ISTANBUL BILGI UNIVERSITY

Forecasting global sales of video games by regression models

MIS 315

11511021 - Hakkı Eren Arkangil

Istanbul,2019

Abstract

Forecasting global sales of video games by regression models

Hakkı Eren Arkangil

January,2019

We review some of the regression models that have been proposed by statisticians and computer scientists.That review will include wide variety of machine learning methods such as Ridge&Lasso,decision trees,bagging and random forests.Then we will compare the accuracy of these methods and try to understand which approach is more efficient to predict our data.The aim of this paper to compare machine learning methods in order to predict global sales of video games.To compare this methods we calculated MSE and RMSE for each regression model.The results shows we have better accuracy with linear models

Keywords: forecasting,predictive modelling,machine learning,regression,sales forecasting.

Table of Contents

ABSTRACT

TABLE OF CONTETS

1. INTRODUCTION

2. METHODOLOGY

2.1 Explanatory Data Analysis

2.2 Correlation Analysis

2.3 Multiple Linear regression

2.4 Ridge and Lasso regression

2.4.a Ridge regression

2.4.b Lasso regression

2.5 Decision tree regression

2.5.a Decision tree pruning

2.6 Bagging and Random Forest

2.6.a Bagging

2.6.b Random forests

3.CONCLUSION

4.REFERENCES

1. INTRODUCTION

Who wouldn't like to predict global sales of video games in the contemporary world.It is key to understand effects of some factors on sales.In the cut-throat marketing era some must know using how to handle with big data which is collected by users.

In the figure 1. Video games sale data set provides us some informations.The sale values which is higher than 0.75 is categorized as high.We also observe higher user scores has positive impanct on sales,where user score is not the only predictor in this dataset.

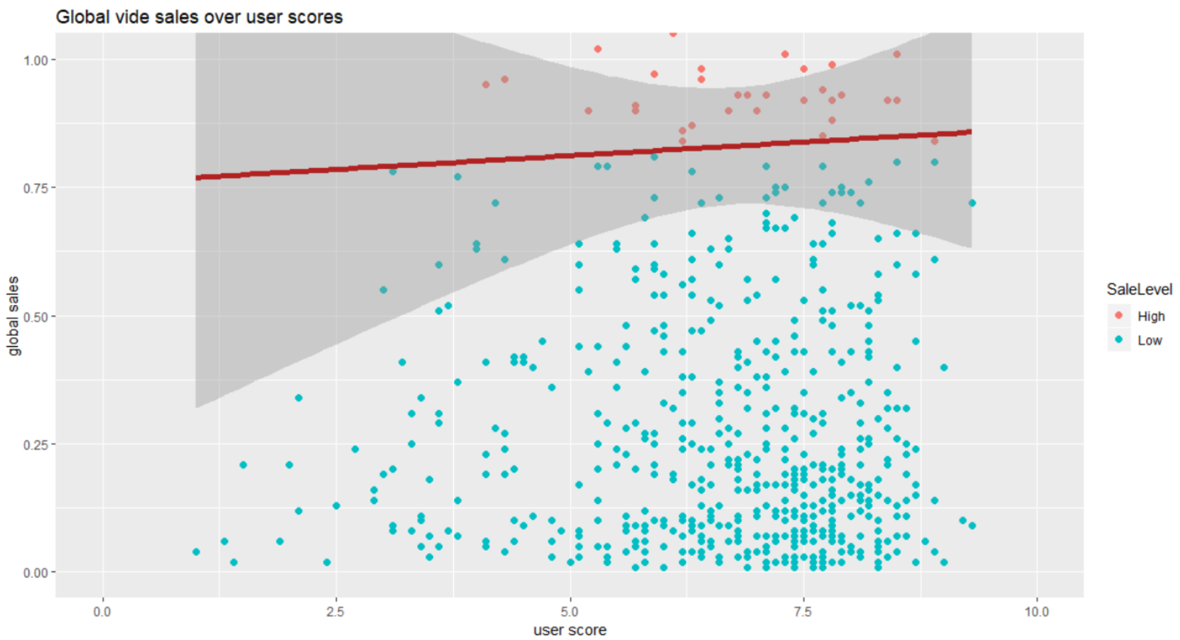

Figure 1. Globalsales over user scores

2. METHODOLOGY

Exploratory data analysis in this study started with observation of features, their types and distributions. To prepare data for modeling,we will look into correlation table to build better models. After the initial data analysis phase, best subset selection method was used to acquire the strongest features in numerical data and narrow down the predictor set. At the modeling stage, regression models were formed. We compared model performances by training set cross-validation errors and test set errors.

2.1 Explanatory Data Analysis

We begin our project by analyzing our dataset which contains 749 rows and 41 columns. While X is giving us the row number directly (from 1 to 749); predictors such as Platform, genre, year of release, publisher and rating are behaving categorically, even if they are not. The means of these binary predictors helps us to see the distribution of data between genre, platform or year.

For example, genresimulation's mean 0.0227 while genresports's has 0.112. This states that sports games are far more popular than simulation games.

SaleLevel (as high and low) is our only categoric predictor. Besides SaleLevel, we have more meaningful predictors (that can take continuous values between 0 to 10) such as user score, count, critic score, count. Global Sales, our dependent variable, is what we try to estimate.

Since variables such as row number and names are not meaningful for our regression analysis, we are going to remove them from our model with SaleLevel because this paper does not include classification analysis.

2.2 Correlation Analysis

Then, we took 2 from continuous similar variables such as genre platform and created a correlation matrix with our meaningful predictors such as critic score, count and user score. This matrix shows us that user score, critic score and count has positive correlation with global sales.

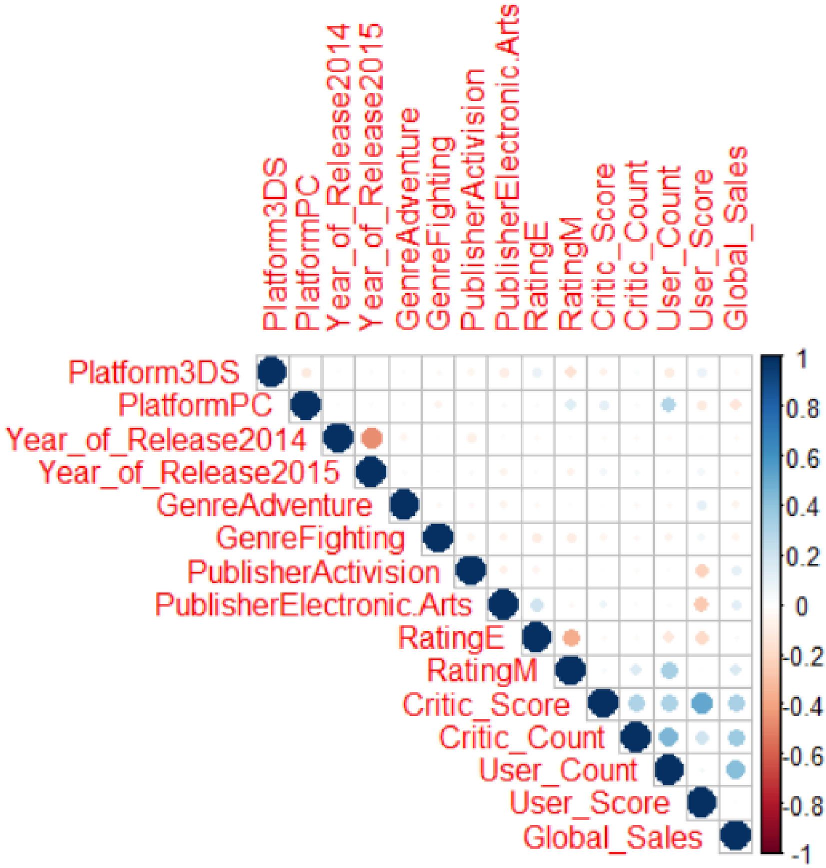

Figure 2:Correlation table

As we claimed user count score critic count score has positive corelation with global sales.In the figure 2 highly correlated values has shown as blue while negative values are red.

2.3 Linear Regression

We trained first 500 rows and tested with the last 249. We fitted our linear regression model by using 38 features selected. After fitting of linear model, R-square of our model turn out be 0.4269, which is not a bad score, but definitely can be improved.

P-value of model is lower than 0.00000000000000022.This means we can reject null hypothesis.
So our model is significant and applicable in real life.Let's examine the coefficients:

Most of the coefficients have high p values and they are not meaningful for our model.However, Critic_Score 0.000000854398296277,

Critic_Count 0.013409051032253660

User_Count 0.000000000006892331 and User_Score 0.0100 does not have high p values therefore they are valid and statistically significant.

But predictors such as Genreplatform is almost 1(0.981) have high p values and they are not good predictors.

When we fit our linear model we obtain estimate and p values. Estimate gives us the estimated coefficients of these predictors. Our constant term is intercept and our intercept value is -1.306

Now let's see marginal effects of predictors on Global Sales by changing them a unit:

∆Global Sales = (even if we increase by a units) a x estimated value of user score from table.

User score values are between 0 and 10.If we increase it by 10 units,10 x coefficient of user score from table will be our predicted global sales.Please note that some of our variables behave as categoric and we can not increase them. We can change their value from 0 to 1 as binary classification.

> coef(summary(lm.fit1))

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.306263224	0.58182989715	-2.245094710	0.025234299871055892
Platform3DS	-0.014664419	0.34490415337	-0.042517374	0.966104630473484161
PlatformPC	-1.282817257	0.30759547048	-4.170468619	0.000036312680413920
PlatformPS3	0.154486641	0.26678008409	0.579078614	0.562818351711444897
PlatformPS4	0.002166646	0.28294520152	0.007657477	0.993893582469817960
PlatformWiiU	-0.228520323	0.35446364849	-0.644693253	0.519445967137337039
PlatformX360	0.083634777	0.30449952013	0.274663083	0.783697873392176558
PlatformXOne	-0.079374356	0.29576683615	-0.268368007	0.788535849584410364
Year_of_Release2014	-0.215866357	0.15877643364	-1.359561691	0.174631839129506777
Year_of_Release2015	-0.449366338	0.18235050331	-2.464299958	0.014091133625783101
GenreAction	0.428565371	0.32277218417	1.327764263	0.184911641401184990
GenreAdventure	0.179130689	0.49112789551	0.364733282	0.715477404514738291
GenreFighting	0.215176116	0.45063996971	0.477490082	0.633238994080425277
GenrePlatform	-0.009345870	0.39692375266	-0.023545757	0.981225105987027835
GenrePuzzle	-0.431185745	0.64415490520	-0.669382071	0.503586159069499217
GenreRacing	0.268094164	0.41113963066	0.652075703	0.514676610682555324
GenreRole.Playing	-0.092374696	0.35081451340	-0.263314921	0.792425254580639504
GenreShooter	0.053963352	0.37102817738	0.145442734	0.884424758758445173
GenreSimulation	0.509440744	0.51338873472	0.992309940	0.321565779490725512
GenreSports	-0.280789440	0.39220447764	-0.715926147	0.474398641866415560
GenreStrategy	-0.130896142	0.49939882760	-0.262107427	0.793355444789412556
PublisherActivision	1.020279091	0.31272745332	3.262518465	0.001186080425990509
PublisherElectronic.Arts	0.648946724	0.26065017894	2.489722919	0.013134888628547246
PublisherNamco.Bandai.Games	-0.031619951	0.31097082384	-0.101681406	0.919053676275598974
PublisherNintendo	0.700156657	0.36672962799	1.909190323	0.056856742232427322
PublisherNippon.Ichi.Software	0.061895511	0.33657526060	0.183897981	0.854174174452801305
PublisherSony.Computer.Entertainment	-0.130105114	0.31992894409	-0.406668782	0.684439593081456854
PublisherTake.Two.Interactive	2.000170306	0.31659390832	6.317778876	0.000000000624831049
PublisherTecmo.Koei	-0.130603955	0.36382250075	-0.358977125	0.719776198181135030
PublisherUbisoft	0.460523879	0.24710041449	1.863711480	0.062996384663342259
PublisherWarner.Bros..Interactive.Entertainment	0.454561867	0.28005469771	1.623118166	0.105246249759086821
RatingE	0.380060700	0.22832961154	1.664526547	0.096685258473642732
RatingM	0.240769042	0.22787068393	1.056603850	0.291244604023918774
RatingT	0.076762685	0.22458699386	0.341794880	0.732660677192782717
Critic_Score	0.033127419	0.00663805962	4.990527461	0.000000854398296277
Critic_Count	0.008936442	0.00360010782	2.482270531	0.013409051032253863
User_Count	0.000503552	0.00007150403	7.042287913	0.000000000006892331
User_Score	-0.157122156	0.06082269537	-2.583281702	0.010092769418722917

Table 1: Linear regression

User score is between 0 and 10, Therefore it's effect on global sales can be calculated as:

Intercept + ß1(assume that is coefficient of User_scores) x 10(we assume given user score is 10).Here's another example:

Let's find out how sales will be affected when a critic gives 10 points to a game.

GlobalsalesΔ = -1.3062 (intercept) + 0.3312 X 10
= -1.306 + 3.312
= 2.006

The predicted global sales increase is 2 million if we change user score value into 10.

Mean Squared Error (MSE) is computed as ≈ 1.538491 while rmse is 1.240359.We will compare this values with other regression m odels later on.First estimated global sales values from our linear model are: 0.9419906 0.6987228 -0.2697085 0.9914792 and 1.9474

2.4 Ridge And Lasso

Ridge regression is an extension for linear regression. It's basically a regularized linear regression model. The λ parameter should be learned as well, using a method called cross validation

Ridge and Lasso techniques are well being used when we have more n umbers of predictors/features than observations. The only difference b etween these 2 techniques are the alpha value .
Lambda is the penalty coefficient and it's free to take any allowed num ber while alpha is selected based on the model you want to try .So if w e take alpha = 0, it will become Ridge and if alpha is 1 our model will be lasso.Now how to decide Lambda?

When we run "Cross Validation" on the data, we get 100 combination of Lambda and their corresponding Mean Squared Error.We can get th e most suitable lambda value by cross validating with bestlam.

2.4.a RIDGE REGRESSION

Ridge shrink the beta coefficient towards zero for unimportant variables .We can use cross validiation to determine our best lambda value which shows how much we decreased the influence of unimportant variables in our model.By bigger lambda values we reduce the weights of variables.We have to fit best lamba value because if we decrease the importance of our predictors too much our model will be more biased with lower varience because we ignore details more and more with bigger lambda values.

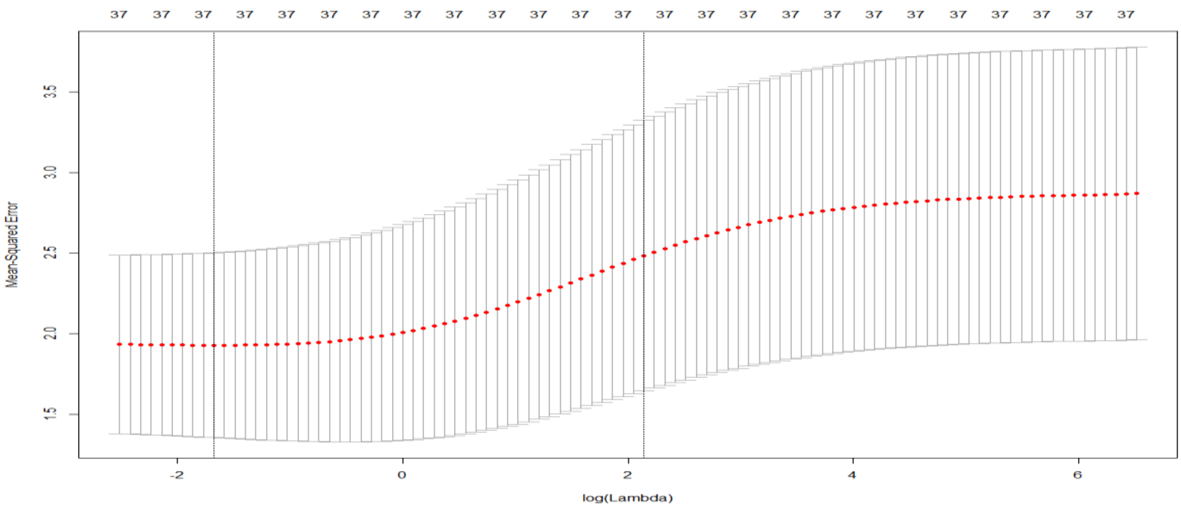

Figure 3: Cross valdation to obtain Lambda for Ridge model

We calculated our best lambda value as 0.18 here.For this lambda we value we get the minimum MSE error in our model.

Ridge

(Intercept)	-1.2448655192
(Intercept)	0.0000000000
Platform3DS	0.0708017225
PlatformPC	-0.9666259019
PlatformPS3	0.2668246905
PlatformPS4	0.1073420496
PlatformWiiU	-0.0917753009
PlatformX360	0.2212165816
PlatformXOne	0.0490160103
Year_of_Release2014	-0.1762864783
Year_of_Release2015	-0.3846073112
GenreAction	0.3108774168
GenreAdventure	0.0187955767
GenreFighting	0.1037499065
GenrePlatform	-0.0842283139
GenrePuzzle	-0.4353056175
GenreRacing	0.1715602920
GenreRole.Playing	-0.1604917022
GenreShooter	0.0190061808
GenreSimulation	0.3181332374
GenreSports	-0.1885416275
GenreStrategy	-0.2153093540
PublisherActivision	0.9161385065
PublisherElectronic.Arts	0.5034534628
PublisherNamco.Bandai.Games	-0.0457702589
PublisherNintendo	0.5496799580
PublisherNippon.Ichi.Software	-0.0059957111
PublisherSony.Computer.Entertainment	-0.1175705375
PublisherTake.Two.Interactive	1.7456894627
PublisherTecmo.Koei	-0.1535115993
PublisherUbisoft	0.3163404528
PublisherWarner.Bros..Interactive.Entertainment	0.3315746720
RatingE	0.2489889296
RatingM	0.1875006253
RatingT	-0.0242800656
Critic_Score	0.0282793979
Critic_Count	0.0092573104
User_Count	0.0004441428
User_Score	-0.1051925615

Table 2: Regression model

In the table shown above we can the values of predictors shrinked into new values in ridge regression.We also calculated Mean Squared Error (MSE) as ≈ 0.944 while rmse is 0.972.

2.4.b Lasso Regression

We can apply same steps for lasso regression.The only difference alpha value will be 1 this time.

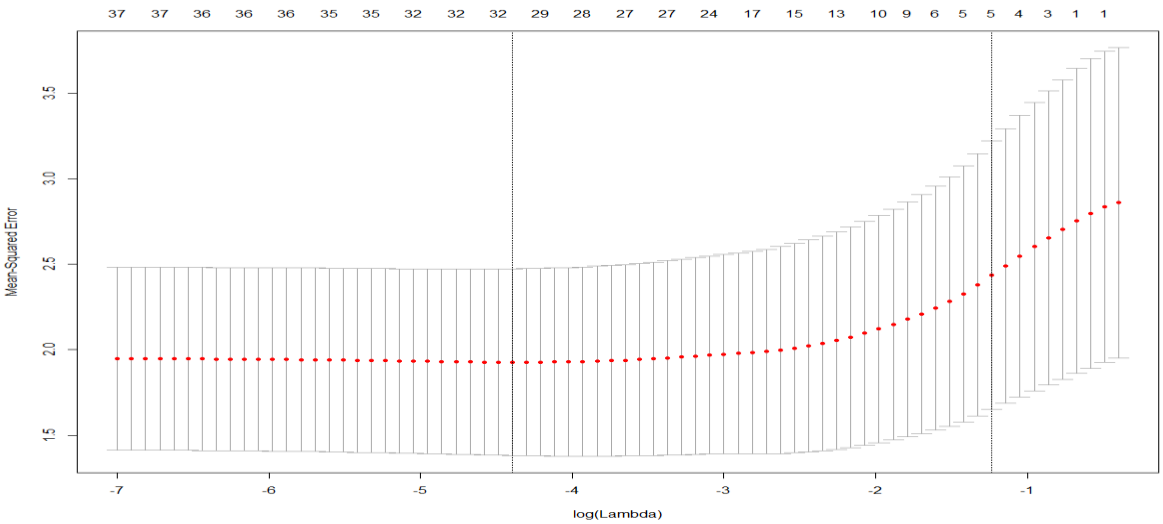

Figure 4: Cross valdation to obtain Lambda for Lasso model

We calculated our best lambda value as 0.012 here. For this lambda we value we get the minimum MSE in our model.

Lasso:

(Intercept)	-1.2171767717
(Intercept)	0.0000000000
Platform3DS	0.0000000000
PlatformPC	-1.1781094056
PlatformPS3	0.1717767679
PlatformPS4	0.0000000000
PlatformWiiU	-0.1231275486
PlatformX360	0.1170921645
PlatformXOne	-0.0050669947
Year_of_Release2014	-0.1732755526
Year_of_Release2015	-0.4007312226
GenreAction	0.3481277904
GenreAdventure	0.0000000000
GenreFighting	0.0867526642
GenrePlatform	-0.0106566060
GenrePuzzle	-0.3209631443
GenreRacing	0.1728869684
GenreRole.Playing	-0.1248482895
GenreShooter	0.0000000000
GenreSimulation	0.3456906677
GenreSports	-0.2183824565
GenreStrategy	-0.1095870797
PublisherActivision	0.9309494855
PublisherElectronic.Arts	0.5428486450
PublisherNamco.Bandai.Games	-0.0061169419
PublisherNintendo	0.5371654940
PublisherNippon.Ichi.Software	0.0000000000
PublisherSony.Computer.Entertainment	-0.0938932874
PublisherTake.Two.Interactive	1.8898716248
PublisherTecmo.Koei	-0.0891837466
PublisherUbisoft	0.3274640547
PublisherWarner.Bros..Interactive.Entertainment	0.3516622931
RatingE	0.2691900931
RatingM	0.1771833178
RatingT	0.0000000000
Critic_Score	0.0312693295
Critic_Count	0.0090359896
User_Count	0.0004909104
User_Score	-0.1362292329

Table 3:Lasso regression

Lasso is eliminating the coefficients while ridge is shrinking them. In the table shown above we can the values of predictors are eliminated into 0 in lasso regression.Thus, lasso is slightly better model than ridge

Mean Squared Error (MSE) is computed as≈0.942 while rmse is 0.970

2.5 DECISION TREE

Since we don’t deal with categorical variables in our regression it is no t so useful to use decision trees.We have 8 terminal nodes before prun ing our tree.

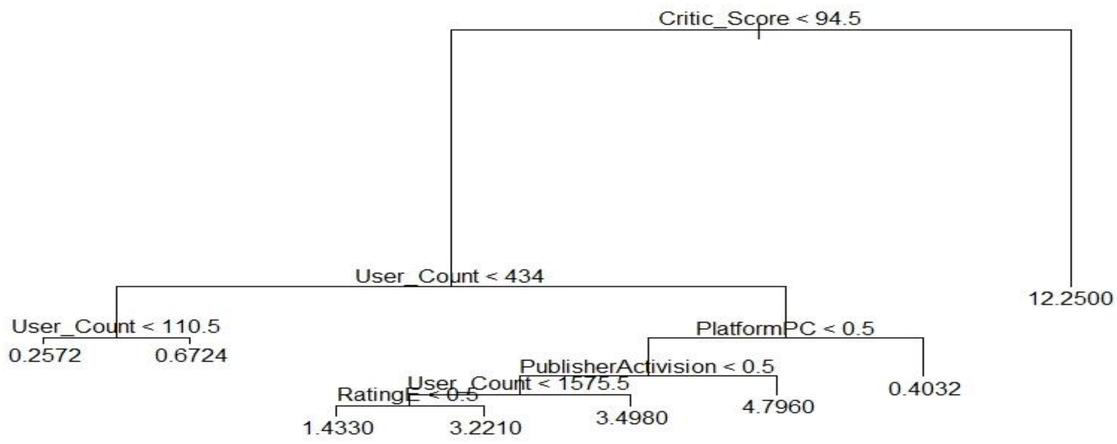

Figure 5:Decision Tree

In the graph related to our decision tree we can see subregions.These t hree regions can be written as ;

R1 = {X | Critic score <94.5},
R2 = {X | Critic score >=94.5, User_count<434},
R3 = {X | Critic score >=94.5, User_count >=434}.
If answer each step as yes or no we can reach terminal nodes.In the en d we will have the mean values of global sales for this subregion .
R7 = {X | Critic score<94.5, User_count<434, User_count<110.5} gives us leftmost terminal node where global sales is predicted as 0.257.

2.5.A DECİSİON TREE PRUNİNG

We decide how many trees should we use by applying cross validation. We can see that we need 8 when we run the cross validation,
but our model already has 8 internal nodes so there is no need to pru ne it

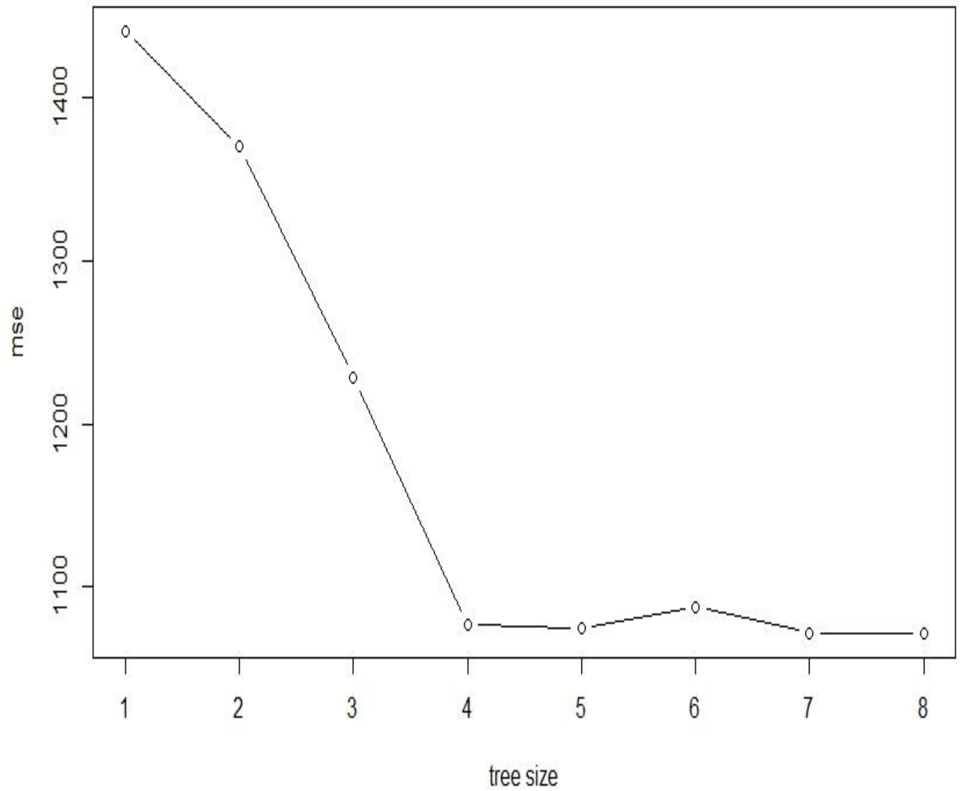

Figure 6: Pruning the Tree

Mean Squared Error (MSE) is computed as ≈ 1.78 while rmse is 1.33 of our decision tree with 8 terminal nodes.

There are some advantages of decision trees it adds visuality and ease of interpretation but still has lower accuracy than linear model,ridge a nd lasso regressions.Therefore, we must use random forest in order to increase efficiency.

2.6 BAGGING AND RANDOM FOREST

2.6.a Baggıng

We can use Bagging (Bootstrap Aggregation) when the variance of a our decision tree is higher than we desire. What we do is create several subsets of data from training sample chosen randomly with replacement. Each set of subset data is used to train their Decision Trees at the moment. We end up with an assembly of different models as an outcome.Average of all the predictions from distinct trees are taking into consideration which has more robustness than a single decision tree. Bootstrapping is choosing random rows, some duplicate while 30% excluded in bootstrapping data to be used as testing set later on.

2.6.b Random forest

Random Forest is an expansion over bagging.It takes one more step where extra to taking the random subset of data, also picks the random collection of features rather than using whole features to build trees. When you lots of random trees, It's a Random Forest. Random forrest is creating lots of trees along random assigned variables.

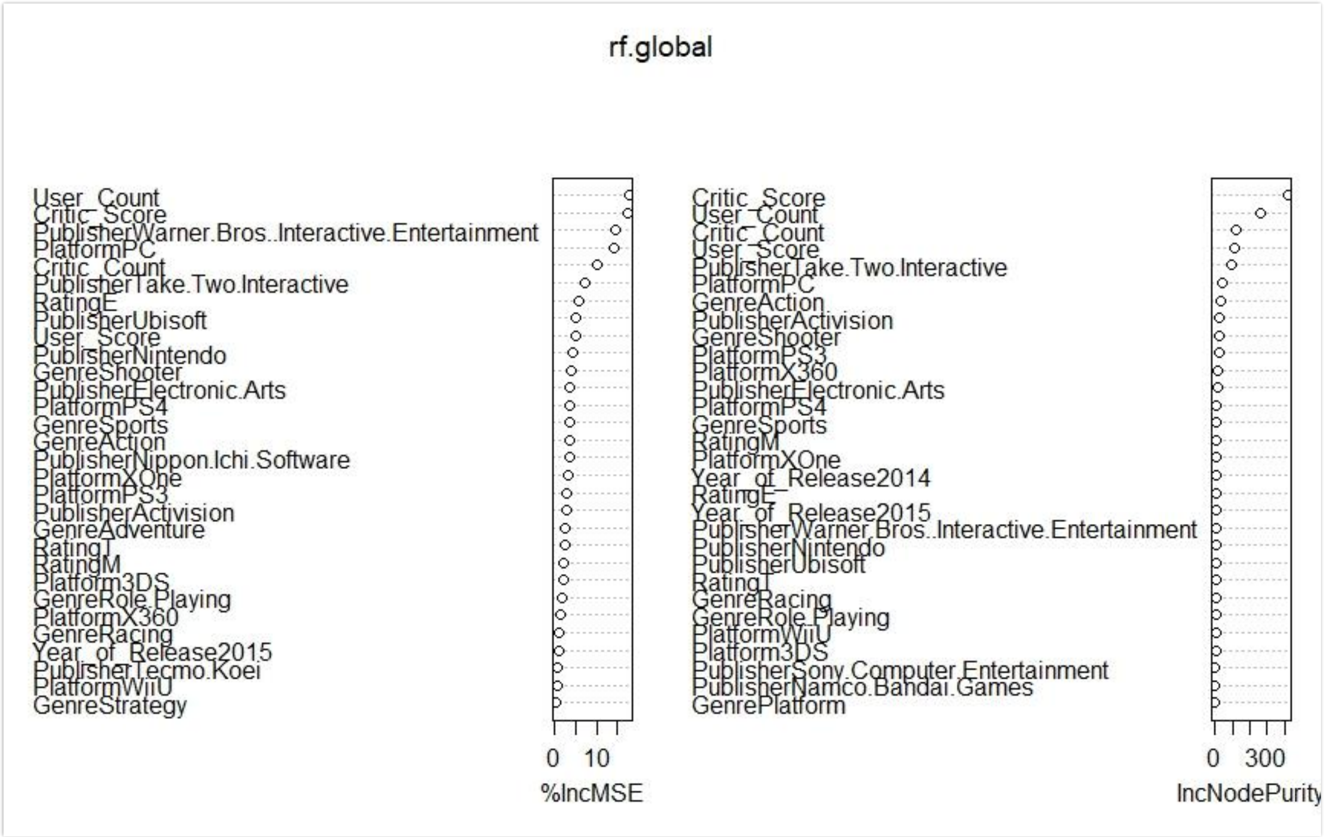

Figure 7:Importance of variables

In the table shown above we can show the importance of the variables and upgrade our model into better one.

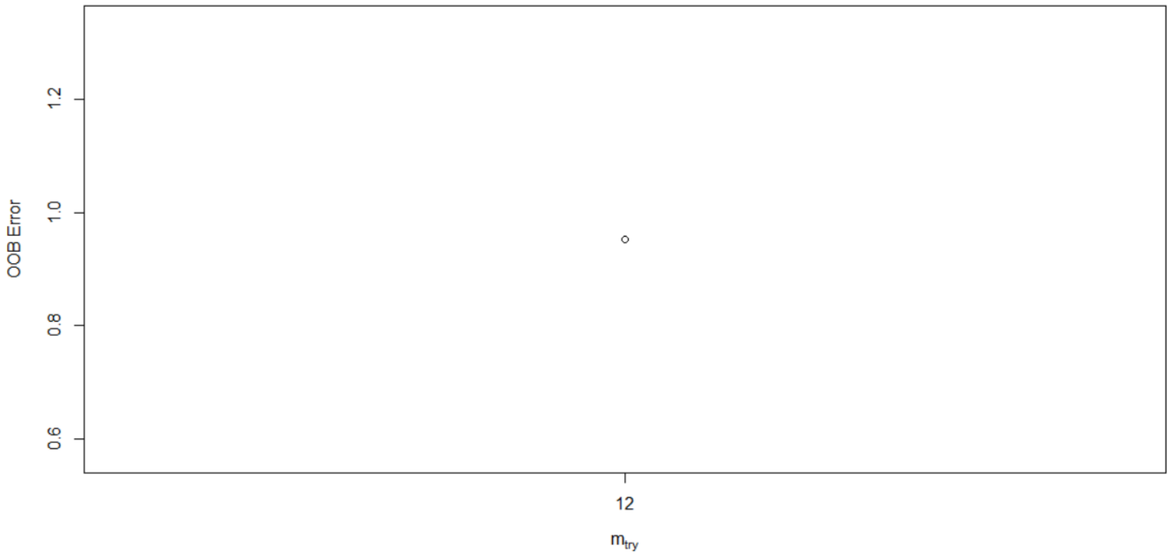

Figure 8: Deciding the mtry

The table shown above show us the best mtry value to use for Random Forest.It is 12 according to our cross-validation results,which is interestingly giving same result with ‘’p=p/3” method where you divide your variable number basically into 3.In our case 38/3 would give us approximately 12.

3. Conclusion

As a conclusion,We can see our dataset give better results with linear regression models.By comparing MSE of linear model with decision tree we can conclude linear model is more desirable.However we made a good progress with using ridge and lasso regressions.On the other hand we reduced MSE by half by using random forest.Lastly to compare our improved models Lasso and Random forest we observe Lasso regression model give us best accuracy amongst all other models.

	LINEAR MODEL	RIDGE	LASSO	DECISION TREE	RANDOM FOREST
MSE	1.53	0.944	0.942	1.78	0.96
RMSE	1.24	0.972	0.970	1.33	0.98

Table 4:Overview of models

4. References:

[1] James,G.,Witten,T.,Hastie,R.,Tibshirani,R. An Introduction to Statistical Learning with Applications in R, (2013)