- Due May 24, 2023 by 11:59pm
- Points 100
- Submitting a file upload
- Available until May 27, 2023 at 11:59pm

This assignment was locked May 27, 2023 at 11:59pm.

A1 Data Exploration in R, Decision Tree Classification

Submission

Submit three files in canvas – A1_I_yourLastName_yourFirstName.png (or another extension created by your screen capture software), A1_II_yourLastname_yourFirstname.Rmd, and A1_II_yourLastname_yourFirstname.html. (Please do not compress these files into one zip file. Submit them as separate files.)

Note: If there is any discrepancy in the task descriptions in the assignment page and the short rubric descriptions, please follow the task descriptions in the assignment page.

Task I (30%)

Objective: Setup your environment to test run R tutorial code

Note: You may choose to use Posit Cloud rather than a local install of R in that case you can ignore step 1 and 2 in the background section below.

Background:

1. Download and install RStudio or use Posit cloud instead.
2. Install R (only needed for a local install)

1. Test mlp_numeric_prediction_tutorial_tictoc.Rmd and install packages required by the program.
    - (a) Download ***mlp_numeric_prediction_tutorial_tictoc.Rmd (https://utah.instructure.com/courses/882385/files/147060918?wrap=1) (https://utah.instructure.com/courses/882385/files/147060918/download?download_frd=1)*** *and* ***insurances.csv (https://utah.instructure.com/courses/882385/files/147060977?wrap=1) (https://utah.instructure.com/courses/882385/files/147060977/download?download_frd=1)***
    - (b) Place these files in your working directory
    - (c) Open the tutorial program in the Source Editor pane by clicking the tutorial program in “Files”.
    - (d) Install packages: The Source Editor pane will alert you to install the packages required. The display of this alert popup may take a minute or so to show up after you open the tutorial program in the source code pane. **It is highly recommended that you wait for this alert popup display, and then click “Install” in this popup to start installing the required packages. The installations may take 10 minutes. To avoid accidentally interrupting the installations, be patient again to monitor the installation progress and wait for their completion.**
    - (e) If necessary, find getwd() and setwd() commands in the tutorial code and replace the working directory inside it with your working directory, e.g.,

      ```
      mydir <- getwd()

      setwd(mydir)
      ```

    - (f) Proceed to knit the tutorial by clicking the “Knit” button above the Source Editor pane. RStudio will prompt you to update some packages required by Rmarkdown. Choose “Yes”, and a popup window will show the installations of these packages.
    - (g) When knitting is completed, the mlp_numeric_prediction_tutorial_tictoc.html will be saved in the “Files” in your project and will be displayed in a new tab of your browser.
    - (h)Take a screenshot of your rstudio screen the timestamp of the mlp_numeric_prediction_tutorial_tictoc.html you generated in your Files pane in the screen shot. Rename this screenshot as A1_I_yourLastName_yourFirstName.png or another extension.
    - (i) For your information, observe the elapsed time at the end of mlp_numeric_prediction_tutorial_tictoc.html so that you have a sense of the amount of time it might take you to run neural network code.

- (*No submission required*) Use rstudio to view, understand and test run the R code in tutorials in Week 1 (and Week 0 if you are just getting started with R and RMarkdown coding). If the tutorials need new packages, e.g., scatterplot3d and psych, please follow the instructions in 3.d above.

Task II (60%)

Packages required: Install the packages: rmarkdown, psych, scatterplot3d and caret.

Input file (for use in code) - *CD_additional_balanced.csv* (https://utah.instructure.com/courses/882385/files/147060840?wrap=1) (https://utah.instructure.com/courses/882385/files/147060840/download?download_frd=1) and data dictionary (not for use in code) - *CD_metadata.xlsx* (https://utah.instructure.com/courses/882385/files/147060897?wrap=1) (https://utah.instructure.com/courses/882385/files/147060897/download?download_frd=1) .

(9 points) Create A1_II_yourLastname_yourFirstname.Rmd that includes code chunks to meet the requirements specified per chunk below. Use the RMD header to include assignment title, author name – you, and the file creation date. Also include header specifications to generate a table of contents and section numbers of your code chunks. The output of aggregate in task 4b2 will NOT be visible until you knit if you have output set to inline.

1. Code chunk 1 (10 points) – Set up, data import and inspection code for the following:

If needed, load packages using library()

- A.
  - Remember to use getwd() and setwd() to set the working directory in your rmarkdown file. For example,

    mydir <- getwd()

    setwd(mydir)

  - Import data using read.csv(). Do not coerce the character variables to factors automatically when loading the data. Examine the overall 'structure' of the input data.
- B. Transform all of the character variables that include categorical values to factor variables. After this transformation, show the overall 'structure' and the 'summary' of the input data.

2. Code chunk 2 (12 points) - For each of these numeric variables - *age*, *duration*, *campaign*, and *pdays*:

- A. Create a histogram and include a title of the histogram.
- B. Create a boxplot and include a title in the plot.
- C. Show deciles of the variable.

3. Code chunk 3 – Explore factor variables (12 points)

Note: Select variable *y* and three other factor variables (e.g, *job*, *education* and *poutcome*) for this task. Do not include additional variables.

- A. (6 points) For each of the selected factor variables, and for each of the variable's levels (e.g., "success", "failure", "nonexistent" of *poutcome*), show the count value and percentage value of instances belonging to that level.
- B. (6 points) For each of the selected variables, show a bar plot of the number of instances (i.e. count) with a level name for each possible value. Show a descriptive title in each plot.

4. Code chunk 4 (17 points)– Explore relationships amongst multiple variables

Note: Do not include additional variables for this task.

- A. (3 points) Use **cor** and **pairs.panels** to display correlations for these seven numeric variables – *age, duration, campaign, pdays, euribor3m, emp.var.rate,* and *nr.employed*.
- B. (8 points) For each of these numeric variables - *duration, emp.var.rate, cons.price.idx, and cons.conf.idx.*
  - (i) Show a **boxplot** of this numeric variable by *y*.
  - (ii) Use the **aggregate** function with 'summary' to aggregate this variable by *y*. The output should be the six number statistics (i.e. min., 1st quantile, median, mean, 3rd quantile, and max.) of the variable (*e.g., duration)*aggregated by "yes" and "no" respectively of *y.*
    - *The output of aggregate in task will NOT be visible until you knit. You can test your code by copying pasting to the console.*
- C. (6 points) Draw a **3d scatter plot** to show *y* values in shapes (e.g. circle for "no", triangle for "yes") for each of the following combinations of numeric variables (along the three axes). Include a main title for the plot and legend for the shapes of *y* in the plot.
  - (i) age, campaign and duration
  - (ii) nr.employed, euribor3m and duration

Suggestions for each chunk:

- Add some simple descriptive text in the text area before the code chunk.
- Add a name or description of each code chunk in {r}. Be sure that you *allow code and output from executing code to be included in the file from rendering A3_yourLastname_yourFirstname.Rmd.*
- Feel free to add comment lines with the requirement item numbers (e.g., # 3.A or # 3.B) to your code cell to help TAs and instructors easily identify your code that addresses a particular requirement.

Task III (10%)

Render A1_II_yourLastname_yourFirstname.Rmd to an HTML output file. Submit these files and A1_I_yourLastname_yourFirstName.xxx to Canvas. Please do not submit a zipped file.

A1: Data Exploration in R, Decision Tree, Classification Performance Evaluation (1)

Criteria	Ratings		Pts
Task I Preparation for coding and running R code in rstudio	**30 pts** **Full Marks**	**0 pts** **No Marks**	30 pts
Task II Rmd header, chunks and formatting.	**9 pts** **Full Marks**	**0 pts** **No Marks**	9 pts
2.1 Set up, data import and inspection code	**10 pts** **Full Marks**	**0 pts** **No Marks**	10 pts
2.2.A Create Histogram for four numeric variables (e.g., age, duration, campaign, and pdays)	**4 pts** **Full Marks**	**0 pts** **No Marks**	4 pts
2.2.B Create Boxplot for four numeric variables (e.g., age, duration, campaign, and pdays)	**4 pts** **Full Marks**	**0 pts** **No Marks**	4 pts
2.2.C Show Deciles for four numeric variables (e.g., age, duration, campaign, and pdays)	**4 pts** **Full Marks**	**0 pts** **No Marks**	4 pts
2.3.A Show the count value and percentage value (four factor variables)	**6 pts** **Full Marks**	**0 pts** **No Marks**	6 pts
2.3.B Plot Bar-plot (Two factor variables)	**6 pts** **Full Marks**	**0 pts** **No Marks**	6 pts
2.4.A Cor and pairs.panels to display correlations for age, duration, campaign, pday, euribor3m, emp.var.rate, and nr.employed.	**3 pts** **Full Marks**	**0 pts** **No Marks**	3 pts
2.4.B.i Boxplot of four numeric variables (duration, emp.var.rate, cons.price.idx, and cons.conf.idx.) by y	**4 pts** **Full Marks**	**0 pts** **No Marks**	4 pts
2.4.B.ii Use the aggregate function with summary to aggregate each of these four variables (duration, emp.var.rate, cons.price.idx, and cons.conf.idx.) by y	**4 pts** **Full Marks**	**0 pts** **No Marks**	4 pts
2.4.C.i Draw a 3d scatter plot to show y values in shapes for age, campaign and duration	**3 pts** **Full Marks**	**0 pts** **No Marks**	3 pts
2.4.C.ii Draw a 3d scatter plot to show y values in shapes for nr.employed, euribor3m and duration.	**3 pts** **Full Marks**	**0 pts** **No Marks**	3 pts

Criteria	Ratings		Pts
Task III Render A1_II_yourLastname_yourFirstname.Rmd to an HTML output file. Submit these files and A1_I_yourLastname_yourFirstName to Canvas. Please do not submit a zipped file.	**10 pts Full Marks**	**0 pts No Marks**	10 pts
Total Points: 100			