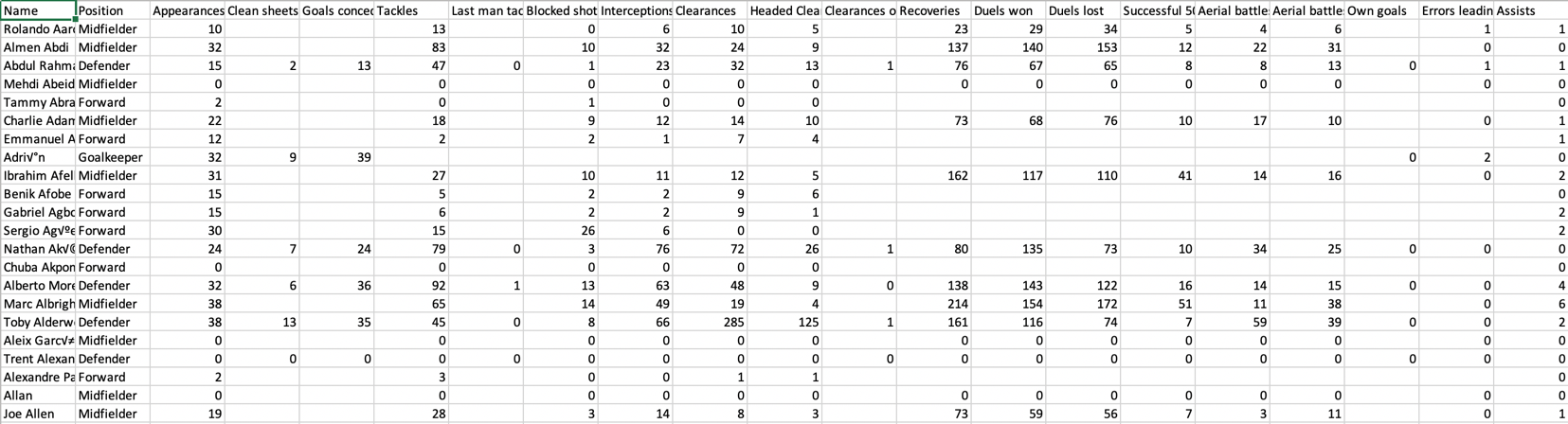
Soccer: Best Goal Scorer Prediction

# 1. Introduction

This report is a detailed assessment on the variables that are needed to predict the best goal scorer in football/soccer. Initially, I want to clear the difference between the best goal scorer and the top goal scorer. The top goal scorer is simply the player that scores the most goals in a season, whereas the best goal scorer is a player that has the best goals per match ratio. For example, in a single season, player x scores 20 goals in 40 games, whereas player y scores 15 goals in 20 games (unequal number of games may be due to injuries, international call up for the player’s national team and family emergencies). Now if you only look at the number of goals, we can easily say that the player x is a better goal scorer than player y, but in order to be labelled the best goal scorer we will need to look at the players goals per match output in terms of their appearances or the number of games they’ve played that season. We calculate this by dividing the goals by the number of appearances made by the player in a single season. Therefore, player y (0.75) has a better goal per match than player x (0.5), and is therefore the best goal scorer in the league.

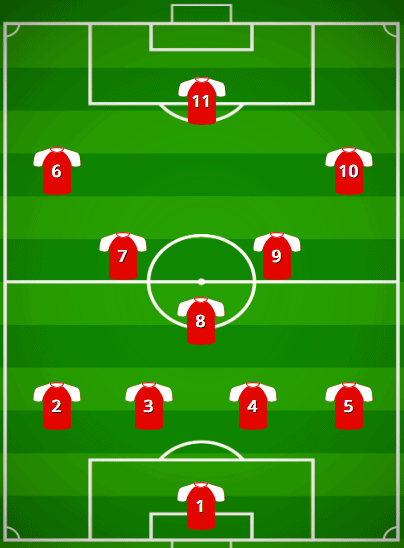
In order to predict the best goal scorer in football, we will look at each players’ goals per match and compare it with all the other players across all the teams in the league.

The data set I used is from Kaggle and is a league data set from the English Premier League season of 2015-2016. I tried to get the latest data, but failed to do so. Nonetheless, it is still a very good data set with numerical variables and also includes all the players from that season with numerous records about their individual qualities and attributes.

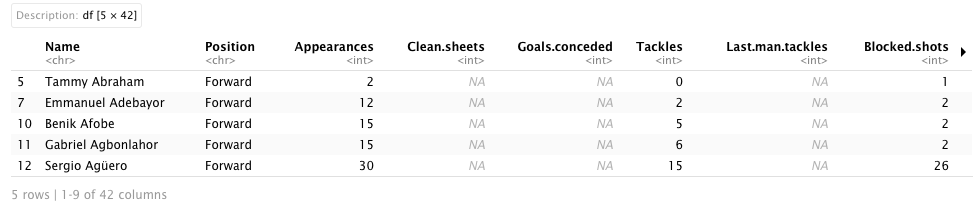
Below is a short screenshot of the dataset I used. 

# 2. Data Cleaning

Now that we’ve read the data, we will clean the data set by filtering the variables we don’t need to predict the goals, per match. In football, the best goal scorers are usually the players that play in a forward position. In the picture below you can see the players circled in red are the players that play in the forward position.

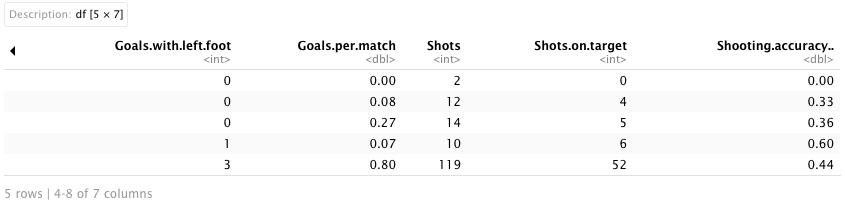


Hence, we eliminate or drop the other positions(including: Goalkeeper, Defender, Midfielder). In this case we create a subset with only the players in the forward position. Therefore after cleaning the dataset to only keep the players in the forward positions, the dataset looked like this.

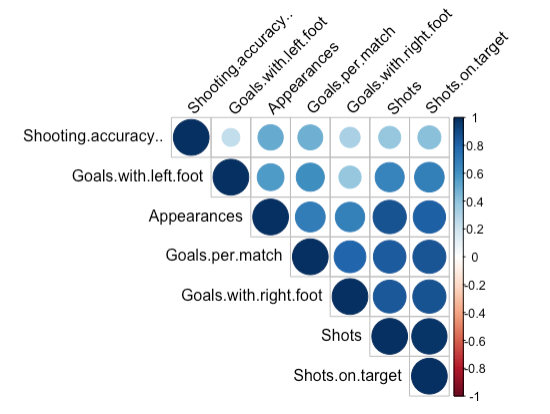


Next, we drop the variables or features that do not have an effect on the goals per match variable. I have been playing football ever since I was a kid and thankfully I can add my knowledge to this report by removing the variables that don’t have an effect on goals per match. For example, for players in forward position, they do not need the Clean.sheets variable to predict their goals per match. Clean.sheets are more o a variable featured towards other positions like Goalkeeper and Defender.

Now, after removing unnecessary variables and removing null values, the dataset looks like this.



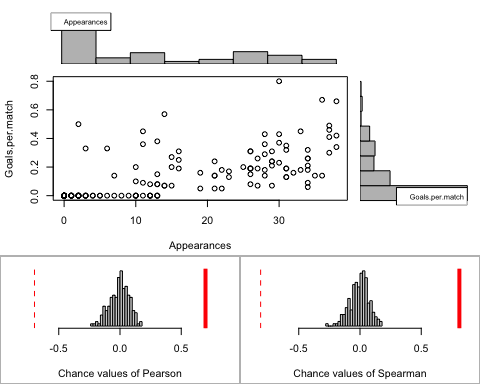
Now, to confirm that I only left the variables that helps in predicting the goals per match, I created a correlation matrix, which shows they are correlated with each other.

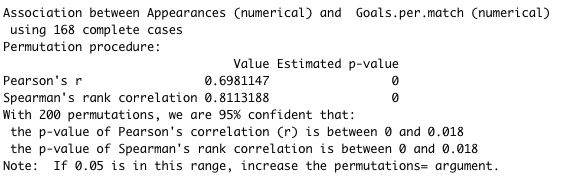


# 3. Association Analysis

For the association analysis, since we’re predicting the goals per match variable, it will be our y variable. Therefore, the other remaining variables will be our x variables. I will be carrying out 4 association analysis tests where Goals per match will always be my y variable.

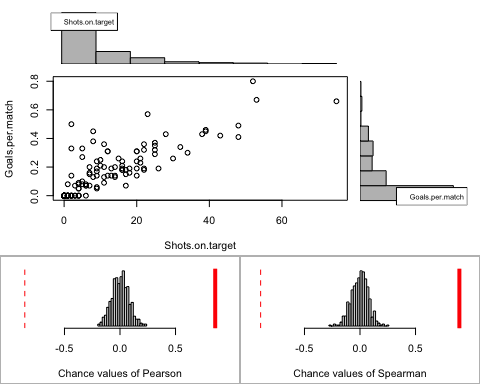
1. Goals.per.match ~ Appearances

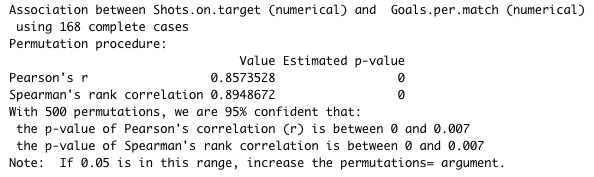




The graph shows that the relationship is not linear, therefore we choose the Pearson’s test. We can see the results and say that the association between Goals per match and Appearances is statistically significant because the p-value is between 0 to 0.018 which is less than 0.05. Therefore appearance (or more games played) is a major factor in predicting the goals per match. the graph also shows that when a player has a higher number of appearances it tends to be that the player has the chance of getting a higher goals per match.

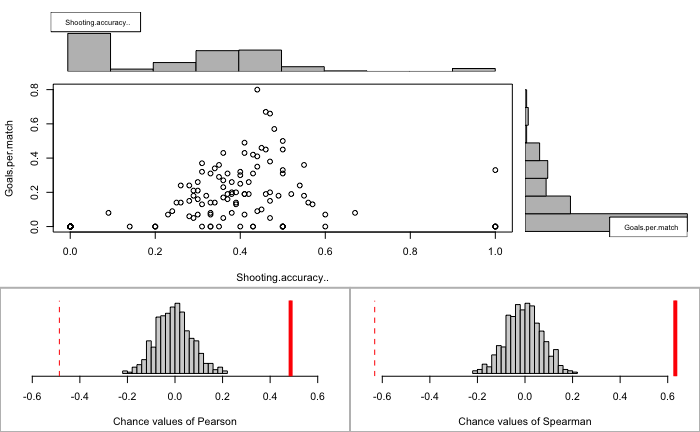
1. Goals.per.match ~ Shots.on.target

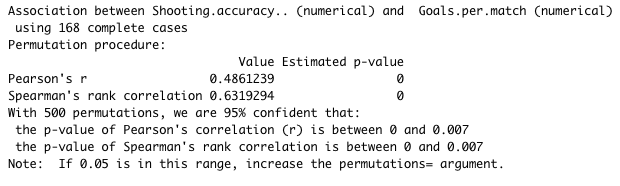




The graph shows that the relationship is linear, therefore we choose the Spearman’s test. We can see the results and say that the association between Goals per match and Shots on target is statistically significant because the p-value is between 0 to 0.007 which is less than 0.05. Therefore shots on target is a major factor in predicting the goals per match. Also from the scatter plot we can see a rise in the goals per match when there is a rise in the shots on target. Therefore when there is an increase in shots on target, the goals per match is likely to increase.

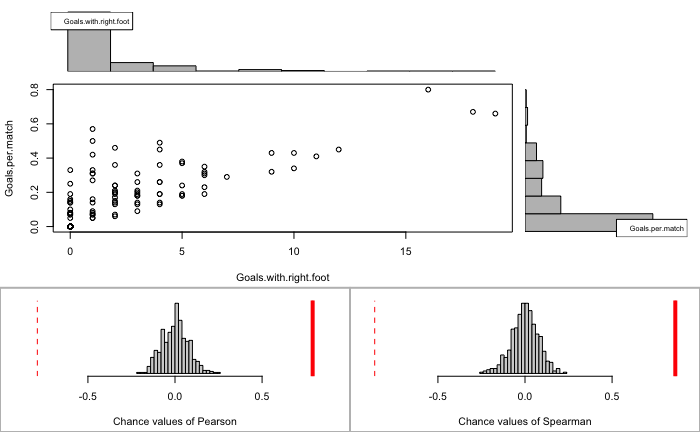
1. Goals.per.match ~ Shooting.accuracy

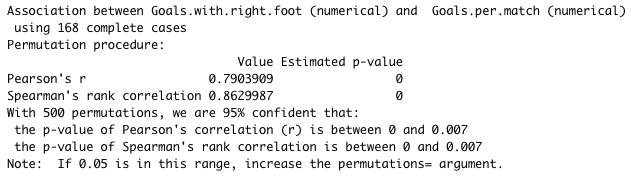




From the results we can say this association between Goals per match and shooting accuracy is a linear relationship and therefore we use the Spearman’s test. The test is statistically significant because the p-value is between 0 to 0.007 which is less than 0.05. Also from the scatter plot, although there are some outliers, we can see a rise in the goals per match when there is a rise in the player’s shooting accuracy.

1. Goals.per.match ~ Goals.with.right.foot





From the scatterplot we can say this association between Goals per match and goals with right foot has a linear relationship. Therefore, we use the Spearman’s test. We can say the test is statistically significant because the p-value is between 0 to 0.007 which is less than 0.05. Therefore we can say there are more chances of players that are right footed with more chances of getting a higher goals per match.

Before we move onto simple linear regression models, we need to confirm that there are no autocorrelated variables, therefore I carried out a test in R to check the autocorrelation of the variables.

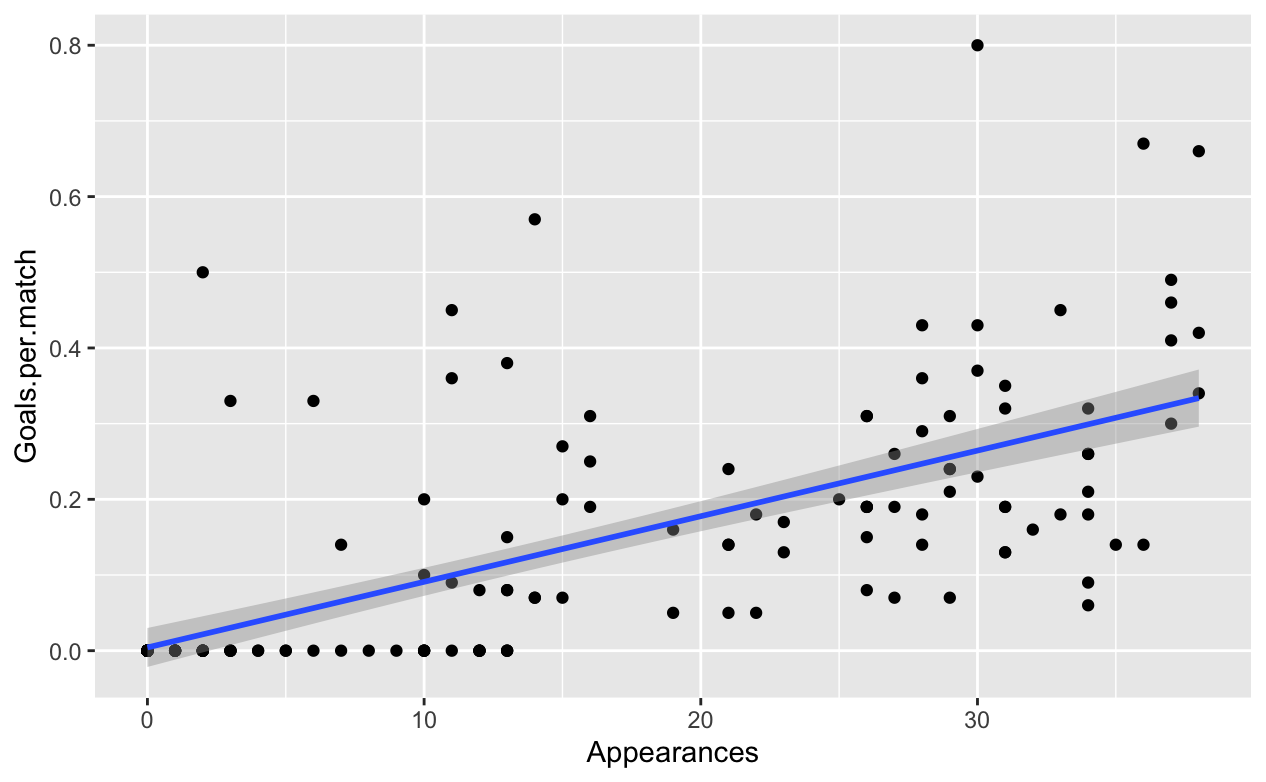
# 

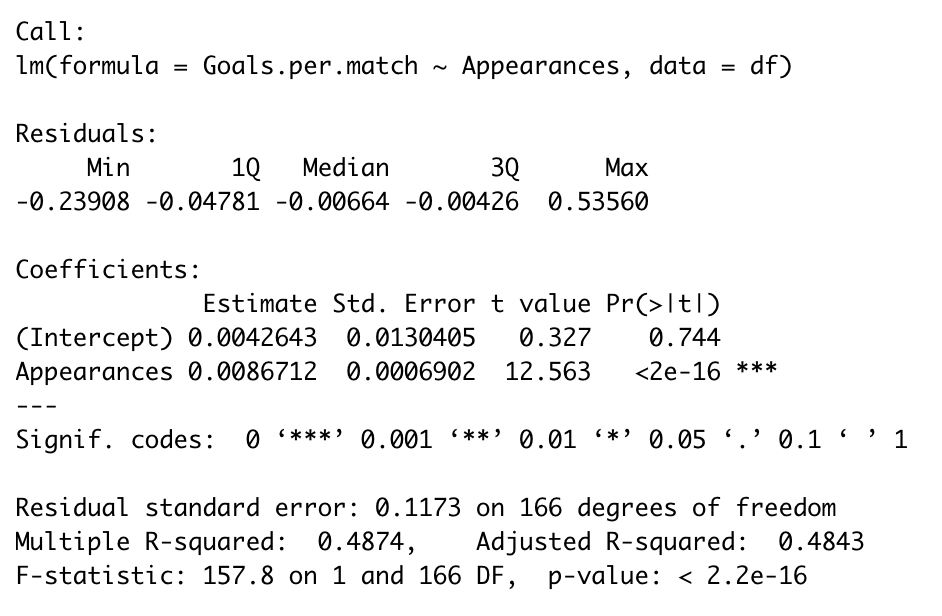
From the results (graphs), we can see the vertical lines drops to 0 after the first line or lag therefore there are no auto correlation in this dataset.

# 4.Simple Linear Regression

Now, we move on to perform the simple linear regression with our variables to describe the relationships between them. We will be able to see the line of best fit through our data by searching for the value of the regression coefficients which minimizes the total error of the model.

The first regression model will be checking the relationship between Goals per match and Appearances.

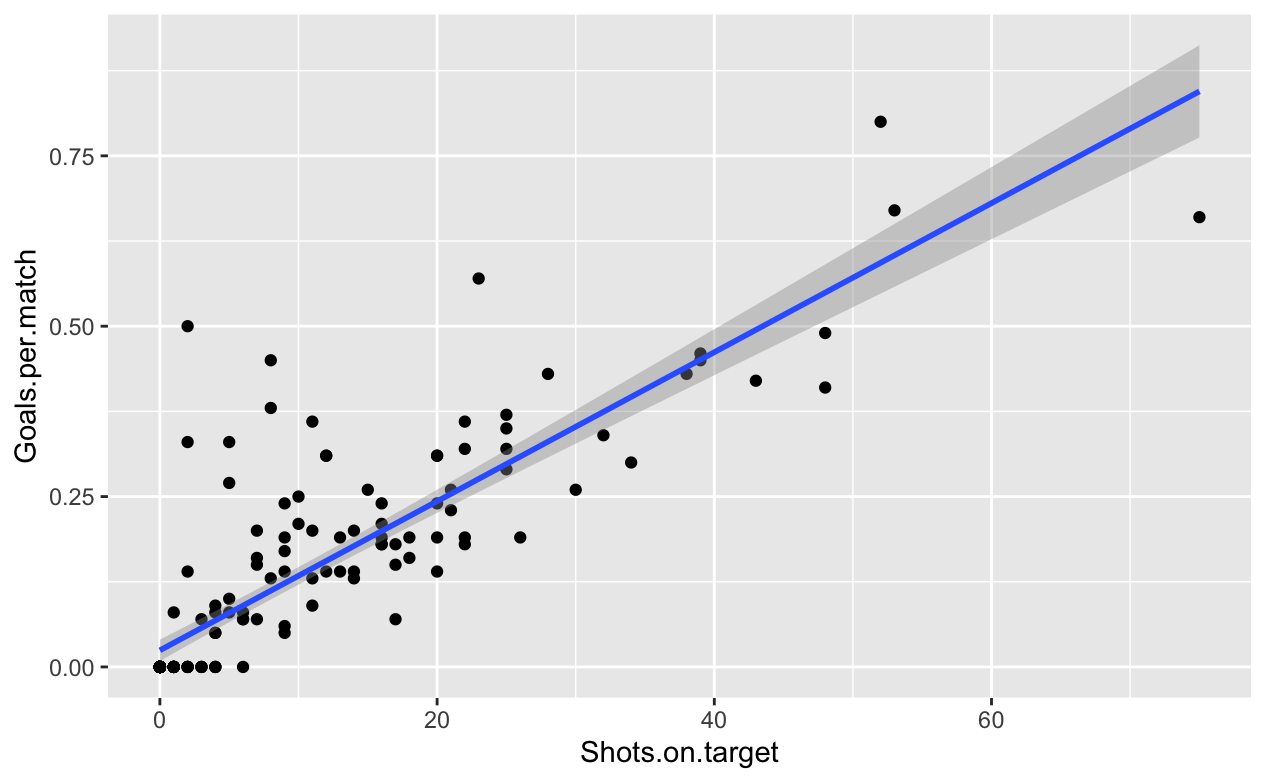


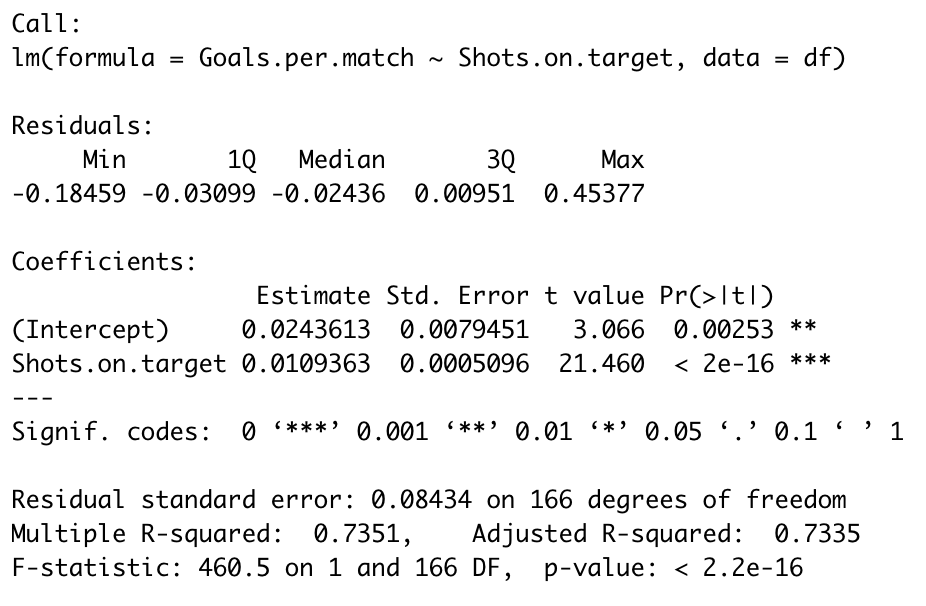


From the test we found a significant positive relationship between goals per match and appearances with p-value < 2.2e-16 or almost 0, which indicates that the model fits well. Also looking at the coefficients, we conclude that

Goals per match = 0.004264 + 0.008671\*Appearances

The next regression model will be checking the relationship between Goals per match and Shots on target.





From this test we found a significant positive relationship between goals per match and shots on target with p-value < 2.2e-16 or almost 0, which indicates that the model fits well. Also looking at the coefficients, we conclude that

Goals per match = 0.024613 + 0.0109363\*Shots on target

Therefore, we can say to predict the best goal scorer, variables (appearances, shots on target, goals with right foot and goals with left foot are important to get accurate results.