Exam of Machine Learning

Master of Management

Spring 2022

Directives:

- “Open documents”.
- No communication allowed (emails, Whatsapp, etc.)!!
- 2 hours (9:00am to 11:00am)
- Steps:
  1. Download the question file (.pdf)
  2. Download the answer booklet (.docx)
  3. Write down your answers in the booklet (save often!!)
  4. At the end of the exam (11:00am), upload your answer booklet on moodle (check it is the latest version).
- No questions related to the exam content. Only for technical reasons.
- You are responsible for technical problems (make sure you have wifi, enough power, a working computer, etc.)

Context

The data set is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. The objective was to promotes a product subscription (bank term deposit).

For this exam, the data set, originally created by Paulo Cortez (Univ. Minho) and Sérgio Moro (ISCTE-IUL) @ 2012, was limited and modified.

The variables are:
1. *age* (numeric)
2. *job*: type of job (categorical: "admin.", "unknown", "unemployed", "management", "housemaid", "entrepreneur", "student", "blue-collar", "self-employed", "retired", "technician", "services")
3. *marital*: marital status (categorical: "married", "divorced", "single")
4. *education* (categorical: "unknown", "secondary", "primary", "tertiary")
5. *default*: has credit in default? (binary: "yes", "no")
6. *balance*: average yearly balance, in euros (numeric)
7. *housing*: has housing loan? (binary: "yes", "no")
8. *loan*: has personal loan? (binary: "yes", "no")
9. *contact*: contact communication type (categorical: "unknown", "telephone", "cellular")
10. *duration*: last contact duration, in seconds (numeric)
11. *campaign*: number of contacts performed during this campaign and for this client (numeric, includes last contact)
12. *previous*: number of contacts performed before this campaign and for this client (numeric)
13. *y* - has the client subscribed a term deposit? (binary: "yes", "no")

Each instance is associated to a client. There are 30907 instances.

The data were pretreated such that all instances are complete (no missing values).

The main objective of the study is to relate *y* (the subscription indicator) to the variables, although the exam questions may be related to sub-objectives or to more specific analysis aspects.

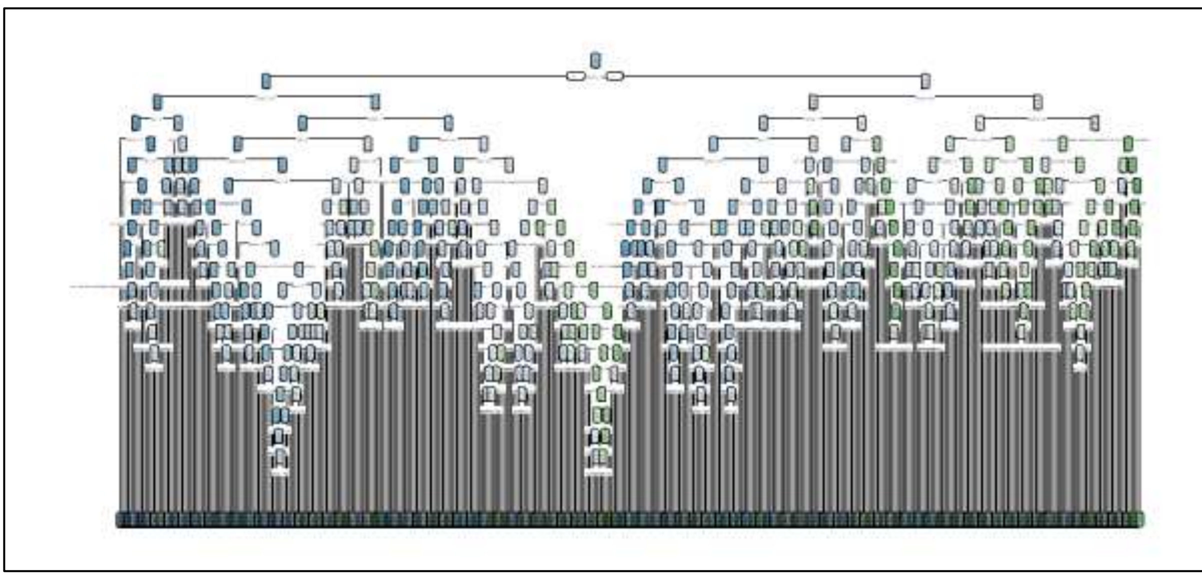
In the following, we use a data partition between training and test set (80/20, i.e., 23181/7726). Below, the complete data is bank, the training set is *bank.tr*, and the test set is *bank.te*.

Problem 1 (9pts)

The following analysis was performed.

Model 1

```
> mod.rp <- rpart(y ~ .,
+    data = bank.tr, control = list(cp=0.000005))
```



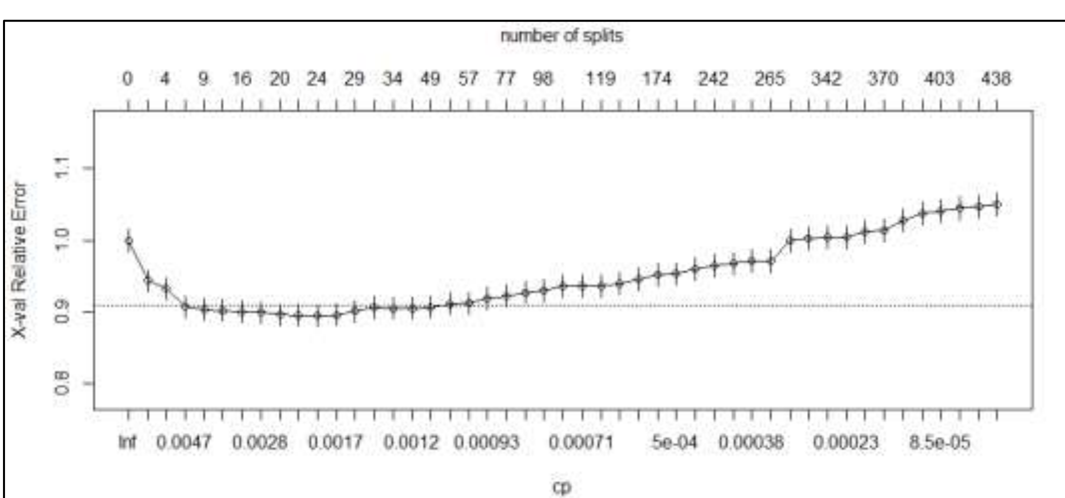
```
Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  19031  1318
       yes   765  2067

               Accuracy : 0.9101
```

Confusion matrix and accuracy on the training set

```
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  6131  714
       yes  467  414

               Accuracy : 0.8471
```

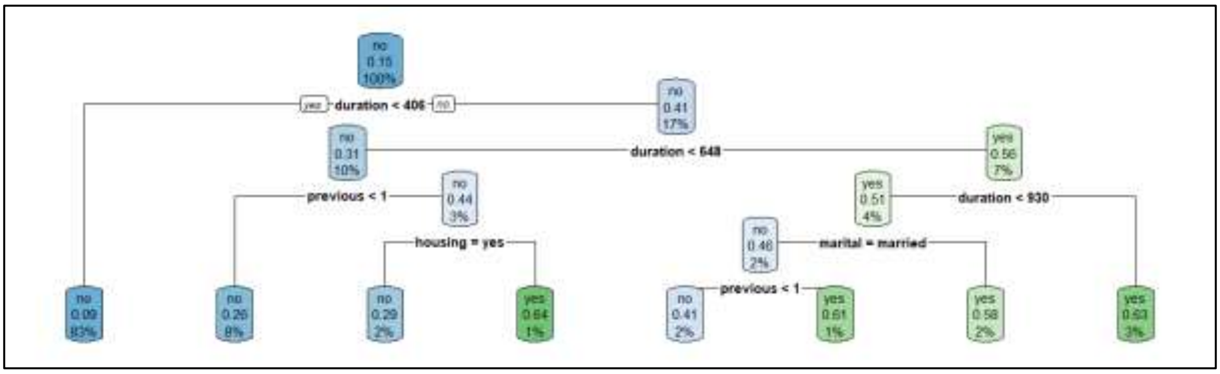
Confusion matrix and accuracy on the test set

```
> plotcp(mod.rp, upper="splits")
```



Model 2

```
> mod.rp.pruned <- prune(mod.rp, cp=0.0035)
> rpart.plot(mod.rp.pruned)
```




```
Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  19211  2418
       yes   585   967

              Accuracy : 0.8705
```

Confusion matrix and accuracy on the training set

```
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  6377  855
       yes  221  273

              Accuracy : 0.8607
```

Confusion matrix and accuracy on the test set

a. Write down what *Model 1* is (name/type of the ML model). (1pt)
Model 1 is a classification tree. [1]

b. Explain if *Model 1* is a good model and, if not, what it suffers from. Justify using the confusion matrix figures of *Model 1*. (2pts)
Model 1 suffers from overfitting [1]: the accuracy in the training set is larger than in the test set. [1]

c. Explain what the difference between *Model 1* and *Model 2* is, and, more precisely, how *Model 2* was build. What is the name of this method? (2pts)
Model 2 was pruned [1] to 8 nodes [0.5] using the 1-SE method [0.5].

d. Additionally, explain what improvement is expected from this method and, justifying using the available confusion matrix figures. (2pts)
Pruning simplifies the model which in turn avoid overfitting [1]. This is successful since the is a small difference between the apparent accuracy and the test set accuracy [1].

e. What is the prediction of the two following instances with *Model 2*? (2pts)

age	job	marital	education	default	balance	housing	loan	contact	duration	campaign	previous
33	entrepreneur	divorced	tertiary	no	37	no	yes	cellular	1082	1	0

age	job	marital	education	default	balance	housing	loan	contact	duration	campaign	previous
30	services	single	secondary	no	148	no	yes	cellular	482	3	0

Instance 1: Duration = 1082 => go right 3 times => predict yes [1]
Instance 2: Duration > 406 (right), Duration < 648 (left), Previous = 0 (left) => "no" [1]

Problem 2 (4pts)

The confusion matrix of *Model 2* on the test set is reported below (the same as in Problem 1).

		Reference (truth)		
		No	yes	Total
Prediction	no	6377	855	7232
	yes	221	273	494
	Total	6598	1128	7726

Given that "yes" is the positive class, compute

a. The sensitivity and the specificity (2pts)
   Sens = 273/1128 = 0.242 [1]
   Spec = 6377/6598 = 0.967 [1] [inversion of specificity and sensitivity is OK…]
b. The Cohen's Kappa. For this, you can use the matrix below that computes the expected frequencies under a random model. (2pts)

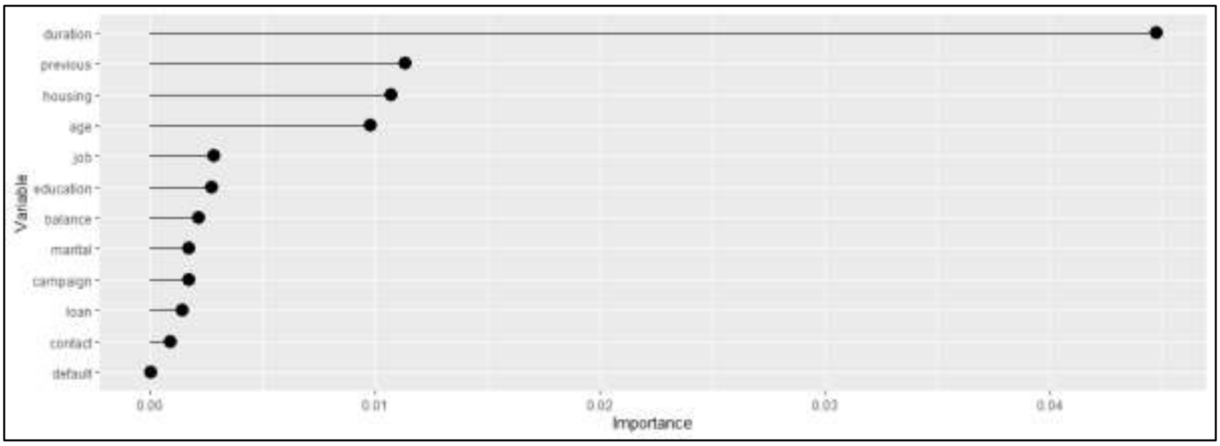
		Reference (truth)		
		No	yes	Total
Prediction	No	6176.1	1055.9	7232
	Yes	421.9	72.1	494
	Total	6598	1128	7726

Ae = (6176.1 + 72.1) / 7726 = 0.809 [1]
Kappa = (A – Ae) / (1 – Ae) = 0.271 [1]

Problem 3 (7pts)

A random forest (see below) was trained on the data and the following analysis was performed.

```
> mod.rf <- ranger(y~., data=bank.tr,
+                  importance = "permutation")
```



a. What can be concluded from this analysis in terms of the link between the outcome *y* (i.e., the subscription to a term deposit) and the observed features? Explain briefly by giving two or three examples. (3pts)
According to this variable importance measures [1], the duration is the most important feature for predicting the outcome y [1]. Mildly important features are housing, age, and job [0.5]. The remaining ones looks less important ones [0.5].

b. By construction, what is the main limitation of this analysis in terms of the links that can be measured? (2pts)
This analysis checks the importance of one variable at a time [1]. It cannot detect cases where two variables are dependent when the model can use either one or the other [1].

c. Briefly explain the main difference between random forests and an ensemble predictor made of bagged trees (that is, combining classification trees and BAGGING). (2pts)
In addition to bagging trees [1], random forests use an additional technique during the construction of each individual tree: each new split can only be made on a subset of the variables that is drawn at random [1].

Problem 4 (8pts)

The following logistic regression was fitted (***Model 3***).

```
> mod.lr <- glm(y~., data=bank.tr, family = "binomial")
> summary(mod.lr)

Call:
glm(formula = y ~ ., family = "binomial", data = bank.tr)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.2062  -0.5056  -0.3475  -0.2219   2.9930

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.567e+00  1.735e-01 -14.794  < 2e-16 ***
age                  3.498e-03  2.601e-03   1.345 0.178648
jobblue-collar      -4.814e-01  8.935e-02  -5.388 7.12e-08 ***
jobentrepreneur     -6.566e-01  1.540e-01  -4.264 2.01e-05 ***
jobhousemaid        -6.172e-01  1.641e-01  -3.760 0.000170 ***
jobmanagement       -3.261e-01  8.551e-02  -3.814 0.000137 ***
jobretired           4.140e-01  1.116e-01   3.708 0.000209 ***
jobself-employed    -4.523e-01  1.322e-01  -3.422 0.000621 ***
jobservices         -3.196e-01  1.008e-01  -3.170 0.001523 **
jobstudent           6.493e-01  1.261e-01   5.150 2.60e-07 ***
jobtechnician       -3.376e-01  8.020e-02  -4.210 2.56e-05 ***
jobunemployed       -2.313e-01  1.289e-01  -1.794 0.072798 .
maritalmarried      -7.240e-02  7.041e-02  -1.028 0.303836
maritalsingle        2.372e-01  8.015e-02   2.960 0.003079 **
educationsecondary   2.584e-01  7.821e-02   3.304 0.000954 ***
educationtertiary    5.782e-01  8.966e-02   6.449 1.13e-10 ***
defaultyes          -8.065e-01  2.446e-01  -3.298 0.000974 ***
balance              2.101e-05  5.949e-06   3.532 0.000413 ***
housingyes          -9.155e-01  4.752e-02 -19.265  < 2e-16 ***
loanyes             -7.293e-01  7.211e-02 -10.113  < 2e-16 ***
contacttelephone    -1.877e-01  8.342e-02  -2.251 0.024408 *
duration             3.816e-03  8.002e-05  47.691  < 2e-16 ***
campaign            -1.495e-01  1.267e-02 -11.795  < 2e-16 ***
previous             1.021e-01  8.078e-03  12.643  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 19275  on 23180  degrees of freedom
Residual deviance: 14904  on 23157  degrees of freedom
AIC: 14952

Number of Fisher Scoring iterations: 6
```

Then the following procedure was applied.

```
> mod.lr.sel <- step(mod.lr)
Start:  AIC=14952.01
y ~ age + job + marital + education + default + balance + housing +
    loan + contact + duration + campaign + previous

            Df Deviance   AIC
- age        1    14906 14952
<none>            14904 14952
- contact    1    14909 14955
- balance    1    14916 14962
- default    1    14917 14963
- marital    2    14937 14981
- education  2    14952 14996
- loan       1    15020 15066
- previous   1    15063 15109
- job       10    15086 15114
- campaign   1    15082 15128
- housing    1    15296 15342
- duration   1    17864 17910

Step:  AIC=14951.81
y ~ job + marital + education + default + balance + housing +
    loan + contact + duration + campaign + previous

            Df Deviance   AIC
<none>            14906 14952
- contact    1    14910 14954
- balance    1    14918 14962
- default    1    14919 14963
- marital    2    14938 14980
- education  2    14952 14994
- loan       1    15023 15067
- previous   1    15065 15109
- campaign   1    15084 15128
- job       10    15108 15134
- housing    1    15312 15356
- duration   1    17870 17914
```

This led to the following new model (***Model 4***)

```
> summary(mod.lr.sel)

Call:
glm(formula = y ~ job + marital + education + default + balance +
    housing + loan + contact + duration + campaign + previous,
    family = "binomial", data = bank.tr)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.2145  -0.5060  -0.3471  -0.2220   2.9854

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.401e+00  1.212e-01 -19.809  < 2e-16 ***
jobblue-collar     -4.844e-01  8.933e-02  -5.422 5.89e-08 ***
jobentrepreneur    -6.521e-01  1.539e-01  -4.237 2.26e-05 ***
jobhousemaid       -6.059e-01  1.638e-01  -3.699 0.000217 ***
jobmanagement      -3.224e-01  8.544e-02  -3.773 0.000161 ***
jobretired          4.827e-01  9.925e-02   4.864 1.15e-06 ***
jobself-employed   -4.500e-01  1.322e-01  -3.405 0.000661 ***
jobservices        -3.233e-01  1.008e-01  -3.209 0.001334 **
jobstudent          6.180e-01  1.239e-01   4.989 6.07e-07 ***
jobtechnician      -3.372e-01  8.018e-02  -4.206 2.60e-05 ***
jobunemployed      -2.272e-01  1.288e-01  -1.764 0.077780 .
maritalmarried     -8.016e-02  7.018e-02  -1.142 0.253373
maritalsingle       2.000e-01  7.521e-02   2.660 0.007825 **
educationsecondary  2.480e-01  7.781e-02   3.187 0.001439 **
educationtertiary   5.641e-01  8.904e-02   6.336 2.36e-10 ***
defaultyes         -8.038e-01  2.444e-01  -3.288 0.001007 **
balance             2.168e-05  5.925e-06   3.659 0.000254 ***
housingyes         -9.236e-01  4.715e-02 -19.590  < 2e-16 ***
loanyes            -7.325e-01  7.209e-02 -10.161  < 2e-16 ***
contacttelephone   -1.710e-01  8.242e-02  -2.074 0.038051 *
duration            3.817e-03  8.000e-05  47.717  < 2e-16 ***
campaign           -1.495e-01  1.267e-02 -11.796  < 2e-16 ***
previous            1.023e-01  8.080e-03  12.666  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 19275  on 23180  degrees of freedom
Residual deviance: 14906  on 23158  degrees of freedom
AIC: 14952

Number of Fisher Scoring iterations: 6
```

a. In *Model 4*, give an interpretation of the coefficients associated with "*duration*" and with "*maritalmarried*". (Note: the reference level for the variable *marital* is "divorced") (3pts)
The coefficient associated to duration is 0.003817. This means that the linear predictor increases by 0.003817 for each unit increase of the duration (everything else being held fixed) [1], and thus probability of yes increases when the duration increases [0.5]. The coefficient of the level married (in marital factor) is -0.08016. This means that the linear predictor increases by -0.08016 if marital changes from divorced to married (everything else being held fixed) [1], and thus probability of yes for married than for divorced customer [0.5].

b. The value of the linear predictor for the instance below is -2.5785,

age	job	marital	education	default	balance	housing	loan	contact	duration	campaign	previous
39	technician	married	secondary	no	22	no	no	cellular	76	2	0

Compute the prediction for the same instance but where default would be "yes" instead of "no". Provide the intermediate calculations (linear predictor, probability, and prediction). (2pts)

Linear predictor = -2.5785 – 0.8038 = -3.3823 [1]

Probability = exp(-3.3823) / (1 + exp(-3.3823)) = 0.03285 [0.5]

Prediction = "no" (0.03825 < 0.5) [0.5]

c. What was the modification brought to *Model 3* in order to build *Model 4*? Briefly explain this method by mentioning its name, its purpose, and how to read the "step method" results. (3pts)

A variable selection based on the AIC (Akaike Information Criterion) was performed [1]. Its purpose is to simplify the model by removing uninteresting variables according to the AIC, making it more robust and less prone to overfitting [1].

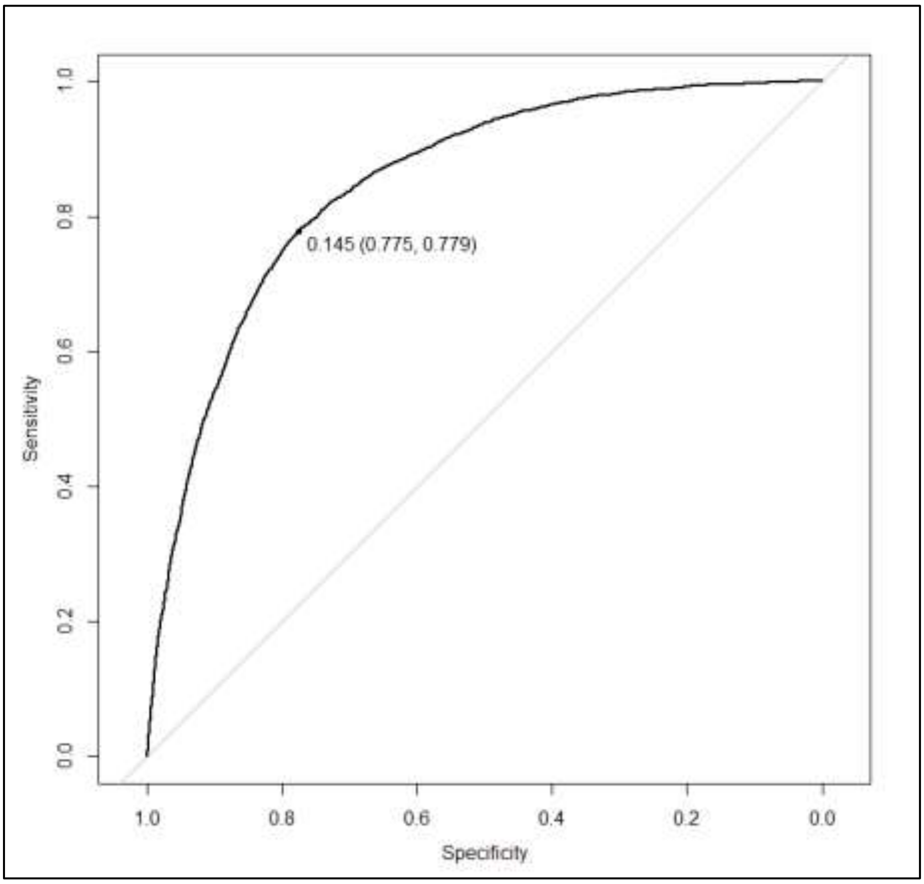
We start with the full model, then, at each step, we select the model with the lowest AIC among the ones with one less variable. This is repeated until the lowest AIC is reached. [1]

Problem 5 (6pts)

Following Problem 4, the metrics of *Model 4* were computed on the test set.

```
> prob.lr.te <- predict(mod.lr.sel, newdata=bank.te, type="response")
> pred.lr.te <- factor(ifelse(prob.lr.te > 0.5, "yes", "no"))
> confusionMatrix(data=pred.lr.te, reference = bank.te$y, positive = "yes")
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  6430  914
       yes  168  214

               Accuracy : 0.86
                 95% CI : (0.852, 0.8676)
    No Information Rate : 0.854
    P-Value [Acc > NIR] : 0.07072

                  Kappa : 0.2263

 Mcnemar's Test P-Value : < 2e-16

            Sensitivity : 0.18972
            Specificity : 0.97454
         Pos Pred Value : 0.56021
         Neg Pred Value : 0.87554
             Prevalence : 0.14600
         Detection Rate : 0.02770
   Detection Prevalence : 0.04944
      Balanced Accuracy : 0.58213

       'Positive' Class : yes
```

Then, the ROC curve was built (on the training set):



a. Analyze the metrics and explain what the issue with these data is and why the accuracy may not be a good metric (Hint: you may also use the EDA available in Appendix). (2pts)
We see that the sensitivity is much lower than the specificity. This is because there are many more "no" than "yes" in the data set (see EDA, e.g.) [1]. The accuracy is not a good metric here since even a model predicting only "no" would have a large accuracy [0.5]. The balanced accuracy is more adapted in this case [0.5].

b. From the ROC curve figure, explain how this problem may be (partially) solved with *Model 4*. To do so, explain what "0.145 (0.775, 0.779)" stands for. (2pts)

This problem may be (partially) solved by using 0.145 as the prediction threshold for the model [1]. That would lead to a specificity of 0.775 and sensitivity of 0.779 (or the inverse…) [1].

c. The same analysis was repeated on another data set (*bank.tr.bl*) built from the training set with the code below. Briefly explain what this method is by explaining how it works, what its purpose is, and, in the case of these data, whether it worked. (2pts)

This method balanced the data by subsampling: a new training data set is built with all the "yes" and a random subsample of "no" of the same size as the "yes" (3385) [1]. This increases the weights of "yes" in the training process of the model. The imbalance of the data is corrected as shown on the balanced accuracy on the test set (0.776) [1].

```
> table(bank.tr$y)

   no   yes
19796  3385
> index.no <- bank.tr$y=="no"
> index.yes <- bank.tr$y=="yes"
> (n.no <- sum(index.no))
[1] 19796
> (n.yes <- sum(index.yes))
[1] 3385
> set.seed(367)
> bank.tr.bl <- rbind(bank.tr[index.yes,],
+                 bank.tr[sample(which(index.no), size=n.yes),])
> table(bank.tr.bl$y)

  no  yes
3385 3385
```

```
> mod.lr.bl <- glm(y~., data=bank.tr.bl, family = "binomial")
> mod.lr.bl <- step(mod.lr.bl, trace=FALSE)
> prob.lr.te <- predict(mod.lr.bl, newdata=bank.te, type="response")
> pred.lr.te <- factor(ifelse(prob.lr.te > 0.5, "yes", "no"))
> confusionMatrix(data=pred.lr.te, reference = bank.te$y, positive="yes")
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  5200  267
       yes 1398  861

               Accuracy : 0.7845
                 95% CI : (0.7752, 0.7936)
    No Information Rate : 0.854
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3895

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7633
            Specificity : 0.7881
         Pos Pred Value : 0.3811
         Neg Pred Value : 0.9512
             Prevalence : 0.1460
         Detection Rate : 0.1114
   Detection Prevalence : 0.2924
      Balanced Accuracy : 0.7757

       'Positive' Class : yes
```

Problem 6 (3pts) (it was written 5pts but the details of points is 3pts; see below)

The following code using caret was run to obtain another model.

```
trctrl <- trainControl(method = "cv", number = 5)#, search="random")
grid <- expand.grid(sigma=c(0.01, 0.02, 0.03, 0.05), C=c(10, 12, 15))
bank.svm.rad <- train(y ~., data = bank.tr, method = "svmRadial",
                      trControl=trctrl,
                      tuneGrid = grid)
```

```
> bank.svm.rad
Support Vector Machines with Radial Basis Function Kernel

4636 samples
  12 predictor
   2 classes: 'no', 'yes'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 3709, 3709, 3708, 3709, 3709
Resampling results across tuning parameters:

  sigma  C   Accuracy   Kappa
  0.01   10  0.8563439  0.2586067
  0.01   12  0.8548341  0.2623804
  0.01   15  0.8556969  0.2722056
  0.02   10  0.8563449  0.3007468
  0.02   12  0.8578553  0.3166945
  0.02   15  0.8578549  0.3245180
  0.03   10  0.8561296  0.3215040
  0.03   12  0.8541883  0.3142318
  0.03   15  0.8550508  0.3221221
  0.05   10  0.8526773  0.3200255
  0.05   12  0.8507351  0.3144661
  0.05   15  0.8507349  0.3206447

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were sigma = 0.02 and C = 12.
```

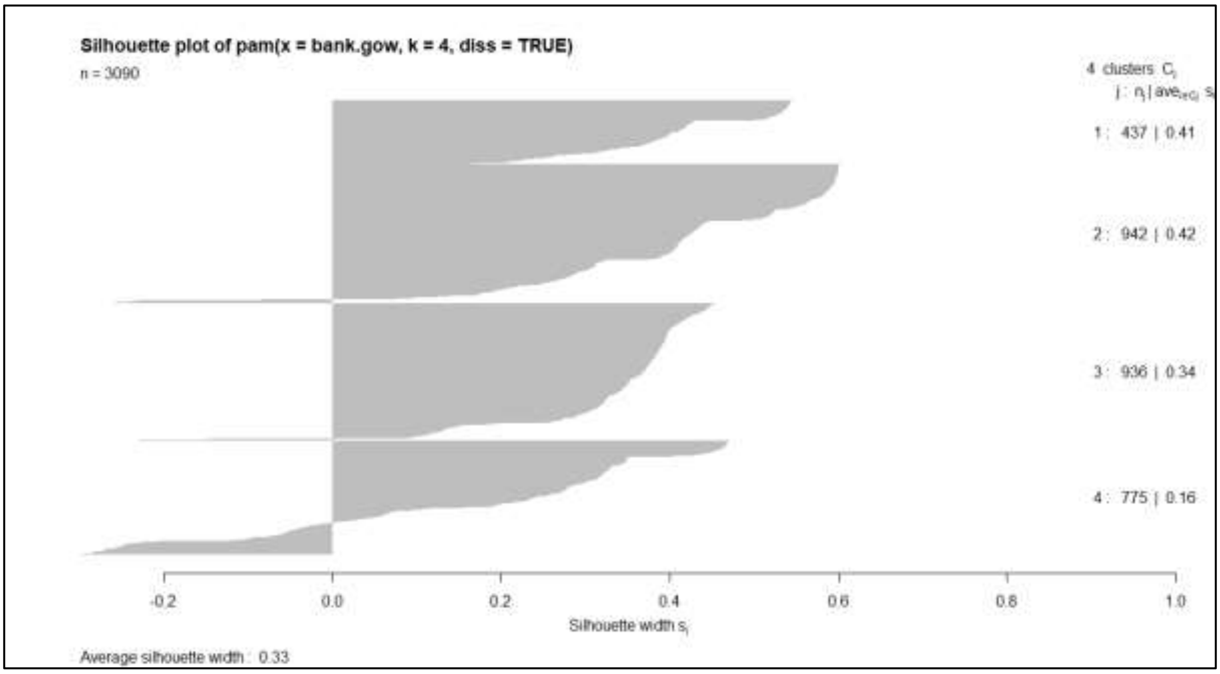
Briefly explain (3pts)

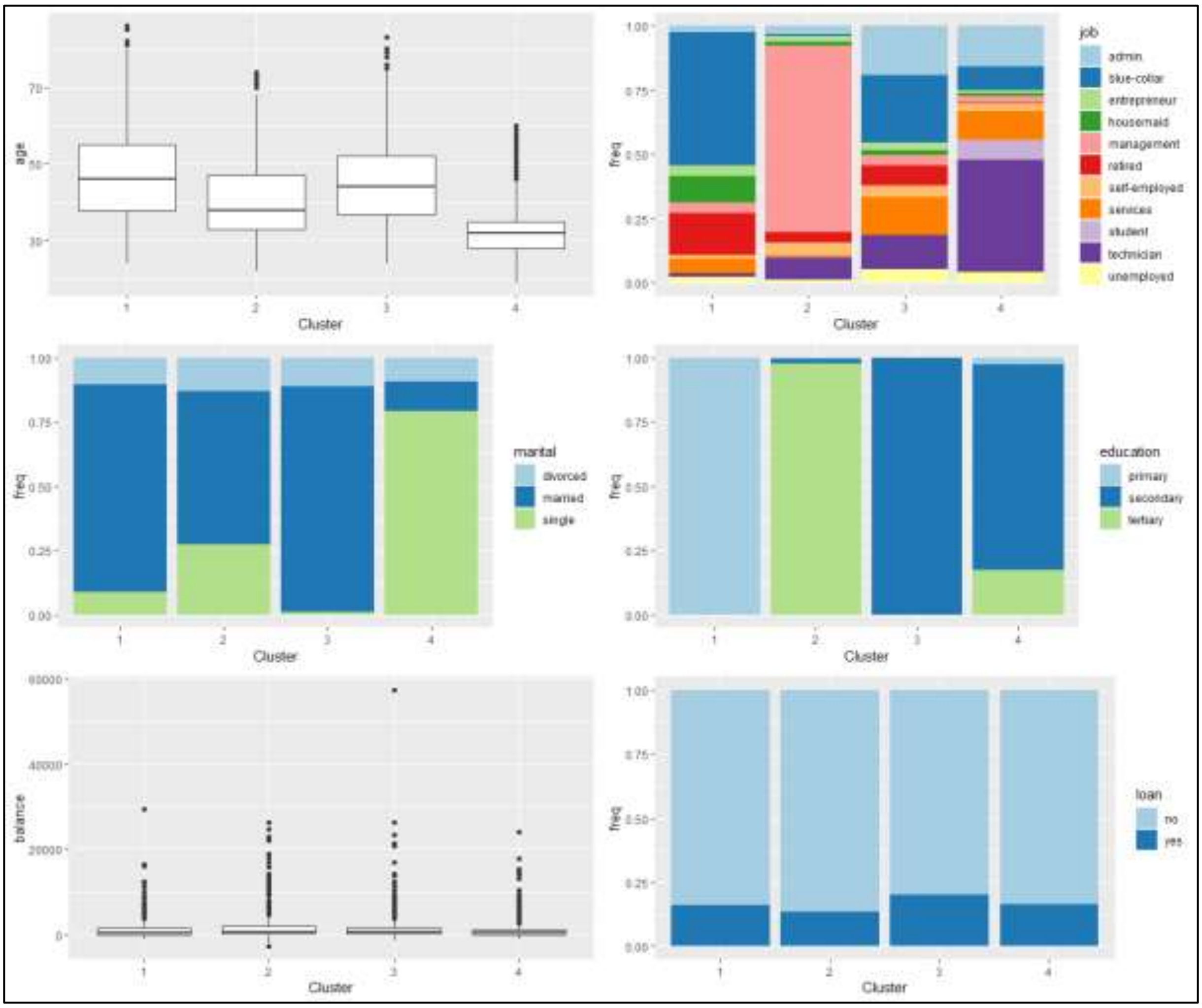
a. What is this model (name)? On what parameters is it tuned?
b. What is the splitting strategy?
c. How are the models evaluated? What is the optimal model?

. a) It is a Support Vector Machine model with a radial kernel [1]. b) The splitting strategy is 5-fold cross-validation [1]. c) The models are evaluated with the accuracy (and kappa). The best model is sigma=0.02 and C=12. [1]

Problem 7 (5pts)

In this problem, we build four clusters of customers based on the variables 1 to 4, 6, and 8 (*age*, *job*, *marital*, *education*, *balance*, *loan*). The cluster are built using PAM with a Gower's distance.





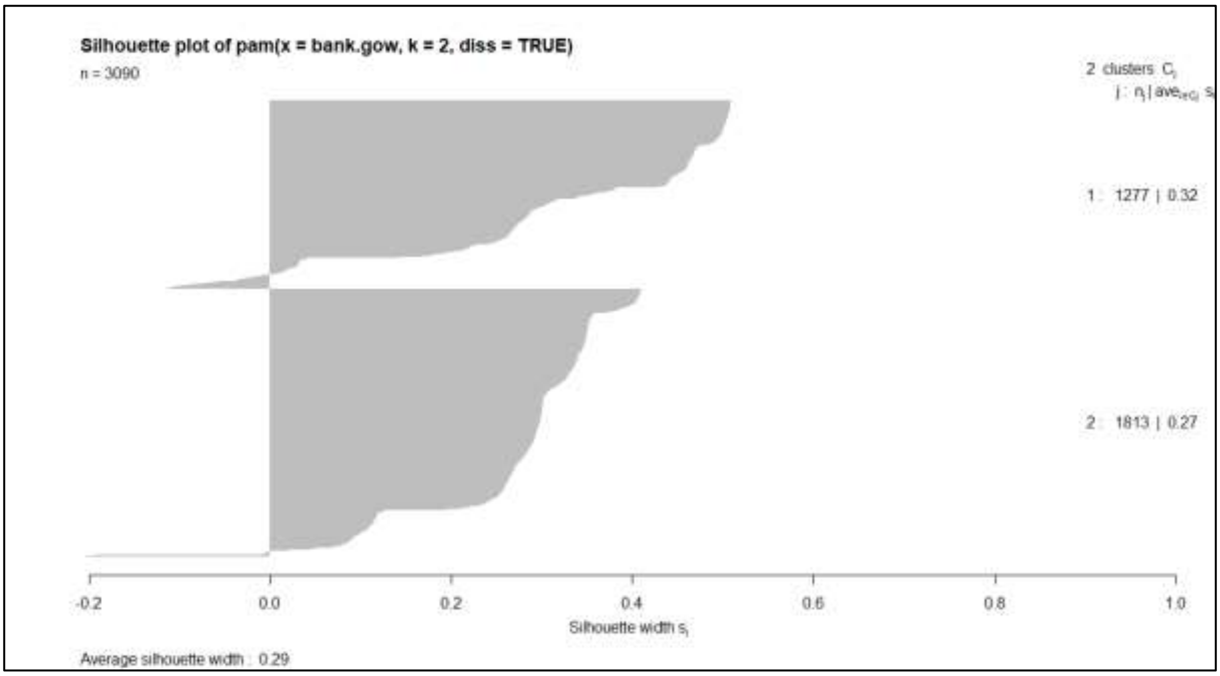
a. What can be said about Cluster 4 compared to Cluster 3? (3pts)
   1) In terms of its homogeneity (i.e., how well/badly their members are clustered).
   2) In terms of their profiles (i.e., the features of their members).

   .1) The silhouette profile shows that Cluster 3 is more homogeneous than Cluster 4 [1] since less instances have a negative silhouette in C3 [0.5].

   .2) Some striking features: age in C3 > age in C4 (in general), C3 has more blue-collars/admins vs C4 has more technicians, C3 has more married people vs C4 has more single people, C3 has only secondary educated people vs C4 has more tertiary educated people. [0.5 each, max. 1.5]

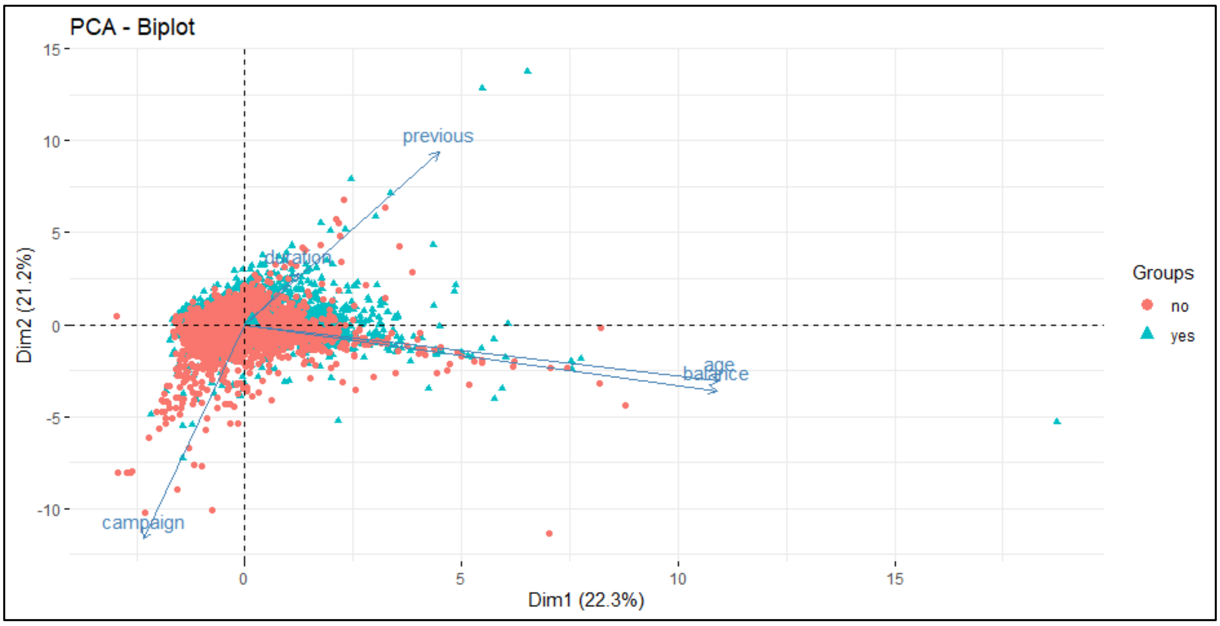
b. Clustering in two clusters with the same method provides the following silhouette plot. With this information, should you prefer to make 2 or 4 clusters? Justify. (2pts)

   The average silhouette of the clustering is 0.33 for k=4 and 0.29 for k=2 [1]. Based on this criterion, we should prefer k=4. [1]



Problem 8 (4pts)

In this problem, we analyze the links between the five variables *age*, *balance*, *duration*, *previous*, and *campaign* using principal component analysis. The biplot below show the result for Dimensions (1, 2) (the two first principal components). On the biplot, the groups (colors) correspond to the outcome y being "yes" or "no", although that variable was not used for the PCA itself.



a. Briefly describe the links between the variables, especially between *age* and *balance*, and additionally between *previous* and *campaign*. (2pts)
From the biplot, we can see that age and balance are positively correlated [1], that previous and campaign are negatively correlated (arrows in same / opposite directions respectively) [1].

b. Is there a link between the outcome (“yes” / “no”) and the five variables that is revealed by the biplot? Is this coherent with the results previously seen from the models especially in Problems 3 and 4. (2pts)
We can see that “yes” is more frequently found when previous (and duration) are large [1]. This is coherent with the previous results like in Problem 3 where we saw that these variables are the most important [0.5], and in Problem 4 where we saw that the coefficients associated with these two variables are positive [0.5].

Data Frame Summary

bank

Dimensions: 45211 x 10
Duplicates: 2581

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	age [numeric]	Mean (sd) : 40.9 (10.6) min ≤ med ≤ max: 18 ≤ 39 ≤ 95 IQR (CV) : 15 (0.3)	77 distinct values		45211 (100.0%)	0 (0.0%)
2	job [character]	1. admin. 2. blue-collar 3. entrepreneur 4. housemaid 5. management 6. retired 7. self-employed 8. services 9. student 10. technician 11. unemployed	5171 (11.5%) 9732 (21.7%) 1487 (3.3%) 1240 (2.8%) 9458 (21.1%) 2264 (5.0%) 1579 (3.5%) 4154 (9.2%) 938 (2.1%) 7597 (16.9%) 1303 (2.9%)		44923 (99.4%)	288 (0.6%)
3	marital [character]	1. divorced 2. married 3. single	5207 (11.5%) 27214 (60.2%) 12790 (28.3%)		45211 (100.0%)	0 (0.0%)
4	education [character]	1. primary 2. secondary 3. tertiary	6851 (15.8%) 23202 (53.5%) 13301 (30.7%)		43354 (95.9%)	1857 (4.1%)
5	default [character]	1. no 2. yes	44396 (98.2%) 815 (1.8%)		45211 (100.0%)	0 (0.0%)
6	balance [numeric]	Mean (sd) : 1362.3 (3044.8) min ≤ med ≤ max: -8019 ≤ 448 ≤ 102127 IQR (CV) : 1356 (2.2)	7168 distinct values		45211 (100.0%)	0 (0.0%)
7	housing [character]	1. no 2. yes	20081 (44.4%) 25130 (55.6%)		45211 (100.0%)	0 (0.0%)
8	loan [character]	1. no 2. yes	37967 (84.0%) 7244 (16.0%)		45211 (100.0%)	0 (0.0%)
9	contact [character]	1. cellular 2. telephone	29285 (91.0%) 2906 (9.0%)		32191 (71.2%)	13020 (28.8%)
10	y [factor]	1. no 2. yes	39922 (88.3%) 5289 (11.7%)		45211 (100.0%)	0 (0.0%)

Generated by summarytools 1.0.0 (R version 4.0.2)
2022-04-27