iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** VI **Month of publication:** June 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Predicting Churn In E-Mall Using Decision Tree

Davinder Paul Singh[1], Vinod Sharma[2]
[1,2]Department of Computer Science & IT, University of Jammu, Jammu

Abstract: ***Different studies and reviews propose that for an organization, economically it is less feasible to connect with another new customer than to hold a current faithful customer. Churn foreseen models are produced by scholastics and professionals to successfully oversee and control customer churning with the aim of holding existing customers. As churn management is an important task for organizations to hold faithful clients, the ability to accurately predict customer churn is important. The present paper proposes a clustering based method to deal with prediction of product churning. In the study to anticipate churning, decision tree has been used to predict churning probability. Comparison has been made out between two Decision Tree algorithms namely C5.0 and Rpart and apart from that Svm, Kernel Svm and Naïve Bayes have also been applied on the dataset. On the basis of performance analysis, conclusion has been made c5.0 suits best in case of having imbalanced Data. Sample dataset provided by Amazon, had been used for the current research work.***
Keywords: ***Customer churn, C5.0, Decision Tree, E-trade industry, Svm.***

I. INTRODUCTION

Churning in E business could be referred to situation when a client purchase item yet return it immediately or return it after a few days from the date of purchase. There could be diverse explanations for the item return. Some of them could be cost issues, size issues, product quality or even feedback from different clients who purchased the same product. If a client immediately returns the item after buying, than the particular kind of churning comes under discrete churning, whereas the returning item after a few days may fall under classification of partial churning.[3]Churning affects the business in the sense that if churning rate is high than company has to spend extra money on either advertising campaigns to engage new customer or giving offers to old customers to retain them .Thus higher rate of churning hinders the growth of organization. The development of current project has been carried out in R programming language. R programming language is most commonly used by statistics and analyst for interpretation and analysis of data. Various data mining techniques have been exploited and proposed to forecast the customer churn. Several techniques are available to predict churning, most widely used of them are: neural networks, support vector machines, classification models and logistic regression models. The present research work considers Decision Tree method to design a model from stored customer data to predict churning. Besides decision tree Svm, Kernel Svm and naïve bayes classifier also have been applied on the dataset. The C5.0 and Rpart algorithm has been applied on the model to get the results along with Svm, Kernel Svm and Naïve bayes .C5.0 algorithms take sample of dataset as its input. This algorithm considers the attribute of the data that will result in effective splitting of its samples into subsets such that each subset belongs to one cluster or the other. Parameter which forms the basis of splitting is the normalized information gain. The working of the algorithm includes the selection of attribute with the highest normalized information gain to take decision. The same algorithm is applied in recursive manner on other subsets. Rpart in R is a package that provides a number of functions which are related to regression tree or classification tree and that is used for classification of problems. Svm is used to treat linearly separable data and when there is no defined boundary to split the data and data is not linearly separable kernel svm is used. Naïve bayes classifier is based on bayes theorem and this is probabilistic classifier because we first calculate the probabilities and then we assign the classes to our new data point based on the probabilities whichever is greater. All algorithms that we have used in our research are supervised learning algorithms. The remaining paper is in written as per the layout. Segment II audits the present literature survey, identified with customer churn and diverse information mining systems used to anticipate churn in distinctive studies. Proposed research methodology is described in section III and Section IV shows the exploratory results on continuous dataset. At long last, conclusion is given in Section VI.

II. LITERATURE REVIEW

Various literature surveys and paper has been published regarding churn management. Most of the previous work is done on data set related to telecommunication industry. However different techniques have been implemented by various authors to make models to predict churning.[1] Anuj Sharma and Dr Prabin Kumar (2011) in their paper have discussed “neural network approach” for churn prediction.Tool used by the author is SPSS clementine 12.0 (Data mining software package) and the Technique being implementing is quick applied method of SPSS. However the technique discussed in paper for preparing churn model is applicable only when

sufficient attributes of data set are available. [2] Chih-Fong Tsai and Yu -Hsin Lu (2009) in their paper have discussed an approach for combining two algorithms of neural networks. The authors have made use of combination of artificial neural networks and Self organized maps.[6]Hussain Rahman (2014) in their paper "churn analysis predicting churn" have made their analysis by focusing on rule based classification for churn prediction in Telecom Company. The tool they have used is Rapid Minor and the technique used by them is Decision Tree. In this paper authors have illustrated the challenges faced in churn. analysis like selecting the decision variable, designing of rules for developing churn model and selection of proper data set. However due to insufficient data, authors have been unable to produce sufficient analysis.[7]Luo Bin,juan (2007) in their paper have made use of Decision tree algorithms like CHAID, CART, C4.5 to predict customer churning. In this paper author have thrown light on benefits of decision tree algorithm such as decision tree require less effort from user for data preparation (no normalization is required) and nonlinear relationship between parameter do not effect

III.RESEARCH METHODOLOGY

Dataset: For the current Research work, the dataset provided by Amazon has been exploited, which concentrates on the assignment of customer churn prediction. Specifically dataset contains information of 74507 clients. Attributes of datasets include age, reason, rating, price, category, customer Id.

Algorithm used: - Usually when Datasets having large number of dimensions is converted into learning model, they produce results which are over fitted. In order to avoid the problem of over fitting, method of feature selection is followed .Feature Selection is a technique in which subset of only relevant features are selected .Feature selection as a rule gives better learning execution. As size of data set is reduced, thus feature Selection also result in minimizing computation cost. In order to reduce the error rate, it is usually advised to remove the parts of the model that depict illegitimate impacts in the training model as opposed to genuine elements. Reduced Error Pruning is such technique which eliminates the section of tree, thus making reduction in the size of tree. Reduced Error Pruning likewise brings about reducing complexity and giving better and accurate results.

A. Algorithm

Step1: A root node is created to make tree.
Step2: Check the base case.
Step3: Apply Feature Selection technique.
Step4: Best Tree = Construct a decision tree using training data.
Step 5: Partition all Training Data into N disjoint set R = R1, R2, …,. RN.
Step 6: For each k = 1,…N, Do
Step 7: Test dataset = Rk.
Step 8: Training dataset = R – Rk.
Step 9: Using Training set, make the decision tree.
Step 10: Decide the performance accuracy Xj.

with the use of Test set predict test set with decision tree and get the prediction error for each data subset. To start with dataset given by Amazon is used. The preprocessing stage involves the removal of columns which were not having so much useful in data analysis like customer Id, product title. The dataset obtained after preprocessing whole is divided into small subset and each subset again divided into two parts namely training dataset and testing dataset.

IV.RESULTS AND ANALYSIS

Tree shown in Figure1 is decision tree obtained after applying C5.0 algorithm. The leaf nodes symbolizes the final conclusion, whether customer will churn or not .The decision tree considers "age" attribute as decision making variable. Based on sample data, Decision Tree considers 36 as the threshold value of age If age of customer is greater than 36, then the observation is made that customer is least likely to churn. Although if age is lower than 36, then age is compared with another threshold value which is selected as 22. So if age of customer is less than 22, again he is least likely to churn, otherwise the age is compared with other threshold value. From the decision tree thus obtained, the conclusion thus made is that customer that will mostly churn falls under age group of <22-24> and <30-36>.

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering

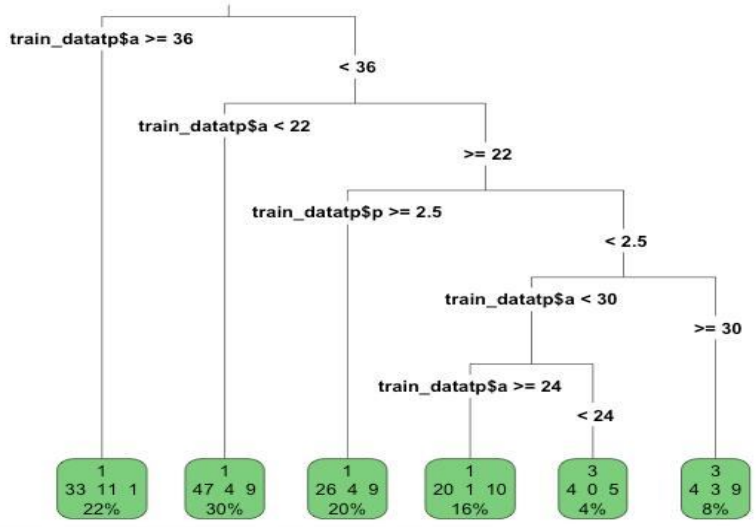


Figure 1. Decision Tree for customer churn.

For the current research work, variable “no churn” is given the corresponding value 1, ”partial churn” is given the corresponding value 2 and “churn” is given the corresponding value 3.The figure shown above is decision tree model for predicting the reason for churning . If value is less than 1.5, the customer is least likely to churn, otherwise value may lie between 2 and 3 for possible churner customers. If value lies in the range of 2 and 2.5 and age of customer is less than 36, then customer will likely churn due to mismatch in product. Similarly if age is more than 36, size would be the primary reason for churning. In same way, if value lies between 2.5 and 3, then customer will definitely churn. However, the reason for churning may vary depending on the age of customer. For the customers having age less than 24, color would be main reason for product churning and size is the main reason for product churning by customers having age greater than 36.

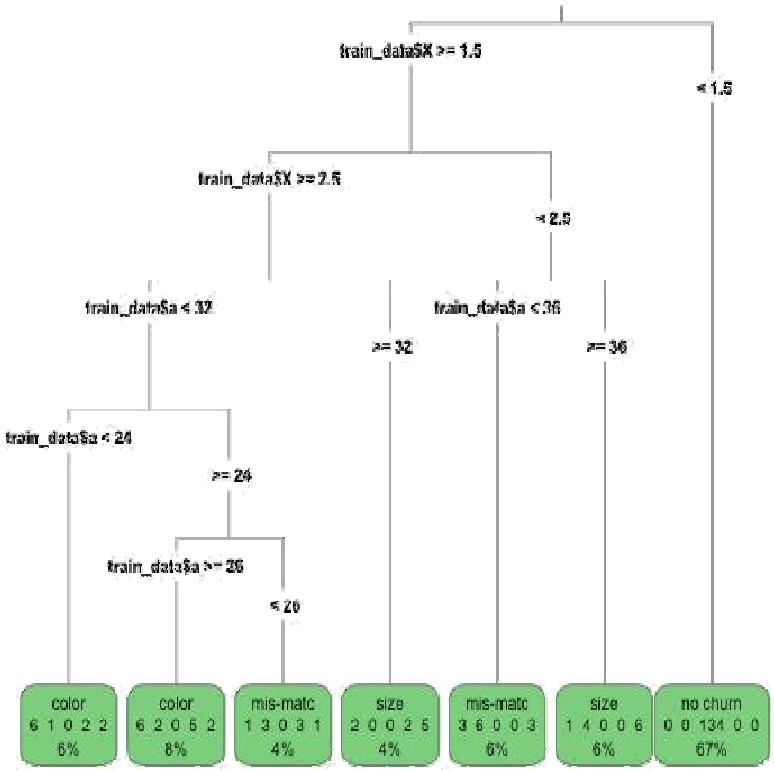


Figure 2. Decision Tree for product churn.

Volume 5 Issue VI, June 2017
ISSN: 2321-9653

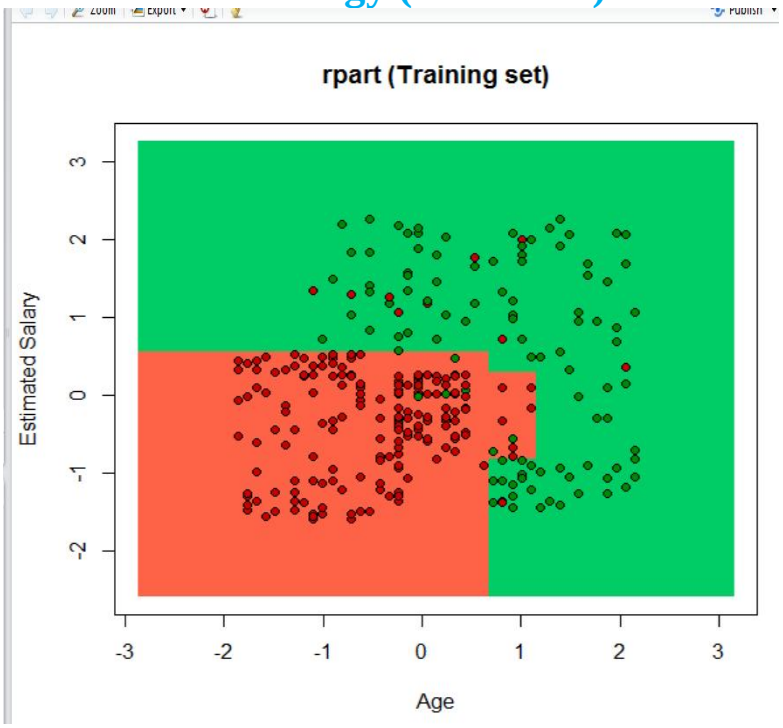


Figure 3. Rpart Training set.

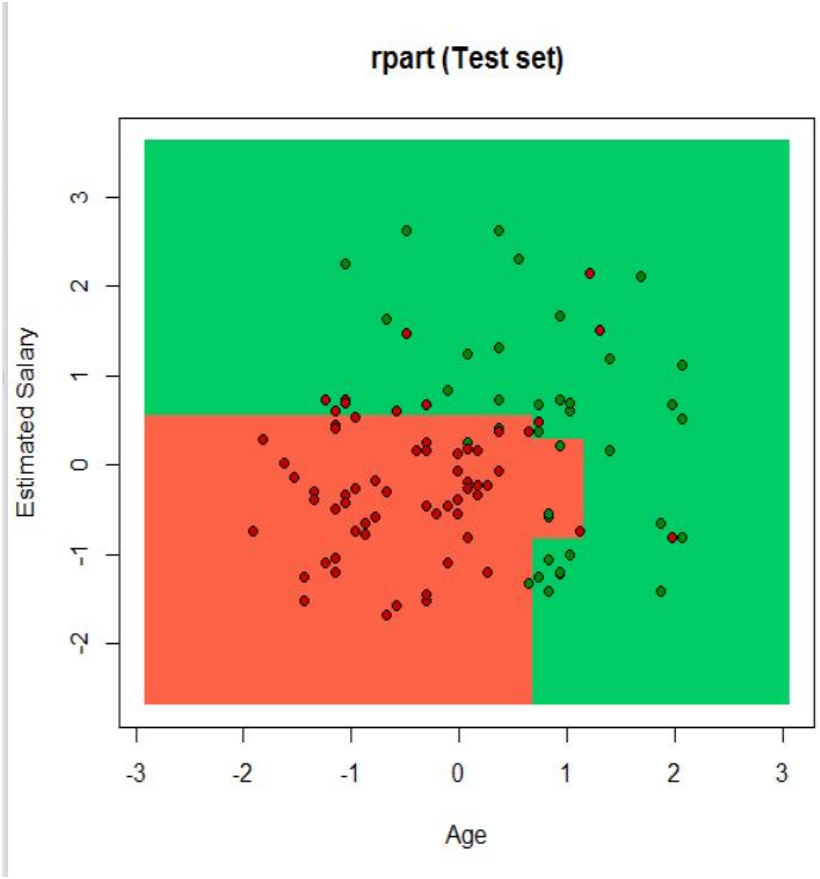
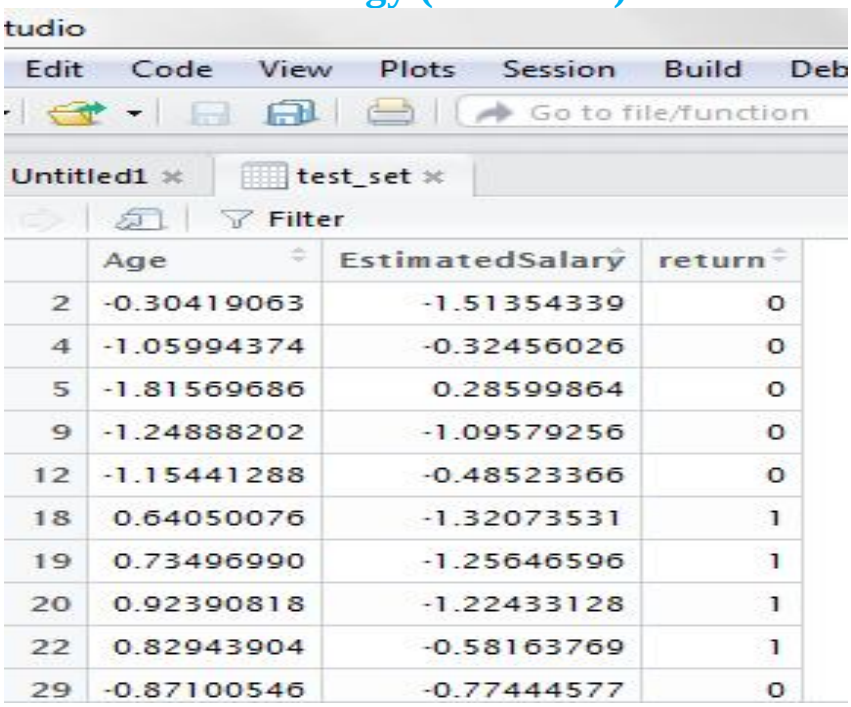


Figure 4. Rpart Test set.

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)



	Age	EstimatedSalary	return
2	-0.30419063	-1.51354339	0
4	-1.05994374	-0.32456026	0
5	-1.81569686	0.28599864	0
9	-1.24888202	-1.09579256	0
12	-1.15441288	-0.48523366	0
18	0.64050076	-1.32073531	1
19	0.73496990	-1.25646596	1
20	0.92390818	-1.22433128	1
22	0.82943904	-0.58163769	1
29	-0.87100546	-0.77444577	0

Figure 5. Test set observations.

```
> View(test_set)
> y_pred
  2   4   5   9  12  18  19  20  22  29  32  34  35  38  45  46  48  52  66  69  74  75  82  84
  0   0   0   0   0   0   1   1   0   0   1   0   1   0   0   0   0   0   0   0   1   0   0   1
 85  86  87  89 103 104 107 108 109 117 124 126 127 131 134 139 148 154 156 159 162 163 170 175
  0   1   0   0   1   1   0   1   1   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0
176 193 199 200 208 213 224 226 228 229 230 234 236 237 239 241 255 264 265 266 273 274 281 286
  0   0   0   0   1   1   1   0   1   0   0   1   1   0   1   1   0   0   1   1   1   1   1   1
292 299 302 305 307 310 316 324 326 332 339 341 343 347 353 363 364 367 368 369 372 373 380 383
  1   0   0   0   1   0   0   1   0   1   0   1   0   1   1   0   0   1   1   0   1   0   1   1
389 392 395 400
  1   1   0   1
Levels: 0 1
> cm
   y_pred
     0  1
  0 53 11
  1  6 30
```

Figure 6. Rpart no. of correct and incorrect observations.

Here in Figure 3 and Figure 4 red region specifies customers which will churn and green region specifies customers which will not churn. These two regions are called as prediction regions. Although we see here that some of the red dots are in green region and some of the green dots are in red region also this tend to happen because we cannot get 100% accuracy. Figure 5 specifies test set observations and we need to compare our incorrect and correct predictions which we get in Figure 6 with our test set observations and this comparison we make on the basis of dependent variable which is return in our dataset because in any machine learning algorithm we need to predict dependent variables on the basis of independent variables. Now when we compare our predictions with our test set observations we get 53+30=83% correct predictions are there and confusion matrix helps us to find these.

Volume 5 Issue VI, June 2017
ISSN: 2321-9653

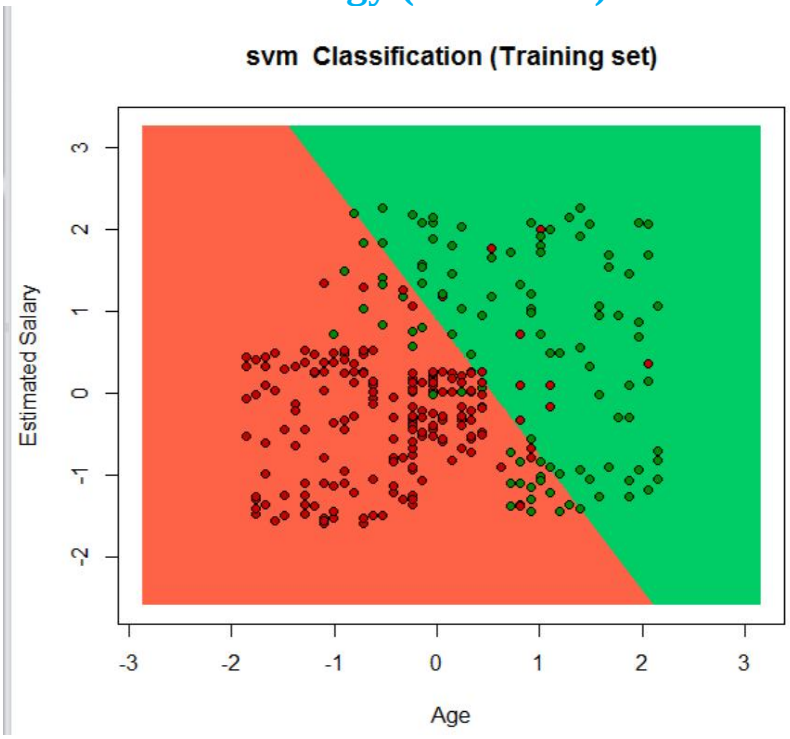


Figure 7. Svm Training set.

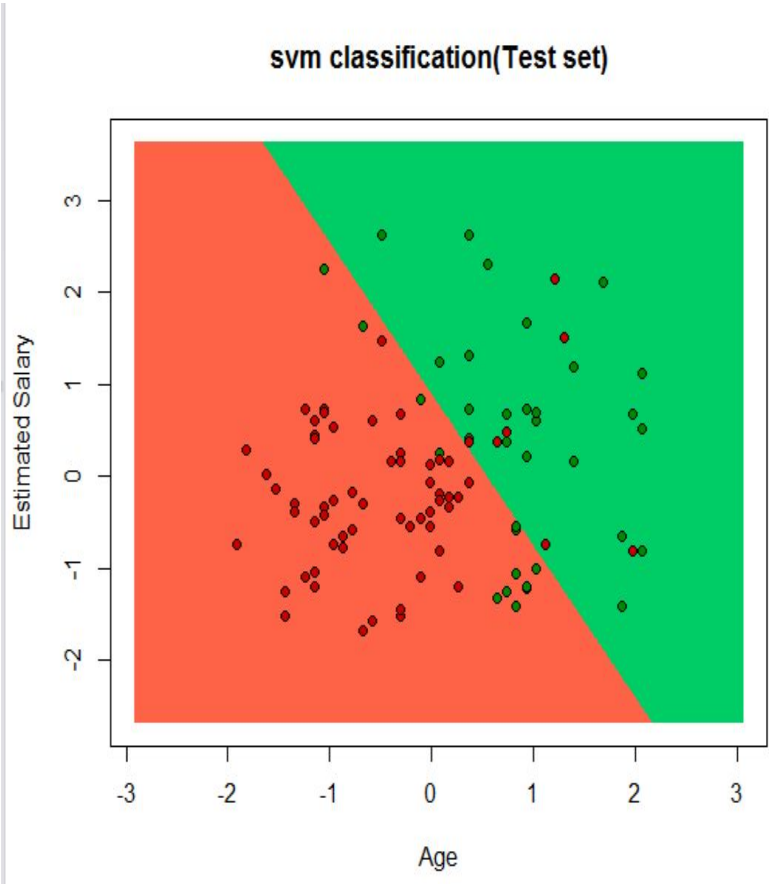


Figure 8. Svm Test set.

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

```
> y_pred
  2   4   5   9  12  18  19  20  22  29  32  34  35  38  45  46  48  52  66  69  74
  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
 75  82  84  85  86  87  89 103 104 107 108 109 117 124 126 127 131 134 139 148 154
  0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0
156 159 162 163 170 175 176 193 199 200 208 213 224 226 228 229 230 234 236 237 239
  0   0   0   0   0   0   0   0   0   0   1   1   1   0   1   0   1   1   1   0   1
241 255 264 265 266 273 274 281 286 292 299 302 305 307 310 316 324 326 332 339 341
  1   1   0   1   1   1   1   1   0   1   1   1   0   1   0   0   0   0   1   0   1
343 347 353 363 364 367 368 369 372 373 380 383 389 392 395 400
  0   1   1   0   1   1   1   0   1   0   1   1   0   0   0   0
Levels: 0 1
> cm
   y_pred
     0  1
  0 57  7
  1 13 23
```

Figure 9 Svm no. of correct and incorrect predictions.

Now when we compare our correct and incorrect predictions of Figure 9 with our test set observations which we see in Figure 5 we find that there are 57+23=80% correct predictions are there in case of Svm.

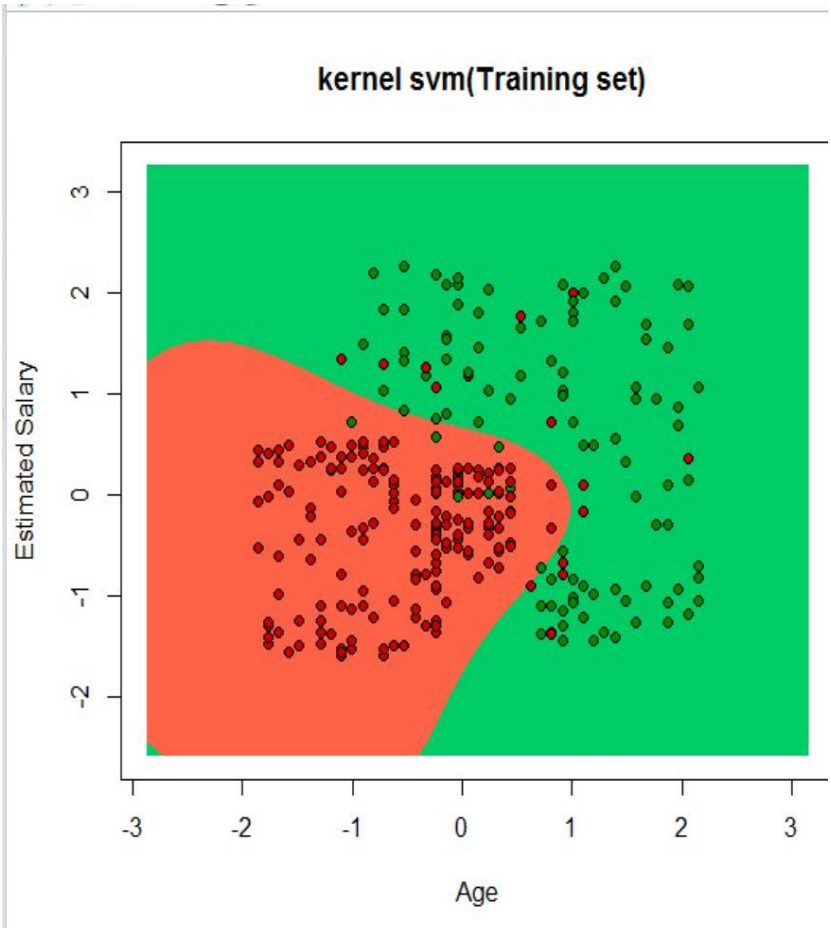


Figure 10. Kernel Svm Training set.

Volume 5 Issue VI, June 2017
ISSN: 2321-9653

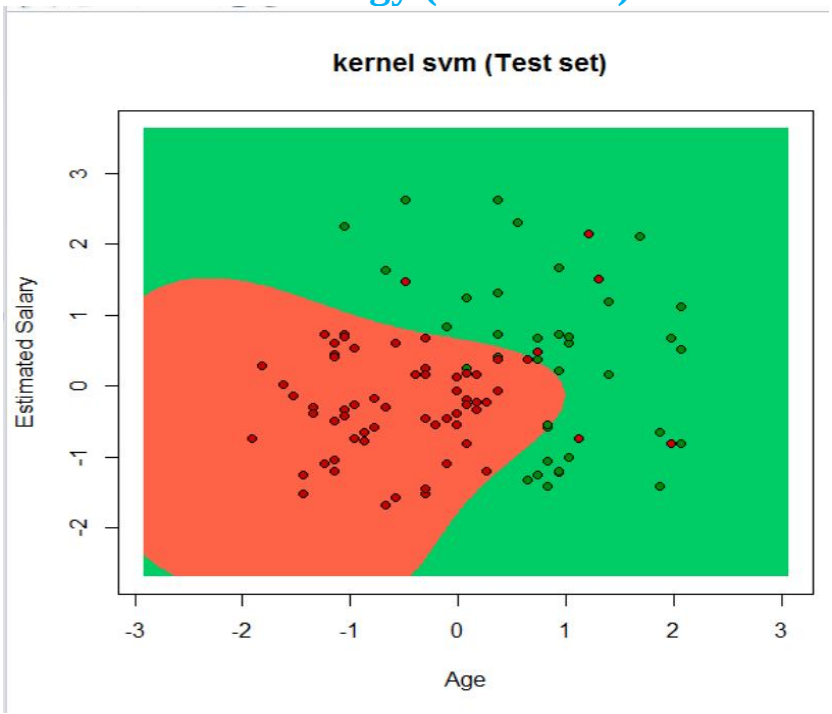


Figure 11. Kernel Svm Test set.

```
> View(test_set)
> y_pred
  2   4   5   9  12  18  19  20  22  29  32  34  35  38  45  46  48  52
  0   0   0   0   0   1   1   1   0   0   1   0   0   0   0   0   0   0
 85  86  87  89 103 104 107 108 109 117 124 126 127 131 134 139 148 154
  0   1   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0
176 193 199 200 208 213 224 226 228 229 230 234 236 237 239 241 255 264
  0   0   0   0   1   1   1   0   1   0   0   1   1   0   1   1   1   0
292 299 302 305 307 310 316 324 326 332 339 341 343 347 353 363 364 367
  1   0   1   0   1   0   0   1   0   1   0   1   0   1   1   0   0   1
389 392 395 400
  1   1   0   1
Levels: 0 1
> cm
   y_pred
     0  1
  0 58  6
  1  4 32
```

Figure 12. Kernel Svm no. of correct and incorrect predictions.

Volume 5 Issue VI, June 2017
ISSN: 2321-9653

On comparing our correct and incorrect predictions of Figure 12 with our test set observations which we see in Figure 5 we find that there are 58+32=90% correct predictions are there in case of Kernel Svm.

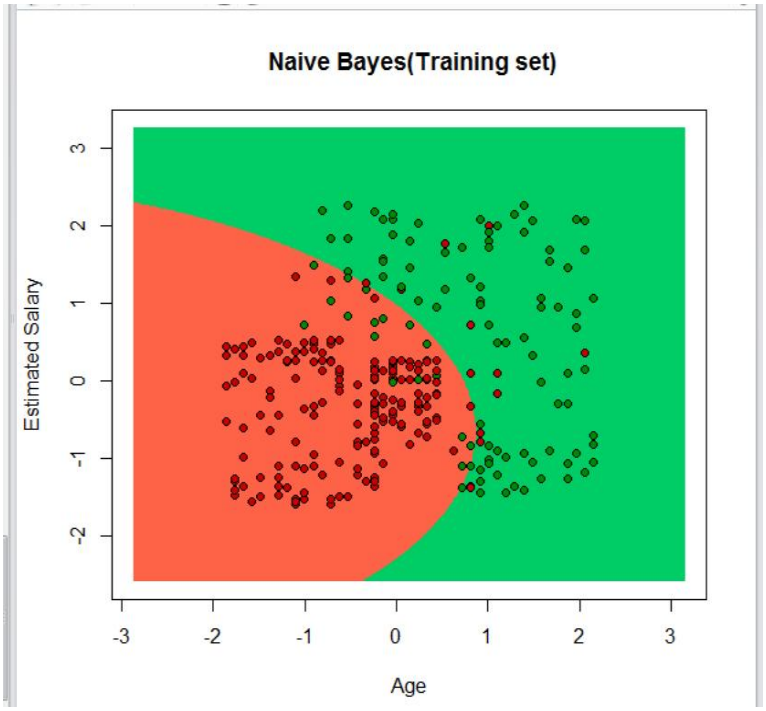


Figure 13. Naïve Bayes Training set.

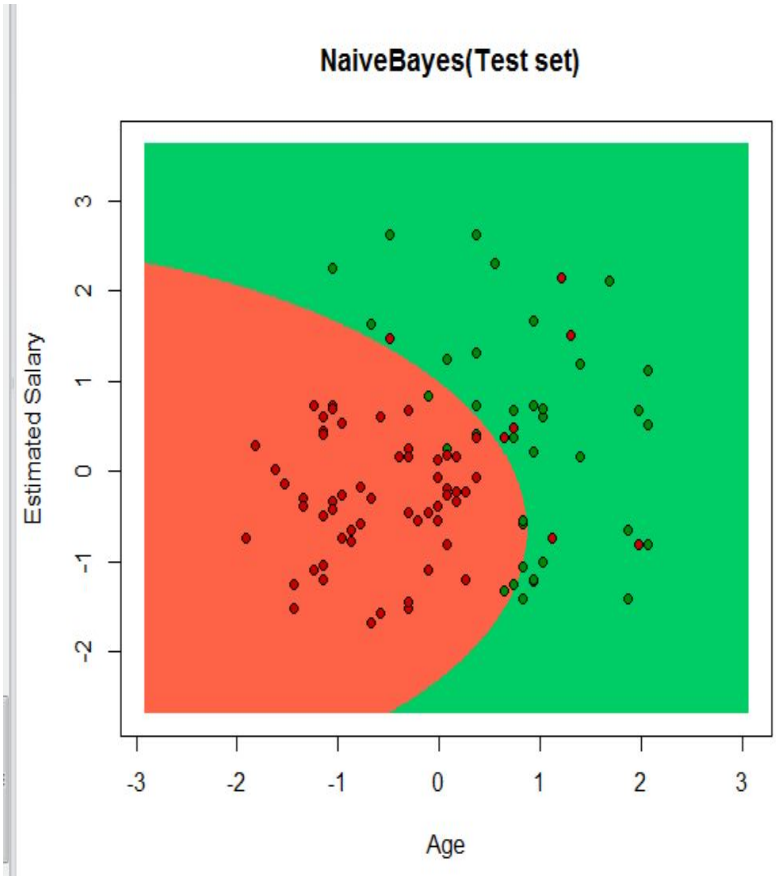


Figure 14. Naïve Bayes Test set.

Volume 5 Issue VI, June 2017
ISSN: 2321-9653


```
> cm
   y_pred
     0  1
  0 57  7
  1  7 29
> y_pred
  [1] 0 0 0 0 0 0 0 1 0 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0
 [40] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 1 0 0 1 1 0 1 1 1 0 1 1 1 1 1 0 1 1 1 0 1 0
 [79] 0 1 0 1 0 1 0 1 1 1 0 0 1 1 0 1 0 1 1 1 1 0 1
Levels: 0 1
>
```

Figure 15. Naïve Bayes no. of correct and incorrect predictions.

Now in case of Naive Bayes we get 57+29=86% correct predictions on comparing with our test set observations of Figure 5.

```
Console ~/

          Size      Errors

            16  173( 6.9%)   <<


           (a)   (b)    <-classified as
          ----  ----
          2134    18    (a): class False.
           155   192    (b): class True.


        Attribute usage:

        100.00% purchaseLimit
         93.76% Color
         19.77% EcommerceRating
          9.16% Frequency_of_shopping
          7.52% Eve.Mins
          2.24% Age.1
          0.88% Size


Time: 0.1 secs
```

Figure 16. C5.0 Error rate.

Volume 5 Issue VI, June 2017
ISSN: 2321-9653

Here as we saw in Fig 16 when we calculate Error rate in case of C5.0 it comes out to be 6.9%. Therefore we get 93.1% prediction accuracy in case of C5.0.

V. TABLE

Name of Classifier	Accuracy
Rpart.	83%
Svm.	80%
Kernel Svm.	90%
Naïve Bayes.	86%
C5.0.	93.1%

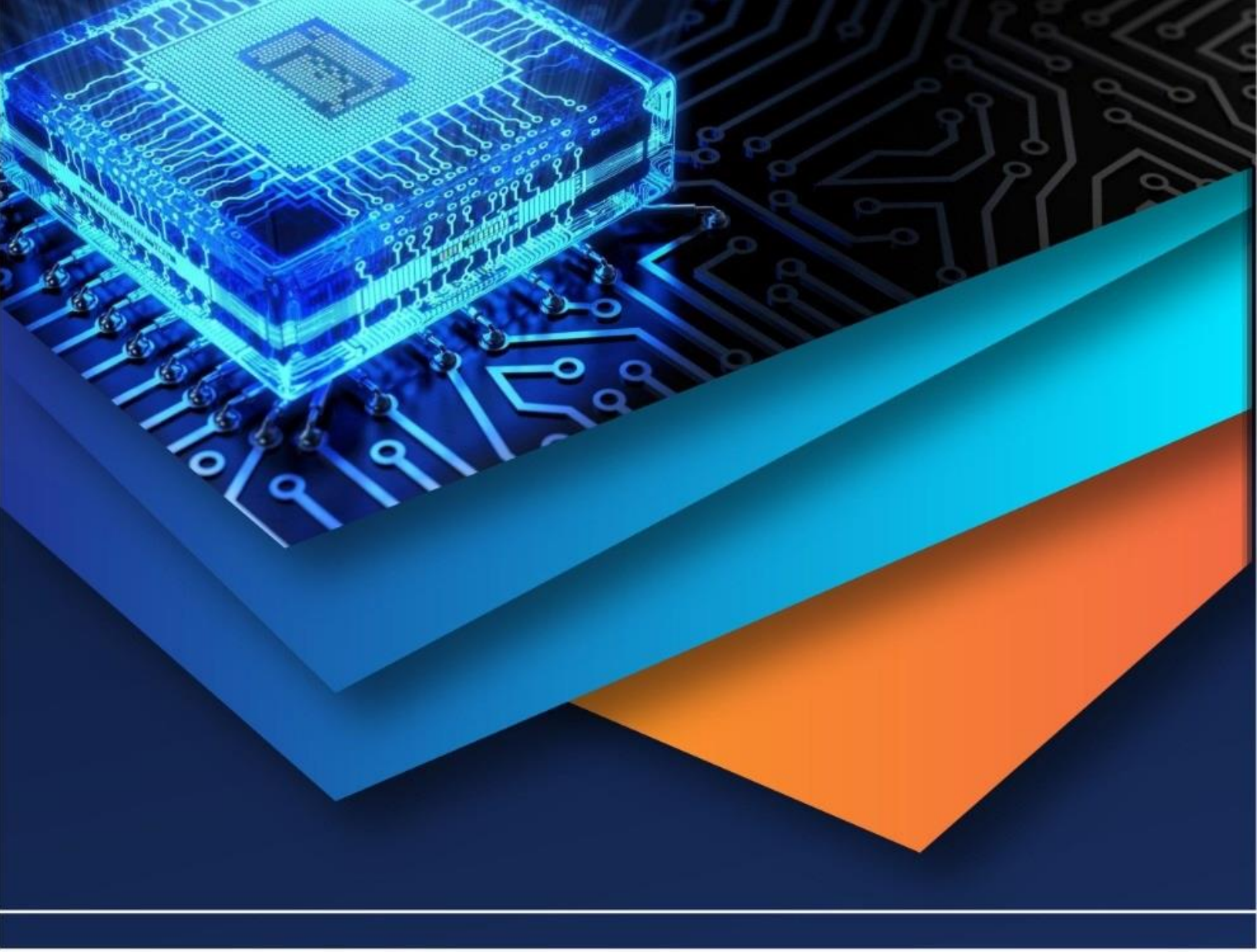
Table 1. Predicted accuracy of different classifiers.

VI. CONCLUSION

In this paper, after applying both C5.0 algorithm, Rpart algorithm on decision Tree and Svm, Kernel Svm and Naïve Bayes following conclusion can be made that C5.0 algorithm works better than its counterparts and it gives better results with less error rate and with more accuracy. C5.0 giving better results with more accuracy than other classifiers because C5.0 makes prediction model by considering more number of attributes but on the other hand we can take only take specified no. of attributes in other classifiers which we have used in our paper. We can take more no. of attributes in order to develop prediction model in other classifiers also instead of C5.0 but all these classifiers will take more time and requires lots of processing in order to produce accurate results. From the future examination perspective, the clear augmentations to be researched is to investigate more powerful estimators to enhance the future prediction by introducing some soft computing technique via some Machine Learning algorithms like Artificial neural network, fuzzy logic, Back propagation algorithm, and many more.

REFERENCES

[1] Anuj Sharma and Dr. Prabin Kumar, "A Neural Network based Approach for Predicting", International Journal of Computer Applications vol. 27, no. 11, pp. 26-29, 2011
[2] Chih-Fong Tsai and Yu -Hsin Lu, "Customer churn prediction by hybrid neural networks Science Direct", vol. Expert Systems with Applications 36, pp.1-7,2009.
[3] Dr. M. Balasubramaniam, M. Selvarani, "Churn prediction in mobile telecom system using data mining techniques", International Journal of Scientific and Research Publications, Volume 4, Issue 4,2014
[4] Gupta, S. and Kim, H.W, "Linking Structural Equation Modeling to Bayesian Networks Decision Support for Customer Retention in Virtual Communities", European Journal of Operational Research, 190, 818-833, 2008
[5] Idris, A., Rizwan, M. and Khan, A, "Churn Prediction in Telecom Using Random Forest and PSO Based Data Balancing in Combination with Various Feature Selection Strategies", Computers & Electrical Engineering, 38, 1808-1819, 2012
[6] M. R. R. Hussain, "Churn analysis Predicting churners", Digital Information Management (ICDIM), vol. 7, no. 3, pp. 237 – 241, 2014
[7] S. Peiji and L. Juan,"Customer Churn Prediction Based on the Decision Tree in Personal Handyphone System Service", Service Systems and Service Management vol. IV, no. 10, pp. 1-5,200
[8] Saad Ahmed Qureshi, Ammar Saleem Rehman, Ali Mustafa Qamar, Aatif Kamal, Ahsan Rehman, "Telecommunication Subscribers' Churn Prediction Model Using Machine Learning", IEEE International Conference on Digital Information Management (ICDIM) pp. 131–136,2013.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY