Explicit Decision Tree for Predicating Impact of Motivational Parameters on Self Satisfaction of the Students

Aniket Muley, Parag Bhalchandra and Govind Kulkarni

Abstract The decision trees are widely accepted as a novel tool for supervised classification and prediction. We have implemented a decision tree over own educational dataset to get insights for predictive academic analytics, which otherwise are invisible. The educational dataset contains personal and socioeconomical variables related to students. It is known that individuality, lifestyle, and responsiveness-related variables have a close association with motivational aspects which together harshly affect student's performance. The decision tree is deployed to escalate the role of motivational variables on self-satisfaction aspects. Analytics were carried out with R software package.

Keywords Educational analytics · Decision trees · Classification · Prediction · Data mining

1 Introduction

This paper is a preliminary endeavor for applied analytics over student's academic data. This analysis is interesting and can augment the quality of the higher education. Student's raw data regarding course preference, results, further progression, are crucial capital for all higher educational organizations. Data mining can be applied to such databases in order to gain challenging outputs [1]. This is called as Educational Data Mining where we primarily investigate analytics for good insights [2]. These insights are in terms of associations, correlations, clusters, etc. Despite the

A. Muley
School of Mathematical Sciences, SRTM University, Nanded, MS, India
e-mail: aniket.muley@gmail.com

P. Bhalchandra (✉) · G. Kulkarni
School of Computational Sciences, SRTM University, Nanded, MS, India
e-mail: srtmun.parag@gmail.com

G. Kulkarni
e-mail: govindcoolkarni@gmail.com

H. S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*, Lecture Notes in Networks and Systems 74,
https://doi.org/10.1007/978-981-13-7082-3_14

percentage of GDP for the year 2013–14 was 1.34% [3], the academic performance of students in India is not improving. There could be many reasons behind it. Self-satisfaction of students regarding the courses they are pursuing can also beone of the reasons. Since these aspects are not visible directly, people have not tackled for the same. This is the right case for data mining explorations. Self-satisfaction is also related to a number of other aspects including academic ambience, easiness in pursuing course, interested areas, family support, etc. These aspects can be categorized as self-related, family-related, motivational, and financial concerns. Since we need to precisely understand the main concern behind self-satisfaction, we have worked out for all these categories. Financially deprived students can secure good ranks. This reflects self-satisfaction of students and also highlights that financial aspects and family support aspects are less significant. Hence, family-related and financial aspects cannot be significantly impactful. Of other remaining aspects, self-related and motivational aspects are more important. The enthusiasm and passion to do something in life is the main motivational catalyst. If we apply analytics for elaboration of this, then it will be very interesting to see the exact role of motivational aspects on self-satisfaction. This work is a multiparty work undertaken together by faculties of educational, computational and statistical sciences. The key idea of this study is to escalate collected datasets through data mining. The minor research objective is to investigate whether motivational parameters significantly contribute to the increase or decrease of self-satisfaction of students? If no, what other variables are related to the same? The below sections brief experimental setup and discussions related to such performance analysis.

2 Research Methodology

Every study from data analytics needs a dataset for implementation of algorithms. This dataset must match with the context of analytical exploration. To do so, we have taken efforts to tailor our own real dataset related to personal, socioeconomical, habitual, and self-related aspects of students of our university. This dataset has 360 records. Every record has 46 fields, each one related to some information/fact about a student. The closed questionnaire is the question with predefined answers, and the method was followed for the creation of this dataset [4]. A standard benchmark illustrated in Pritchard and Wilson [4] was adopted for the same. Some contemporary works [5–8] were also reviewed discover important variables from our dataset/questionnaire. Some preselected students were interacted during finalization of the questionnaire and in total; four trial testing have been made to formulate the concluding edition of the questionnaire. Excluding name details, class details, and roll number details, the concluding questionnaire consisted of 43 variables. The MS Excel 2010 software was used to testimony the dataset. Standard policies for data preparation and cleansing were done [1] as per the requirements of data mining experiments. We cannot study all questions in the questionnaire and test for their association with other questions. This will be exhaustive and divert us from the

main rationale [9]. Our applied work must lead to a new dimension [10–12]. That is why, we tried to examine the relationship between student's self-satisfaction and motivational aspects.

All implementations were carried out on R data mining platform. The supervised learning methods called decision trees and their algorithms were implemented using R open source software [13]. The *Rattle* is a free graphical user interface for Data Science, developed using R. The decision tree model exploits a recursive partitioning approach. Its traditional algorithm and default part of *rpart* package of R. This ensemble approach tends to produce complication models than a single decision tree. Classification definitely needs a potential data which can be classified by setting some rules [1]. There are two crucial steps to do so, first learning and second is classification itself. The system is trained first with sample data and once the system learns, test data is supplied to the system to check accuracy. If the accuracy is good enough, the rules can be practical for the new data [1]. This workout can be frequently done via a decision tree or a neural network for the actual workout [1, 2]. A decision tree has branches and nodes [1, 14]. It contains split, which tests the value of an appearance of the attributes. The outcome of the test is labeled on edges. A class label linked with every leaf node [1, 14]. The given entropy as appraise of the impurity is then defined. This calculation is called information gain. The section two below summarizes all these considerations.

3 Experimentations and Discussions

To devise the decision tree, we tried out some set of experiments. Our aim is to evaluate the decision tree induction method. To draw the decision tree, *rattle* uses a command *rpart* from *rpart* package. We obtained the rules associated to the tree and the summary of the Decision Tree model for Classification is given in Table 1. Table 1 represents the root node, level of trees, split of the observations, and loss of frequency at each level split and corresponding probability values. The * denotes terminal node. The total number of population selected was n = 251.

The terminal node variables actually used in the tree construction are: CAREERDREM, F_T_FRIEND, F_T_STUDY, PLACELVING, REGION, and SELF.LIB. Initially, the root node error is 0.25498 (n = 251). Some of the generated tree rules can be illustrated as follows:

1. Rule number: 3 [PER_SATISF = 1 cover = 17 (7%) prob = 0.94], PLACELIVNG >= 3.5
2. Rule number: 19 [PER_SATISF = 1 cover = 25 (10%) prob = 0.92], PLACELVING < 3.5, F_T_STUDY >= 2.5, CAREERDREM < 3.5, F_T_FRIEND >= 2.5
3. Rule number: 37 [PER_SATISF = 1 cover = 7 (3%) prob = 0.86], PLACELVING < 3.5, F_T_STUDY >= 2.5, CAREERDREM < 3.5, F_T_FRIEND < 2.5, F_T_FRIEND < 1.5

Table 1 Details of classification of nodes

Variable	Nodes	Children generated	Probability
Root	251	64, 1	(0.25498008 0.74501992)
PLACELVING < 3.5	234	63, 1	(0.26923077 0.73076923)
F_T_STUDY >= 2.5	76	27, 1	(0.35526316 0.64473684)
CAREERDREM >= 3.5	16	6, 0	(0.62500000 0.37500000) *
CAREERDREM < 3.5	60	17, 1	(0.28333333 0.71666667)
F_T_FRIEND < 2.5	35	15, 1	(0.42857143 0.57142857)
F_T_FRIEND >= 1.5	28	14, 0	(0.50000000 0.50000000)
SELF.LIB < 0.5	20	8, 0	(0.60000000 0.40000000)
REGION < 1.5	13	4, 0	(0.69230769 0.30769231) *
REGION >= 1.5	7	3, 1	(0.42857143 0.57142857) *
SELF.LIB >= 0.5	8	2, 1	(0.25000000 0.75000000) *
F_T_FRIEND < 1.5	7	1, 1	(0.14285714 0.85714286) *
F_T_FRIEND >= 2.5	25	25, 2, 1	(0.08000000 0.92000000) *
F_T_STUDY < 2.5	158	158, 36, 1	(0.22784810 0.77215190) *
PLACELVING >= 3.5	17	17, 1, 1	(0.05882353 0.94117647) *

4. Rule number: 5 [PER_SATISF = 1 cover = 158 (63%) prob = 0.77], PLACELVING < 3.5, F_T_STUDY < 2.5
5. Rule number: 73 [PER_SATISF = 1 cover = 8 (3%) prob = 0.75], PLACELVING < 3.5, F_T_STUDY >= 2.5, CAREERDREM < 3.5, F_T_FRIEND < 2.5, F_T_FRIEND >= 1.5, SELF.LIB >= 0.5.

To judge the accuracy of the classifier, we worked out for generation of the error matrix. This matrix is generated after validating the proportions. It is evident that the overall error is 22.6% and averaged class error is 41.1% (Table 2).

Table 2 Error matrix

		Predicted		Error
		0	1	
Actual	0	5.7	15.1	72.7
	1	7.5	9.5	9.5

Table 3 Accuracy of results

Train	Test	Validate
0.6894	0.6157	0.5325

The area under the ROC curve for the *rpart* model is computed for 70% training set data, 15% test data and balance 15% validation data. The accuracy of the result was summarized in below Table 3. The results validated simply state that the random selection is responsible for the accuracy level of the data.

The output decision tree is shown in Fig. 1. In order to interpret Fig. 1, we look at the dataset variables and determine crucial and important variables, and then comes up with a node, and so on. The tree is shaped by splitting data up by variables and then counting to see how many are in each bucket after each split.

The decision tree model explores that the satisfaction of student is directly dependent on his/her place of living, and then in the next stage, it depends on free time available for study. Further, it classifies free time available to spend with friends, and then splits in terms of availability of having a self-library with the student. Further splitting is done through regional classification. Finally, there is no classification. This means the final induction of decision tree has been reached. All splitting parameters are playing an important role in the identification of the satisfaction of students. Further, it is suggested that if we can properly focus only on these parameters of individuals, significant improvement in their satisfaction level could be witnessed.

4 Conclusions

It was overtly unspoken that numerous societal, routine and many other aspects are linked with the satisfaction level of students and mere good performance cannot be judged by studious nature alone. In order to escalate these, the study took it as confront and drew a decision tree for selected parameters. Using the interdisciplinary approach, we found that, besides studies, other aspects like the student's satisfaction did depend on self-motivation. This work highlights that the place of living significantly matters related to the satisfaction of the students. Second such variable boosting self-motivation is the quantum of the free time for study. The self-career dream is the third major motivational variable. If prognostication methods are implemented, then change in place of living, motivational canceling can boost student's performance.

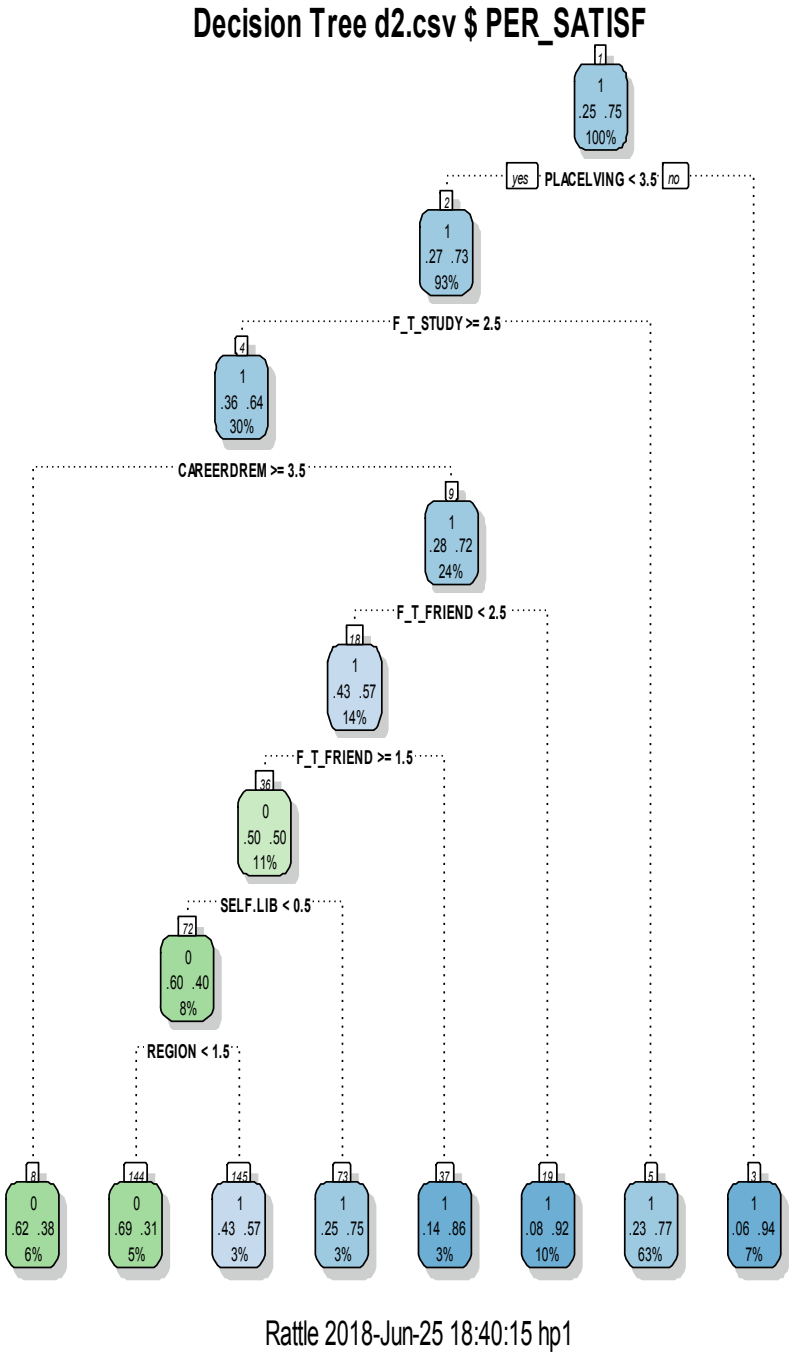

Fig. 1 Final Decision Tree

References

1. Dunham M (2002) Data mining: introductory and advanced topics, by Margaret H. Dunham, Pearson Publications
2. Han J, Kamber M (2006) Data mining: concepts and techniques, 2nd edn. The Morgan Kaufmann Series in Data Management Systems, Jim Gray
3. https://economictimes.indiatimes.com
4. Pritchard ME, Wilson GS (2003) Using emotional and social factors to predict student success. J Coll Stud Dev 44(1):18–28
5. Ali S et al (2013) Factors contributing to the students academic performance: a case study of Islamia University Sub-Campus. Am J Educ Res 1(8):283–289
6. Graetz B (1995) Socio-economic status in education research and policy in John Ainley et al., socio-economic status and school education DEET/ACER Canberra. J Pediatr Psychol 20(2):205–216
7. Considine G, Zappala G (2002) Influence of social and economic disadvantage in the academic performance of school students in Australia. J Sociol 38:129–148
8. Bratti M, Staffolani S (2002) Student time allocation and educational production functions. University of Ancona Department of Economics Working Paper No. 170
9. Field A (2000) Discovering statistics using R for Windows. Sage Publications
10. Joshi M, Bhalchandra P, Muley A, Wasnik P (2016) Analyzing students performance using academic analytics. In: IEEE international conference ICT in business industry and government (ICTBIG), pp 1–4
11. Muley A, Bhalchandra P, Joshi M, Wasnik P (2016) Prognostication of student's performance: factor analysis strategy for educational dataset. Int J 12
12. Muley A, Bhalchandra P, Joshi M, Wasnik P (2018) Academic analytics implemented for students performance in terms of Canonical correlation analysis and chi-square analysis. Inform Commun Technol. Springer, Singapore, pp 269–277
13. Web resource. https://cran.r-project.org/web/packages/rattle/vignettes/rattle.pdf
14. Rokach L, Maimon O (2014) Data mining with decision trees: theory and applications. World Scientific