STA5104 2022/23 2nd term Assignment 2

The dataset "bank-market1.csv" contains the cleaned dataset from assignment 1.

Column	Name	Description
1	age	continuous: age of the customer
2	marital	categorical: "married","divorced","single" (divorced include widowed)
3	education	categorical: "unknown","secondary","primary","tertiary"
4	balance	continuous: average yearly balance, in euros
5	housing	has housing loan? (binary: "yes","no")
6	loan	has personal loan? (binary: "yes","no")
7	duration	continuous: last contact duration, in seconds
8	campaign	no of contacts performed during this campaign (numeric)
9	pdays	no of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
10	previous	no of contacts performed before this campaign and for this client (numeric)
11	poutcome	outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")
12	deposit	has the client subscribed a term deposit? (binary: "yes","no")

Note that the last column deposit is the binary target variable. Furthermore, we define deposit="yes" as the positive group and deposit="no" as negative. That is,

Predict \ deposit	0	1
0	True negative	False negative
1	False positive	True positive

Q1 (knn)

(a) Read in the data and save them in *dc*. Change the last column in *dc* to numeric: *deposit=1* or *0* for *deposit="yes" or "no"* respectively and save them in *d*.

(b) Use the last 5 digits of your student id as random seed, select 80% of records from *d* as training dataset and save them in *d0*. The rest of the records are saved in *d1* as testing dataset. Change the target variable *deposit* in *d0* into a factor object and save them in *y0*.

(c) Select columns 1, 4 and 7 to 10 from *d0* and *d1* and scale them using *scale.con()* function and save them in *z0* and *z1* respectively.

(d) Use the improved *k_nn()* function to perform knn with *v=10*. Save the result in *bank.knn*. Which value of *k* gives the best result? Produce the classification table for this *k* and save it in *tab* and compute its error rate as well.

(e) Write the following function *f1sc(tab)* to compute and output the error rate, precision, recall and F1 score of the input 2x2 table *tab*.

```
f1sc<-function(tab) { # assume the input tab is 2x2 with 1st row an column as negative
   *** fill in your R codes here **
   cat('erate =',er, 'precision =', prec, ' recall =', recall, ' F1 score =', f1, '\n')
}
```

Use this *f1sc()* function to compute the F1 score of tab in part(d).

(f) Repeat part (c) to (e) using *stand()* function instead of *scale.con()* function.

Q2 (naive Bayes)

(a) Load the library *e1071*. Use *naiveBayes()* function in this library to perform naive Bayes on columns 1 to 11. Save the result in *bank.nb*.
(b) Compute the predicted *Pr{deposit=1|x}* and save them in *pr*. [Hint: use *type='raw'* in the *predict()*].
(c) Use the threshold value *c=0.5*, (i.e., predict *deposit=1* if *pr>c*). Produce the classification table and save it in *tab*.
(d) With the 2x2 table in part (c), compute the error rate and F1 score using the *f1sc()* function.
(e) Using a loop to repeat part (d) for *c=0.1,…, 0.9*. Which value of *c* gives the best F1 score?

Q3 (Logistic regression)

(a) Fit a logistic regression of deposit in *d0* with columns 1, 4, and 7 to 10. Use *step()* function to the model. Save your final model in *bank.lreg*.
(b) Find the probability of success *Pr{deposit=1|x}* in *d0* and save them in *pr0*. Using *c=0.5* as the threshold value, produce the classification table, compute the error rate and the F1 score for *d0*.
(c) Find the probability of success *Pr{deposit=1|x}* in *d1* and save them in *pr1*. Using *c=0.5* as the threshold value, produce the classification table, compute the error rate and the F1 score for *d1*.
(d) Using a loop to repeat part (c) for *c=0.1,…, 0.9*. Which value of *c* gives the best F1 score?

Q4 (Summary)
Summarize, compare and comment on the best error rate and F1 score in Q1, Q2 and Q3.

Submit your assignment via **blackboard** system on or before **March 26, 2023.**

You have to save and submit all the R commands in asg2.r. Your R commands should be commented as clearly as possible. Save your answers and outputs in asg2.doc or asg2.pdf.