CIS417 – Introduction to Business Analytics

Fall 2017

Homework 5

Churn Part I

Use the DSBchurn.CSV for this homework. This file describes customers of a cell phone company and is used to predict churn, i.e., whether a customer will leave the cell company or stay with it. The file is available on BlackBoard.

Variable	Values	Meaning
COLLEGE	{zero, one}	one = college grad, zero = not
INCOME	numeric	Annual income
OVERAGE	numeric	Avg. overcharges per month
LEFTOVER	numeric	Avg. percentage leftover minutes per month
HOUSE	numeric	Value of dwelling
PHONE_PRICE	numeric	Cost of phone
LONG_CALLS	numeric	Avg. number of long calls (>15 min) per month
AV_DURATION	Numeric	Average call duration in minutes
REP_SAT	{very_unsat, unsat, avg, sat, very_sat}	Reported level of satisfaction
REP_USAGE	{very_little, little, avg, high, very_high}	Self-reported level of usage
REP_CHANGE	{ never_thought, no, considering, perhaps, actively_looking_into_it}	Was customer considering change his/her plan?
LEAVE	{LEAVE, STAY}	Whether the customer left or stayed (class variable)

This file is from Provost and Foster, the authors of our textbook. Some examples in the textbook use this dataset. Feel free to look up more details from the textbook. Churn is listed in the index in the back of the book. Note: This is different from the churn data set we used in our lectures.

You will be using this for the next few assignments.

Write R code to:

1. Read in the CSV file using read.csv(file.choose()) and save it into churn data frame.
2. Examine the structure of the churn data frame. It should look like the figure below:

```
> str(churn)
'data.frame':20000 obs. of  12 variables:
 $ college    : Factor w/ 2 levels "one","zero": 2 1 1 2 1 2 2 1 2 2 ...
 $ income     : int  31953 36147 27273 120070 29215 133728 42052 84744 38171 105824 ...
 $ overage    : int  0 0 230 38 208 64 224 0 0 174 ...
 $ leftover   : int  6 13 0 33 85 48 0 20 7 18 ...
 $ house      : int  313378 800586 305049 788235 224784 632969 697949 688098 274218
153560 ...
 $ phone_price: int  161 244 201 780 241 626 191 357 190 687 ...
 $ long_calls : int  0 0 16 3 21 3 10 0 0 25 ...
 $ av_duration: int  4 6 15 2 1 2 5 5 5 4 ...
 $ rep_sat    : Factor w/ 5 levels "avg","sat","unsat",..: 3 3 3 3 5 3 5 5 4 4 ...
 $ rep_usage  : Factor w/ 5 levels "avg","high","little",..: 3 3 5 4 3 2 3 3 3 3 ...
 $ rep_change : Factor w/ 5 levels "actively_looking_into_it",..: 4 2 5 2 3 4 1 2 1
3 ...
 $ stay       : Factor w/ 2 levels "LEAVE","STAY": 2 2 2 1 2 2 2 2 2 1 ...
```

3. If the column types do not match, the use the conversion functions to fix it.
4. Fix the order of levels of factors to match that in the table on the first page. See http://www.r-bloggers.com/reorder-factor-levels-2/ for help.
5. Save the data frame as DSBchurn.Rda for later reuse.
6. Create randomly sampled training and test data sets with about 66.7% and 33.3% of the observations, respectively. Use the seed 3478 so that it is repeatable across the groups.
7. Grow a tree using the training dataset to explain the stay class variable. Use minsplit=100 to keep the tree small for now.
8. Display fit (type fit and hit return).
9. Explain rows numbered 1, 10, and 3. Which node is the parent node. What was the immediate split that created it? What is the count of stay and leave at this node? (put these as comments in the R file)
10. Plot and label the tree. (save the pdf)
11. Print the confusion matrix for the test data set.
12. Determine the accuracy, error rates, recall, specificity, and precision for this tree and the test data set.
13. Write a one page report (No more than 300 words) explaining the tree to the CEO of the company. It must be written at the appropriate level for the senior manager and explain the main findings. Save this as a pdf document with the word count at the bottom of page. Pages with more than 300 words will not be accepted.
14. Upload the R script, the one page report, and the pdf of the plot.