Istanbul Bilgi University

FORESEEING EURO VALUES OF SOCCER PLAYERS BY REGRESSION MODELS

TURHAN KILIÇCIOĞLU

MIS 315

TABLE OF CONTENTS

Table of Contents

1. INTRODUCTION
2. DATA AND ANALYSIS
   1. Multiple Regression Model
   2. Ridge & Lasso Regression Model
   3. Multiple Regression Model vs. Lasso Regression Model
   4. Regression Tree
   5. Random Forest
3. CONCLUSION

ABSTRACT

This project paper aims to conclude the prediction of the “Euro Value” of soccer players using various regression models. The full dataset was deducted to a smaller quantity of dataset and therefore, we used 35 predictors and 1500 observations (which 3 of them are categorical and the rest are numerical). The project goal is to predict which regression model is the best option to use and compose the dataset according to it. After implementing the regression models and predicting each value of them after the analysis, we foresee the possible “Euro Value” of each player varied to a different set of variables. All in all, in order to foresee the “Euro Values” of each soccer player, we used “Multiple Regression”, “Ridge and Lasso Regression”, “Regression Tree” and “ Random Forest” models for our dataset. The conclusion has been formed by comparing the Mean Squared Error (MSE) values of each model. The “Random Forest” model gives the optimal result in the end with the least MSE value.

1. INTRODUCTION

FIFA (International Federation of Association Football) is using player attributes in its videogames to forecast the player values in virtual reality. Although these values are fictional, it’s measured by how players perform in the field in real life. In order to foresee the optimal values of players, we use a set of regression models to exclude undesirable variables and variables with minimal effect.

A more compact FIFA 18 Ultimate Team dataset is accounted for with different variables to see how those variables behave in different regression models. The dataset consists of a few categorical and many numerical variables.

In the beginning, we have a look at the summary of the dataset which shows us the general status of the categorical and numerical variables, coefficients and how those variables intercept with the first variable which is “Euro Value”. We deduct the first two variables which are “ID” and “name” which are unwanted in our dataset as they do not affect the prediction of our models. Also, the last variable which is called “gk”(goalkeeper) attribute and goalkeepers are removed due to the instruction given by our assistant professor. The interception will help the models to delete unnecessary variables later on.

During the analysis of the dataset, we begin to see the deduction of the quantity of MSE value which tells us we are nearing the final objective of this project. After the Multiple Regression model, there is a huge deduction on the quantity of the MSE value and there is a slight increase after the Ridge Regression model but after the Lasso Regression model, the deduction is ongoing until the final model which is called Random Forest. In order to make this analysis healthy, we use cross-validation and prediction function which show us which variables are unwanted or that have minimal effect to foresee the “Euro Value” of a player.

The first model that we use is called the Multiple Regression Model. We apply this model to all variables and observations and we then apply Ridge and Lasso Regression models to eliminate the unwanted variables and to see the difference between the first model (Multiple Regression). After the elimination of the unwanted variables, the MSE value has made a huge deduction for the dataset. After the first model, we continue by applying the Regression Tree model. When we use the cross-validation, the tree happens to be trimmed to a shorter one. In the final step, the Random Forest is applied and 10 most important variables are found. Upon implementing the cross-validation to the Random Forest, we achieve the least MSE value of a variable and mtry (number of variables available for splitting at each tree node) number. The models seem to work as there are differences between each model and by how their prediction methods work.

1. DATA AND ANALYSIS

In this dataset, we have 35 variables and 1500 observations. 3 of them are categorical and the rest are numerical variables.

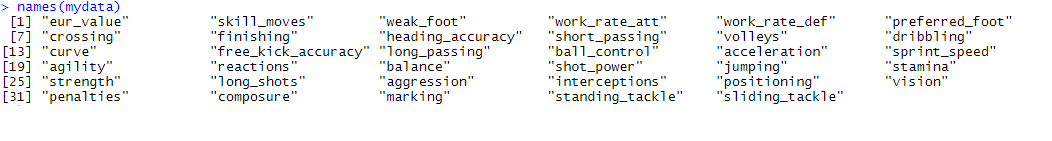


Figure 1: Final Variables of Fifa 18 Ultimate Team Dataset

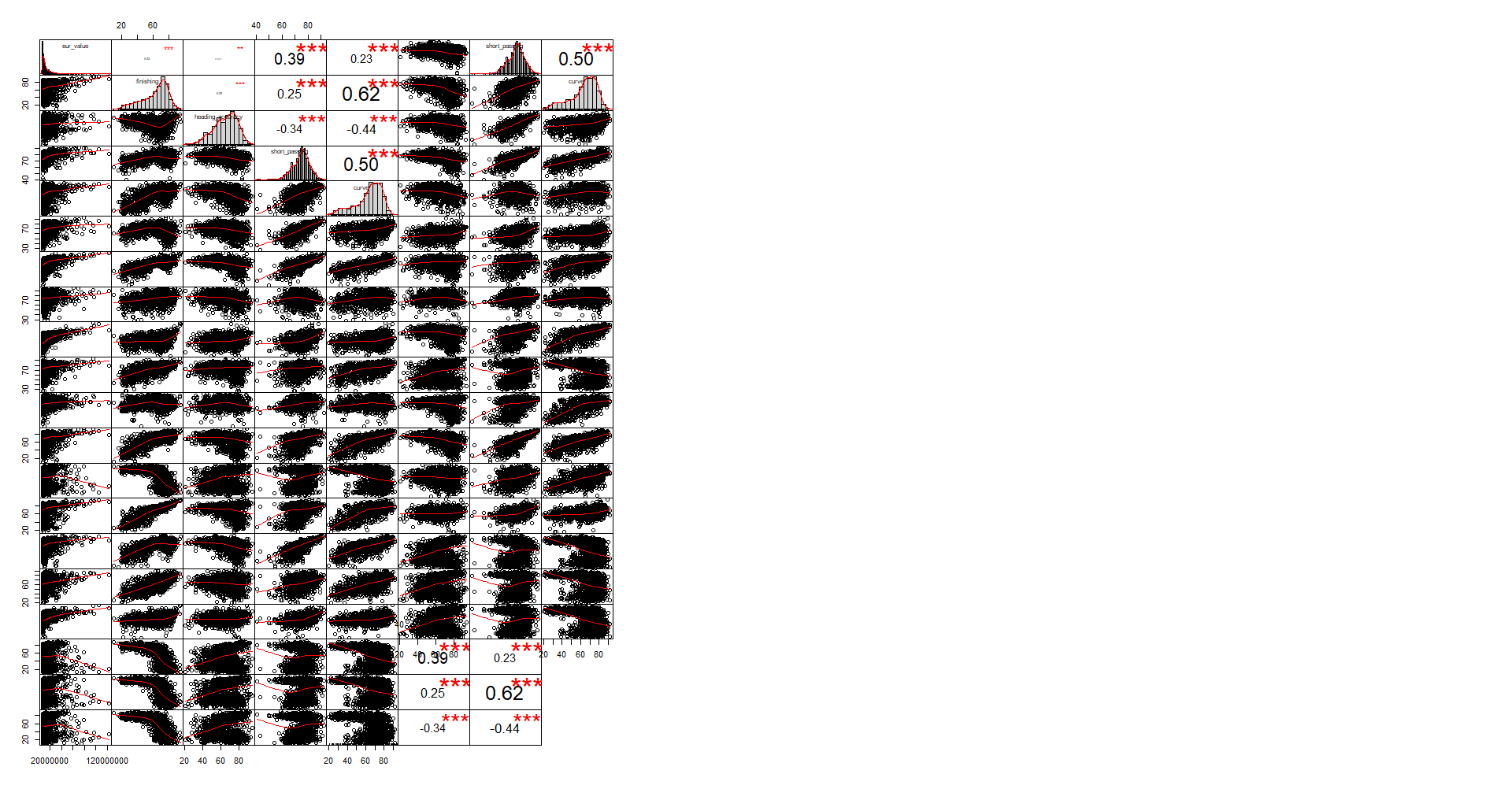


Figure 2: Correlation Chart for Euro Value and 3 different variables

This chart shows us positive correlations between the 1st (“Euro Value”) variable and rest of the variables.

Analyzing the “Euro Value” by taking “finishing” into account, we see the correlation as 0.23 which means it's positive and somewhat bigger than zero, plus the 3 stars, so it can be said that “finishing” changes the value of players and it's an important variable. Most of the players have 60+ finishing skill and as the finishing skills of players go up, also the value of those players' go up. The red line is also not linear.

Analyzing the “Euro Value” by taking “heading accuracy” into account, we see the correlation as 0.077 with 2 stars, which means that it's not an important variable as much as “finishing” or “penalties” variables that affect the value of a player. Most of the players have 65+ heading accuracy skills. The red line is linear.

Analyzing the “Euro Value” by taking “penalties” into account, we see the correlation as 0.21 with 3 stars, which means that an important variable that affects the value of a player. Most of the players have 60+ penalties skills. The red line is not linear.

However, “finishing” and “heading accuracy” variables have negative correlation between them and so does the “heading accuracy” with “penalties” variables. Although “finishing” and “penalties” variables have positive correlation between them and all correlations between these 3 variables have importance for the dataset, the negative correlations show us that if a variable increases, the other one decreases.

We also realize that some variables have negative correlation relationships with eachother whereas some have positive but we can also take into account that all variables have high importance as they are bigger than zero and they have more than two stars.

After checking for the multicollinearity problem, we see there is no issue in the dataset. The next part is to split the dataset into two parts as training and test. We take the first 1000 observations as training and the rest as test set.

2.1 Multiple Regression Model

At this stage, we fit the multiple regression model with all the variables in Figure 1. By taking a look at the Figure 3, we realize that some of the independent variables have positive or negative correlations between “Euro Value” (the dependent variable). For example, the variance of «aggression» has no effect in predicting the “Euro Value” of a player whereas «free kick accuracy» has a medium effect and «reactions» has a large effect. Adjusted R-squared is 0.5275 which equals 52.8% so we can define this result as variables in this model is useful as it shows the variability of the “Euro Value” in this model with the variables and it also shows us how the independent variables affect the dependent variables. The p-value is so close to 0 so it means our model is meaningful.

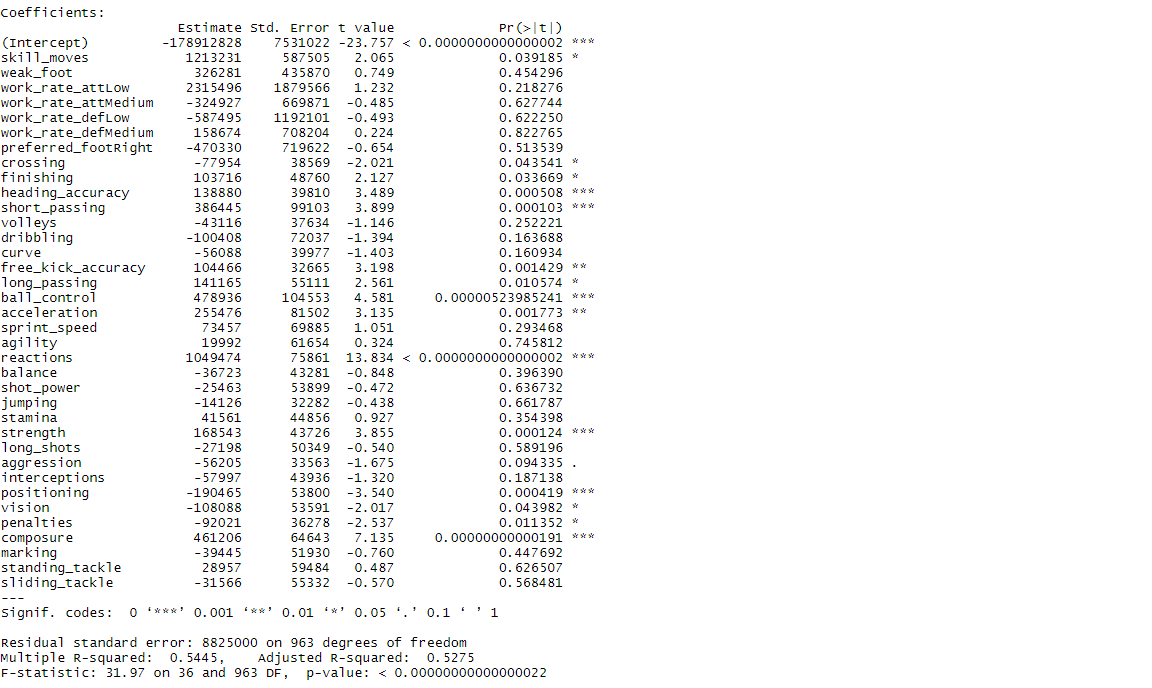


Figure 3: Summary of Multiple Regression Model

The MSE value of this model has been calculated as 4.583674e+13 after predicting the “Euro Value” by using the test data. 2.2 Ridge & Lasso Regression Model

In the next part, we apply the Ridge and Lasso Regression model. After the calculation, we find the best lambda values for each model by using cross-validation. The best lambda value for Ridge model is 572,237 and for Lasso it is 46,416. We can interpret these best lambda values as the least MSE values. If we analyze the Figure 4, we can see how many of the coefficients are non-zero in this model by looking at the top of the model. In the Figure 6, in the Lasso Regression, as the lambda increases, approximately at the lambda value of 10, coefficients become 0 and the error increases. At the lambda value of 11, it becomes exactly zero.

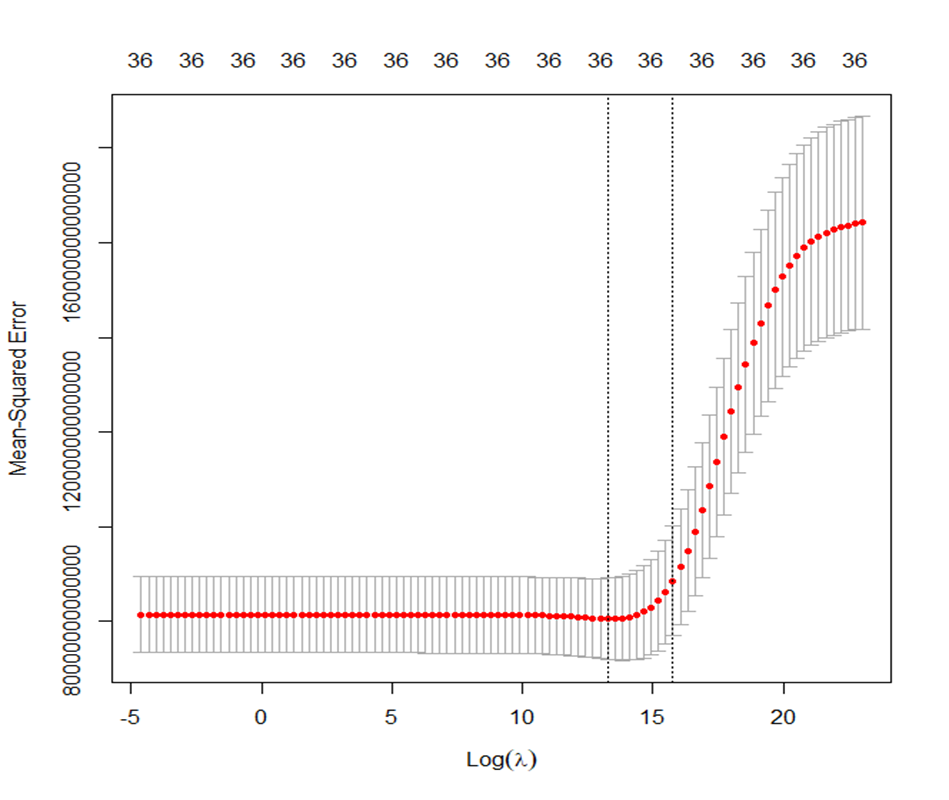


Figure 4: MSE vs log(lambda) – Ridge Regression

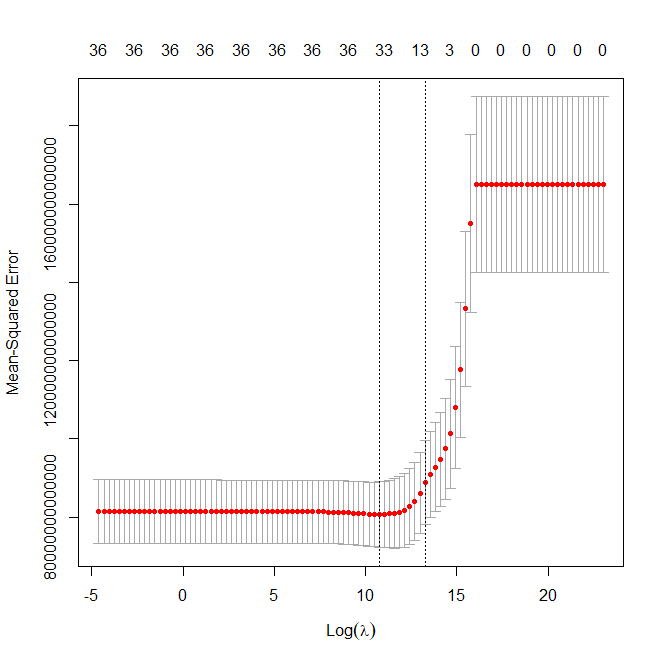


Figure 5: MSE vs log(lambda) – Lasso Regression

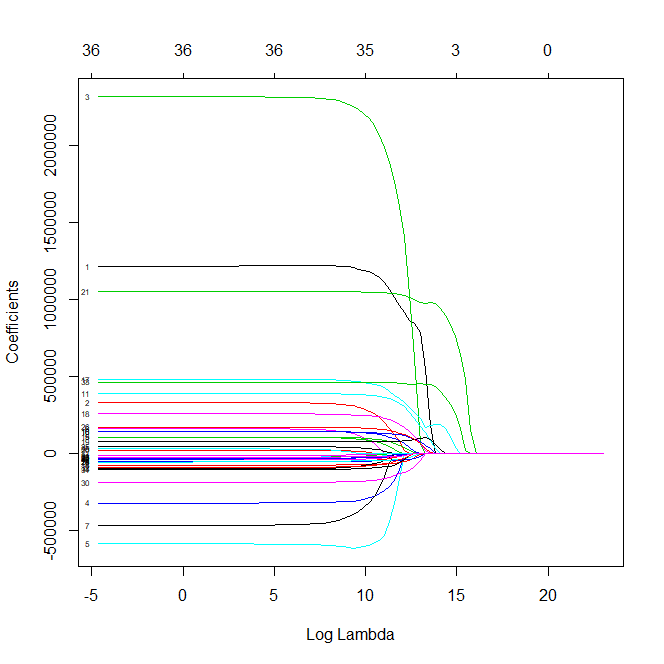


Figure 6: Coefficients vs log(lambda) – Lasso Regression

Before we move to the next model, we predict the “Euro Value” with these models and test data to calculate the second MSE for both regression models. For the Ridge Regression model, the MSE value is 4.208114e+13 and for the Lasso Regression model, the MSE value is 4.406923e+13.

2.3 Multiple Regression Model vs. Lasso Regression Model

The coefficients of both models will be compared in order to see the difference between these models. In the Figure 7, we can see that some of the variables are eliminated from the model. Those variables have higher p-values which make them unwanted in this model. By using the Lasso Regression model, we make their coefficients zero and therefore we get a better MSE value than the Multiple Regression model.

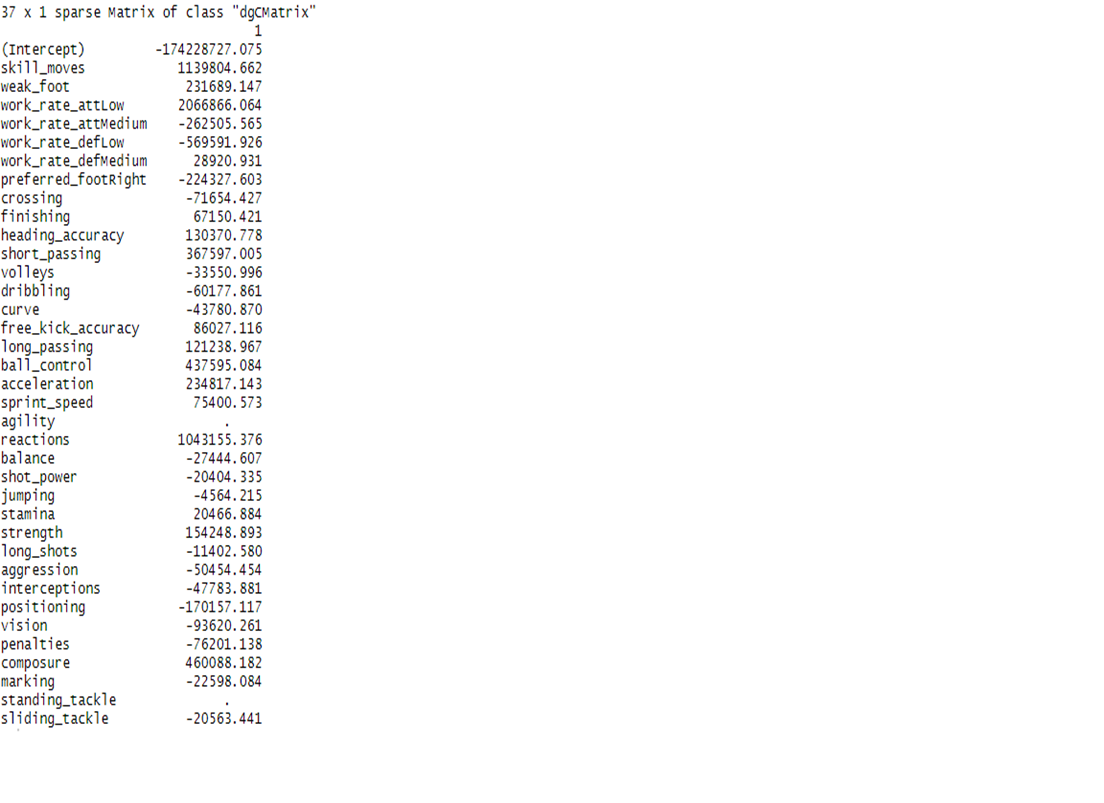


Figure 7: Summary of the Lasso Regression Model

2.4 Regression Tree

Now we apply the Regression Tree with all the variables. Due to the cross-validation, we find the terminal nodes as 11 in order to predict the “Euro Value”. The optimal level is 11 in the cross-validation. The Figure 8 shows us the result.

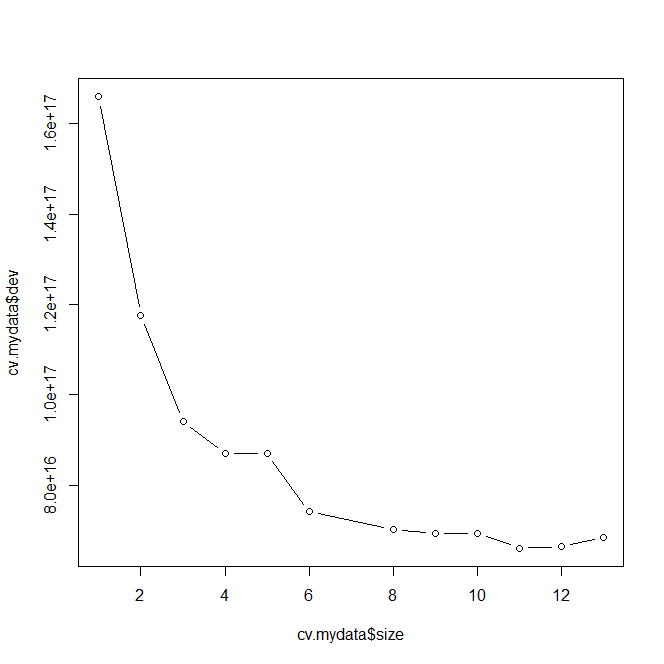


Figure 8: Cross Validation Result – Regression Tree

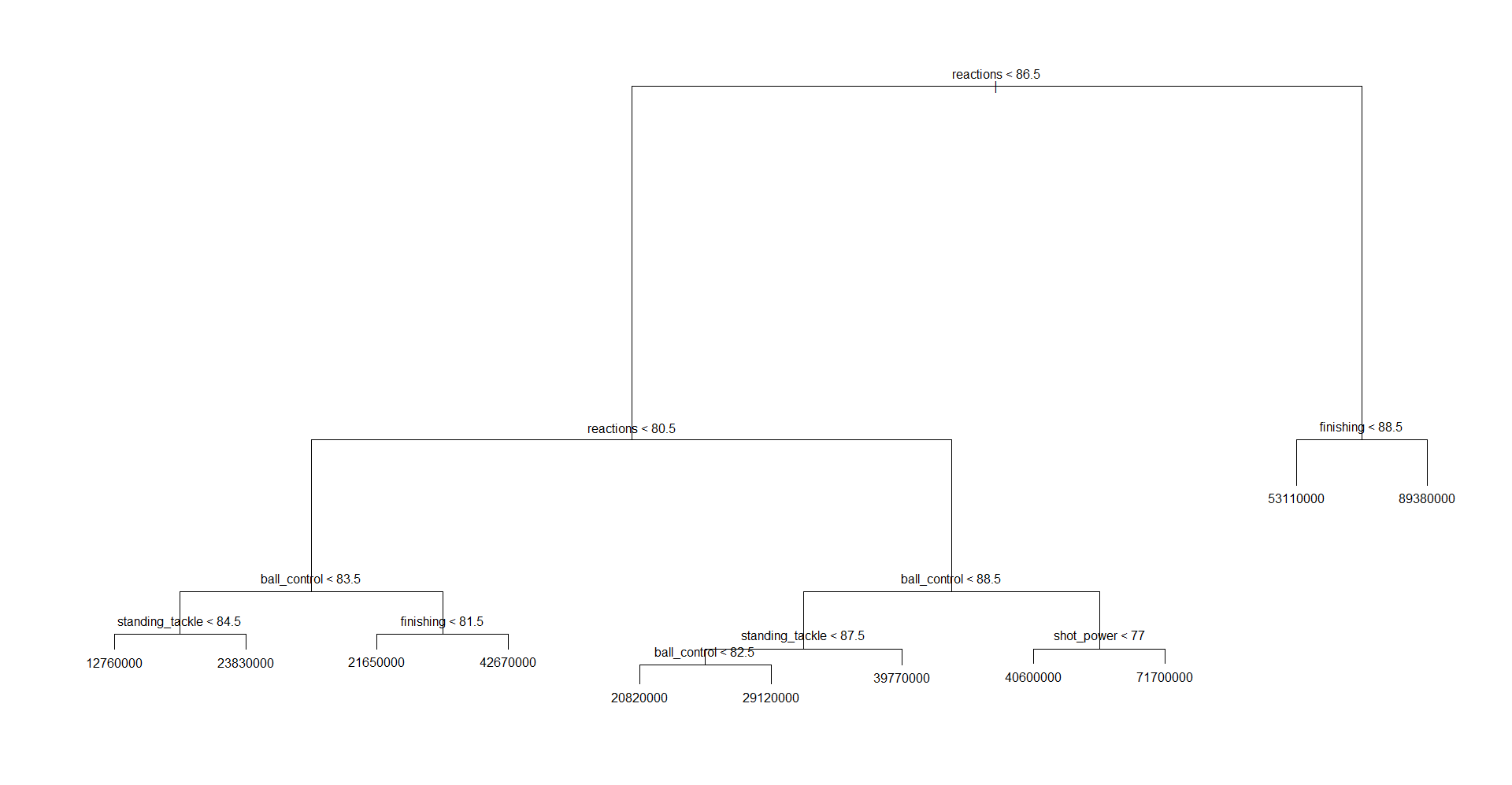
After cross-validation, we prune the tree to have only 8 splits and the result is shown in the Figure 9. In the figure, we can easily see that the most important variable is reactions at the first level. As the final step for the Regression Tree model, we predict the “Euro Value” with this model and we find the MSE value as 3.014445e+13.

Figure 9: Regression Tree Plot

2.5 Random Forest

For the final step, we apply the Random Forest model. In the last model, the most important 10 variables are “reactions”, “ball control”, “standing tackle”, “short passing”, “finishing”, “sliding tackle”, “vision”, “dribbling”, “marking” and “sprint speed”. The mtry is 15 with the least variable MSE value of a variable which is 3.999819e+13. After implementing the Random Forest model, we predict the “Euro Value” with this model and the test data and we get the final MSE value as 1.584264e+13.

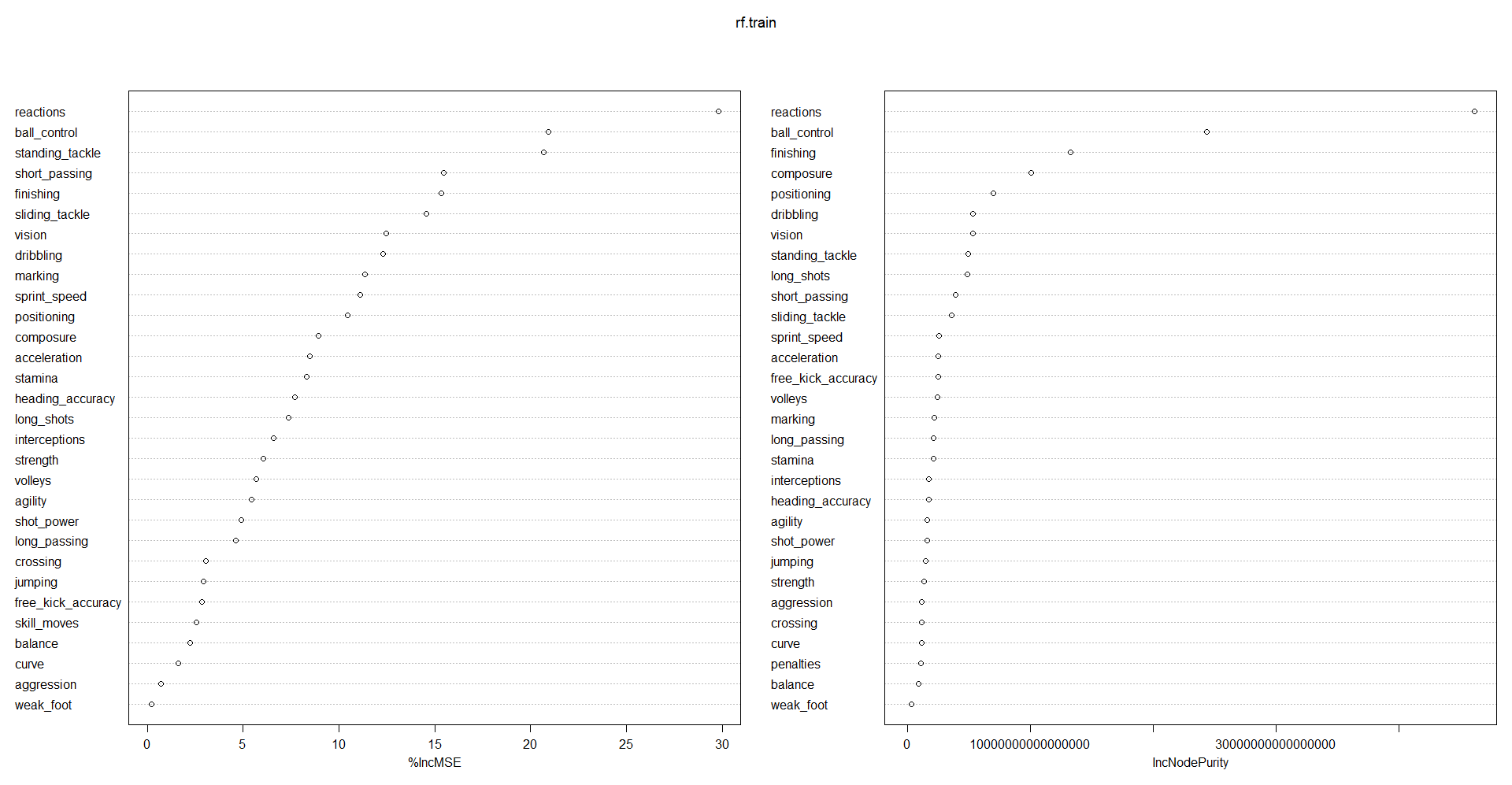


Figure 10: %IncMSE Values for Variables - Random Forest

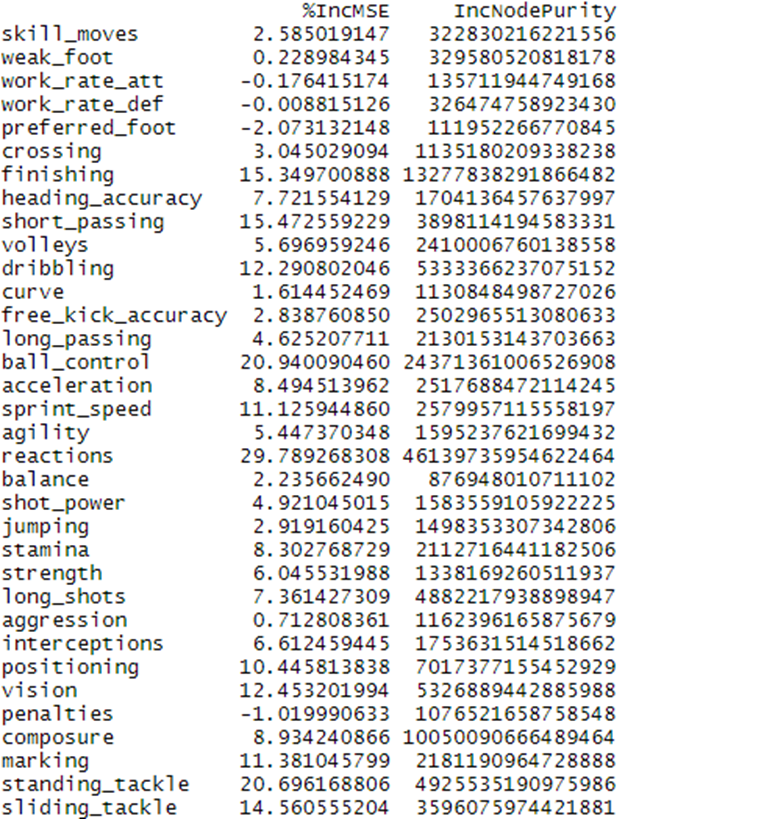


Figure 10: Summary of %IncMSE Values for Variables - Random Forest

1. CONCLUSION

All in all, we have compared the MSE values of all models and we have found out that the Random Forest model gives us the least MSE value. This means that after some models, the coefficients of the variables become close to zero or some of the variables are eliminated after cross-validations. Especially after the Lasso Model, some variables are non-existent. The Random Forest model has the best prediction power among all models.

|  |  |
| --- | --- |
| **MODEL NAME** | **MSE VALUES** |
| Multiple Regression Model | 4.583674e+13 |
| Ridge Regression Model | 4.208114e+13 |
| Lasso Regression Model | 4.406923e+13 |
| Regression Tree | 3.014445e+13 |
| Random Forest | 1.584264e+13 |