Prediction of the value of the football players

İlhami Yılmaz  
*Middle East Technical University*Ankara, Turkey  
ilhamiyilmaz.iy@gmail.com

*Abstract*—In this paper, by using FIFA 20 Player Dataset, the models are conducted to predict the value of the players by other variables. Before conducting the model, data preparation is made. Missing values are imputed, some variables are dropped, min-max transformation is used. After cleaning, Multiple Linear Regression, Artificial Neural Network, Support Vector Machine, Random Forest and XGBoost models are conducted. The models are compared by their Root-Mean-Square Errors (RMSE) and Mean Absolute Errors (MAE). All procedures are executed on R-Studio

Keywords—Multiple Linear Regression, ANN, SVM, Random Forest, XGBoost, Modelling, FIFA

# INTRODUCTION

The dataset is retrieved from Kaggle.com and it contains the details of the football players from the game FIFA 20. This data contain 18278 players and 104 variables.

Our target variable will be the *value* variable, which shows the value of the player in Euro, and we will be focusing on how value changes when other variables change. After cleaning the data, Multiple Linear Regression, Artificial Neural Network, Support Vector Machine, Random Forest and XGBoost models are conducted. The models are compared by their Root-Mean-Square Errors (RMSE) and Mean Absolute Errors (MAE). All procedures are executed on R-Studio

# DATA DESCRIPTION AND PRE-PROCESSING

## Dataset

The dataset is retrieved from Kaggle.com and it contains the details of the football players from the game FIFA 20. This data contain 18278 players and 104 variables. The data has 1.294.332‬ missing points. Dealing with missing points are going to be covered in the following part of the analysis. The variable description is given below.

* *sofifa\_id*: Shows the id of the player (we will not use this variable since we are not interested in the ID numbers of the players.)
* *player\_url*: It includes the url of the player which includes all the information about the player. We will drop this column too.
* *short\_name*: The short name of the player
* *long\_name*: The long name of the player
* age: Age of the player
* dob: Date of birth of the player
* *height\_cm*: The height of the player in centimeters
* *weight\_kg*: The weight of the player in kilograms
* *nationality:* The nationality of the player
* *club:* The team that the player is playing for
* *overall:* The overall rating of the player (?/100)
* *potential:* The potential rating of the player (?/100)
* *value\_eur*: The value of the player in Euro
* *wage\_eur*: The wage of the player in Euro
* *player\_positions*: The positions that the player can play in with no problem.
* preferred\_*foot:* The foot that the player prefer to use (Left/Right)
* *international\_reputation*: The international reputation of the player (?/5)
* *weak\_foot*: The power of the weak foot of the player. It is also a scale of 5 variable. Having a 5-star weak foot rating means that a player's weaker foot shot is identical to their preferred foot shot.
* *skill\_moves*: The ability that a player can perform technical moves. To all skill moves, it is assigned a number between 1 (least complex) and 5 (most complex). Players with higher skills are able to perform more complex moves
* *work\_rate*: First one shows the Attacking Work Rate of the player, and after the “/”, it shows the Defensive Work Rate of the player. They both rated between low, medium and high, which defines how a player puts effort to participates in attacks and defenses even when they are out of position.
* *body\_typ*e: It shows the body type of the players. They are shown as “Lean”, “Normal”, “Stocky”, but there are some exceptions like “Messi”, “C. Ronaldo”, “Neymar” etc.
* *real\_face*: It shows if their real face is used or not (Yes/No)
* *release\_clause\_eur*: Any team can buy the player without negotiating with the club and start directly talk to the player if they pay for that amount of money.
* *player\_tags:* Some tags of the player (it can be empty) (i.e: “Dribbler”,”Tactician”,”Playmaker”)
* *team\_position*: The position that the player plays in his club. (SUB is for substitute players, meaning they are not on the first eleven.)
* *loaned\_from:* Shows the original club of the player for the loaned ones. If the players are not loaned, the line is empty.
* *joined*: The date of the player joins the club.
* *contract\_valid\_until*: The year which the contract of the player ends
* *nation\_position*: The position that the player plays in his national club. (If the player does not play for the national club, the line is empty.)
* *nation\_jersey\_number*: Shows the jersey number of the player on his national club.
* *pace, shooting, passing, dribbling, defending, physic, gk\_diving, gk\_handling, gk\_kicking, gk\_reflexes, gk\_speed, gk\_positioning*: They are all scaled from “0” to “100”. The ones with start with gk, shows the goalkeeper attributes. For the goalkeeper’s other attributes are NA, or vice versa.
* *player\_traits*: The traits of the player has. It can be empty.
* *attacking\_crossing, attacking\_finishing, attacking\_heading\_accuracy, attacking\_short\_passing, attacking\_volleys, skill\_dribbling, skill\_curve, skill\_fk\_accuracy, skill\_long\_passing, skill\_ball\_control, movement\_acceleration, movement\_sprint\_speed, movement\_agility, movement\_reactions, movement\_balance, power\_shot\_power, power\_jumping, power\_stamina, power\_strength, power\_long\_shots, mentality\_aggression, mentality\_interceptions, mentality\_positioning, mentality\_vision, mentality\_composure, defending\_marking, defending\_standing\_tackle, defending\_sliding\_tackle, goalkeeping\_diving, goalkeeping\_handling, goalkeeping\_kicking, goalkeeping\_positioning, goalkeeping\_reflexes*: Again they are all scaled from “0” to “100”. Except from the first ones these columns do not include NA’s.
* *ls, st, rs, lw, lf, cf, rf, rw, lam, cam, ram, lm, lcm, cm, rcm, rm, lwb, ldm, cdm, rdm, rwb, lb, lcb, cb, rcb, rb*: These are the names of the positions and these variables shows how overall of the player changes by variables. Since we are not interested in these variables, we will drop them too.

## Descriptive Statistics

Since the data have 104 variables, showing the descriptive statistics for all variables is a bit unnecessary. Descriptive statistics helps us to understand the behavior of the data. To have an insight of the data, Table 1 is shown for some of the variables.

1. Descriptive Statistical Summary of Numerical Data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Age** | **Overall** | **Value** | **Shooting** |
| **Min.** | 16 | 48 | 0 | 15 |
| **1st Qu.** | 22 | 62 | 325000 | 42 |
| **Median** | 25 | 66 | 700000 | 54 |
| **Mean** | 25.3 | 66.24 | 2484038 | 52.3 |
| **3rd Qu.** | 29 | 71 | 2100000 | 63 |
| **Max.** | 42 | 94 | 105500000 | 93 |
| **NA's** | 0 | 0 | 0 | 2036 |

In this table, Tukey’s five number summary statistics are shown, such as minimum value, 1st quantile, median, 3rd quantile and the maximum vale of the variables. Also, mean of the variables and the number of NA’s in the variables are shown too.

In the target variable, which is Value, mean and median are so away from each other. It is because there are so many players are low in value, that causes non-normality.

In this data, there are 12 categorical variables (such as club, nationality, weak foot, work rate), 3 string variables and 89 numeric variables.

## Exploratory and Confirmatory Data Analysis

In this part of the study, 3 research questions have been set and with the proper graphs and tests, conclusion has made. These research questions help us to understand our dataset. Also, some problems are in the dataset cannot be seen before trying to work on the dataset.

## C.1 Correlation matrix

By correlation matrix, correlation coefficients between variables can be seen. It is useful when seeing how variables affect each other. Also, interpretation of the graph is easy.

Since creating the correlation matrix with 104 variables makes the graph almost invisible, some variables are dropped. Dropped variables are the ones who are highly correlated with each other and can be expected easily. For example, *gk\_handling* and *gk\_reflexes* are highly correlated with each other and can be expected so. Correlation matrix is below.

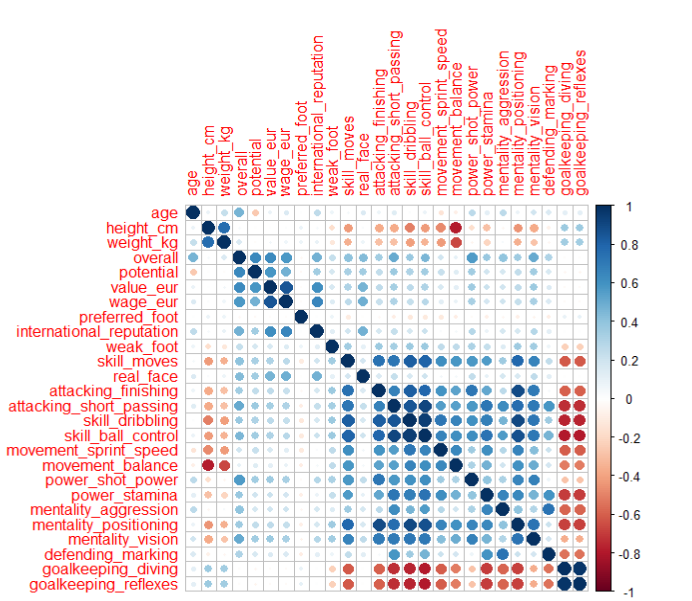


Figure 1 Correlation Matrix

In this figure, there are some valuable information such as

* Balance is negatively correlated with both height and weight.
* Height and weight are slightly affecting goalkeeping skills in a good way.
* Age and overall are positively correlated, but age and potential are negatively correlated. That means that the players of new generation of the football will be more powerful than the players of this generation.
* When weight and height increase, skill moves, finishing, passing, dribbling, control, sprint speed, stamina, and mental attributes decrease.
* Preferred foot of the player is not correlated for any of the variables.
* Value and the wage of the player are highly and positively correlated by the variables overall, potential, international reputation.

## C.2 Distribution of variables

Some of the distributions of the variables are shown below.

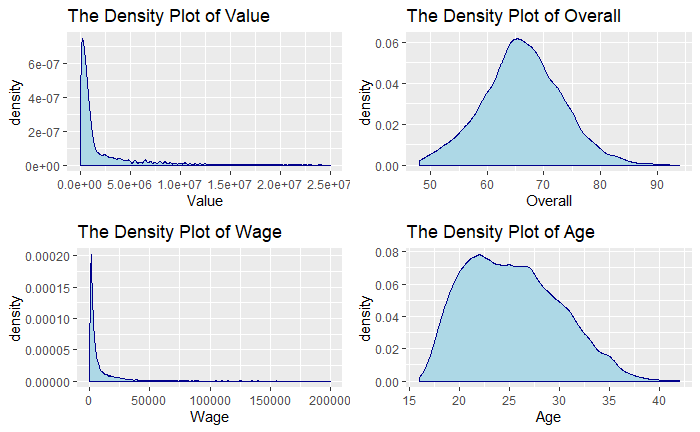


Figure 2 Distribution of Variables

In this figure, we can see that the distributions of value and wage are right-skewed. Also, these both density plots are the x-axis limited ones to see the density. If x-axis is not limited, density becomes almost invisible because there are some world class players which are paid so much better than the others.

On the other hand, distribution of the variable overall looks like symmetric. Also, age is so close to be a symmetric distribution, but it is a right skewed distribution with skewness = 0.38.

The densities of value and wage looks similar. Already have been found that value and wage are highly correlated in *Figure 1*. For the modelling part, dropping the variable wage could be a good idea because they look affiliated.

Also, the distribution of the value is not normal obviously. Before conducting the linear model, necessary transformations need to be done to make the variable value distributed normally.

## C.3 Is there a relationship between the position and the value of the player?

In this research question, the relationship between the position of the player and the value of the player are compared. People usually think that forward players are more valuable than the others because they have to be creative and talented. Also, in this research questions all positions are compared with each other.

However, the team position variable in this dataset originally contains 29 levels. The problem was most of the players can play in different positions, and in this variable all the positions are written, and in FIFA 20, positions of the players are considered in depth.

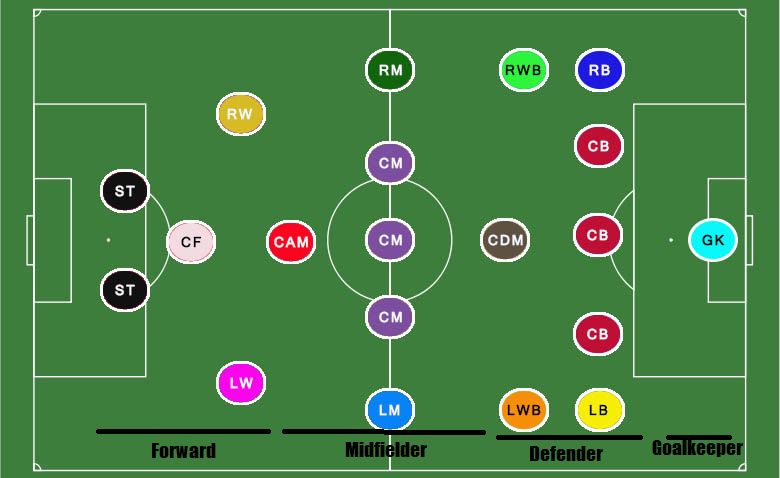


Figure 3 Positions of the Players

To not to work with so many levels and see it easily, a new variable added named team position new. This new variable has only 6 levels, which are goalkeeper, defender, midfielder, forward, sub, res (sub is for substitute players and res is for reserves).

The violin plot is below.

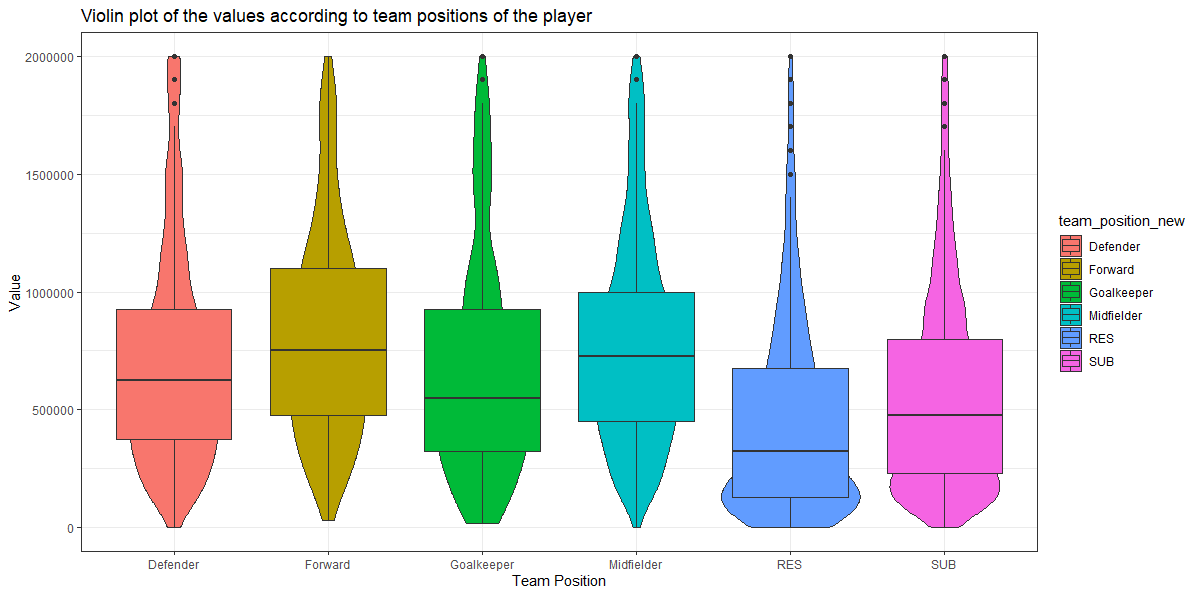


Figure 4 Violin plot of the values according to team positions of the players

Y-axis is limited to see boxes clearly. As can be expected we can see that res < sub < first 11. When comparing first 11 with median values, we can roughly see forward > midfielder > defender > goalkeeper. However, by *Figure 4,* it seems there is not a significant difference between them. To be sure, a test needs to be conducted.

Already known that the variable value is not distributed normally. So, ANOVA test cannot be conducted. However, we can conduct a Kruskal-Wallis H Test. The result is below.

1. Kruskal-Wallis Rank Sum Test

|  |  |  |
| --- | --- | --- |
| **Kruskal-Wallis Chi-Squared** | **df** | **p-value** |
| 2197.2 | 5 | 2.2e-16 |

Since p-value 2.2e-16, which is less than 0.05, it can be concluded that there are significant differences between the team position. To see which one differs from each other, multiple pairwise-comparison between groups need to be done.

1. Paırwıse Comparısons Usıng Wılcoxon Rank Sum Test

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Defender** | **Forward** | **Goalkeeper** | **Midfielder** | **RES** |
| Forward | 2e-16 | - | - | - | - |
| Goalkeeper | 0.79 | 1.7e-15 | - | - | - |
| Midfielder | 2e-16 | 7.1e-05 | 9.3e-10 | - | - |
| RES | 2e-16 | 2e-16 | 2e-16 | 2e-16 | - |
| SUB | 2e-16 | 2e-16 | 2e-16 | 2e-16 | 2e-16 |

By the Table III, all the variables are significantly different from each other except goalkeeper and defender. In other words, there is not a significant difference between goalkeeper and defender.

By the Table III and *Figure 4,* it can be concluded that by the positions there can be a ranking in values such as, forward > midfielder > goalkeeper = defender > sub > res.

## Missingness

Missingness in the data makes our models problematic. In this part of the study, missingness of the data is tried to be handled by doing necessary imputations. Before making the imputations, mechanism of the missingness needs to be understood. To find if there is a pattern or missing values are just random, the aggregation plot of missing is shown below.

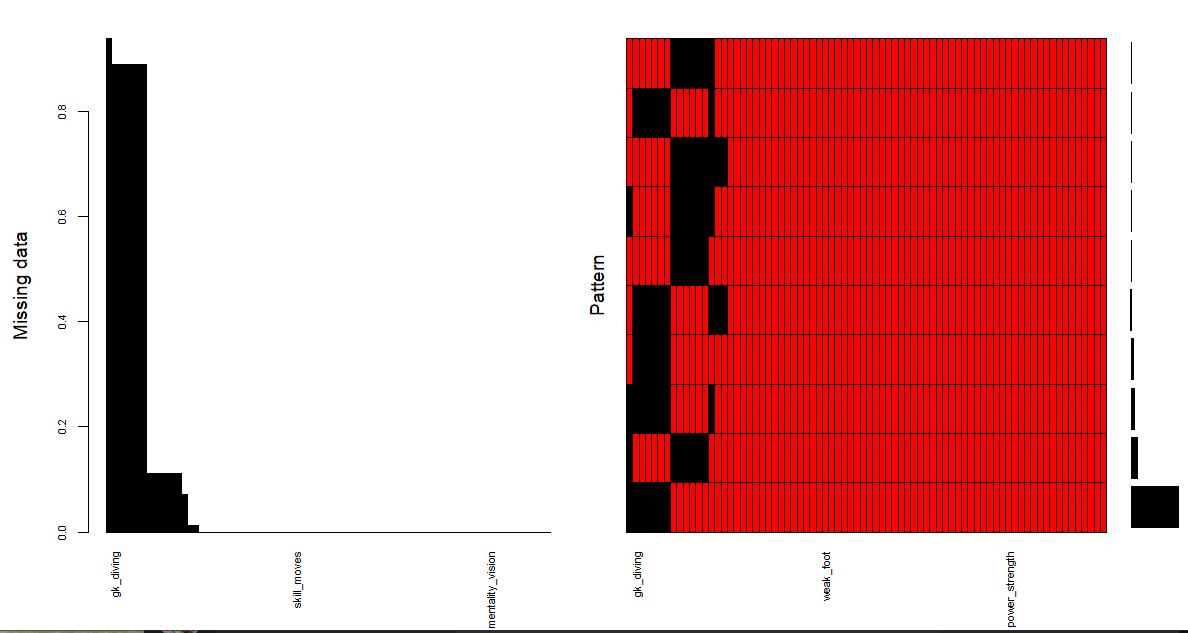


Figure 5 Aggregation Plot of Missing

Probably, the missingness is not random, so each variable needs to be check separately.

* *release clause* (1298 NA): They are the players who do not have the release clause option. The team which wants the player cannot give a certain amount of money to the club and start negotiating with the player. To negotiate with the player, the team must negotiate with the team of the players first. For this variable, regression imputation is used to remove NA’s. The model which is used is *release clause* ~ *value*.
* *team jersey number* (240 NA): After checking the summary for these players, it is seen that their *contract valid until* values are also NA’s. Also, when looking at their clubs, there are no club names, just nations. So, it can be said that they are the players who are not in a club. Missing points of *team jersey number* are removed by replacing them with the value “0”. Also, for *contract valid until* variable, missing points are removed by replacing them with “2019”. The reason behind choosing 2019 is the necessity a year which is in past to emphasize they do not have a contract in today.
* *nation jersey number* (17152 NA): These are the players who do not play for their nations now. Missing points are removed by replacing them with “0”.
* *pace, shooting, passing, dribbling, defending, physic* (2036 NA): In this dataset, there are 2036 goalkeepers, and these missing points for those goalkeepers. Replaced by “0”.
* *gk\_diving, gk\_handling, gk\_kicking, gk\_reflexes, gk\_speed, gk\_positioning* (16242 NA): These are the goalkeeper attributes, and these are the missing points for the players who are not goalkeepers. Replaced by “0”.

Until now, descriptive statistics are checked, research questions are answered to have an insight about the data. Missing points are replaced by the suitable values.

After these steps, min-max normalization is used for all numerical features. Since all categorical variables in this data are nominal, dummy variables are created. However, since there are 698 clubs, 162 nationalities, creating dummy variables with these two variables will make the data too crowded. So, dummy variables are not created for these two variables.

Since *wage* looks so much affiliated to *value,* *wage* will not be used in the modelling part.

Also, to check the RMSE of the models, the data are divided into two parts as training and test dataset, and they contain 80% and 20% of the observations, respectively.

# MODELLING

Since missing imputation is done and necessary variables are chosen, the data are ready for the modelling. The data are split into two parts as training and test. Training dataset contains 80% of the observations, which equals to 14625 observations. Likewise, test dataset contains the other 20% of the observations, which equals to 3653 observations. Model will be conducted by training dataset and test dataset will be used to check errors of the model.

## Multiple Linear Regression

Multiple linear regression is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. In the first part of the study, it is seen that *value* is not distributed normally. orderNorm transformation, which is in bestNormalize package is used to make the distribution of *value* normal.

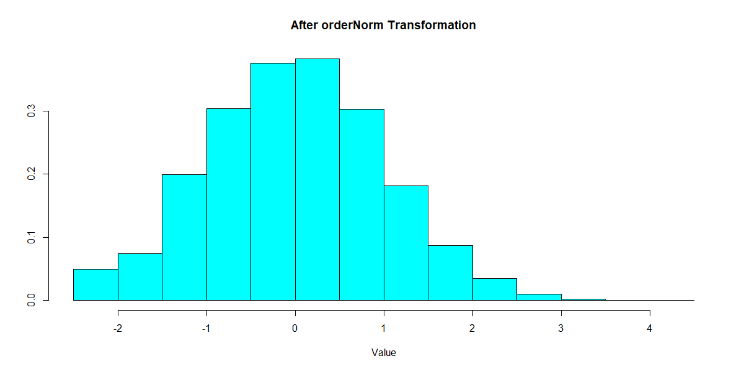


Figure 6 Histogram of the Value after orderNorm transformation

After encoding, the number of variables increased to 83. A linear model is set with these variables, then according to their VIF values, eliminations made. After the elimination, number of variables dropped to 59. By conducting the linear model, the summary below can be obtained.

1. The Summary Of Multıple Lınear Regressıon

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Estimate** | **Std. Error** | | **z value** | **Pr(>|z|)** |
| **Age** | 0.245 | 0.33 | | 7.420 | 1.2e-13 |
| **Long Pass** | 0.312 | 0.038 | | 8.237 | 2e-16 |
| **Goalkeeper** | 0.144 | 0.024 | | 6.018 | 1.8e-09 |
| **Positioning** | -0.24 | 0.044 | | -5.461 | 4.8e-08 |
| Adjusted R-Squared: 0.8204 | | | p-value: < 2.2e-16 | | | |

Since showing all 59 of the variables is going to be a waste of space, just some of the variables are shown above. Since the p-value of the model is less than 0.05 (2.2e-16), the model is significant. Also, adjusted R-Squared is equal to 0.82, it means that 82% of the change in the variable *value* can be explained by regressors.

## Artificial Neural Networks

In this method, again the variable value is tried to be predicted by the same regressors which are used in part A, which is multiple linear regression. The only difference is, since there is not an assumption about normality, the original scaled *value* is used except from the transformed one.

The model with one layer and one neuron is conducted. The plot of the model is given below.

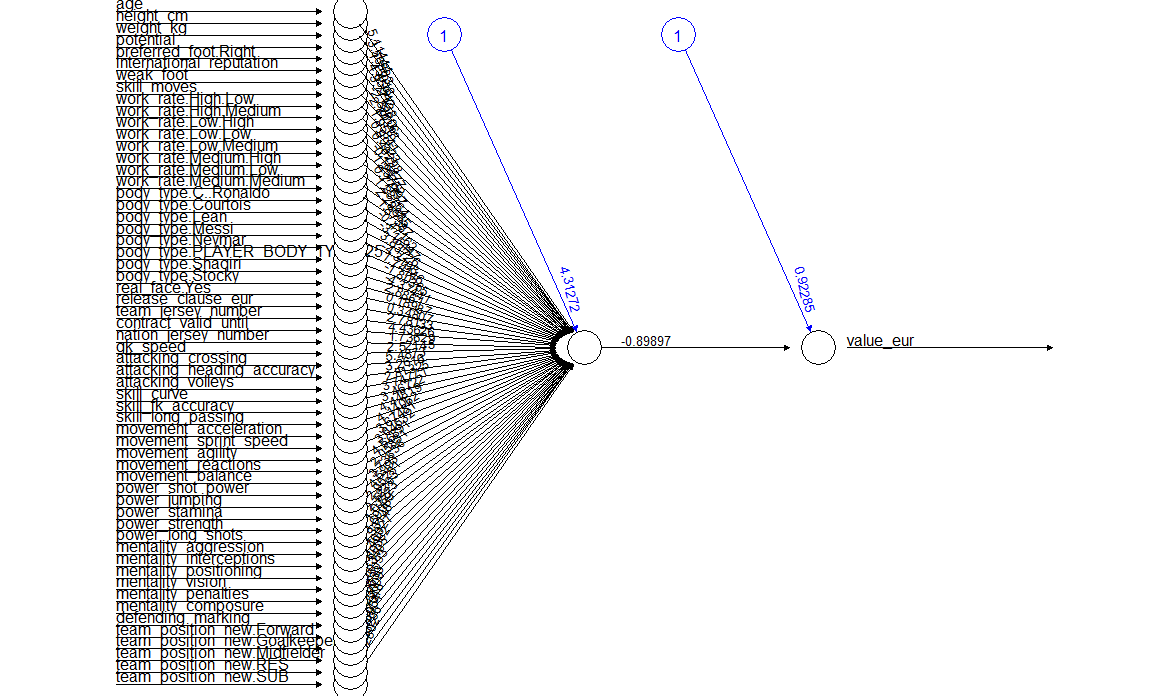


Figure 7 Plot of Neural Network with One Hidden Layer and One Neuron

Since the number of variables is equal to 59, seeing all variables on the figure is a bit hard.

## Support Vector Machine

In [machine learning](https://en.wikipedia.org/wiki/Machine_learning), support vector machines are [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) models with associated learning [algorithms](https://en.wikipedia.org/wiki/Algorithm) that analyze data used for [classification](https://en.wikipedia.org/wiki/Statistical_classification) and [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis). In our situation regression type is used. The comparison between default and tuned SVM models are given below.

1. The Comparıson Between SVM Models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Cost** | **Gamma** | **Epsilon** | **MSE** |
| **Default** | 1 | 0.01724 | 0.1 | 0.00139 |
| **Tuned** | 1 | 0.001 | 0.05 | 0.00109 |

By Table V., tuned model worked well than the default model.

## Random Forests

Random forest is another Supervised Learning algorithm which uses ensemble learning method for classification and regression. After tuning, mtry value is chosen as 10. After conducting the model, actual versus predicted graph will be like below.

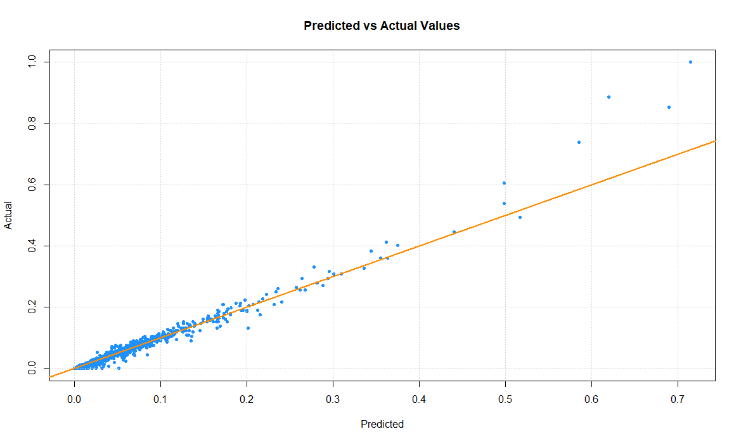


Figure 8 Scatter plot of the actual versus predicted values by Random Forests

By *Figure 8*, can be seen that the model is useful when predicting values lower than 0.5, which implies the players whose values are lower than the mean of the values.

## XGBoost

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. Can be used to solve regression, classification, ranking, and user-defined prediction problem. To check the importance of variables, xgb.plot.importance function is used. Just the most important variables are shown to see the variables clearly.

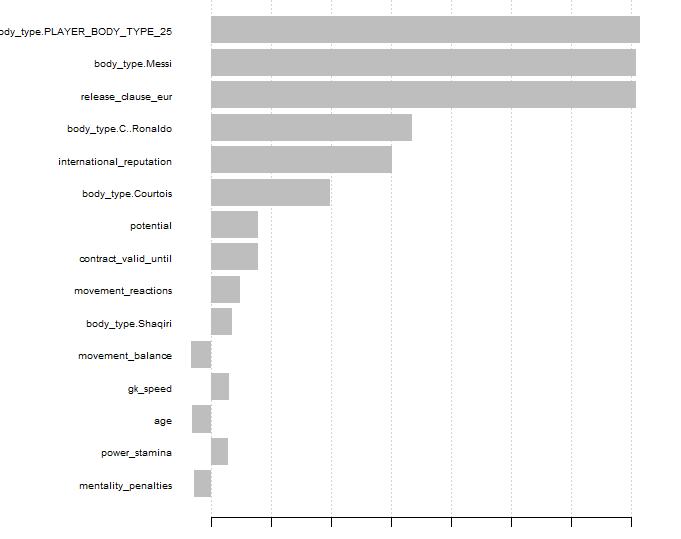


Figure 9 XGBoost Importance Plot

By *Figure 9*, body type is the one of the most important variables for the model. Also, *international reputation, potential, contract valid until, reactions, balance, gk\_speed, age, stamina, penalties* are the important variables for the model.

When default XGBoost gives RMSE as 0.026, tuned model gives 0.0077. Final values used for the model were nrounds = 1000, lambda = 1, alpha = 0 and eta = 0

## Performance Comparison Between Train and Test Data

In the modelling part, 5 different models are conducted, such as Multiple Linear Regression, ANN, SVM, Random Forests and XGBoost. To compare these models with each other, model performance scores are compared.

To compare models, RMSE and MAE values are used. RMSE means Root-mean-square error, and MAE means mean absolute error. The model in the lowest RMSE and MAE value are searched.

1. The Comparıson Between Models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Train Data** | | **Test Data** | |
|  | **RMSE** | **MAE** | **RMSE** | **MAE** |
| **Linear Model** | 0.02694 | 0.01505 | 0.02654 | 0.01466 |
| **ANN** | 0.05357 | 0.02742 | 0.05036 | 0.02569 |
| **SVM** | 0.03298 | 0.02027 | 0.03306 | 0.01982 |
| **Random Forest** | 0.00447 | 0.00133 | 0.01196 | 0.00305 |
| **XGBoost** | 0.00088 | 0.00057 | 0.00776 | 0.00264 |

Since we are interested in how useful our predictions, we will look for the model with the lowest RMSE and MAE values for the test data. By the table it can be said that, XGBoost is the best model among these models. Then it is followed by Random Forest, Linear Model, SVM, ANN, respectively.

# CONCLUSION

In this paper, FIFA 20 Player Dataset is analyzed. Research questions are asked to understand the behavior of data. Missing points are imputed. Data is scaled. After preparing the data, models are conducted. With the lowest error points, XGBoost is the best model among these models.

Most of the variables are very valuable for predicting the *value* of the players. However, some variables explain the response value better than the others. The variables can be count as *body type, release clause, international reputation, potential, contract valid until, reactions, balance, speed, age, stamina, penalties,* respectively.

##### References

1. D. Specht, “A general regression neural network,” *IEEE Transactions on Neural Networks*, vol. 2, no. 6, pp. 568–576, 1991.
2. R. Lopes, “FIFA Ultimate Team Positions and Formations,” 31-Jul-2016. [Online]. Available: https://www.fifauteam.com/fifa-ultimate-team-positions-and-tactics/. [Accessed: 20-Jun-2020].
3. W. Kenton, “How Multiple Linear Regression Works,” *Investopedia*, 29-Jan-2020. [Online]. Available: https://www.investopedia.com/terms/m/mlr.asp. [Accessed: 20-Jun-2020].
4. V. Morde, “XGBoost Algorithm,” Medium, 08-Apr-2019. [Online]. Available: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d. [Accessed: 20-Jun-2020].
5. A. Chakure, “Random Forest and its Implementation,” Medium, 07-Apr-2020. [Online]. Available: https://towardsdatascience.com/random-forest-and-its-implementation-71824ced454f. [Accessed: 20-Jun-2020].