BSHC4 DAD / BSHTM4 P4BD
CA1 (30%) Sem. 1 2019/20

CA Instructions

- Download the CA dataset and starting R code from Moodle.
- Add your **student ID** to the set.seed() function, and run the starting code.
- Do the **first 3** foundation questions (F4 is optional).
- Do **at least 4** other questions (the best 4 will be used to determine your score).
- Copy + paste the output noting the columns you lost into a Word document.
- For each question, provide a screenshot of output, and a brief discussion of your approach in Word doc.
- You are required to upload a **.zip archive** with the following to Moodle:
  - The Word document with the answers
  - The R script file with the code you wrote for answering the questions
  - The sample data that you saved to the disk with the write.csv function: **my_ca_dataset.csv**

Questions

Foundations – up to 9 points (11 including the optional show-off question).
[F1] Ensure that each attribute is appropriately encoded, i.e. as a character, numeric, factor (with correct levels).
[F2] Identify which attributes have missing values, and how many. Identify which attributes have 'unknown' values, and how many.
[F3] Deal with an attribute that has missing values (show the outcome of your approach by redoing F2). Options to do this are:

[Basic] – removing the column
[Intermediate] – assigning the mean/median or mode (if categorical) value, or
[Advanced] – imputing the missing values

OPTIONAL Show-Off Foundations Question:
Check for outliers; if you find some, decide if you want to keep them in the dataset or not (defend your choice!) – you don't need to remove them though, just show how you would find them and discuss their presence.

Hint: some advanced/show-off questions can't deal with missing values. Once you've done F3, be pragmatic.

Basic:
[B1] Pick a numeric attribute and compute the mean, median, standard deviation, min, and max values.
[B2] Pick one of the categorical attributes and determine the most and least common values.
[B3] Does your dataset suggest if more customers subscribed a term deposit?
[B4] Visualise an attribute with at least 3 distinct values. Briefly explain what your plot shows.

Intermediate:
[I1] Propose a question concerning the deposit attribute and at least one numeric attribute and answer it. Briefly note if the answer is what you expected.
[I2] Propose a question concerning the median age and one categorical attribute and answer it.
[I3] Visualise one numeric attribute against a categorical attribute (not deposit) and interpret the results.
[I4] Visualise one numeric attribute against another numeric attribute and interpret the results.

Advanced:
[A1a/b] Visualise and correctly interpret the relationship of at least one categorical **and** one numeric attribute on deposit. You may do this question twice but **cannot** reuse attributes.
[A2] Which attribute(s) appear to be most influential on deposit? Be careful with missing values!
[A3] Convert Age into a categorical that captures different types of customers (explain your choice of Age ranges). Identify which type of customers seems to stay more on the phone (e.g., in terms of mean, median, etc.)?

Show-Off:
[S1] Build a logistic regression to predict deposit rate and comment on its performance.
[S2] Build a machine learning model, explain what it does and if applicable how it performs. Best (easiest) options here would be a C5.0 tree, Random Forest, Naïve Bayes, kNN, or SVM. Be careful with missing values!
[S3] Run a clustering algorithm either on all numerical (e.g. k-means), **OR** all categorical (e.g. k-medoids) attributes.

Data Description

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

Attribute information:

Input variables:
# bank client data:
1 - age (numeric)
2 - job : type of job (categorical: "admin.","blue-collar","entrepreneur","housemaid","management","retired","self-employed","services","student","technician","unemployed","unknown")
3 - marital : marital status (categorical: 1 - "divorced", 2 - "married", 3 - "single", 0 - "unknown"; note: "divorced" means divorced or widowed)
4 - education (categorical: "basic.4y","basic.6y","basic.9y","high.school","illiterate","professional.course","university.degree","unknown")
5 - default: has credit in default? (categorical: 1 - "no", 2 - "yes", 0 - "unknown")
6 - housing: has housing loan? (categorical: 1 - "no", 2 - "yes", 0 - "unknown")
7 - loan: has personal loan? (categorical: 1 - "no", 2 - "yes", 0 - "unknown")
# related with the last contact of the current campaign:
8 - contact: contact communication type (categorical: "cellular","telephone")
9 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
10 - day_of_week: last contact day of the week (categorical: "mon","tue","wed","thu","fri")
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y="no"). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

# other attributes:
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14 - previous: number of contacts performed before this campaign and for this client (numeric)
15 - poutcome: outcome of the previous marketing campaign (categorical: "failure","nonexistent","success")

# social and economic context attributes
16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
17 - cons.price.idx: consumer price index - monthly indicator (numeric)
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):
21 - deposit - has the client subscribed a term deposit? (binary: "yes","no")

Missing Attribute Values: There are several missing values in some categorical attributes, all coded with the "unknown" label. These missing values can be treated as a possible class label or using deletion or imputation techniques.

Marking Scheme

Expected value of questions

Question	Value	Notes
F1	up to 4	Based on level of correctness
F2	1 or 2	
F3	1, 2, or 3	Based on level attempted
F4	up to 2	There are many ways to do this, some are better than others
Basic	1.1	
Intermediate	2.5	
Advanced	4.5	
Show-Off	6.5	

Possible combinations of questions in reference to potential grades

	Foundations				4 Questions					
	F1	**F2**	**F3**	**F4**	**B**	**I**	**A**	**SO**	**Total (of 30)**	**Total (as %)**
Around Pass	4	2	1		4				11.4	38
	4	2	1		3	1			12.8	42.7
H2-2	4	2	2		2	2			15.2	50.6
	4	2	2		1	3			16.6	55.3
H2-1	4	2	2			4			18	60
	4	2	2			3	1		20	66.7
H1-1	4	2	3			2	2		23	76.7
	4	2	3			1	3		25	83.3
Exceptional	4	2	3				4		27	90
Max Score	4	2	3	2			1	3	35	116