A Comparative Analysis Using RStudio for Churn Prediction

Vani Kapoor Nijhawan, Mamta Madan, Meenu Dave

Abstract: ***With the availability of numerous data, in each and every sphere, it has become significant to analyze the voluminous data, and utilize the generated patterns for the future predictions. This is what we refer to as data mining. This paper exploits, decision tree technique, to predict churning trends of telecom users. For this study, authors are making use of R and its GUI Rattle.In this paper, the focus is, to compare the variations in churning patterns of a number of users, based on the reflections made by different variables or factors and then make the predictions thereafter.***

Keywords:Data mining, Decision Tree, Customer churn, RStudio, Rattle

I. INTRODUCTION

Churning of customers is an area of concern in almost all spheres. Telecommunications, is one such field. Here, the users of mobile phones, keep on switching their providers every now and then. There are enormous reasons for the same. Out of the many factors, in the dataset, some have a strong impact on the churning and others have least or no effect. In this paper, the authors would like to explore the same and would like to highlight, a few factors, responsible, for the churning of telecom customers and identifying the least affecting one. This analysis would be done in RStudio using Rattle, as the tool, for generating the decision trees.

Decision Tree

Decision tree is that the flow diagram representing a tree in top- down fashion, ranging from root node and moving towards the terminal nodes. Here, the inner nodes represent a take a look at or call and also the branches represent the outcomes. it's a well-liked technique attributable to its simplicity. [1].

Software Tools Being Used R

R is a very popular open source statistics and graphics including hundreds of additional packages available for free, which are very useful for providing help in data mining, machine learning and statistics. [w1]

R: Rattle package

Rattle stands for, R Analytic Tool To Learn Easily (Rattle). Rattle serve as an interface to R in helping the user to load data directly from a CSV file (or using ODBC), transform it (if required), explore the data, apply various techniques, build models and analyse the results. [w2]

II. RELATED WORK

S. Hussain et al.[2], have used data processing within the field of education to predict the performance of under-graduates and used R for analysis. the explanation for exploitation R is that, R provides heaps several applied mathematics techniques within the style of varied packages for modeling, analysing, clustering, classification etc. Also, the programmers having data of C, C++, Java, .NET or Python will write their own code to govern the R objects. apart from this, R contains graphical packages conjointly to supply sensible quality graphs. [3]
Wonhee Cho et al.[3] in their paper on huge knowledge have compared and analysed interactive internet packages with R visual image packages. They additional that R has been improved for a giant knowledge analysis and mining tool. it's supported with multiple packages for various targets with visual image.

J. Rao, R. Kelappan, P. Pallath [4] in their paper tried to use, data processing ideas like regression, classification and hybrid approaches to effectively set up the work allocations within the designing section of package development exploitation R- scripts. These results can be accustomed predict the additional designing of the task and facilitate the management to achieve higher visibility.
Authors in [5] conjointly found that R is one among the foremost widespread languages within the knowledge science, applied mathematics and machine learning community. They additional that several knowledge scientists ar hindered by its limitations of obtainable functions to handle giant datasets with efficiency. So, the authors mentioned here regarding the solutions like providing a public code repository that attendees are ready to access and adapt to their own observe.
Manpreetkaur and Dr Prerna [6] in their paper, aforementioned that if the info already obtainable with the medium firms if analysed rigorously, will throw some light-weight on churning patterns of the purchasers. This info are often used for the present and approaching customers to style the retain policies.

III. METHODOLOGY

The process of mining, the gathered dataset in rattle needs to undergo some basic steps of data cleansing and pre-processing.

Revised Manuscript Received on May 28, 2019.
Vani Kapoor Nijhawan, Assistant Professor, VIPS, GGSIPU Delhi
MamtaMadan, Professor, VIPS, GGSIPU Delhi
Meenu , VIPS, GGSIPU Delhi

Data Collection and Description

For collecting the dataset, a questionnaire designed online and circulated amongst a number of mobile phone users, belonging to different age groups, states, having different service providers and different types of payment plans. With this survey, around 300 records were gathered and analysed in this paper.

Data Pre-processing

Pre-processing of data means to clean the data by removing missing values and also removing the irrelevant data from the collection..

IV. IMPLEMENTATION AND OBSERVATIONS

The data file collected from survey, was given as input to RStudio, after pre processing steps. In RStudio, package rattle is being used as a GUI for generating the corresponding decision tree and then analyzing it.

Attributes in Dataset

The attributes in the dataset used are the following:-

Table 1: List of attributes in the used dataset.

S.No.	Attribute Name
1	Age
2	Gender
3	State
4	No. of connections
5	Provider
6	Payment plan
7	Duration (how long the connection is being used)
8	Type of services Used
9	Churn

Steps in Rattle to generate Decision Tree with all attributes:

Step 1: Browse the data file in source tab.

Step 2: Go to execute tab on the top left corner , to see the file variables and verify the attribute types . Make sure that churn should be selected as the Target attribute.

Step 3: Then click on model tab and select tree option and click on execute to see all the decision rules for the tree.

Step 4: Last step is to click on draw button towards the middle right , and view the generated decision tree in RStudio.

Generated Decision Tree

Here, all the fields have been set as input fields. The tree generated is shown below.

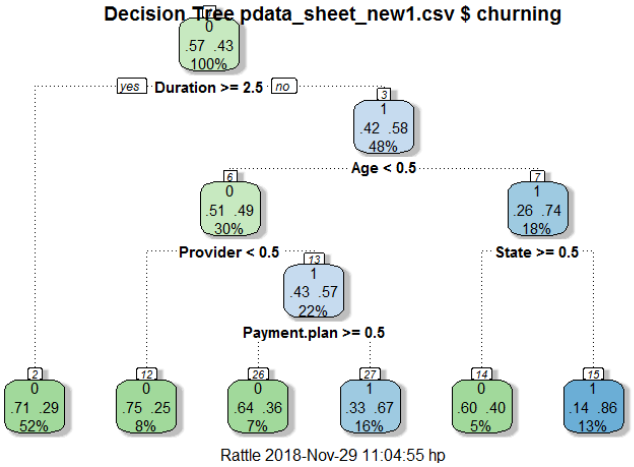


Figure 1: Decision tree generated by Rattle.

Here, two types of nodes can be seen. 0 represents the non-churn and 1 represents the churners. Above generated tree, can help in predicting the churning behavior of any customer.

Error Matrix corresponding to the tree:

Table 2: Error matrix for the Decision Tree.

	Predicted	
Actual	0	1
0	23	4
1	7	11

Steps in Rattle To generate Decision Tree (with Gender as ignore attribute):

Firstly, repeat the step1 and step2 of 4.2.1 and then continue with the following steps.

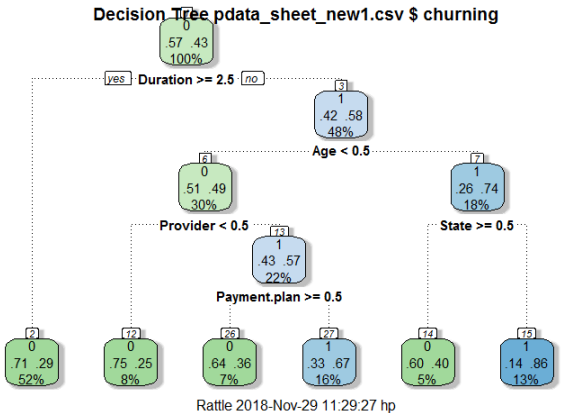
Step 1: In the execute tab of data tab, set the gender attribute as ignore and then execute.

Step 2: Then click on model tab and select tree option and click on execute to see all the decision rules for the tree.

Now, repeat the step4 of to get the desired tree.

Generated Decision Tree (Gender Ignored)

Here, all the gender field have been set as ignore field. The tree generated is shown below.



Fiure. 2: Decision tree generated by Rattle (Gender as ignore).

Error Matrix corresponding to the tree:

Table 3: Error matrix for the Decision Tree (Gender as ignore).

Predicted		
Actual	0	1
0	23	4
1	7	11

Comparative Analyses

On comparing the decision trees, in section 4.2.2 and section 4.3.2, it can be clearly seen, that both the trees are, nothing but a replica of each other. So, this shows that even if the gender attribute is ignored, the tree remains unaffected.
With this, the author has made an observation that whether the customer is a male or a female, it has nothing to do with the churning behavior of that person, which is not the case with other attributes. If the duration or provider is put in the ignore category, the tree changes.
Same is the case with the error matrices of both the trees, in sec.4.2.3 and sec. 4.3.3. This means, the accuracy of the tree is also not getting influenced.

V. CONCLUSION AND FUTURE SCOPE

This paper analyses a dataset of 300 records in rattle, for generating different versions of decision tree by ignoring some attribute or the other and finally, making an inference of which attribute is the not affecting the basic tree structure. The generated trees in sec.4.2.2 and 4.3.2, show that churning behaviour of customers is free from the gender biasness. In other words, it can be said that, if we remove the gender field from our dataset, even then the churning trends would not be showing any change and so would be the decision tree.
From the above observation, authors would like to bring to the notice, that gender field is the least important one, out of all fields in the dataset. Now, the future work on the same can be carried out, by deleting the gender attribute from the data file. It would help the authors to focus towards the other dimensions.

REFERENCES

1. Han J.and Kamber M., (2011). Data Mining: Concepts and Techniques, Morgan Kaufmann Publish.
2. SadiqHussain, JitenHazarika, PranjalBuragohain, G.C. Hazarika, “Educational Data Mining on Performance of under Graduate Students of Dibrugarh University using R”, International Journal of Computer Applications (0975 – 8887) Volume 114 – No. 11, March 2015.
3. Wonhee Cho, Yoojin Lim, Hwangro Lee, Mohan Krishna Varma, Moonsoo Lee, Eunmi Choi, “Big Data Analysis with Interactive Visualization using R packages”, In Proc. of the 2014 International Conference on Big Data Science and Computing, Article No. 18, ISBN: 978-1-4503-2891-3,ACM..
4. J. Rao, R. Kelappan, P. Pallath, “Recommendation system to enhance planning of software development using R”, In the Proc. of the 4th International Workshop on Recommendation Systems for Software Engineering, June 2014, ACM.
5. J. Agosta, D. Thakurta, R. Horton, M. Inchiosa, S. Kumar, M. Zhao, “Scalable Data Analytics Using R: Single Machines to Hadoop Spark Clusters”, In the Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
6. Manpreet Kaur, Dr. Prerna Mahajan. “Churn Prediction in Telecom Industry Using R”, International Journal of Engineering and Technical Research (IJETR), ISSN: 2321-0869, Volume-3, Issue-5, May 2015.
7. T.Miranda Lakshmi, A.Martin, R.Mumtaj Begum, Dr.V.PrasannaVenkatesan, “An Analysis on Performance of Decision Tree Algorithms using Student’s Qualitative Data”, I.J.Modern Education and Computer Science, 2013. Published Online June 2013 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijmecs.2013.05.03.
8. V.Umayaparvathi, K. Iyakutti. “Applications of Data Mining Techniques in Telecom Churn Prediction”, International Journal of Computer Applications,(0975 – 8887) Volume 42– No.20, March 2012.