BUAN 6356 (Johnston)
Homework 4A(20231004)
Due: 7 October 2023 (6PM)

Points available: 130

This assignment is about running linear models and recursive partitioning (tree) models on data with binary outcomes. You will need the data.table and partykit packages.

As always, the first commands of your code MUST include:

setwd("c:/data/BUAN6356/HW_4"); source("prep.txt", echo=T)

and the last command of your code MUST include:

source("validate.txt", echo=T)

The required code CAN be set up for conditional execution. (E.g.: set a Boolean variable and then use it in an if() to execute these statements.)

Be careful with the quote characters as they must ALL be the same at the beginning and end of a string. (Use the single or double quote character from the key next to "Enter".) Inclusion of these lines is required BEFORE your code will be tested. I hope that most of you understand this by now … :-)

The data for this assignment is "student_default_ISLR.csv" from the *UTDbox>data* directory. Use data.table::fread() to import the data. You may need to install "partykit" or some other package but DO NOT include any install.packages() code in the code you submit.

A 10% testing (validation) sample will be used with this assignment. Seed the RNG with 646609930.

Your outcome variable (dependent) is "default" and the independent variables (covariates) are everything else in the dataset

Submit the code to eLearning as an ASCII file with file extension “.txt” which can be copied directly into R. (That is, the same way you have done this process for the earlier homework assignments.)

You may submit this assignment as many times as needed until you get full credit.

Deliverables (all names are case sensitive; models are result of fit functions):

1. seed (vector) random number seed
2. raw (data.table) the original data (no data modifications)
3. wk (data.table) raw with “default” as indicator, “student” as factor
4. mbrCutoff (vector) MBR cutoff value
5. nTst (vector) number of elements in testing set
6. tstFrac (vector) testing set proportion
7. tst (vector) index values for testing set
8. lBase (glm) training set logistic model using all variables (no interactions)
9. lBaseTst (vector) testing set logistic model predictions using lBase
10. lBaseTstCls (vector) testing set logistic model classifications
11. tBase (constparty) training set ctree() model using all variables
12. tBaseTst (vector) testing set ctree() model predictions
13. tBaseTstCls (vector) testing set ctree() model classifications

Notes/Hints:

- “factor” variables can be converted using as.numeric() but watch your resulting codes. Conversions can be examined by (e.g.) the table() function.
- Outcome variables (dependent) should only be on one side of the “~”
- See “UTDbox>demo>01f_classification_titanic” for example 2-category classification code.
- When generating predictions for the logit model, use the parameter ***type=”response”*** with the appropriate quote characters.
- “wk” should have the same variable names and order as “raw” with the variable “default” transformed from a string to a numeric indicator and variable “student” transformed to a factor. Be sure to validate the transformation of “default” such that “Yes” becomes 1.0 .
- for this assignment, the step() function should not be used.