Years of Education and Vocabulary Aptitude Decision Tree Analysis

Charles Adkins

Table of Contents

Years of Education and Vocabulary Aptitude Decision Tree Analysis

Abstract

This report presents the analysis of a data set that includes values for years of education, vocabulary test scores and a unique student identification number, which links the three together. The scope of the analysis is to determine whether or not years of education impacts vocabulary aptitudes, which will be conducted utilizing a decision tree model. The analysis will rely on a Decision Tree technique invoked by the R platform and Rattle package.

Introduction

The purpose of this report is to discuss and present an analytical study that utilizes a Decision Tree analysis technique. A Decision Tree analysis will be utilized to develop the relationship between the variables to display the thresholds of the collective outcomes.

The analysis will be performed on data gathered from the years 1974-2016 during a census that is conducted, on average, every two years. The census has been carried out as a social science survey, as a mere means to uncover a trending occurrence within society. These results can be utilized by anyone who warrants them valuable and is also interested in understanding the correlation between the variables.

The purpose of the analysis to determine if students vocabulary aptitudes improve with more years of education. The respective data frame is quite robust in nature, including 30,351 observations, accompanied by three columns. The columns are represented accordingly: Column one - unique student identification numbers, column two - years of education, and column three - number of questions each student answered correctly on a 10-question word test. The 10-

question test evaluates the students ability to comprehend a given word and select the synonym of that word.

The analysis will answer the question: “Do more years of education result in a higher vocabulary test score?”. The hypothesis is that the results will show a direct positive correlation between years of education and vocabulary test scores.

The results of the analysis can be used for informative purposes, trend theory or program evaluation, development and implementation. By understanding the relationship between the two variables, the reader will be able to conceptually perceive how they compliment and influence one another.

Data Appraisal

This particular data has been collected since 1974 and continues to be the focus of a social science census conducted by the National Opinion Research Center, with 2016 being the most recent collection. The dataset contains a limited amount of variables, which makes the analysis straightforward. To establish the relationship between the years of education and the vocabulary test values, the variables will be made input and target variables, respectively. By forming the variables in this manner, the analysis will uncover how the years of education influence the vocabulary test results value, if at all. The limitations of the dataset, due to its simplicity, are that the years of education cannot be definitively be linked to the causation of particular vocabulary test results. Although there may be some orientation of correlation between the two variables, without including other relevant variables into the analysis, we are merely displaying the correlation between the two factors, not a specific causation relationship.

Technique

The data set originally contained five columns and 30e+3 observations. The columns were labeled with the following: Column One: Unique Student ID number, Column Two: Observation year, Column Three: Sex of the student, Column Four: Years of education, Column Five: Vocabulary Test Score. For the purpose of this analysis, the Year and Sex columns were deleted, as they were not revenant to the posed research question. The data had no missing values and contained no unusual outliers.

Once the data was loaded into Rattle, the data set was partitioned into three proportioned groups, 70/15/15. The data was separated in this way to allow for 70% to be sampled as the training dataset and the remaining 30% to be split between the validation set and the testing set. The training set contained 21,245 observations (70% of original) and is represented by the Decision Tree output in Appendix A, while the optimized Decision Tree is in Appendix B.

To optimize the results, the tuning parameters will be altered to present the model in a clear and concise fashion. Such parameters include Minimum split, Minimum bucket, Maximum depth and complexity. Minimum split determines how many observations are required before an additional branch is created, while minimum bucket determines how many observations are needed in each node to justify its creation. Maximum depth regulates how many nodes and branches are created in terms of distance away from the root node, while complexity determines how detailed the tree will be, to include as many differing threshold values as the complexity parameter approaches closer to the value “0”.

The use, preparation and manipulation of this particulate data set is legal and ethical in all cases and standards. The data was obtained from an open source data storage site, of which was

procured from the National Opinion Research Center database, set to be viewed and utilized for public purposing, but not to be done so without due source credit.

Evaluation

Utilizing the decision tree model technique and the Rattle interface coincides with the characteristics of the data set and the established purpose and scope of the analysis. By establishing the input and target variables, Rattle can calculate their relationship and determine the degree to which the variables impact one another and the thresholds of each nodal impact. These methods help support organizational decision-making by developing and articulating the relevancy to which the research question is supplemented by the thresholds of structural results. The granular level of detail communicated by the decision tree model clearly identify a trend of correlation between the included variables.

The agility of this analysis can be fully realized by incorporating all of the original variables or tuning the variable classifications and the sensitivity parameters to create several variations of decision tree model. Due to the robust nature of the data set and the selection of original and current variables, various research questions can also be determined and implemented.

The data information is vague in terms of personal identification information (PII) and simply represents an educational observation. The interpretation of such data and the associated results are but a representation of behavioral and intellectual social patterning. The data set, analysis and results are being used for educational purposes only and has the freedom to be reproduced and interpreted by any party who finds the information useful.

Decision Tree Model

The optimized decision tree has produced some interesting results, of which clearly display the relationship between the two variables. The structure summary of the decision tree is included in Appendix C.

Decision Tree Step-By-Step Process

-To begin the Decision Tree Analysis, initialize RStudio and open the Rattle package by typing:

```
library(rattle)
rattle()
```

which opens up the Rattle GUI.

-Once Rattle is opened, within the Data tab, select the CSV file from its appropriate location in the filename drop-down menu button.

-Ensure that you partition the dataset to allow for a training, validation and testing set, in their own specific percentage of whole.

-Then click the Execute button.

-The interface will display the variables within the associated dataset and will assign roles to each variable.

-Ensure that the Student ID is marked as Ident, Education is marked as Input and Vocabulary is marked as Target.

-Then click Execute.

-Navigate to the Model tab, where Rattle will display model options and tuning parameter options.

-Ensure Tree is selected and the tuning parameters are representative of the output that is desired.

-Then click Execute.

-The interface will display the textual form of the tree, the code required to implement the tree in Rscript and errors.

-Click the Rules button, then the Draw button.

-The Rules button will print the textual rules that the tree abides by, while the Draw button will plot the tree within the R IDE Plots window.

-Depending on the output, the tree can be considered complete, or alterations and tuning actions can be taken in order to create a more suitable tree. Considerations should include size, error, simplicity, function, analysis scope, overfitting, etc.

Model Results

The rules of the decision tree show that various relationships and thresholds between the two variables (Appendix D). On average, as each education value increases by 2.95 years, the vocabulary score value increases by one. As proposed with the initial hypothesis, the decision tree displays how various values for education years correspond to vocabulary aptitude scores and how the correlation is positively integrated. These conclusions are reasonable in the fact that we make the assumption that with more education and exposure to increasingly more complex subjects and surrounding text, the more a student will be familiar and confident in identifying and understanding the words presented to them.

Common errors made during the creation of decision tree models can potentially distort the conclusions that an accurate decision tree would communicate to its viewers. Some examples of those errors might be overfitting the model, of which displays certain characteristics indicative of the statistically insignificant values within the dataset itself. When the model begins to overfit the data, the model becomes less accurate and the message of the data becomes lost in the noise. For this reason, this particular decision tree was restricted to displaying specific

areas of decision node values, corresponding to tiers of education levels (Middle school, High school, College and Beyond). Although the dataset has minimal variables, the robust nature of the observations offers its conclusions with a high-level of certainty and a statistically insignificant level of error. As the decision tree displays (Appendix E), after five decision splits the cross-validation error was reduced from 1.000 to 0.7704, while the Root Node Error was 4.4546, which is an objective measure of predictive accuracy. The variance table displays the p-value statistical significance value, which communicates how significant the relationship between the variables is, of which reads '< 2.2e-16 ***', signifying a high level of significant association between years of education and vocabulary test scores (Appendix F).

Conclusion

The decision tree model and the associated evaluations indicate and clearly identify a positive correlation between the years of education and vocabulary test score variables. By incorporating all observations and structuring the relative variables in a manner that can be reproduced, the results are verifiable and conclusive. As a years of education increase, the overall vocabulary aptitude score increases. As stated previously, the initial hypothesis for the analysis, coincides with the results of the study, and as such, can be utilized for various forms of organizational decision-making processes.

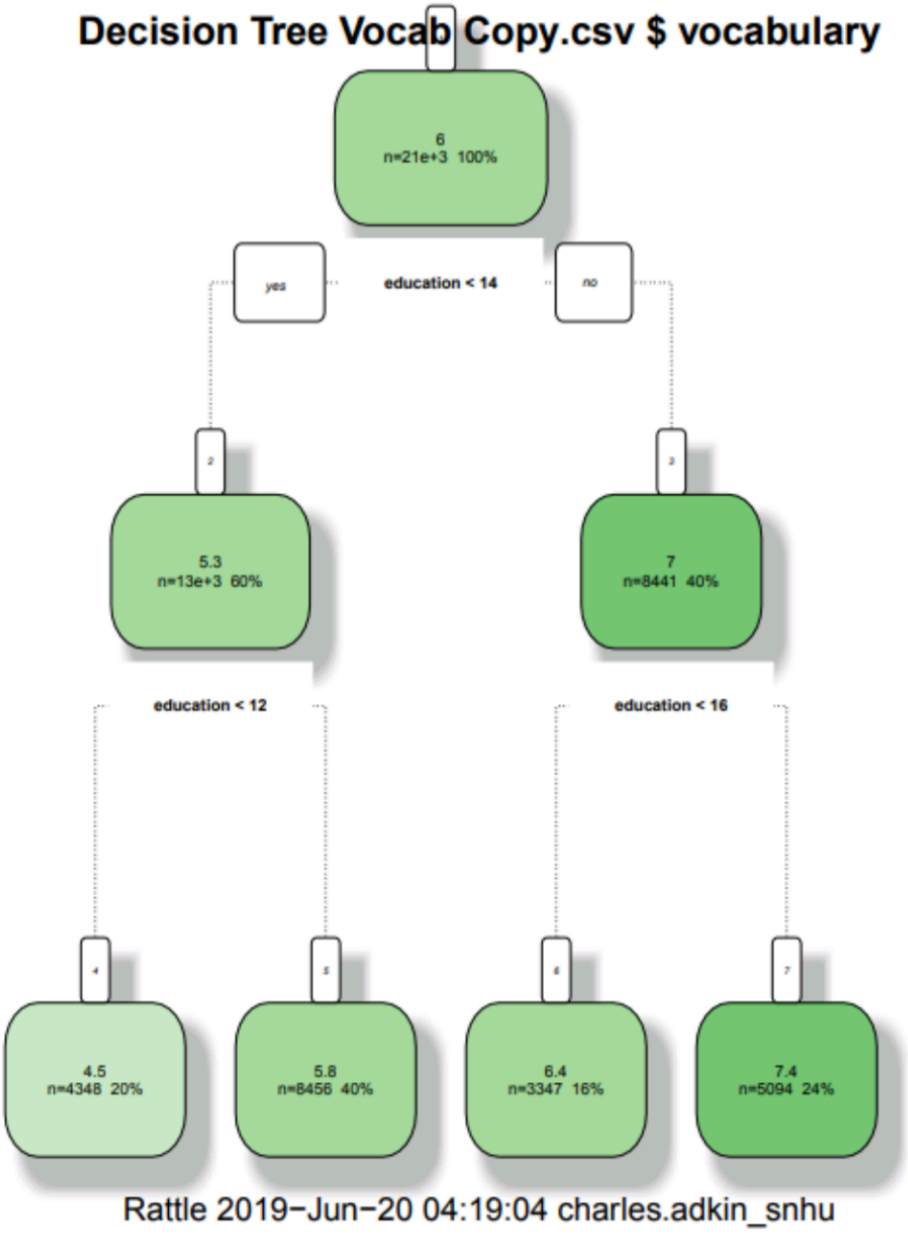
References

General Social Survey (GSS) (2016). National Opinion Research Center Datasets. Vocabulary and Education. Retrieved June 18, 2019 from https://vincentarelbundock.github.io/Rdatasets/datasets.html

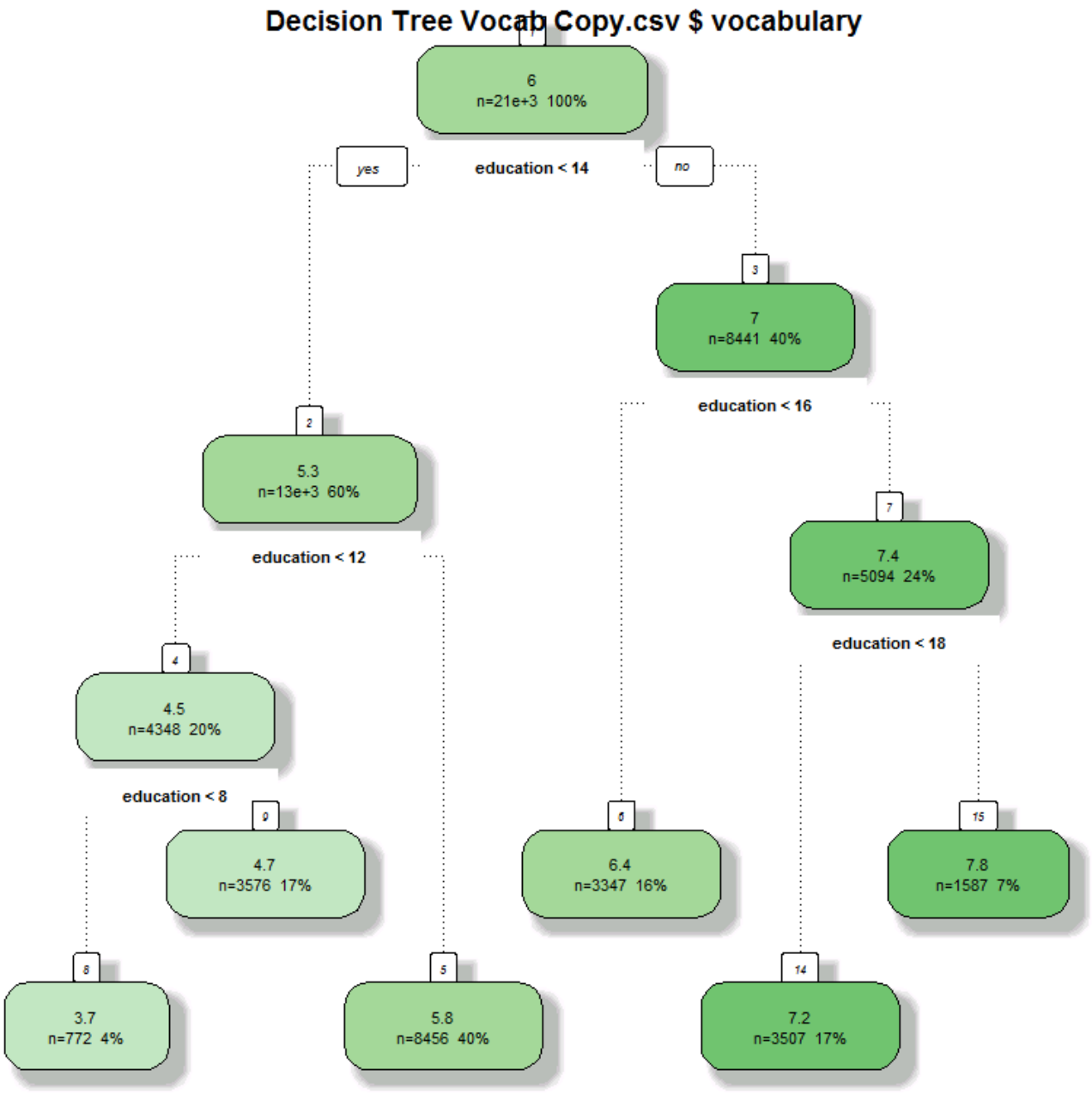
Williams, G. (2011). Data Mining with Rattle and R. The Art of Excavating Data for Knowledge Discovery. Chapter Eleven - Decision Trees. Pages 205-243. Retrieved June 28, 2019.

Appendices

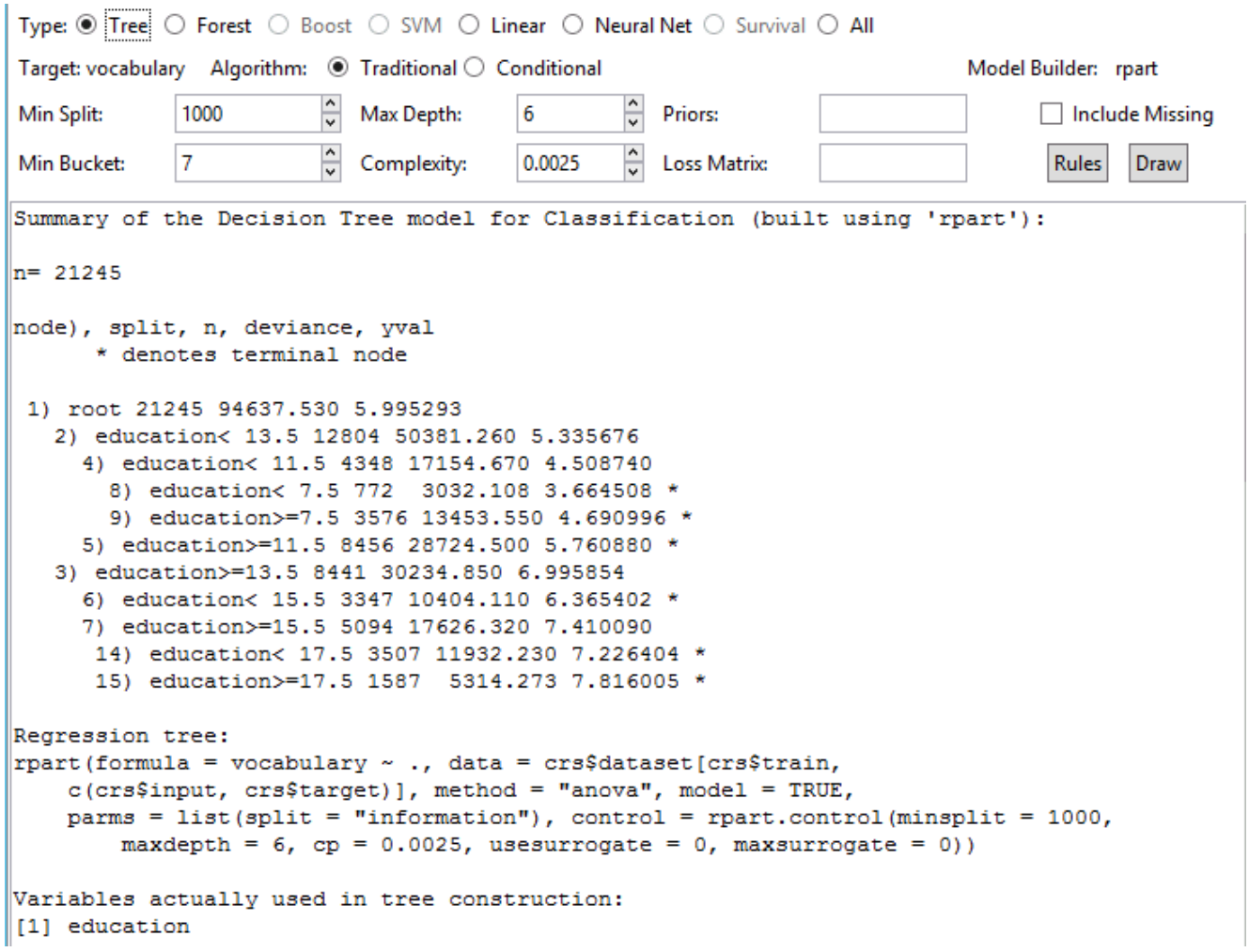
Appendix A: Original Decision Tree (Prior to Optimization)



Appendix B: Decision Tree (After Optimization)



Appendix C: Decision Tree Summary



Appendix D: Decision Tree Rules

```
Tree as rules:

 Rule number: 5 [vocabulary=5.76087984862819 cover=8456 (40%)]
   education< 13.5
   education>=11.5

 Rule number: 9 [vocabulary=4.69099552572707 cover=3576 (17%)]
   education< 13.5
   education< 11.5
   education>=7.5

 Rule number: 14 [vocabulary=7.22640433418877 cover=3507 (17%)]
   education>=13.5
   education>=15.5
   education< 17.5

 Rule number: 6 [vocabulary=6.36540185240514 cover=3347 (16%)]
   education>=13.5
   education< 15.5

 Rule number: 15 [vocabulary=7.81600504095778 cover=1587 (7%)]
   education>=13.5
   education>=15.5
   education>=17.5

 Rule number: 8 [vocabulary=3.66450777202073 cover=772 (4%)]
   education< 13.5
   education< 11.5
   education< 7.5
```

Appendix E: Decision Tree Error Values

```
Root node error: 94638/21245 = 4.4546

n= 21245

         CP nsplit rel error  xerror      xstd
1 0.1481591      0   1.00000 1.00008 0.0093480
2 0.0475720      1   0.85184 0.85452 0.0083554
3 0.0232933      2   0.80427 0.80363 0.0080057
4 0.0070692      3   0.78098 0.78442 0.0079371
5 0.0040133      4   0.77391 0.77455 0.0078982
6 0.0025000      5   0.76989 0.77044 0.0079034

Time taken: 0.01 secs

Rattle timestamp: 2019-07-19 23:25:13 charles.adkin_snhu
=====================================================================
```

Appendix F: Analysis Variance Table

```
Analysis of Variance Table

Response: vocabulary
             Df Sum Sq Mean Sq F value
education     1  21621 21621.2  6290.4
Residuals 21243  73016     3.4
             Pr(>F)
education < 2.2e-16 ***
Residuals
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
[1] "\n"
Time taken: 0.03 secs

Rattle timestamp: 2019-06-20 03:08:23 charles.adkin_snhu
================================================================================
```