01.31.2022

**MACHINE LEARNING FOR MANAGERIAL DECISION MAKING**

**FINAL PROJECT**

SEMA MENTE

11811030

**Contents**

**Introduction………………………………………………………………………………………………………………………1**

**Analysis…………………………………………………………………………………………………………………………….2**

**Conclusion………………………………………………………………………………………………………………………12**

1

**Introduction**

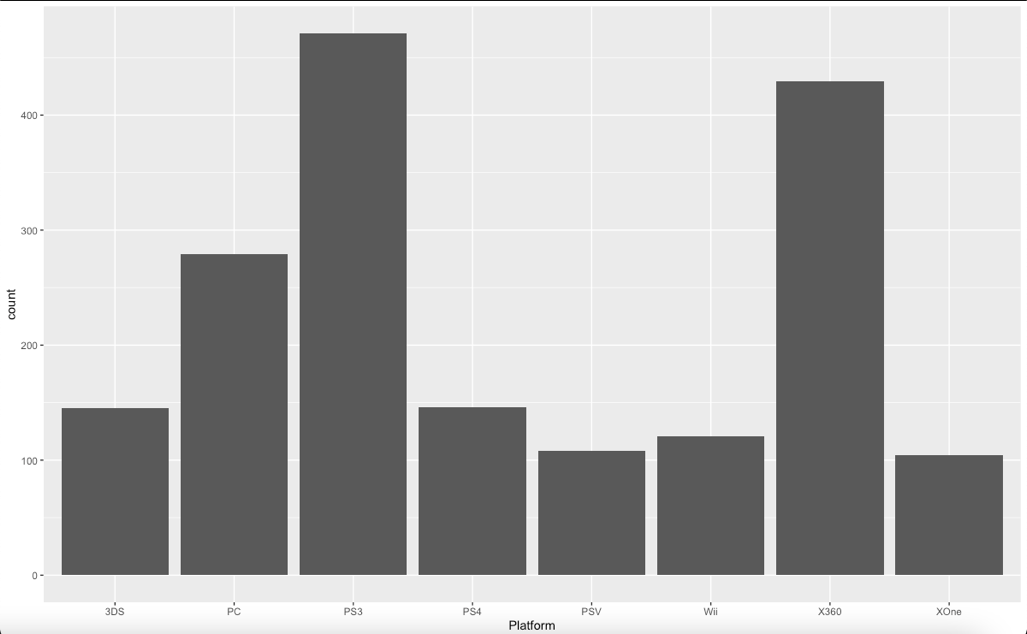
The Dataset contains other attribution of video games such as sales, year, platform, variant, and more. There are 1803 observations and 9 variables in the data set. This project analysis the linear regression model, analyzes predicters and our independent variables, uses lasso and ridge models, makes predictions on the model, and includes analysing and interpreting by performing various operations using models such as decision tree models and random forest models.

2

**Analysis**

The purpose of using the bar graph in this project is to determine which platform the video games were played in which year. In order to make this analysis, a bar graph is made containing the number of uses of the Platform and the names of the platform. Likewise, in order to get correct information from the dataset, the Year bar graph is created to show which years the data in the dataset belongs to. With the information given, it is determined and interpreted in which year the game was played on which platform.

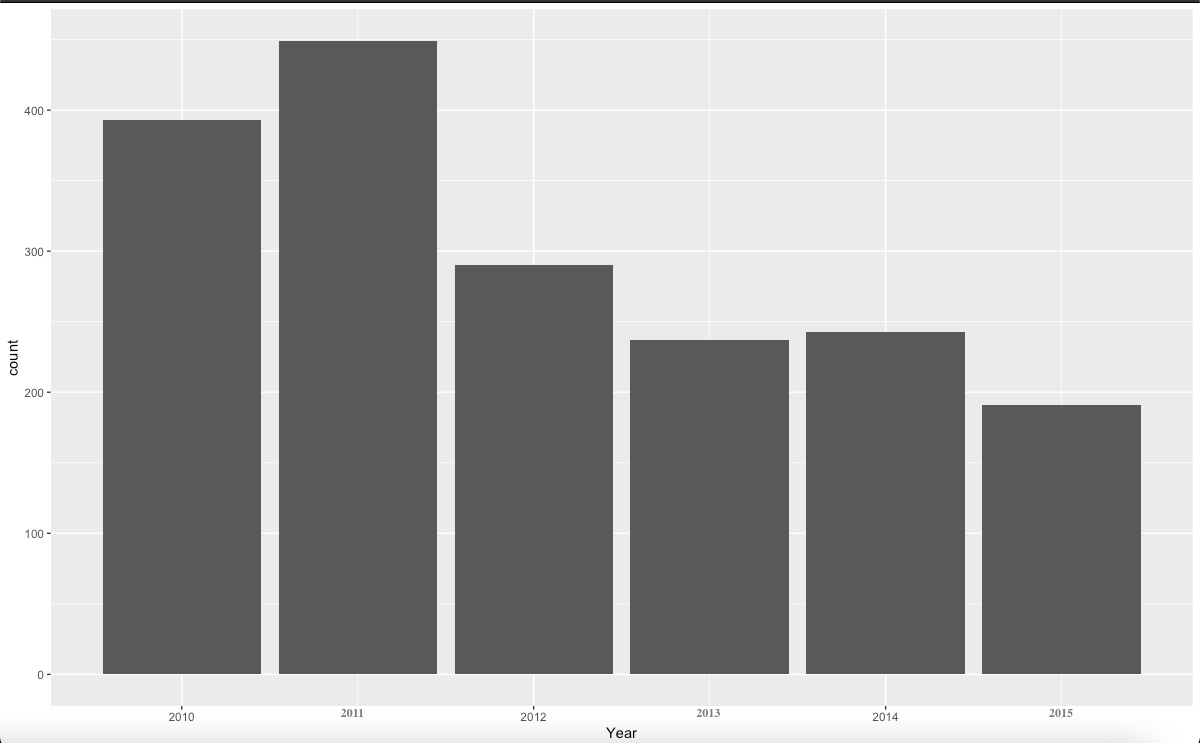
a) Analyze your predictors/independent variables using techniques shown in the class.



1.1 Platform bar chart

Looking at the 1.1 bar chart, according to the data set, the most used platform is PS3, with more than 400 preference, followed by the X360. The least used platforms are PSV and XOne.

3



1.2 Year bar chart

According to the year information given to us in the data set, the year 2011 is the most followed by the year 2010 with more than 400 data in the bar graph showing the year. By looking at the 1.2 bar chart, we can say that as the years progressed, the use of video games in the dataset gradually decreased. There may be two reasons for this; First of all, the information given in the dataset may not make sense because it is randomly selected. Secondly, the rate of playing video games has decreased over the years and people may be oriented towards different types of games or different categories of games. In short, the popularity of video games may have decreased and new alternatives may have emerged.

4

köpek, beyaz içeren bir resim

Açıklama otomatik olarak oluşturuldu

* 1. Year and Platform

The purpose of the 1.3 point graph is, by looking at the year of the data in the data set, we can tell which platform it is using and the amount of usage. By looking at the dotted graph, we can say that in 2010 and 2011, the most games were played on PS3 and x369 platforms. After the PS4 was released in the last month of 2013, we can see from the table that the use of PS3 decreased. The platform that is closest to the constant according to the years is the PC. There have been minor changes over the years, but compared to other data, the least change has been in this platform.

5

b) Take the first 1000 observation as the training set and use the rest as the test set.

1000 observation has taken by using train and dataset.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

c) Fit a multiple regression model to predict “Global\_Sales”. Use only the training set to fit the regression model

Lineer model is used for predict ‘Global\_Sales’.



6

d) Analyze your estimated regression models. Comment on coefficients, adjusted R square and F statistic of the model.

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

By looking at the residual values, we can say that except for our maximum value, the other values are fine, close to 0.

The sign of the linear coefficient indicates the direction of the linear relationship between x and y. When r (the coefficient) is near 1 or −1, the linear relationship is strong; when it is near 0, the linear relationship is weak. If we look at the Estimates of our Coefficients values, the fact that they are all close to 0 indicates that the estimates made are close and the estimated figures in our system are in good shape. The given stars (\*) indicate how well the prediction was made.

7

When we look at the f-statistics part, we see p-value: < 0.00000000000000022. Since this means less than 0.5, our model is a meaningful model at 95% confidence level.

We can say these variable significantly affect the our value of the video games.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Lastly, for r squared we can say we can explain almost half of the variation and it is not bad.



e) Predict “Global\_Sales” in the test set using the regression model obtained in (c). Calculatethe root mean square error of the test set (RMSE)

We take MSE and RMSE values.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

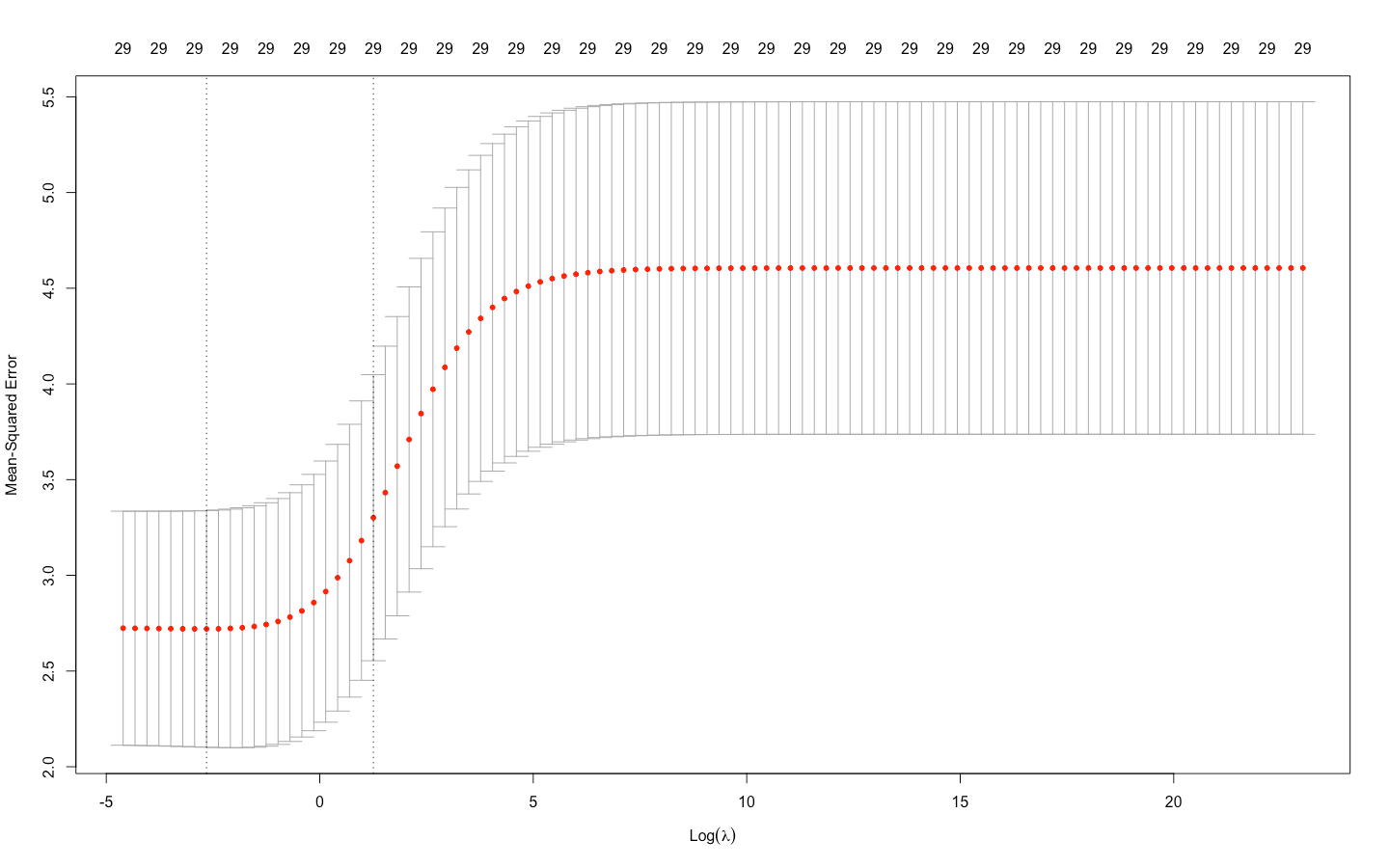
f) Fit a Ridge model and a Lasso model to predict “Global\_Sales”. Use only the training set to fit these regression models. Determine the lambda parameter using cross-validation.

Find the Ridge model we create train and test matand we change alpha number to 1 and we create Lasso model. We use glmnet library.

metin içeren bir resim

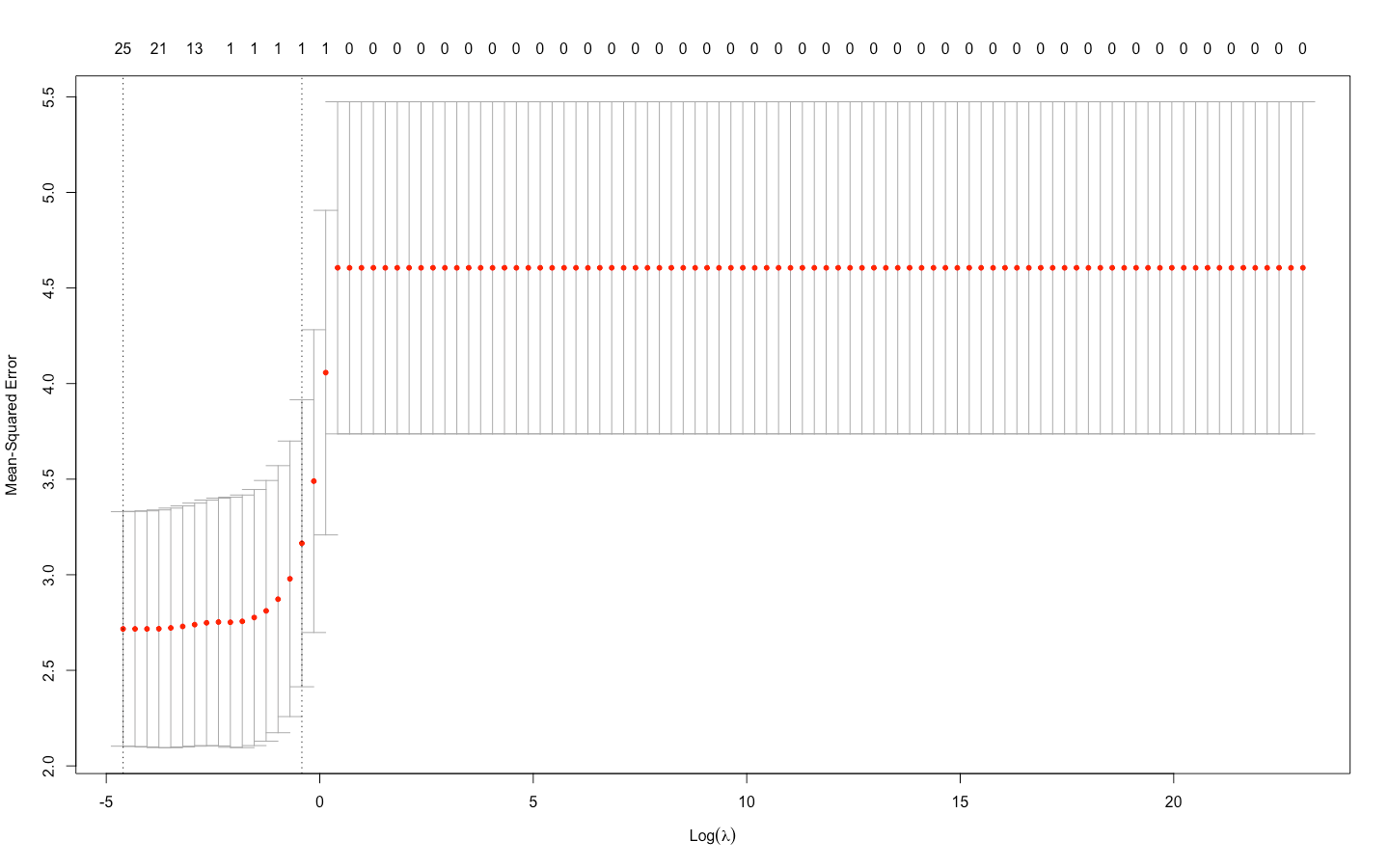
Açıklama otomatik olarak oluşturuldu

8



1.4 Graph

According to 1.4 graph ridge parameter and cross validation error stay stabil until optimal point. After optimal point cross validetaion is increases. (Optimal error)



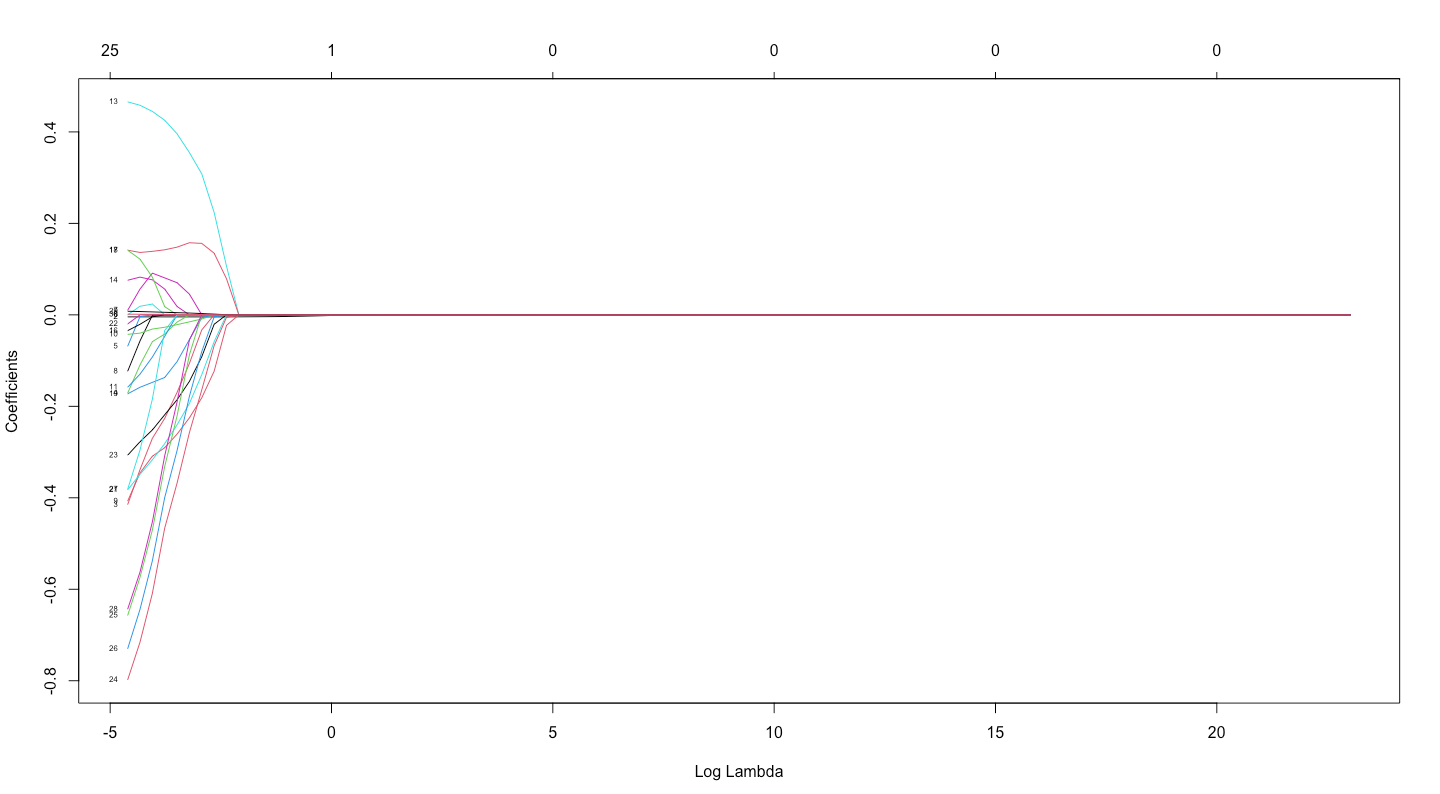
* 1. Graph

Our optimal point is very close to the beginning. After the optimal point, the parameter rises and remains constant after a while.

9

If we compare the 1.4 and 1.5 graphs, we can say that the optimal point of the 1.5 graph is very close to the beginning, but in the 1.4 graph, the optimal point is located just before the value increase. Graph 1.5 increased sharply from constant to higher and then stabilized again. In Graph 1.4, on the other hand, after the optimal point, it rose harmoniously and remained stable again. Graph 1.4 looks more reassuring.

g) Analyze your Lasso Model. Compare your Lasso Model with the multiple regression model estimated in (c).



Graph 1.6

In graph 1.6 we read the graph from right to left. All coeffecients came from 0 and coeffecient values changed as the ridge parameter increased. As seen in the graph 1.6 we can say there are 25 non-zero variables.When the Log Lamda increase coeffecients values are decreasing.

h) Predict “Global\_Sales” in the test set using the Ridge model and the Lasso model obtained in (f). Calculate RMSEs of these models only using the test set.

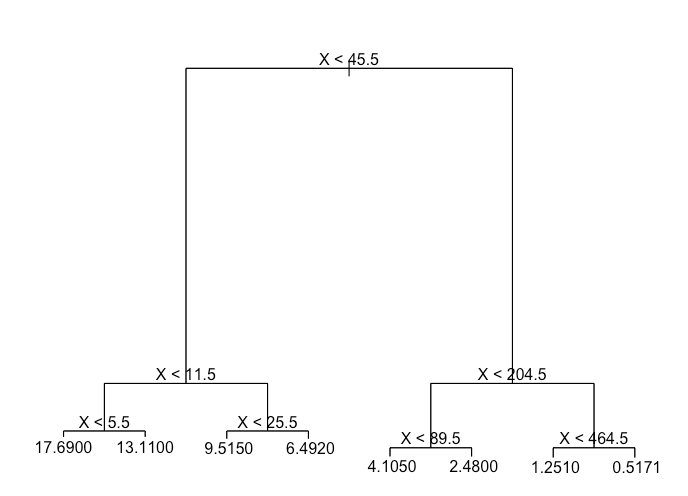
We calculate RMSEs models. RMSE2 for ridge and RMSE3 for lasso. We can say Ridge perform better than Lasso.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

10

i) Fit a regression tree to predict “Global\_Sales”. Use only the training set to fit the regression model. Determine the number of terminal nodes using cross-validation.



1.7 Graph

Tree methods have high accuracy, stability, and ease of interpretation. Unlike linear models, they can also map nonlinear relationships quite well. Classification or regression can be adapted to solve any given problem. We use Tree library to get out models. 1.7 Graph is our best model for the dataset.

j) Predict “Global\_Sales” in the test set using the regression tree model obtained in (i). Calculate the RMSE of the regression tree only using the test set.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

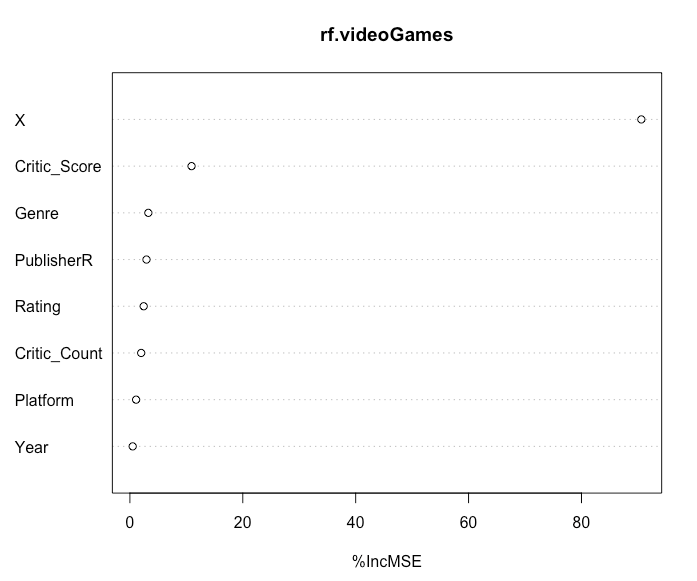
We calculate Root Mean Square Error ysing the set.

11

k) Fit random forests to predict Global\_Sales. Use only the training set to fit the regression model. Determine the number of variables used in each split using the cross-validation. Grow 500 trees for random forest.

Random forest is used as a common classification in algorithms. We use random forest library.

l) According to the random forest obtained in (k), which variables are import? Comment.



1.8 Graph

According to 1.8 graph we can say most important variables is critic\_score and less importans is year. Except for Critic\_Score, the other values are close to each other.

m) Predict “Global\_Sales” in the test set using the random forest model obtained in (k). Calculate the RMSE of the random forest only using the test set.

RMSE5 is the random forest root mean square error.

n) Compare RMSEs obtained in (e), (h), (j) and (m).



We can say best RMSE is RMSE5 randomforest. The worse variables is RMSE1.

12

**Conclusion**

Since RandomForest has the lowest RMSE value, it is the best model. The good sides of this model; It reduces overfitting in decision trees and helps improve accuracy, is flexible to both classification and regression problems, works well with both categorical and continuous values, automates missing values found in data, and does not require data normalization as it uses a rule-based approach. The negative aspects of this model are; It requires a lot of computational power as well as resources, as it creates a large number of trees to combine its outputs. It also requires a lot of time for training as it combines multiple decision trees to determine the class, and suffers from interpretability due to the collection of decision trees and cannot determine the importance of each variable.