PREDICTING THE WAGE OF A SOCCER PLAYER USING HIS GAME STATISTICS

# Executive summary

Attributes for players such as potential rating, crossing rating, finishing rating, heading accuracy rating, dribbling rating, stamina rating, strength rating, long shots rating, aggression rating, positioning rating, penalties rating and sliding tackle rating are claimed to contribute to the player’s wage. Players with higher rating are claimed to be earning higher wages. To investigate how each of those variables are related to the wage of the player, it is necessary to perform correlation analysis and regression analysis to investigate the relationship between each of those attributes and the response variable wages.

The null hypotheses that were tested in this project were:

1. There is no relationship between potential and wages.
2. There is no relationship between crossing and wages.
3. There is no relationship between finishing and wages.
4. There is no relationship between heading accuracy and wages.
5. There is no relationship between dribbling and wages.
6. There is no relationship between stamina and wages.
7. There is no relationship between strength and wages.
8. There is no relationship between long shots and wages.
9. There is no relationship between aggression and wages.
10. There is no relationship between positioning and wages.
11. There is no relationship between penalties and wages.
12. There is no relationship between sliding tackle and wages.
13. The independent variables potential, crossing, finishing, heading accuracy, dribbling, stamina, strength, long shots, aggression, positioning, penalties and sliding tackle are predictors of the response variable wage

From the results of results of multiple linear regression that was conducted, a significant relationship was found to exist between the independent variable potential, crossing, heading accuracy, dribbling, stamina, strength, long shots, aggression, positioning and penalties and the response variable wages. There was no significant relationship between the predictor variables finishing and sliding tackle and the response variable wages. From the correlation matrix and scatter plot matrix that were plotted, we observed that there was a positive correlation between the response variable wages and the independent potential, crossing, finishing, heading accuracy, dribbling, stamina, strength, long shots, aggression, positioning, penalties and sliding tackle. However, some correlation appeared to be very small, nearly zero.

Also, we built three regression models that can predict the wage of a player based on his/her attributes. To build the model, we used multiple linear regression, decision tree and random forest algorithms to fit the training set which was 70% of the data. The model was then tested using the remaining 30% of the data and the model with the highest prediction accuracy selected. The models from the multiple linear regression, decision tree and random forest algorithms had a prediction accuracy of 51.87%, 68.79%, and 82.47% respectively. The model that was selected was thus the random forest model because it had the highest prediction accuracy.

# Summary of learning

From this analysis, I learnt that correlation matrix is the best way to know the correlation between the variables in the data set as opposed to the pair plot matrix. This is because the pair plot matrix was not very visible when having many variables and may data points. Also, the scatter plot matrix was great in visualizing scatter plot between variable in the data set because there is no need to visualize it pair by one pair. The density plot was used to investigate if the response variable is normal distributed. It was found to be right skewed.

Also, I learnt that multiple linear regression is great for hypothesis testing of the relationship between the response variable and the predictor variables while random forest and decision tree are great in building regression models to predict wage based on player statistics.

# Data used in the project

The data used in this project was from the website <https://www.kaggle.com/karangadiya/fifa19> which contains detailed features for every player registered in FIFA 2019 database. The data was originally scraped from the website <https://sofifa.com/> . The variables of interest were in the data was :

1. Wage - current wage
2. Potential - potential rating
3. Crossing - rating on scale of 100
4. Finishing - rating on scale of 100
5. Heading accuracy - rating on scale of 100
6. Dribbling - rating on scale of 100
7. Stamina - rating on scale of 100
8. Strength - rating on scale of 100
9. Long shots - rating on scale of 100
10. Aggression - rating on scale of 100
11. Positioning - rating on scale of 100
12. Penalties - rating on scale of 100
13. Sliding tackle - rating on scale of 100

# Exploratory Data Analysis

##### Method used in the data analysis

To establish the association between the response variable wage and the predictor variables potential, crossing, finishing, heading accuracy, dribbling, stamina, strength, long shots, aggression, positioning, penalties and sliding tackle, correlation matrix and scatterplot matrix were used. To investigate the relationship between each of the predictor variable and the dependent variable, and whether the relationship was significant, t-test was used. Multiple linear regression was then used to investigate if the predictor variables are a significant predictor of wage.

To build a regression model that can be used to predict wages, multiple linear regression, decision tree and random forest algorithms were used. The multiple linear regression used took the form:

The models were evaluated using the testing set prediction accuracy. The model with the highest testing set prediction accuracy was picked as the optimum model.

##### Variable identification

The variable wage was the response/target variable while the variables potential, crossing, finishing, heading accuracy, dribbling, stamina, strength, long shots, aggression, positioning, penalties and sliding tackle were the predictor variables. All the variables were numeric and continuous.

##### Summary statistics of the variables in the data

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Wage** | **Potential** | **Crossing** | **Finishing** | **HeadingAccuracy** | **Dribbling** | **Stamina** | **Strength** | **LongShots** | **Aggression** | **Positioning** | **Penalties** | **SlidingTackle** |
|
| **mean** | 9.75 | 71.32 | 49.73 | 45.55 | 52.30 | 55.37 | 63.22 | 65.31 | 47.11 | 55.87 | 49.96 | 48.55 | 45.66 |
| **sd** | 22.02 | 6.13 | 18.36 | 19.53 | 17.38 | 18.91 | 15.89 | 12.56 | 19.26 | 17.37 | 19.53 | 15.70 | 21.29 |
| **median** | 3.00 | 71.00 | 54.00 | 49.00 | 56.00 | 61.00 | 66.00 | 67.00 | 51.00 | 59.00 | 55.00 | 49.00 | 52.00 |
| **minimum** | 0.00 | 48.00 | 5.00 | 2.00 | 4.00 | 4.00 | 12.00 | 17.00 | 3.00 | 11.00 | 2.00 | 5.00 | 3.00 |
| **maximum** | 565.00 | 95.00 | 93.00 | 95.00 | 94.00 | 97.00 | 96.00 | 97.00 | 94.00 | 95.00 | 95.00 | 92.00 | 91.00 |
| **s.size** | 18159.00 | 18159.00 | 18159.00 | 18159.00 | 18159.00 | 18159.00 | 18159.00 | 18159.00 | 18159.00 | 18159.00 | 18159.00 | 18159.00 | 18159.00 |

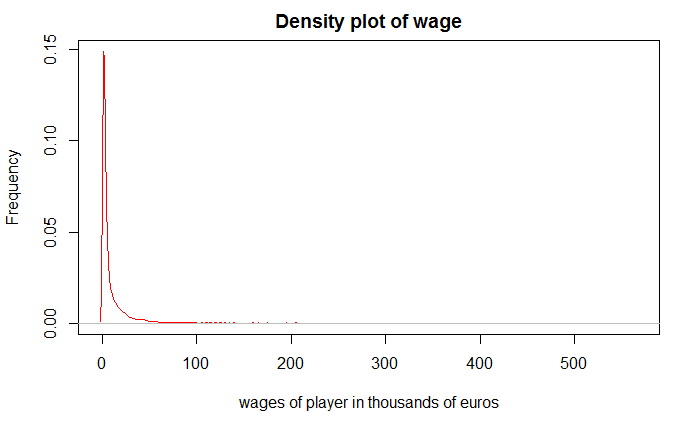
The average wage of player was € 9.75 (*SD =* 22.02) thousand. The wage of the players appears to be right skewed since the mean and the median does not correspond € 3.00. The minimum was wage was found to be € 0.00 and the maximum wage was found to be € 565.00. The average potential rating was 71.32 (*SD =* 6.13), average crossing rating was 49.73 (*SD =* 18.36), average finishing rating 45.55 (*SD =* 19.53), average heading accuracy rating 52.30 (*SD =* 17.38), average dribbling rating 55.37 (*SD =* 18.91), average stamina rating 63.22 (*SD =* 15.89), average strength rating 65.31 (*SD =* 12.56), average long shoot rating 47.11 (*SD =* 19.26), average aggression rating 55.87 (*SD =* 17.37), average positioning rating 49.96 (*SD =* 19.53), average penalties rating 48.55 (*SD =*15.70), average sliding tackle rating 45.66 (*SD =* 21.29).

##### Correlation matrix showing association between variables in the data frame

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Wage** | **Potential** | **Crossing** | **Finishing** | **HeadingAccuracy** | **Dribbling** | **Stamina** | **Strength** | **LongShots** | **Aggression** | **Positioning** | **Penalties** | **SlidingTackle** |
|
| **Wage** | 1.00 | 0.49 | 0.23 | 0.22 | 0.19 | 0.24 | 0.18 | 0.14 | 0.25 | 0.19 | 0.23 | 0.22 | 0.11 |
| **Potential** | 0.49 | 1.00 | 0.25 | 0.24 | 0.20 | 0.32 | 0.20 | 0.08 | 0.27 | 0.17 | 0.25 | 0.22 | 0.13 |
| **Crossing** | 0.23 | 0.25 | 1.00 | 0.66 | 0.47 | 0.86 | 0.67 | -0.03 | 0.74 | 0.47 | 0.78 | 0.65 | 0.41 |
| **Finishing** | 0.22 | 0.24 | 0.66 | 1.00 | 0.47 | 0.82 | 0.51 | -0.01 | 0.88 | 0.24 | 0.89 | 0.84 | -0.07 |
| **HeadingAccuracy** | 0.19 | 0.20 | 0.47 | 0.47 | 1.00 | 0.55 | 0.63 | 0.49 | 0.51 | 0.69 | 0.53 | 0.55 | 0.53 |
| **Dribbling** | 0.24 | 0.32 | 0.86 | 0.82 | 0.55 | 1.00 | 0.69 | -0.03 | 0.84 | 0.44 | 0.90 | 0.77 | 0.27 |
| **Stamina** | 0.18 | 0.20 | 0.67 | 0.51 | 0.63 | 0.69 | 1.00 | 0.26 | 0.60 | 0.65 | 0.64 | 0.52 | 0.54 |
| **Strength** | 0.14 | 0.08 | -0.03 | -0.01 | 0.49 | -0.03 | 0.26 | 1.00 | 0.05 | 0.47 | 0.01 | 0.05 | 0.30 |
| **LongShots** | 0.25 | 0.27 | 0.74 | 0.88 | 0.51 | 0.84 | 0.60 | 0.05 | 1.00 | 0.39 | 0.86 | 0.81 | 0.13 |
| **Aggression** | 0.19 | 0.17 | 0.47 | 0.24 | 0.69 | 0.44 | 0.65 | 0.47 | 0.39 | 1.00 | 0.38 | 0.34 | 0.72 |
| **Positioning** | 0.23 | 0.25 | 0.78 | 0.89 | 0.53 | 0.90 | 0.64 | 0.01 | 0.86 | 0.38 | 1.00 | 0.80 | 0.12 |
| **Penalties** | 0.22 | 0.22 | 0.65 | 0.84 | 0.55 | 0.77 | 0.52 | 0.05 | 0.81 | 0.34 | 0.80 | 1.00 | 0.07 |
| **SlidingTackle** | 0.11 | 0.13 | 0.41 | -0.07 | 0.53 | 0.27 | 0.54 | 0.30 | 0.13 | 0.72 | 0.12 | 0.07 | 1.00 |

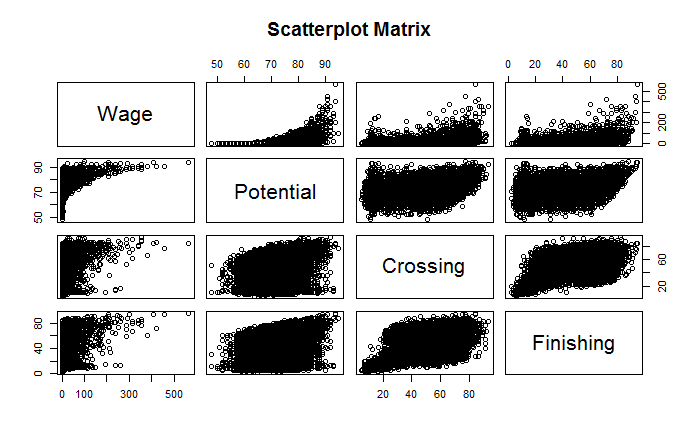
From the correlation matrix, we observed that the variables wage and potential (r = 0.49), wage and crossing (r = 0.23), wage and finishing (r = 0.22), wage and heading accuracy (r = 0.19), wage and dribbling (r = 0.24), wage and stamina (r = 0.18), wage and strength (r = 0.14), wage and long shots (r = 0.25), wage and aggression (r = 0.19), wage and positioning (r = 0.23), wage and penalties (r = 0.22), wage and sliding tackle (r = 0.11) are positively correlated.

##### Density plot indicating the distribution of wages of players



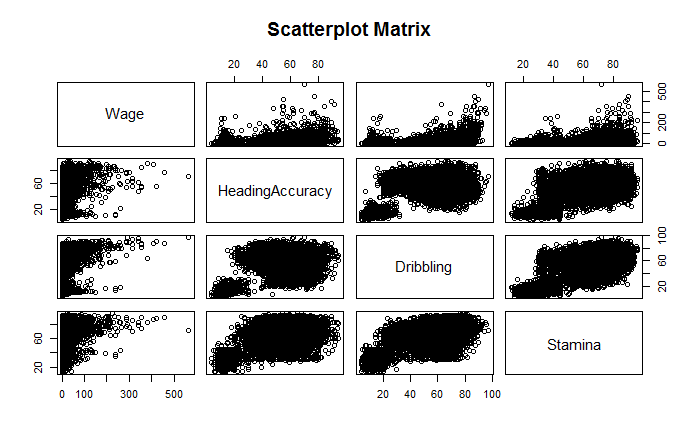
The density plot was used to investigate if the variable current wage of players is normally distributed. It was established that it is a right skewed distribution thus not normal.

##### Scatterplot matrix showing linear relationship the between the variables wages potential, crossing and finishing



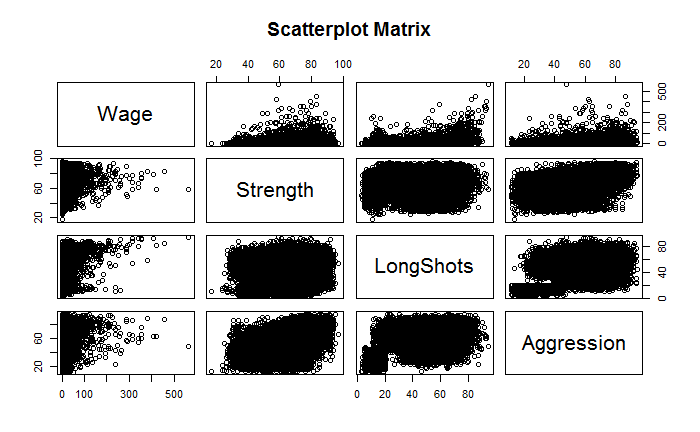
The variables potential rating, crossing rating and finishing rating appears to be having a linear relationship with the variable wage.

##### Scatterplot matrix showing linear relationship the between the variables wages, heading accuracy, dribbling and stamina



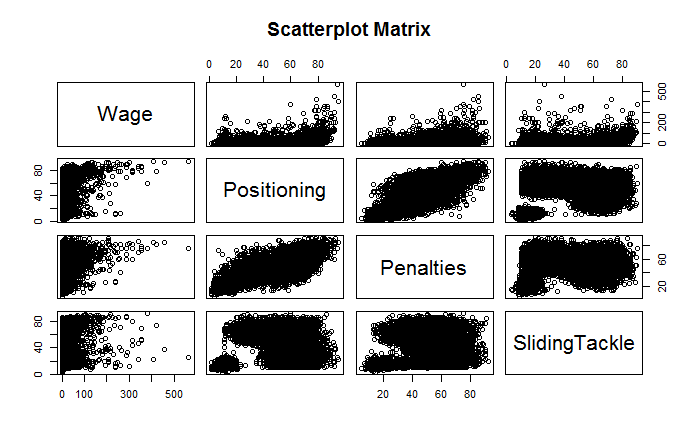
The variables potential head accuracy, dribbling rating and stamina rating appears to be having a linear relationship with the variable wage.

##### Scatterplot matrix showing linear relationship the between the variables wages, strength, long shots and aggression



The variables strength rating and long shots rating and aggression rating, appears to be having a linear relationship with the variable wage.

##### Scatterplot matrix showing linear relationship the between the variables wages, positioning, penalties and sliding tackle



The variables positioning rating and penalties rating and sliding tackle rating, appears to be having a linear relationship with the variable wage.

##### Results output from the multiple linear regression

The data was split into 70% training set and 30% testing set. The testing set was then fitted using least square method. The following results was found.

Call:

lm(formula = Wage ~ ., data = training\_set)

Residuals:

Min 1Q Median 3Q Max

-41.39 -8.20 -2.01 4.34 511.72

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -130.74375 2.23553 -58.485 < 2e-16 \*\*\*

Potential 1.70030 0.02997 56.736 < 2e-16 \*\*\*

Crossing 0.19050 0.02106 9.044 < 2e-16 \*\*\*

Finishing 0.01848 0.02793 0.662 0.508213

HeadingAccuracy -0.06310 0.01853 -3.405 0.000664 \*\*\*

Dribbling -0.21596 0.02825 -7.645 2.23e-14 \*\*\*

Stamina -0.06561 0.01873 -3.503 0.000462 \*\*\*

Strength 0.18733 0.01846 10.146 < 2e-16 \*\*\*

LongShots 0.07305 0.02252 3.243 0.001184 \*\*

Aggression 0.08388 0.01831 4.582 4.65e-06 \*\*\*

Positioning 0.09259 0.02607 3.552 0.000384 \*\*\*

Penalties 0.08294 0.02218 3.739 0.000186 \*\*\*

SlidingTackle -0.01328 0.01623 -0.818 0.413325

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.39 on 12703 degrees of freedom

Multiple R-squared: 0.2746, Adjusted R-squared: 0.2739

F-statistic: 400.6 on 12 and 12703 DF, p-value: < 2.2e-16

The multiple year model can thus be described using the formula:

As the potential rating increase by 1 rating, the wage increase by € 1.70 thousand holding other attributes constant. This change is significant (*t* = 56.74, *p* < .05). We thus reject the null hypothesis that there is no relationship between wages and potential rating.

As the crossing rating increase by 1 rating, the wage increase by € 0.19 thousand holding other attributes constant. This change is significant (*t* = 9.04, *p* < .05). We thus reject the null hypothesis that there is no relationship between wages and crossing rating.

As the finishing rating increase by 1 rating, the wage increase by € 0.02 thousand holding other attributes constant. This change is not significant (*t* = 9.04, *p* < .05). We thus fail to reject the null hypothesis that there is no relationship between wages and finishing rating.

As the head accuracy rating increase by 1 rating, the wage decrease by € 0.06 thousand holding other attributes constant. This change is significant (*t* = -3.41, *p* < .05). We thus reject the null hypothesis that there is no relationship between wages and head accuracy rating.

As the dribbling rating increase by 1 rating, the wage decrease by € 0.22 thousand holding other attributes constant. This change is significant (*t* = -7.65, *p* < .05). We thus reject the null hypothesis that there is no relationship between wages and dribbling rating.

As the stamina rating increase by 1 rating, the wage decrease by € 0.07 thousand holding other attributes constant. This change is significant (*t* = -3.50, *p* < .05). We thus reject the null hypothesis that there is no relationship between wages and stamina rating.

As the long shot rating increase by 1 rating, the wage increase by € 0.19 thousand holding other attributes constant. This change is significant (*t* = 10.15, *p* < .05). We thus reject the null hypothesis that there is no relationship between wages and long shot rating.

As the aggression rating increase by 1 rating, the wage increase by € 0.08 thousand holding other attributes constant. This change is significant (*t* = 4.58, *p* < .05). We thus reject the null hypothesis that there is no relationship between wages and crossing rating.

As the positioning rating increase by 1 rating, the wage increase by € 0.09 thousand holding other attributes constant. This change is significant (*t* = 3.55, *p* < .05). We thus reject the null hypothesis that there is no relationship between wages and crossing rating.

As the penalty rating increase by 1 rating, the wage increase by € 0.08 thousand holding other attributes constant. This change is significant (*t* = 3.74, *p* < .05). We thus reject the null hypothesis that there is no relationship between wages and crossing rating.

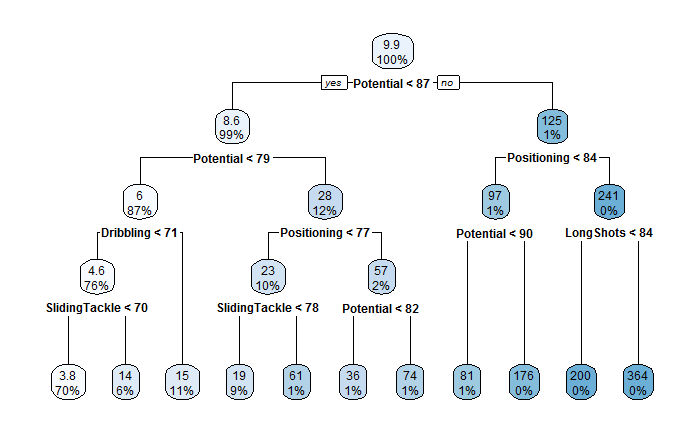
As the side tackle rating increase by 1 rating, the wage decrease by € 0.01 thousand holding other attributes constant. This change is significant (*t* = -0.81, *p* < .05). We thus reject the null hypothesis that there is no relationship between wages and crossing rating.

Also, the predictor variables potential, crossing, finishing, heading accuracy, dribbling, stamina, strength, long shots, aggression, positioning, penalties and sliding tackle used in the model are significant predictors of the current wage of a player (*F(*12,12703*)* = 400.6, *p* < .05).

To evaluate the multiple linear regression model, the testing set was used to predict the new predicted response variable. The new predicted response variable was then compared with the real testing set values and prediction accuracy computed. The Prediction accuracy of the multiple linear regression model was found to be 51.87%.

##### Results output from the Decision Tree model

In attempt to improve the prediction accuracy of the multiple regression model, decision tree algorithm was fitted on the training set. The decision tree model was formed based on the following decision tree.



To evaluate the decision tree model, the testing set was used to predict the new predicted response variable. The new predicted response variable was then compared with the real testing set values and prediction accuracy computed. The Prediction accuracy of the decision tree model was found to be 68.79%. It was better than the multiple linear regression in predicting wages of player.

##### Results output from the Random Forest Model

In improving the results of the decision tree, an ensemble of 500 decision trees were deployed to predict wages using the random forest algorithm. The random forest model improved the prediction accuracy to 82.47%. We thus concluded that random forest model is the best in predicting the wages of players using the predictor variables potential, crossing, finishing, heading accuracy, dribbling, stamina, strength, long shots, aggression, positioning, penalties and sliding tackle.