Example Exam of Machine Learning

Master of Management – Business Analytics

Adapted from Exam 2024

Context

A retailer is facing an important return rate in one of its businesses: customers are ordering products online and once the order is delivered, the customer can decide to return it. This issue increases delivery costs while not generating sales. It also creates inefficiencies regarding stock level management as returned products travel back to the warehouses without being planned.

The retailer wants to reduce the return rate to avoid or reduce the previously mentioned issues.

The retailer has lots of data (product ordered, payment method, region of order, lead time for delivery, etc.) for each order and return customers make, which allows various analyses to be performed. Some of these data points are used for the exam.

The data set has the following features:

- Return: binary indicator of returning product, *Keep* or *Return*.
- Payment_Method: categorical, 4 levels, *Cash on delivery, Credit Card, Free, Invoice.*
- Price: price of the delivery, numerical (currency).
- Subscription_type: categorical, 6 levels, (confidential), *S1, ..., S6*.
- Delivery.frequency: ordinal, 6 levels, *2 week, 1 month, 2 month, 3 month, 4 month, 6 month*.
- Category_level_1: categorical 8 levels (confidential) *C1, ..., C8*.
- Number_of_orders_already_passed_by_the_customer: numerical.
- Number_of_subsciptions_for_the_customer: binary, *One* or *More*.
- Age: numerical (years).
- Gender: categorical, 3 levels, *Male, Female, Not Specified*.
- log_sales_amount: logarithm of the total amount already bought, numerical (log-currency).

You can find an EDA of the data at the end of the exam.

Problem 1: concepts (12pts)

a. To what task and sub-task does the problem of predicting the Return from the other features refer? (1pt)
It is a supervised learning task [0.5]. More specifically, binary classification [0.5].

b. Write down a list of at least three models that can be used to predict the Return from the other features (name/type of the models). (1pt)
Possible models: logistic regression, classification tree, neural networks, random forest. [1 for any three models]

c. In few sentences, and in broad terms, explain the concept of overfitting:
   (i) what overfitting is and why it is a problem. (1pt)
   (ii) how overfitting can be detected (mention one method). (1pt)
   (iii) how overfitting is related to model complexity. (2pts)
(i) Overfitting occurs when the model's prediction performance cannot be generalized outside the training set [0.5]. The model is useless to predict unseen instances [0.5]. (ii) One method is to compare the metrics on the training set and on the test set. They should be close [1]. (iii) Complex models are more prone to overfit the training set [2].

d. Cite two methods that can be used to solve overfitting. You can explain a general method and/or a model-specific method. (2pts)
Generally, hyperparameter tuning can be used to diminish overfitting. To a lesser extent, bagging can moderate overfitting. For regression models (logistic or linear) AIC based variable selection or penalized loss (LASSO, Ridge, and elastic net) can be used. For CART, pruning can be used. For neural networks, penalization can also be used. [1+1; any two methods].

e. In broad terms and few sentences, explain the issue of imbalanced classes in classification:
   (i) what it is and why it is a problem. (1pt)
   (ii) how it can be detected (mention one method). (1pt)
   (iii) how it can be solved (mention one method). (2pts)
(i) Imbalanced class issue arises when some classes of the response are overrepresented in the training set. [0.5]. It is a problem because models will tend to predict only the most common class [0.5]. (ii) It can be detected when specificity and sensitivity are very different. Also, when the balanced accuracy is lower than the accuracy [1]. (iii) To solve it, one can either use up- or down-sampling. One can also optimize the probability threshold on the ROC curve [2; any method].

Problem 2: calculations (9pts)

a. A model was fitted on a part of the data. The confusion matrix on the training set is shown below. Compute the apparent balanced accuracy. Justify your calculation by providing all the intermediate calculations. (4pts)

		Reference (truth)		
		Keep	Return	Total
Prediction	Keep	83022	8841	91863
	Return	906	1627	2533
	Total	83928	10468	94396

The balanced accuracy is the average of the specificity and the sensitivity [1pt].
Sens = 83022 / (83022 + 906) = 0.989[1pt]
Spec = 1627 / (8841 + 1627) = 0.155 [1pt]
Bal. Acc. = (0.989 + 0.155) / 2 = 0.5723 [1pt]

b. The summary of a logistic regression fitted on the data is shown below (note: the positive class is "Return").

Summary of logistic regression

```
Call:
glm(formula = Return ~ Payment_Method + Price + Delivery_frequency +
    Age, family = "binomial", data = dat_tr)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -1.1e+00    5.9e-02   -18.9   <2e-16 ***
Payment_MethodCredit Card -1.9e+00    2.4e-02   -77.8   <2e-16 ***
Payment_MethodFree        -2.3e+00    9.1e-02   -24.9   <2e-16 ***
Payment_MethodInvoice     -1.2e+01    7.6e+01    -0.2    0.874
Price                      1.3e-04    4.0e-05     3.2    0.001 **
Delivery_frequency2 month  9.6e-01    4.2e-02    22.9   <2e-16 ***
Delivery_frequency2 week  -1.1e+00    3.6e-01    -3.1    0.002 **
Delivery_frequency3 month  6.5e-01    4.7e-02    13.9   <2e-16 ***
Delivery_frequency4 month  7.2e-01    5.6e-02    12.8   <2e-16 ***
Delivery_frequency6 month -9.1e-01    1.0e-01    -8.9   <2e-16 ***
Age                       -1.4e-02    8.2e-04   -16.5   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Compute the prediction of the following instance (Instance 1). Provide intermediate calculations. (3pts)

```
Payment_Method                                   "Credit Card"
Price                                            "817"
Subscription_type                                "S1"
Delivery_frequency                               "1 month"
Category_level_1                                 "C4"
Number_of_orders_already_passed_by_the_customer  "5"
Number_of_subsciptions_for_the_customer          "One"
Age                                              "49"
Gender                                           "Female"
log_sales_amount                                 "6.705639"
```

Linear predictor: -1.1 – 1.9 + 0.00013*817 + 0 – 0.014*49 = -3.58 [1]

Predicted Probability: exp(-3.58) / (1 + exp(-3.58)) = 0.0275
Predicted class: 0.275 < 0.5 => "Keep"

c. Consider an instance (Instance 2) that would be the same as Instance 1, except for the Payment_Method being "Free" instead of "Credit Card". Answer by TRUE or FALSE to the statements below, each time, briefly justifying your answer.
   (i) The predicted class of Instance 2 would be different than the one of Instance 1. (1pt)
   (ii) The prediction of Instance 2 is more certain (the model has higher confidence) than the one of Instance 1. (1pt)

   Because the coefficient of Free is more negative than the one of Credit Card, the probability would be lower for Instance 2. Thus, (i) is FALSE because the probability being smaller, it will also be below the threshold of 0.5 [1]. And (ii) is TRUE because the probability being lower, it is even further than 0.5 and considered certain for "Keep". [1]

Problem 3: model quality (9pts)

After having trained a random forest on 80% of the data set, the following results are obtained. On the left, the results on the training set, and, on the right, the results on the test set.

Training set

```
Confusion Matrix and Statistics

          Reference
Prediction  Keep Return
    Keep   83482   3270
    Return   446   7198

               Accuracy : 0.9606
                 95% CI : (0.9594, 0.9619)
    No Information Rate : 0.8891
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7736

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.68762
            Specificity : 0.99469
         Pos Pred Value : 0.94165
         Neg Pred Value : 0.96231
             Prevalence : 0.11089
         Detection Rate : 0.07625
   Detection Prevalence : 0.08098
      Balanced Accuracy : 0.84115

       'Positive' Class : Return
```

Test set

```
Confusion Matrix and Statistics

          Reference
Prediction  Keep Return
    Keep   20659   1592
    Return   323   1024

               Accuracy : 0.9188
                 95% CI : (0.9153, 0.9223)
    No Information Rate : 0.8891
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4774

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.39144
            Specificity : 0.98461
         Pos Pred Value : 0.76021
         Neg Pred Value : 0.92845
             Prevalence : 0.11086
         Detection Rate : 0.04339
   Detection Prevalence : 0.05708
      Balanced Accuracy : 0.68802

       'Positive' Class : Return
```

a. What are the two main issues that can be diagnosed from these confusion matrices? Explain in a few sentences by mentioning the issues and justifying using the results from the two confusion matrices. (3pts)

   The difference between specificity and sensitivity [0.5] in both cases reveals a problem of imbalanced data [1]. In addition, the model overfits the data [1] as shown by the difference in balanced accuracy [0.5].

b. A diagnostic was posed by the analyst and the following adaptation/code was performed. In a few sentences, explain which strategy was used and if it did solve the issues identified previously. (3pts)

```
dat_tr_bal <- downSample(y = dat_tr$Return, x=dat_tr[,-1],
                         yname = "Return")
mod.rf_bal <- ranger(Return~., data=dat_tr_bal)

rf.pred_tr <- predict(mod.rf_bal, data=dat_tr_bal, type="response")
confusionMatrix(reference=dat_tr_bal$Return, data=rf.pred_tr$predictions,
                positive="Return")
rf.pred_te <- predict(mod.rf_bal, data=dat_te, type="response")
confusionMatrix(reference=dat_te$Return, data=rf.pred_te$predictions,
                positive="Return")
```

Note: dat_tr and dat_te are the training and test set respectively.

Training set

```
Confusion Matrix and Statistics

          Reference
Prediction  Keep Return
    Keep    9862    446
    Return   606  10022

               Accuracy : 0.9498
                 95% CI : (0.9467, 0.9527)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8995

 Mcnemar's Test P-Value : 9.478e-07

            Sensitivity : 0.9574
            Specificity : 0.9421
         Pos Pred Value : 0.9430
         Neg Pred Value : 0.9567
             Prevalence : 0.5000
         Detection Rate : 0.4787
   Detection Prevalence : 0.5076
      Balanced Accuracy : 0.9498

       'Positive' Class : Return
```

Test set

```
Confusion Matrix and Statistics

          Reference
Prediction  Keep Return
    Keep   16889    437
    Return  4093   2179

               Accuracy : 0.808
                 95% CI : (0.803, 0.813)
    No Information Rate : 0.8891
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3958

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

            Sensitivity : 0.83295
            Specificity : 0.80493
         Pos Pred Value : 0.34742
         Neg Pred Value : 0.97478
             Prevalence : 0.11086
         Detection Rate : 0.09234
   Detection Prevalence : 0.26579
      Balanced Accuracy : 0.81894

       'Positive' Class : Return
```

The strategy was to down sample [1] the data to equalize the two classes [1]. It solved the issue of imbalance data (specificity and sensitivity are now closer) but not the problem of overfitting as shown by the difference in the apparent and test balanced accuracy [1].

c. A further analysis was performed and is shown below. Explain in a few sentences what is the purpose of the code and what can be concluded from the results. (3pts)

Code:

```
control <- trainControl(method='cv',
                        number=5,
                        search="random")
set.seed(1)
rf_random <- train(Return ~ .,
                   data = dat_tr_bal,
                   method = 'rf',
                   metric = 'Accuracy',
                   tuneLength  = 10,
                   trControl = control)
```

Result:

```
Random Forest

5000 samples
  10 predictor
   2 classes: 'Keep', 'Return'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 3999, 4000, 4000, 4000, 4001
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   3    0.7670071  0.5340141
   4    0.7672087  0.5344181
   7    0.7714097  0.5428162
   8    0.7716079  0.5432135
  12    0.7692075  0.5384087
  14    0.7622067  0.5244068
  18    0.7598055  0.5196049
  19    0.7586047  0.5172037
  20    0.7596043  0.5192020

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 8.
```

This is an attempt to tune the “mtry” hyperparameter of the random forest [1]. The objective is to fight against overfitting [1]. We expect that mtry=8 will be the best choice [1].

Problem 4: interpretation (9pts)

The random forest that was fitted in Problem 3 is used by the analyst to produce the following graphs.

Figure 1:

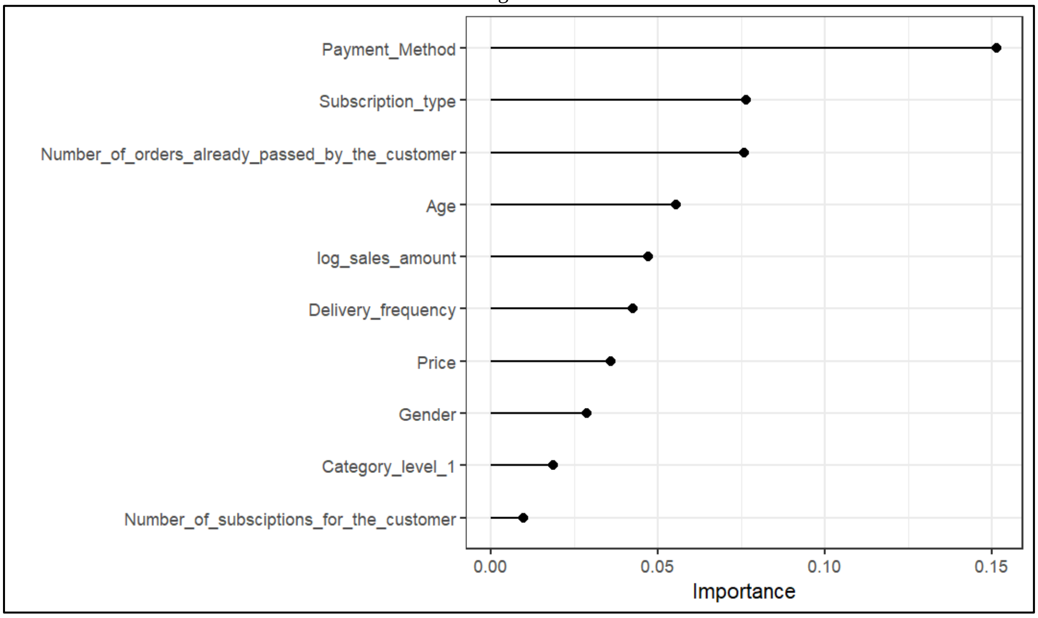


Figure 2:

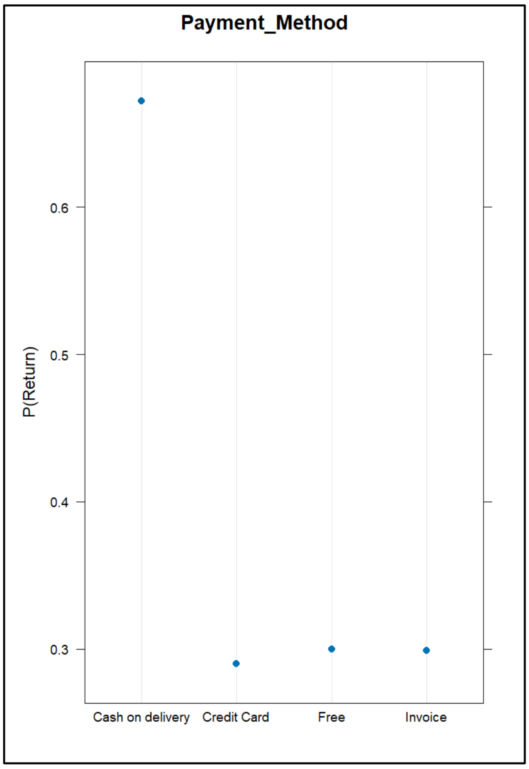
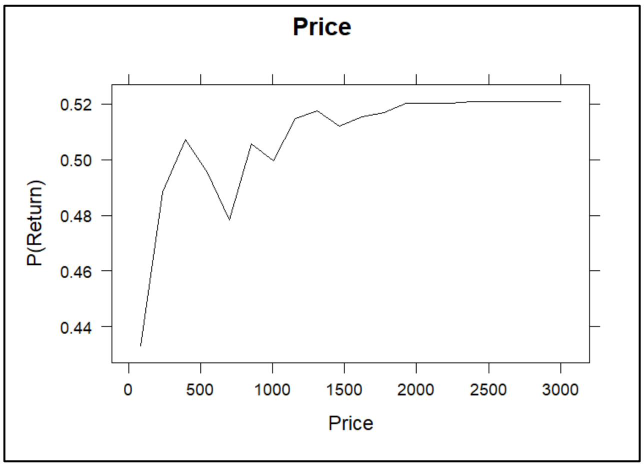


Figure 3:



Note: Pay attention to the scale on the y-axis.

a. To which method does Figure 1 refer? What can be concluded from it (explain for two or three features)? (3pts)
The method is called "variable importance" [1]. We can conclude that the most important feature is the payment method [1] and the least important is the number of subscriptions [1].

b. To which method do Figures 2 and 3 refer? What can be concluded from Figure 2 and from Figure 3? In particular, you must explain how the conclusion matches those of Figure 1, and what complement of information we can get to Figure 1. (3pts)
Figures 2 and 3 are Partial Dependence Plots [1]. From Fig 2, we can conclude that the probability of return is the highest for "Cash on delivery" [1]. Fig 3 shows that the probability of return increases on average with the price [1]. However, this is a small effect (see scale). Thus, like in Figure 1, payment method appears to be important and price not very important [1].

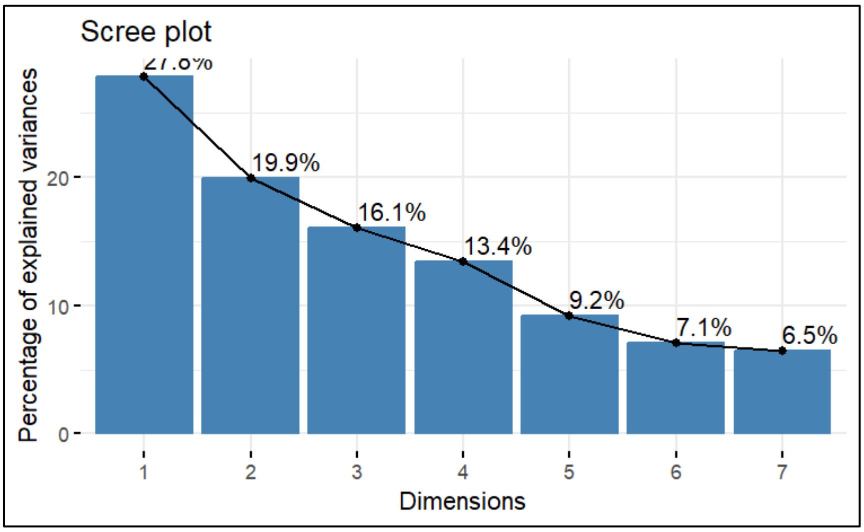
c. How do the results of Figures 2 and 3 agree with or contradict the logistic regression results shown in Problem 2.b.? Justify. (3pts)
Like in logistic regression, the payment method is important and cash on delivery is associated with the largest probability of return [1]: the coefficients vary a lot with payment type and the coefficients other than cash on delivery are negative [1]. For the price, the coefficient is positive but quite small (in agreement with Figure 3) [0.5]. The association is thus positive but weak [0.5].

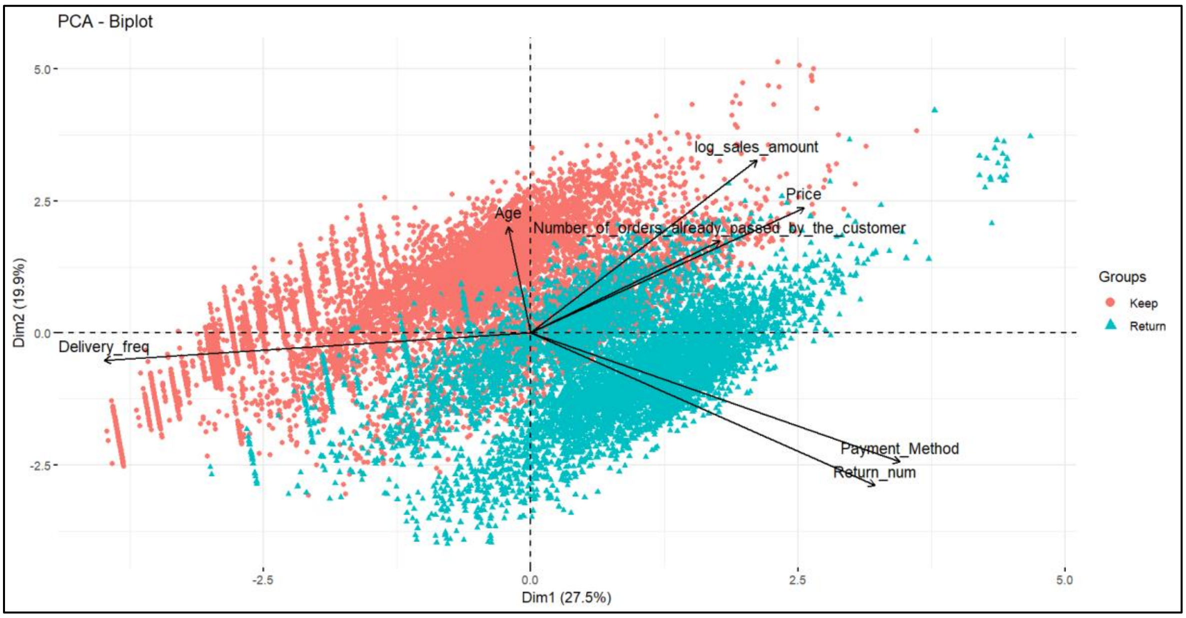
Problem 5: dimension reduction (9pts)

Below, the variables were modified as follow:

- *Return_num* is 0 if “Return=Keep” and 1 if “Return=Return”
- *Delivery_freq* is a numerical version of Delivery_frequency, in months. E.g. 0.5 if Delivery_frequency is “2 week”, 1 if Delivery_frequency is "1 month", etc., up to 6.
- *Payment_Method* is 1 if “Cash on delivery” and 0 otherwise.

The following analysis was then produced using features *Return_num*, *Payment_Method*, *Price*, *Number_of_orders_already_passed_by_the_customer*, *Age*, *log_sales_amount*, *Delivery_freq*, *Payment_Method*.





Note: colors are the *Return* (Keep/Return). *Return_num* is the numerical version of *Return*.

a. How can we measure the quality of this 2-dimensional representation of the data (biplot)? Explain briefly what this measure represents. (1pt)

The quality is measured by the proportion of variance explained by the principal components [0.5]. Here the two first components explain 27.8%+19.9%=47.7% of the variance [0.5].

b. What can be concluded regarding the link between the return status of a command and with:
   (i) the payment method. (2pt)
   (ii) the delivery frequency and the return status of a command. (2pt)
   (iii) the price. (2pt)

   For each, briefly justify and give a “business” interpretation.

   (i) The two features are strongly positively associated as shown by the two confounded arrows [1]. Most returns are associated with payment by cash [1]. (ii) The association is mild. Largest delivery frequencies are associated with “Keep” [1]. The returns are more common when one orders more often [1]. (iii) The link is mild. The higher the price the more there is a chance of return [1]. More expensive orders have more chance to be returned [1].

c. The subscription type was not included in the analysis. Could it be included? If yes, explain how. If not, explain why. (2pt)
   The subscription type is categorical nominal (not ordinal) [0.5] and thus cannot be included in the PCA [1] that is limited to numerical features [0.5].

Problem 6: advanced questions (5pts)

The analyst wants to understand what drives the return beyond the payment method. To this aim,

- She removes the instances with “Payment_Method = Cash on delivery”.
- She down-samples the data to balance classes “Return” and “Keep”.

She makes an 80/20 data splitting and trains a random forest on the following variables:

- Payment_Method (categorical)
- Price (numerical)
- Number_of_orders_already_passed_by_the_customer (numerical)
- Age (numerical)
- log_sales_amount (numerical)
- Delivery_freq (numerical; 0.5 for “2 week”, 1 for “1 month”, etc.)

The results are shown below (Analysis 1):

Training set

```
          Reference
Prediction Keep Return
    Keep   2932    110
    Return  104   2926

               Accuracy : 0.9648
                 95% CI : (0.9598, 0.9693)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9295

 Mcnemar's Test P-Value : 0.7325

            Sensitivity : 0.9638
            Specificity : 0.9657
         Pos Pred Value : 0.9657
         Neg Pred Value : 0.9638
             Prevalence : 0.5000
         Detection Rate : 0.4819
   Detection Prevalence : 0.4990
      Balanced Accuracy : 0.9648

       'Positive' Class : Return
```

Test set

```
          Reference
Prediction Keep Return
    Keep    520    163
    Return  239    596

               Accuracy : 0.7352
                 95% CI : (0.7122, 0.7572)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4704

 Mcnemar's Test P-Value : 0.0001835

            Sensitivity : 0.7852
            Specificity : 0.6851
         Pos Pred Value : 0.7138
         Neg Pred Value : 0.7613
             Prevalence : 0.5000
         Detection Rate : 0.3926
   Detection Prevalence : 0.5501
      Balanced Accuracy : 0.7352

       'Positive' Class : Return
```

In a second analysis, she takes the five numerical features, performs a principal component analysis from which she extracted the first two principal components, PC1 and PC2. Then, she binds Payment_Method, PC1, PC2 into a data frame (thus, the final data set has three features).
Again, she makes an 80/20 data splitting and trains a random forest. The results are shown below (Analysis 2):

Training set

```
          Reference
Prediction Keep Return
    Keep   1293    297
    Return 1743   2739

               Accuracy : 0.664
                 95% CI : (0.652, 0.6759)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3281

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9022
            Specificity : 0.4259
         Pos Pred Value : 0.6111
         Neg Pred Value : 0.8132
             Prevalence : 0.5000
         Detection Rate : 0.4511
   Detection Prevalence : 0.7381
      Balanced Accuracy : 0.6640

       'Positive' Class : Return
```

Test set

```
          Reference
Prediction Keep Return
    Keep    306     83
    Return  453    676

               Accuracy : 0.6469
                 95% CI : (0.6223, 0.671)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2938

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8906
            Specificity : 0.4032
         Pos Pred Value : 0.5988
         Neg Pred Value : 0.7866
             Prevalence : 0.5000
         Detection Rate : 0.4453
   Detection Prevalence : 0.7437
      Balanced Accuracy : 0.6469

       'Positive' Class : Return
```

a. In a few sentences, explain the purpose of combining a dimension reduction technique (here PCA) with a supervised learner (here random forest). In this case, did it reach the purpose? Explain the advantages and drawbacks in this case. (3pts)
The PCA reduces the dimension and thus helps to fight against overfitting [1]. This can be seen here as Analysis 2 shows no overfitting, unlike Analysis 1 [1]. The drawback is a loss of information: the accuracy is much lower in Analysis 2, even on the test set [1].

b. What alternative could use the analyst to PCA? Mention and explain briefly what the advantages could be. (2pts)
PCA could be replaced by an auto-encoder [1]. The advantages could be a better fit due to non-linearity [0.5] and a the fact that it can adapt to categorical variable and thus include Payment_Method [0.5].

EDA of the data:

Data Frame Summary

Return_data

Dimensions: 117994 x 11
Duplicates: 50602

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	M
1	Return [factor]	1. Keep 2. Return	104910 (88.9%) 13084 (11.1%)		117994 (100.0%)	
2	Payment_Method [factor]	1. Cash on delivery 2. Credit Card 3. Free 4. Invoice	32515 (27.6%) 60253 (51.1%) 25169 (21.3%) 57 (0.0%)		117994 (100.0%)	
3	Price [numeric]	Mean (sd) : 664.2 (322.3) min ≤ med ≤ max: 84 ≤ 817 ≤ 3000 IQR (CV) : 405 (0.5)	64 distinct values		117994 (100.0%)	
4	Subscription_type [factor]	1. S1 2. S2 3. S3 4. S4 5. S5 6. S6	33627 (28.5%) 10705 (9.1%) 1780 (1.5%) 38310 (32.5%) 31442 (26.6%) 2130 (1.8%)		117994 (100.0%)	
5	Delivery_frequency [factor]	1. 1 month 2. 2 month 3. 2 week 4. 3 month 5. 4 month 6. 6 month	10838 (9.2%) 54807 (46.4%) 436 (0.4%) 17713 (15.0%) 6973 (5.9%) 27227 (23.1%)		117994 (100.0%)	
6	Category_level_1 [factor]	1. C1 2. C2 3. C3 4. C4 5. C5 6. C6 7. C7 8. C8	1183 (1.0%) 8552 (7.2%) 152 (0.1%) 64559 (54.7%) 204 (0.2%) 25092 (21.3%) 15194 (12.9%) 3058 (2.6%)		117994 (100.0%)	
7	Number_of_orders_already_passed_by_the_customer [numeric]	Mean (sd) : 7.9 (9.4) min ≤ med ≤ max: 1 ≤ 4 ≤ 80 IQR (CV) : 9 (1.2)	73 distinct values		117994 (100.0%)	
8	Number_of_subsciptions_for_the_customer [factor]	1. One 2. More	102366 (86.8%) 15628 (13.2%)		117994 (100.0%)	
9	Age [numeric]	Mean (sd) : 47 (12.5) min ≤ med ≤ max: 13 ≤ 47 ≤ 100 IQR (CV) : 16 (0.3)	86 distinct values		117994 (100.0%)	
10	Gender [factor]	1. Female 2. Male 3. Not Specified	63288 (53.6%) 45394 (38.5%) 9312 (7.9%)		117994 (100.0%)	
11	log_sales_amount [numeric]	Mean (sd) : 6.9 (0.8) min ≤ med ≤ max: 0 ≤ 6.8 ≤ 10.1 IQR (CV) : 0.7 (0.1)	1232 distinct values		117994 (100.0%)	

Generated by summarytools 1.0.1 (R version 4.3.2)
2024-04-25