1

There are total 14 variables in FIFA18 dataset. Overall rating is the target variable.

There are 4 categorical variables: Name, Position, International rep and Preferred foot. The rest 9 are numeric continuous variables.

Name variable will not be considered for model building as it is not contributing to any significant information that can impact model building.

Summary Statistic of 14 variables:

**Categorical variables:**

1. Name: The categorical variable shows characters in each string of player names.
2. Position: Shows different positions on field where the player plays. Summary shows the count of players in each sub-category of the position. There are enough number of observations in each category. For example: 695 at ST, 391 at CB position etc.
3. Intl\_rep : There are 6 sub-categories from 0-5. And summary shows count of players belonging to each of them. There are enough number of values in each one of them. For example: 20 in 0, 745 in 1, 973 in 2, 822 in 3, 284 in 4, 55 in 5.
4. Pref\_foot: shows count of preferred foot of a player, left or right. There are 694 players who players whose preferred foot is left and 2205 have right foot as preferred foot.

**Continuous Variables:**

1. Overall: This is the target variable of the dataset where the model is run against it to predict the overall rating given other parameters of a player. Minimum Rating observed is:75, Maximum is 99. 25% players have rating below 79 whereas, 25% players have rating above 87. Mean is 83.17 and Median is 82.0. Histogram is right skewed.
2. Age: Represents age of the player. Minimum Rating observed is:17 that’s the youngest player age, Maximum is 77. 25% players have age below 25 whereas, 25% players have rating above 31. Mean is 28.37 and Median is 28. As mean is almost equal to median, values are normally distributed. But as the max value is way far from the 3rd quartile, it shows right skewness.
3. Height: Represents height of a player. Minimum height observed is:158, Maximum is 201. 25% players have height below 175 whereas, 25% players have rating above 186. Mean is 180.7 and Median is 180.0. As mean is almost equal to median, values are normally distributed and symmetric.
4. Weight: Represents weight of a player. Minimum weight observed is:57, Maximum is 110. 25% players have weight below 70 whereas, 25% players have rating above 80. Mean is 75.6 and Median is 75.0. As mean is almost equal to median, values are normally distributed and symmetric.
5. Pace: Represents pace of a player. Minimum pace observed is:33, Maximum is 99. 25% players have pace below 73 whereas, 25% players have rating above 87. Mean is 79.73 and Median is 81.0. As mean lesser than the median, values are normally distributed and left skewed.
6. Dribbling: Represents dribbling ability of a player. Minimum value for dribbling observed is:43, Maximum is 99. 25% players have value for dribbling below 76 whereas, 25% players have rating above 86. Mean is 80.18 and Median is 81. As mean lesser than the median, values are normally distributed and left skewed.
7. Shooting: Represents shooting ability of a player. Minimum value for shooting observed is:19, Maximum is 99. 25% players have value for shooting below 68 whereas, 25% players have rating above 84. Mean is 74.26 and Median is 77. As mean lesser than the median, values are normally distributed and left skewed.
8. Passing: Represents passing ability of a player. Minimum value for passing observed is:36, Maximum is 99. 25% players have value for passing below 70 whereas, 25% players have rating above 83. Mean is 76 and Median is 77. As mean lesser than the median, values are normally distributed and left skewed.
9. Defending: Represents defending ability of a player. Minimum value for defending observed is:18, Maximum is 98. 25% players have value for defending below 37 whereas, 25% players have rating above 78. Mean is 56.9 and Median is 53. As mean greater than the median. There are two major peaks in the graph showing two mountain like structure.
10. Physicality: Represents ability of a player to recover quickly from the injury. Lesser the recovery period, better is the physicality. Minimum value for physicality observed is:43, Maximum is 98. 25% players have value for physicality below 70 whereas, 25% players have rating above 83. Mean is 75.71 and Median is 77. As mean lesser than the median, values are normally distributed and left skewed.

2

The pairs function generates scatterplot of all continuous variables against the target variable **Overall**. From the matrix of graphs we can observe that, **Overall** has a clear positive linear relationship with Pace, Dribbling, Shooting, Passing and Physicality. This means Overall rating of a player increases with the increase in Pace, Dribbling , shooting, passing and physicality.

Apart from the target variable, there are some other relationships observed in the plot.

1. Height and weight have positive linear relationship.
2. Height and pace have negative linear relationship.
3. Height and dribbling have negative linear relationship.
4. Weight and pace have negative linear relationship.
5. Weight and dribbling have negative linear relationship.
6. Pace and dribbling have positive linear relationship.
7. shooting and dribbling have positive linear relationship.
8. passing and dribbling have positive linear relationship.
9. shooting and passing have positive linear relationship.

Positive Linear relationship means, as the x value increases, y value increases linearly. (Directly proportional)

Negative linear relationship means, as the x value increases, y value decreases linearly.(Inversely proportional)

**Histograms of continuous variables:**

1. Histogram of **Overall** isright skewed.
2. Histogram of **Age** isright skewed.
3. Histogram of **Height** issymmetric bell shaped and normally distributed.
4. Histogram of **Weight** is symmetric but slightlyright skewed.
5. Histogram of **Pace** isleft skewed.
6. Histogram of **Dribbling** isleft skewed.
7. Histogram of **Shooting** isleft skewed.
8. Histogram of **Passing** isalmost symmetric but left skewed.
9. Histogram of **Defending** has two maxima and looks like two symmetric bell shapes having minima at mean.
10. Histogram of **Physicality** isleft skewed.

3

Correlation matrix explains if the variables have positive or negative relationships. Also, it tells about the strength of the relationship whether its strong, moderate or weak.

* If the absolute value is above 60% it is considered as strong.
* If the absolute value is above 40% and below 60%, it is considered as moderate.
* If the absolute value is below 40%, it is considered as weak.

Correlation with **Overall** (target variable):

* Positive weak relation with age (25.5 %)
* Positive weak relation with height (1.8 %)
* Positive weak relation with weight (5.5 %)
* Positive weak relation with pace (26.1 %)
* Positive moderate relation with dribbling (57.3 %)
* Positive moderate relation with shooting (46.2 %)
* Positive strong relation with passing (63.9 %)
* Positive weak relation with defending (12.2 %)
* Positive weak relation with physicality (37.5 %)

Other observed correlations:

* Height has strong positive relation with weight( 78.9%)
* Weight moderate positive relation with physicality (55.9%)
* Pace has strong positive relation with dribbling( 61.1%)
* Height has moderate positive relation with physicality ( 51.9%)
* Dribbling has strong positive relation with shooting( 77.5%)
* Dribbling has strong positive relation with passing( 81.1%) <- Highest
* Shooting has strong positive relation with passing ( 61.7%)
* Shooting has moderate positive relation with defending ( 55.2%)

The strong interdependency of these variables cause multi-collinearity in the model.

4

**Boxplot of Overall Vs Position**

There are 14 boxplots generated representing distribution of each position of player corresponding to overall rating.

* For position CAM, the mean is closer to 1st quartile and the distribution is right skewed with one unusual entry at the right side.
* For position CB, the mean is exactly at the centre but skewed towards right. And there are two outliers.
* For position CDM, the mean is closer to 1st quartile and the distribution is right skewed with no outliers.
* For position CF, the distribution is centre aligned as the box is exactly in the middle but the mean is closer to 1st quartile causing distortion of the symmetric nature.
  + Similarly, the description goes for other 10 values.
* Specific point to hightlight are:
  + Outliers can be observed for positions Cam, CB, LM, LW, LWB, RB and RM
  + For position LWB, the mean coincides with the 1st quartile value
  + No position shows left skewness

**Boxplot of Overall Vs Intl\_rep**

There are 6 boxplots generated representing distribution of international reputation of player corresponding to overall rating.

* Outliers can be observed for the plots of international reputation as 0, 1, 2 and 3.
* For Int\_rep as 3, min value is almost at equal distance from 1st quartile as the distance of maximum from 3rd quartile. It shows almost symmetric shape hence symmetric normal distribution.
* For Int\_rep as 0, mean is closer to the 1st quartile. Also, minimum is closer to the 1st quartile than the maximum from 3rd quartile.
* For Int\_rep as 1, Minimum is very close to the first quartile, indicating that there are very less values in this region. Maximum is far from 3rd quartile hence the distribution is right skewed.
* For Int\_rep as 2, mean is closer to1st quartile. Also, Minimum is closer to 1st quartile than the maximum from 3rd quartile. The distribution is right skewed.
* For Int\_rep as 4, mean is closer to the 3rd quartile. Minimum and Maximum are almost equally spaced from 1st and 3rd quartile respectively.
* For Int\_rep as 5, mean is closer to the 3rd quartile. Also, the maximum is very close to the 3rd quartile as compared to the distance of minimum from 1st quartile.

**Boxplot of Overall Vs pref\_foot**

* Outliers can be observed in left foot boxplot.
* In both plots, mean is closer to the 1st quartile.
* In both plots, maximum is far from the 3rd quartile, showing right skewness

5

Since forward stepwise regression is suitable for choosing between a large number of potential predictors we will consider all predictors in the FIFA data set. The variables added or dropped at each stage are detailed in the output. The F and p-values for each variable at the step they were added to the model are supplied at the end of the output along with some model fit statistics.

The stepwise regression (or stepwise selection) consists of iteratively adding and removing predictors, in the predictive model, in order to find the subset of variables in the data set resulting in the best performing model, that is a model that lowers prediction error.

The model building starts with no predictors in the model, iteratively adds the most contributive predictors, and stops when the improvement is no longer statistically significant

In case of **FIFA** dataset, the model start with nothing i.e. **fitnull** =1

The model builds by taking one predictor variable at a time during each step.

**Fitfull** = all predictor variables (both conitinuous and categorical).

At every step, there is stepwise addition of one predictor variable from full model and it is then compared with the **extractAIC** value, by listing them in descending order.

At the next step again depending upon the AIC, the max AIC value is added to the solution.

The process stops when the AIC of variables is lesser than the AIC of none and resultant models given.

When stepwise forward regression applied to FIFA dataset, the resultant model was built with all variables except preferred foot.

Hence, the new model contains target variable regressed with 12 other predictor variables.(One variable less)

6

I have used **scatterplotMatrix** function from package **car** to find if there exist any interaction between continuous variable and categorical variables.

The identification criteria of existence of interaction is by observing the parallel nature of lines.

If the lines are **parallel**, it implies that there **does not exist any interaction** between the continuous and categorical variable.

If the lines are **non-parallel**(intersect each other), there **exists interaction** between the continuous and categorical variable.

According to scatterplotMatrix, there exist interaction between all continuous variables and categorical variables except **Dribbling**. (As the lines are parallel only in case of dribbling vs position and dribbling vs int\_rep).

To handle the interactivity between these variables, I have added multiplicative terms in my model.

For example, there exists interaction between age and position, then, in my model I will also add a term age\*position. Likewise, I have added all combinations except the ones involved dribbling.

7

Does your modified "best" model have multicollinearity? What test did your use to assess if your model had multicollinearity? What were the results of the test? What are the implications of the model having multicollinearity? In R modify your model to remove the multicollinearity if it is present. Describe what you did to remove the multicollinearity if it was present?

Summary statistic of the model has lots of terms having p-value much higher than 5%.

From the correlation matrix, I came to know that there exists multicollinearity in the model.

To test the presence of multicollinearity, I performed F-test on my model by using **summary.aov(model3) .**

All terms having p-value less than 5% are significant in the model and should be retained. The other terms are causing multi-collinearity. Such terms

There exists a strong positive or negative relationship among some of the variables.

Such variables are not adding any more significant information. it can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable and difficult to interpret. Multicollinearity saps the statistical power of the analysis, can cause the coefficients to switch signs, and makes it more difficult to specify the correct model. The confidence intervals tend to be wider. Therefore, we may not reject the “zero null hypothesis”. Model developed using the variables having multi-collinearity may not provide as accurate results because we are missing out on relevant variables/information in the first set.

I used the parameter to calculate variance inflation factor which is causing variation in the data.

vif(model3).

The terms having VIF >10 are dropped from the model. These are the terms causing inflation in the variance. Hence, need to be removed.

By doing this, I dropped predictor variables age, height, weight, pace, intl\_rep\*defending, intl\_rep\*physicality and shooting from the model.

After building the new model, I again tested it with F-test to be sure if the terms are significant or not.

After dropping these variables, the result of F-test gave all terms as significant (meaning the p-value is lesser than 5%)

8

I have performed regression again on the new model which was obtained in the previous step. And then computed summary statistics for it.

The latest model now contains

* Position
* Intl\_rep
* Dribbling
* Passing
* Defending
* Physicality
* And the interaction terms

The p-value is 2.2 e^-16 and hence it is highly significant. Also the Residual Standard Error is 1.273. The lower the error is better.

The Multiple R squared value is 94.76% and the adjusted R Squared value is 94.42%.

9

First I calculated leverage and plotted it. Then found the data items having leverage > 0.9 as all other values had leverage significantly lower.

This could be the potential influencing elements.

I found row element **1310** and **2654** having leverage > 0.9

To be sure of this, I computed cooks distance. These two points were passed to cooks distance. It was observed that both the points were actually not influencing the model in the cooks plot.

There is another value having cooks distance > 0.25 which is an outlier.

The point 2498 is an outlier element.

I created a new dataset Fifa\_data2 as Fifa\_data[-2498,]. This newly formed dataset which is free of outliers will now be used for further computations.

After that, I used my previously created latest model but used new data set and performed regression again on it.

Perfomed VIF to check if there is still any element causing variance inflation.

And ran F-test to be sure if all items are significant.

Next, I plotted residual plot against all variables and fitted vs residual to check if the residuals do not exhibit any significant pattern or not. All graphs are random and there is no significant pattern observed.

After that, I calculated studentized residual (r). and plotted it against all predictor variables in my model.

My Normal QQ plot is almost a straight line which closely resembles a perfect line.

Next, I plotted boxplot of r to check the distribution of it. It is perfectly symmetric with mean at center. 1st and 3rd quartiles are equally spaced from minimum and maximum respectively.

Histogram of r is bell shaped, symmetric and normally distributed.

10

The Multiple R squared value is 94.93% and the adjusted R Squared value is 94.6

The p value is 2.2\*e^-16 therefore the model is highly significant and the residual standard error is 1.252

The performance parameters of the model have been slightly improved after removing outliers from the model.

Residual standard error: 1.252 on 2723 degrees of freedom

Multiple R-squared: 0.9493, Adjusted R-squared: 0.946

F-statistic: 292.8 on 174 and 2723 DF, p-value: < 2.2e-16

R squared value can be misleading when we assess the goodness of fit of the linear regression model.

The addition of predictors increases the R squared value and appears to be a better fit. However, a model with too many predictors and higher order polynomials begins to train the random noise in the data, leading to overfitting of the model. To address this, Adjusted R square is used which adjusts the R2 for having too many variables in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance and decreases when a predictor improves the model by less than expected by chance.

Adjusted R squared is always lower than the R-squared.

AIC stands for Akaike's Information Criteria. The basic idea of AIC is to penalize the inclusion of additional variables to a model. It adds a penalty that increases the error when including additional terms. The lower the AIC, the better the model. AIC estimates the quality of each model, relative

to each of the other models for the same dataset. It provides a means for model selection. The basic idea behind AIC is the model using less number of parameters is better.

T-test is used to estimate the population parameter, i.e. population mean, and is also used for hypothesis

testing for population mean. It provides an insight whether the difference between the means of two groups

is due to chance or is reliable. T test answers the question

"Would the difference between these two groups be the same in a new sample from the same population?".