JIGSAW ACADEMY
THE ONLINE SCHOOL OF ANALYTICS

Decision Tree Case study

Refer to the file "BH.csv" at "**/Data Science R/Assignments/Graded Assignments/Topic 11.2 Decision Trees**". This file has data on pay-day loans. There is a variable called "cus_employername". This variable has around 1600+ unique values. You are working in a team of analysts who are trying to build a classification model that will predict who will be a good customer and who will be a bad customer. Your team wants to use this variable "cus_employername" in their model. Since this variable has a lot of unique values so creating dummy variable for each of the levels will be very time consuming. Your task is to use decision tree algorithm to come up with a way to club the levels of this variable into two groups:

(1) Group 1-> Consisting of all employer names where the bad rate is high

(2) Group 2 -> Consisting of all the employer names where bad rate is low

You will also need to come up with an efficient code/procedure to create a dummy variable denoting which rows of the data correspond to "Group 1" and which rows correspond to "Group 2"

Hints:

(i) You will need to run a decision tree algorithm here.

(ii) Once you run the decision tree, the algorithm will automatically divide the employernames into two groups. (You can control the number of groups to be created by controlling the depth of the tree using control.rpart() parameter)

(iii) Once you have the tree with two groups of employernames make sure that you check the bad rate in each of the groups (You will find one group having a very high bad rate and one group having a very low bad rate)

(iv) Post this you will need to figure out a way to assign to each row of the data the group label (Group 1 or Group 2). Read the documentation for objects of class rpart to figure this out.