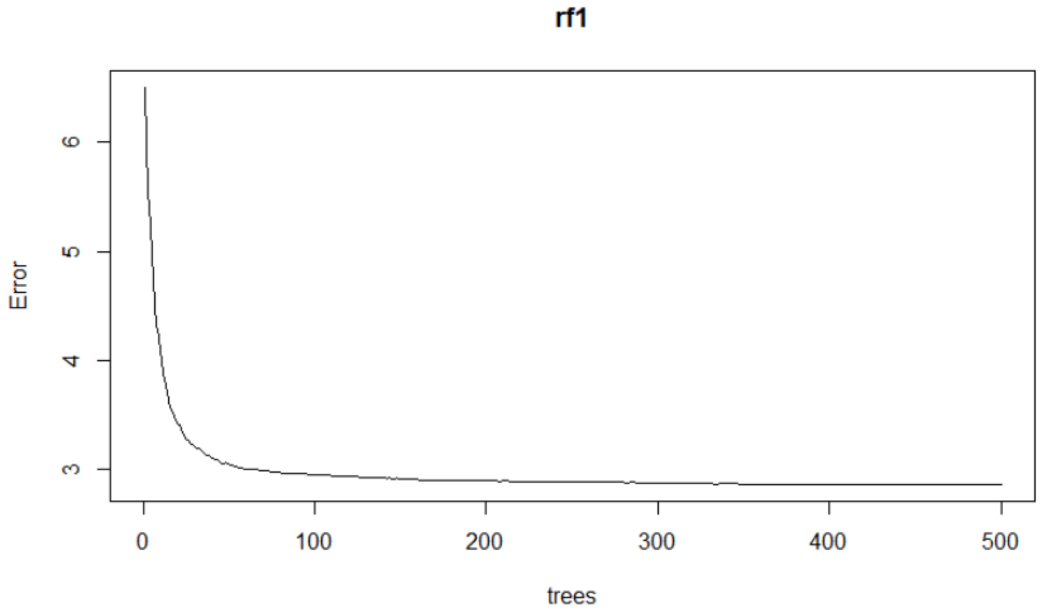
As one can see, the randomForest model did do a better job at achieving a higher accuracy of 95.88% approximately in comparison to 91.44% of the linear regression model.



From our randomForest model, we can also see the most important variables that are correlated to our outcome variable: Potential ratings of a player. The graph above shows the top 15 variables that are most important to Potential ratings. IncNodePurity is in essence the indicator of the importance of that variable in predicting Potential ratings. It is no surprise that Overall ratings of a player has the highest importance since a player's overall ratings should be quite close to their potential ratings. Age is something that is under looked here, since it also plays an important role in predicting the Potential ratings of a player. Players of a young age in football are considered to have a very high potential whilst as they get older, their potential ratings might not be as high and that's when they have reached their full potential, or maximum potential. Other important attributes are Ball Control, Reactions, Position, Dribbling, Standing Tackle ratings and many more.

Tree Tuning

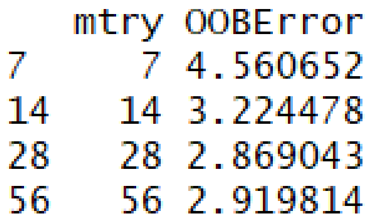
The next question is how do we know if R have used the most accurate model in predicting Potential ratings of players? The answer is we don't know but we can test it out. First, we can plot our model in terms of OOB error estimates against the number of trees used (Note: OOB error estimates, also known as out-of-bag error estimates are estimations about the error in the model when it tests against the data that wasn't included in the bootstrapping dataset):



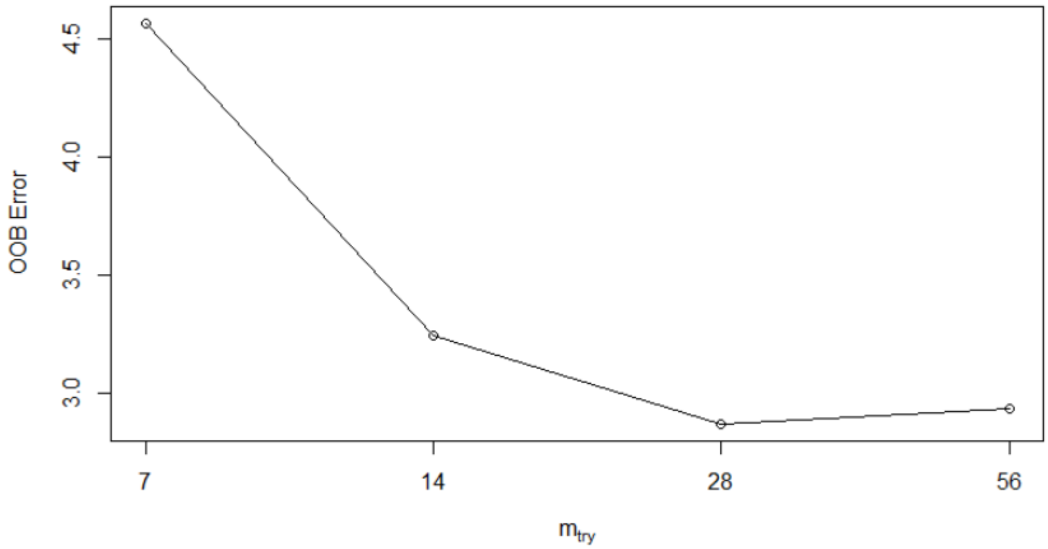
We could see that as we grow the number of trees, our error term significant goes down. Therefore, 500 trees are a good number to use in our model.

The last step in finalizing our randomForest model is to tune it. Before when we ran our first randomForest model, R uses its own number of trees to grow and the number of variables randomly sampled as candidates at each split. The tuneRF function in RandomForest package allows us to see if there are better number of trees to grow and the number of variables to randomly sampled at each split in accordance to the OOB error estimates. OOB error estimates, also known as out-of-bag error estimates are estimations about the error in the model if the model were to use data that it has not seen before. We could see that at mtry of 14, the OOB error estimate is at 3.22% approximately.

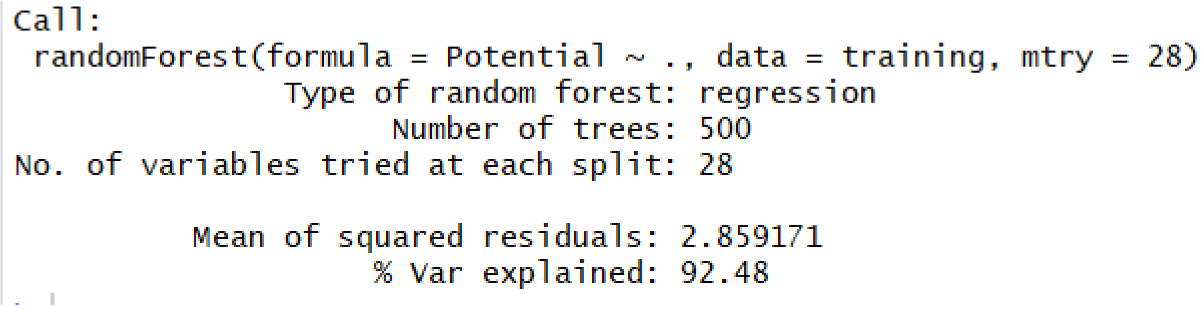
After tuning the tree, R found 4 instances of mtry at 7,14,28 and 56 where there is a significant difference in OOB error:



```
   mtry OOBError
7     7 4.560652
14   14 3.224478
28   28 2.869043
56   56 2.919814
```

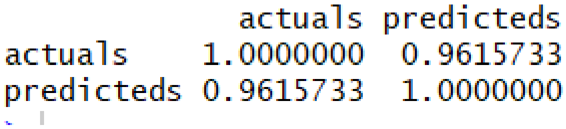


Surprisingly, as we can see the number of variables tried at each split that yields the lowest OOB error is actually 28 and not 14. Hence, we can now add constraints to our first randomForest model and hope to improve our accuracy rates:



```
Call:
 randomForest(formula = Potential ~ ., data = training, mtry = 28)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 28

          Mean of squared residuals: 2.859171
                    % Var explained: 92.48
```

We could see that the model has a lower OOB error at 2.86 and yields a higher R-squared value in comparison to the previous un-modified model. The last step is to use our model to test its performance for unseen data:



```
             actuals predicteds
actuals    1.0000000  0.9615733
predicteds 0.9615733  1.0000000
```

Finally, we have the best accuracy for our modified randomForest model, which is at **96.16%**. We are happy that our model accurately predicts a potential rating of players in FIFA19 96% of the time.

Conclusion

Comparison of Regression Models			
Model	**R-Squared Value (%)**	**Accuracy (%)**	**Standard Error/OOB Error**
Multiple-Regression	84.29	91.44	2.44
Unmodified randomForest	91.53	95.88	3.22
Modified randomForest	92.48	96.16	2.86

In conclusion, after using three different types of models with modifications, our team decides to use the modified randomForest model as our model to predict the potential ratings of a player in FIFA19, since it yields the lowest error and the highest R-squared value and also has the highest accuracy out of all models.