**MIS 315 Final Project - Report**

# **FORECASTING FOOTBALL PLAYERS MARKET VALUE WITH MACHINE LEARNING MODELS**

Ferhat Eren Dalçık, 11752009  
eren.dalcik@bilgiedu.net

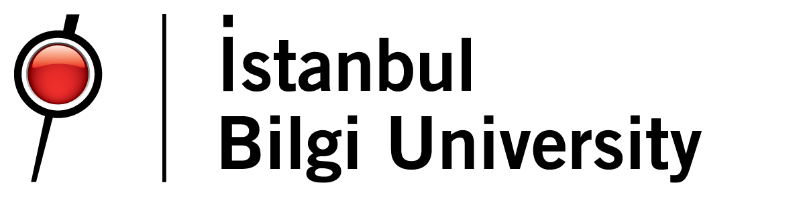
Supervised by Assistant Professor Barış Soybilgen

January 20, 2021

Istanbul Bilgi University

Department of Business Administration

Management Information Systems



# Abstract

With the widespread use of technology and the internet in daily life, people all over the world have started to create unstoppable data stacks and data traffic. In business life, after the effect of data on decision-making mechanisms and forecasting is seen, the need for decision-making tools and data scientists that can play an important role has increased. Likewise, this science, which can be applied to business life, can be applied to many areas of life. Sport is one of them. This study aims to predict football player value using FIFA 18 Ultimate Player. After reviewing the dataset and making the necessary arrangements, some regression models were built using R (a language and environment for statistical computing and graphics). Comparing the results of these models, it was decided which one gave the best result. Among Random Forest Model, Multiple Regression, Lasso and Ridge Regression, and Regression Tree; Random Forest Model gave us the best result with the lowest MSE. These models have shown us what abilities can help us to do the forecast. Machine learning algorithms have a big potential to make predictions such as in this work.

**Keywords:** Machine Learning, Regression Models, Forecasting, Football Player Value

# Table of Contents

[Abstract 2](#_Toc61445083)

[Table of Contents 3](#_Toc61445084)

[List of Figures 4](#_Toc61445085)

[1. Introduction 5](#_Toc61445086)

[1.1 Project Outline 5](#_Toc61445087)

[2. Data Exploration and Analysis 6](#_Toc61445088)

[3. Regression Models 7](#_Toc61445089)

[3.1 Multiple Regression Model 7](#_Toc61445090)

[3.2 Ridge and Lasso Regression Models 8](#_Toc61445091)

[3.3 Regression Tree Model 10](#_Toc61445092)

[3.4 Random Forest Model 11](#_Toc61445093)

[4. Conclusion 11](#_Toc61445094)

[References 13](#_Toc61445095)

# List of Figures

Figure 1 Variable Names of FIFA 18 Ultimate Team Dataset 6

Figure 2 Pairs Plot with 4 Predictors of FIFA 18 Ultimate Team Dataset 7

Figure 3 Correlation Plot with 4 Predictors of FIFA 18 Ultimate Team Dataset 7

Figure 4 Multiple Regression Model Summary 8

Figure 5 Log(Lambda) Graph of Ridge Regression 9

Figure 6 Log(Lambda) Graph of Lasso Regression 9

Figure 7 Coefficients and Lamda 10

Figure 8 Lasso Model Summary 10

Figure 9 Cross Validation Result 11

Figure 10 Regression Tree 12

Figure 11 Increase in MSE for Predictors 13

Figure 12 All RMSE Values of Models 14

# 1. Introduction

Football is a game where two teams of eleven players play in the same field using all parts of their body except their hands and arm, trying to score a goal with the ball in the game. The team with more goals over the opponent team wins the game. There are also other rules and regulations to dictate the game preventing the game to be move out of the aspect of the football. Club managers, owners, and coaches trying to come up with better strategies to win the game with the resources and players they have. As I stated in the abstract, data scientists have a huge impact on developing models and analyzing the data for the organizations they take part in using machine learning models. Also, football clubs already using other types of equipment like heatmaps, information system solutions to gather up deep insights about their players and how they act during a game. Choosing the best player with optimal values is always something football clubs try to achieve since getting a contract with good players is expensive or giving higher salaries to a player that not having the attributes that can fit their organization is too risky. With having the correct tools and data, it is possible to make these predictions. Predictions that all clubs can benefit from statistically.

This study is an example of how a person can analyze a player's value or which skills can affect a player's value with the dataset having the value of a player and other attributes of a football player. FIFA 18 Ultimate Team Dataset that has categorical and numerical variables, used in this project to predict football player values. Random Forest Model, Multiple Regression, Lasso and Ridge Regression, and Regression Tree models are fitted in this project to do our prediction. Prediction results will be evaluated with their RMSE where it stands for Root of the Mean Square Errors. MSE is the most commonly used measure type of the regressions, it is computed by training data that was used to fit the model. [1] In other words, it tells us how concentrated the data is around the line of best fit. It can be computed with a certain method and a model with the best RMSE is chosen. Regression models are built using R statistical language and RStudio development environment.

# 1.1 Project Outline

The project is organized as follows: In part 2 analysis of our dataset is given. Predictors and dependent variables are evaluated. In part 3 regression models written in the introduction part are built and analyzed. In part 4 there is a conclusion, discussion, future work, and other implementation examples are provided.

# 2. Data Exploration and Analysis

In FIFA 18 Ultimate Team Dataset we have 1599 observations and 37 predictors. After dropping goalkeepers from our dataset we have 1500 observations and 37 predictors. So dataset had 99 players as goalkeepers. All models are fitted with non-goalkeeper players. To have a better model with better results non-numerical and non-binomial variables are removed from the dataset. So the final dataset has “eur\_value” which is the value of a player in Euro as its first column (predicted variable) and other skills of a player in other columns

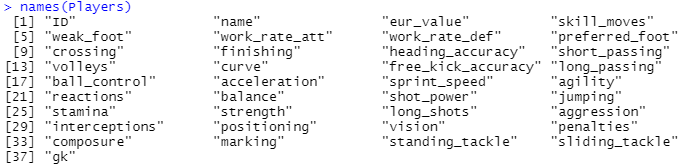


Figure 1 Variable Names of FIFA 18 Ultimate Team Dataset

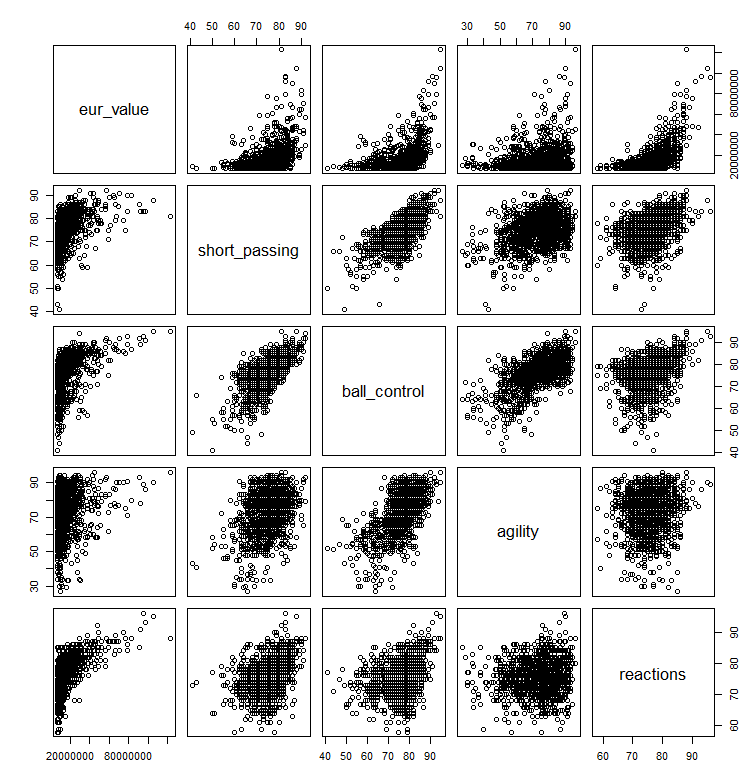
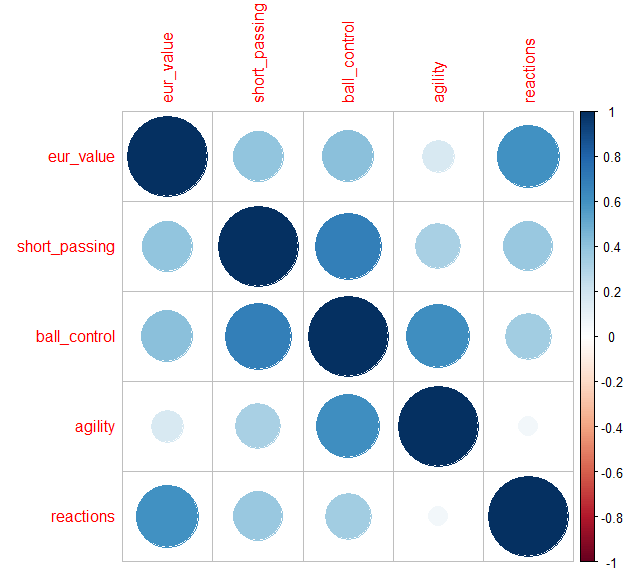
There are several plots shown in this part where we can see the positive or negative relationships between our predictors and predicted variables without doing any regression models. In Figure 2 it can be interpreted from the Correlation Plot of Euro Value of a Player (eur\_value) has a positive relationship between reactions and agility. Since it is too hard to interpret all predictors together I took some parts of it to analyze. Other predictors also had no significant problem regardings our predicted value. A similar interpretation can be done with the Pairs Plot too.

Figure 2 Pairs Plot with 4 Predictors of FIFA 18 Ultimate Team Dataset

Figure 3 Correlation Plot with 4 Predictors of FIFA 18 Ultimate Team Dataset

# 3. Regression Models

As stated in the previous part dataset has 1500 variables and 37 predictors without any players that have goalkeeper ability. After subsetting categorical variables from our dataset to have better fits in regression models, the new dataset now has 31 predictors and 1500 variables. Dataset is splitted into two parts as a training and test set. Train dataset has 1000 variables and the test set has 500 variables.

# 3.1 Multiple Regression Model

Before starting to fit regression models it is important to explain what regression is. In summary, regression refers to; determining the relationship between a dependent variable and one or several variables [1]. In the multiple regression model, we fit our model with all 30 predictors. This regression is fitted R and gave this result represented in Figure 4.

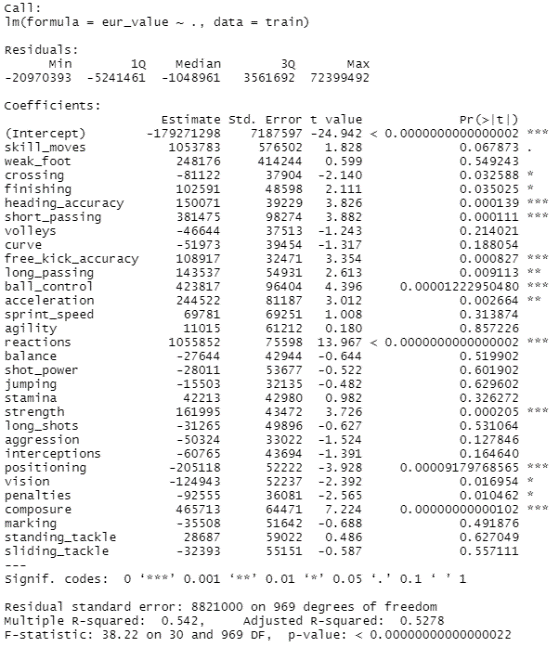
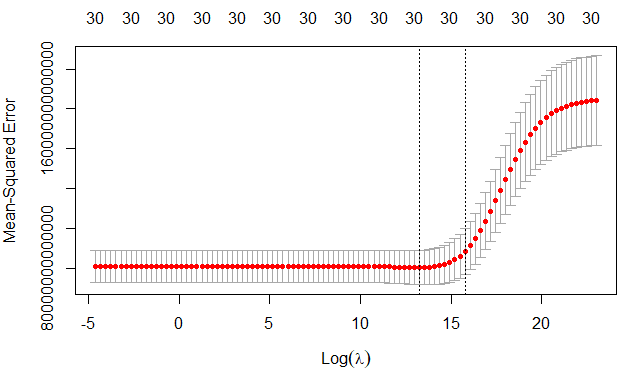
In Figure 4 we can see which variables have a negative and positive relationship between our dependent variable (eur\_value). Residuals in summary are the distance of the data to the fitted line. Ideally, they should be symmetrically distributed, for example, minimum (min) and maximum (max) values should be similar to each other. Also, the model has some significant variables. heading\_accuracy has a p-value lower than 0.05 and a positive effect on the value of a player and it is significant. We can make the same analysis for short\_passing, finishing, long\_passing, ball\_control, acceleration, reactions, strength, composure, and free\_kick\_accuracy. The model also has insignificant variables that have a p-value is higher than 0.05. Those variables are; skill\_moves, volleys, curve, sprint\_speed, agility, balance, shot\_power, jumping, stamina, aggression, long\_shots, interceptions, marking, standing\_tackle, sliding\_tackle. Multiple R-squared is 0.542 which is somewhat close to 1. The adjusted R-squared value is 0.5278 variability of the “eur\_value” can be explained 52.78% with our predictors. Degrees of freedom are 30 and 969 with 0.05 probability level critical F-value is 1.47096083 which is lower than 38.22. Finally calculated RMSE value for this model is 6830993.

Figure 4 Multiple Regression Model Summary

# 3.2 Ridge and Lasso Regression Models

After applying the Multiple Regression model, we have Lasso and Ridge Regression models. Lasso Regression can exclude useless variables from equations it is a little better than Ridge Regression at reducing variance in models that contains a lot of useless variables but when the majority of variables are useful, Ridge Regression appears to do a bit better. With regardings to this, both of the models are applied and their RMSE values are compared. In contrast we in both models we try to find the best fitting line with Ridge and Lasso's Regressions having their Penalty values in their equation.

We found the best lambda values for Ridge and Lambda Regression applying cross-validation. The best lambda value for Ridge is 13.25731 and for Lasso, it is 10.4663. We can see the graphs of the best lambda values in Figures 5 and 6.



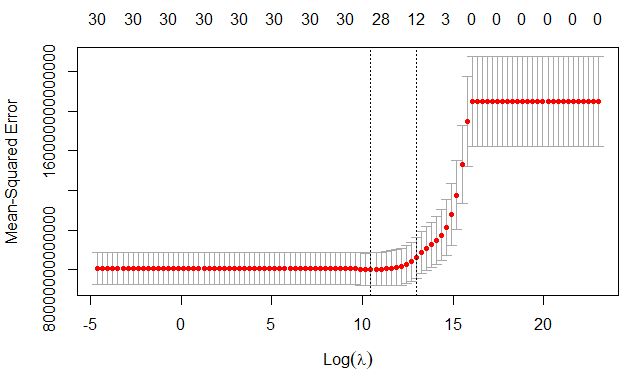


Figure 5 Log(Lambda) Graph of Ridge Regression

Figure 6 Log(Lambda) Graph of Lasso Regression

In Figure 7 we can see the relationship between coefficients and lambda value of Lasso. As lambda value increases coefficient values become zero. Also in Figure 8, we can see which predictors affected the linear model (positive or negative). Since Ridge Regression can make our predictors coefficients so close to zero with cross-validation Lasso Regression can make these insignificant variables zero, meaning it can eliminate them from our model. We should keep in mind that these are the same insignificant predictors found in Multiple Regression (Figure 4).

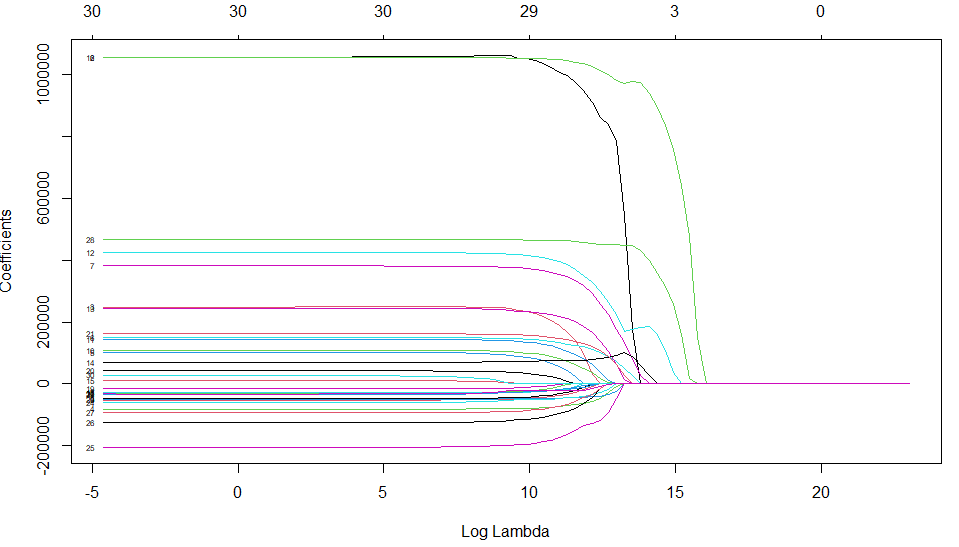
Finally calculated RMSE values for Ridge Regression are 6542752 and for Lasso 6716368.

Figure 7 Coefficients and Lamda

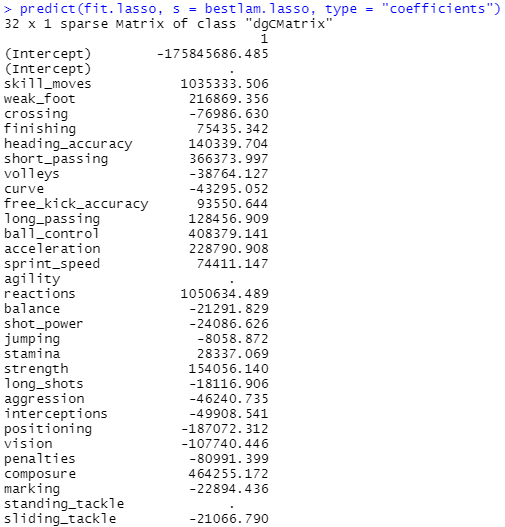


Figure 8 Lasso Model Summary

# 3.3 Regression Tree Model

For the next model, we applied a Regression Tree with all predictors. The cross-validation number of terminal nodes is found as 13 but the optimal one is 12 and it is applied to the regression tree model. In Figure 9 we can see the number of terminal nodes and in Figure 10 Regression Tree is represented. According to Regression Tree, the most important predictor in our set is reactions. After reactions, we can see the ball\_controll is the important predictor in the second level.

RMSE value for this model is 5198323.

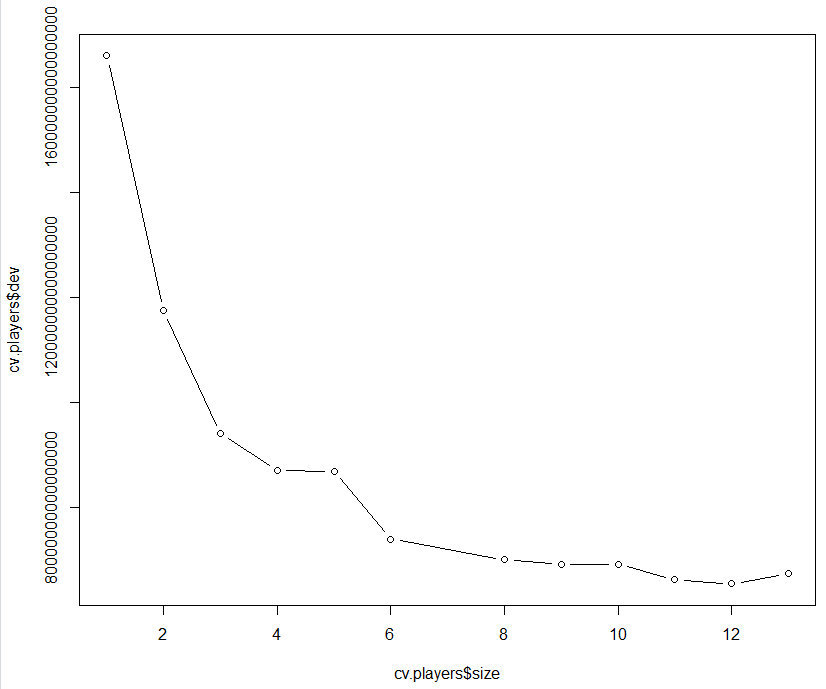


Figure 9 Cross Validation Result

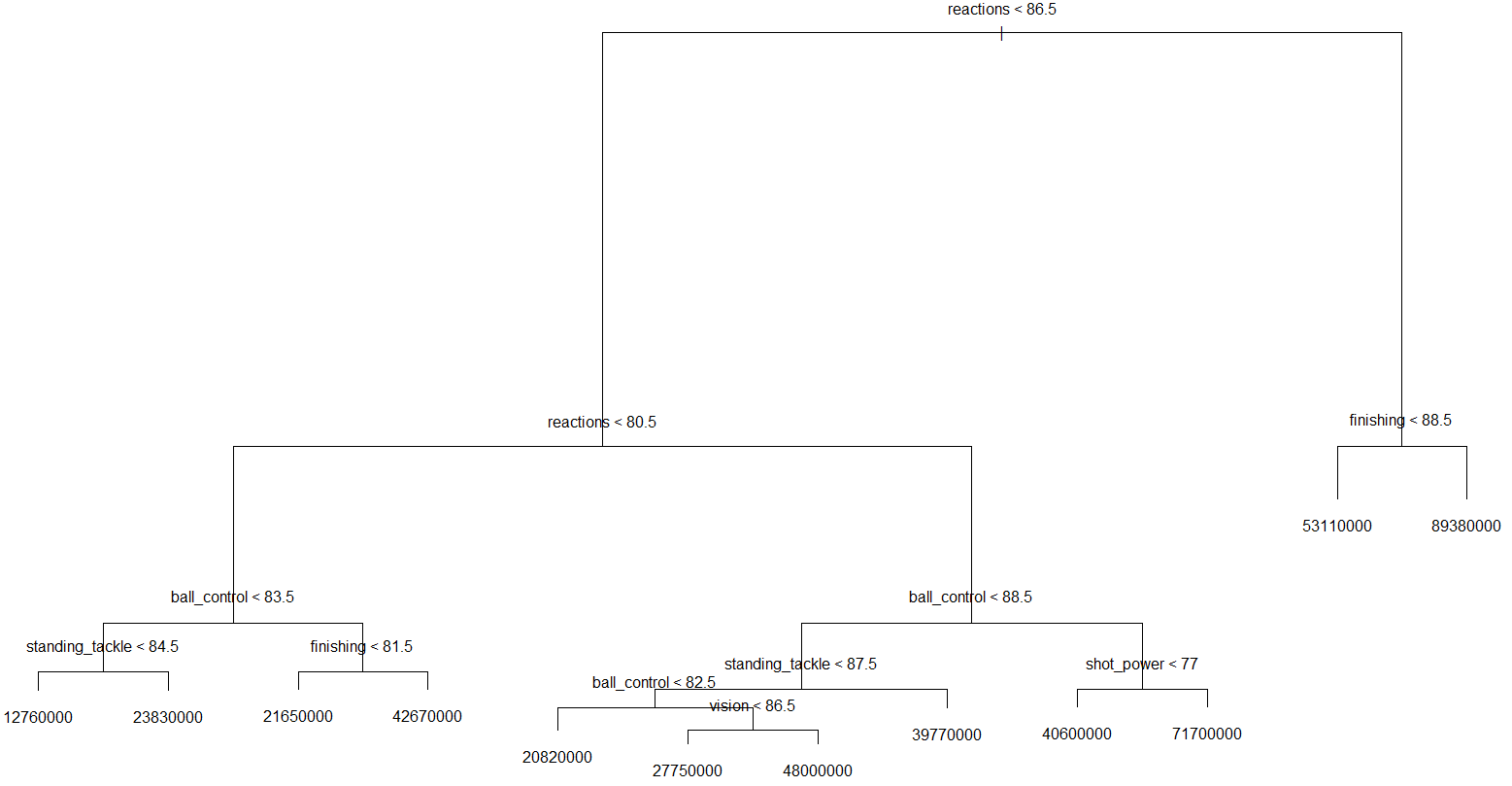
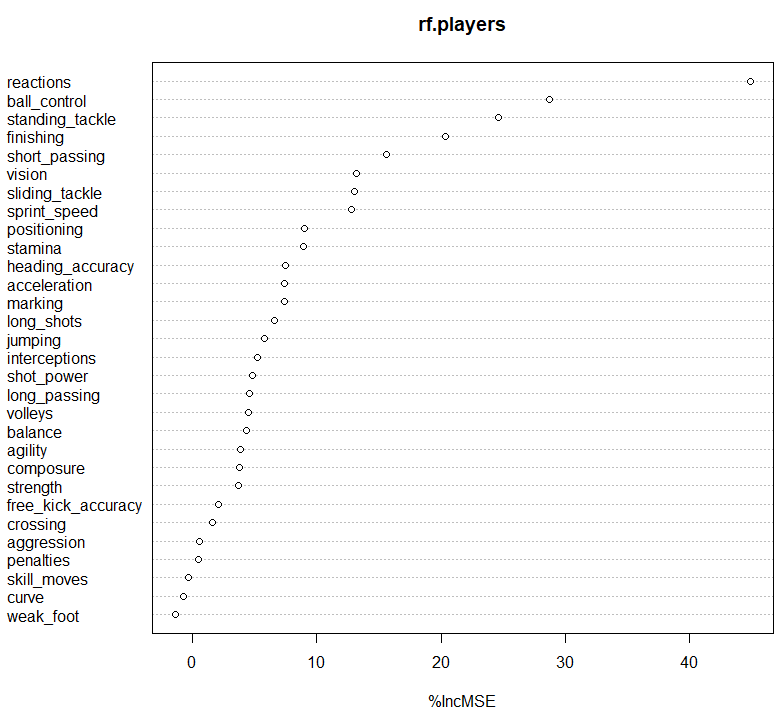


Figure 10 Regression Tree

# 3.4 Random Forest Model

In our last regression model, Random Forest is applied. Random Forest is fitted with the number of variables from 1 to 30 which the model tries all predictors to make the regression with 5 fold cross-validation to find the lowest MSE that is 18. So, using 18 variables in each split gives us the best result. In Figure 11 we can see predictors are sorted according to their increase in MSE (%IncMSE). This output shows that what happens to MSE if we remove them from our model. So if a variable has a higher (%IncMSE) it has more importance to our model. The top important variables in our model are reactions, ball\_control, stading\_tackle, finishing. Removing reactions from the model will have the most increase in MSE. We can think the other way around. Removing the bottom variables like weak\_foot, curve, skill\_moves from our model does not affect MSE.

The calculated RMSE value for Random Forest Model is 3979884.



# 4. Conclusion

Figure 11 Increase in MSE for Predictors

The aim of oıf this project is to classify and predict football player value with different machine learning models. Random Forest Model, Multiple Regression, Lasso and Ridge Regression, and Regression Tree models are used to analyze FIFA 18 Ultimate Team Dataset. Results are interpreted with their RMSE. Our model with the lowest RMSE is Random Forest Model which means prediction accuracy is better than the others.

Random Forest Models are models that easy to evaluate can give important insights into our dataset.

In Figure 12 we can see all RMSE of models together.

These regression models can also be applied to different sports such as basketball. Similar forecasting can be done with basketball players dataset containing player skills. Likewise, these machine learning models can be applied to a different type of studies that requires predictions. Machine learning models are extremely important for making predictions and their reputation will increase significantly in the upcoming years.

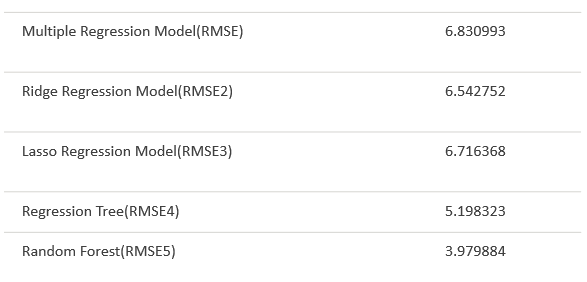


Figure 12 All RMSE Values of Models

# References

|  |  |
| --- | --- |
| [1] | Gareth James, Daniela Witten, Robert Tibshirani, Trevor Hastie, An Introduction to Statistical Learning with Applications in R, 2013. |