CE/CZ4073 : Data Science for Business

2017-2018, Semester 2 | Nanyang Technological University, Singapore

Assignment 2

Posted on 1 March 2018 · Clarify by 13 March 2018 · Submit by 15 March 2018

Submission : Submit a single `R` file for the assignment – `assign2_FullName.R` – containing all the steps and comments in connection with the exploratory data analysis and model building for both the problems, where `FullName` is your full name, as per the official records at NTU.

Your submission should be emailed to `sg.sourav@ntu.edu.sg` by midnight of 15 March 2018 from your official NTU Email Address (do not use your personal email address). Late submissions will be penalized with lower grades, and submissions after 19 March 2018 will not be graded.

Problem 1 [10 points]

Background : Consider the attached dataset `assign2_ChurnData.csv`, which has 20 variables:

`"Gender"`	`"SeniorCitizen"`	`"Partner"`	`"Dependents"`
`"Tenure"`	`"PhoneService"`	`"MultipleLines"`	`"InternetService"`
`"OnlineSecurity"`	`"OnlineBackup"`	`"DeviceProtection"`	`"TechSupport"`
`"StreamingTV"`	`"StreamingMovies"`	`"Contract"`	`"PaperlessBilling"`
`"PaymentMethod"`	`"MonthlyCharges"`	`"TotalCharges"`	`"Churn"`

The target is to fit an *optimal* tree model and a random forest to predict the chance of `"Churn"`.

Task : Import the dataset, perform exploratory data analysis on the variables, and construct the optimal *cross-validated* decision tree to predict the response variable `"Churn"` in case of the given dataset `assign2_ChurnData.csv`. Follow this up by building an *optimal* random forest model to predict `"Churn"` in case of the given dataset. Briefly comment (within the code) on your observations and on the choices you make in the process of building the *optimal* models. Also mention (within the code) which variables you think are the most important in this case.

Problem 2 [10 points]

Background: The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. In the following online challenge, Kaggle asks you to analyse what sorts of people were likely to survive, as a case-study of binary classification.

Task : Take part in the Kaggle Competition – "Titanic: Machine Learning from Disaster" – available at `https://www.kaggle.com/c/titanic`. You are only allowed to use optimal decision trees or random forests as your classification model. Submit your code within the submission file `assign2_FullName.R`, and if you wish, you may also submit your predictions online at Kaggle. Briefly comment (within the code) on the choices you make in the process of building the *optimal* models, as well as which variables you think are the most important for *survival* in this case.

This is an individual assignment. Properly acknowledge every source of information that you refer to, including discussions with your fellow students, if any. Verbatim copy from any source is strongly discouraged, and plagiarism will be heavily penalized. It is strongly recommended that you write the codes completely on your own. Feel free to write the codes in Python if you want.