DEPARTMENT OF ECONOMICS
AND BUSINESS ECONOMICS
AARHUS UNIVERSITY

EXAM ASSIGNMENT

Study Programme and level	**MSc Business Intelligence + elective**						
Term	Winter 2024/25 – ordinary exam						
Course name and exam code(s)	Machine Learning for Business Intelligence 1				460202E004		
Exam form and duration	Written onsite exam, NO internet allowed				4 hours		
Date and time	3 January 2025				14:00-18:00		
Supplementary material/aids	All	X	Specified		No		
Anonymous exam	Yes	X	No		Please do **not** write your name or student ID number anywhere.		
Use of generative AI (GAI) allowed	Yes		No	X			
Hand-in of handwritten material allowed	Yes		No	X			
Hand-in of extra material (appendix) in WISEflow allowed	Yes	X	No				
Other relevant information	**Avoid being suspected of exam cheating** Remember to state references and use quotation marks, if you copy text from other sources or re-use parts of a previously submitted exam paper (plagiarism and self-plagiarism). Students must answer the exam assignment **individually**. All submitted exam papers are checked for plagiarism, so cheating and collaboration between students will be detected.						
Number of pages (incl. front page)	4 pages						

Other instructions:
It is important that you start uploading your exam paper well in advance – at least 10 min. before end of exam.

Practical information

To submit your exam, you must upload **a blank PDF file** together with a .ZIP or similar archive file containing all the source code files necessary to reproduce your results. In addition, **the R script(s) must contain the required comments to your code and answers**. It must be clear which question you are answering in your script(s). You do NOT have to produce a separate report as PDF.

In order to obtain points for an exercise, your script should produce the correct result directly without any alteration of your code. Any mismatch will result in a reduction of points. If you are convinced that a small bug that you cannot fix is causing your code to fail at any point in your script, you may be able to salvage a few points by clearly and concisely explaining where it occurs and how you think it should be solved. The length of your comments and discussions should be appropriate in regard to each question.

This exam contains ***2 sets of questions***. There are ***70 points*** on this exam.

Good luck!

Data and context

The description of the case and dataset was provided during the Applied Data Work session and will not be repeated here. Please refer to the files on Brightspace for any additional details if needed. The dataset attached to this exam is the same one uploaded on Brightspace.

Problem 1 (AAT) – 35 points

The goal of this task is to predict the likelihood of individuals purchasing food online (variable PROD12) based on socio-demographic and economic factors. Proceed with the following:

a) Load the provided dataset into R. Select columns 1 to 18, and column 57. Your selection should include the following variables:

1. SEXO (factor)
2. EDAD (numeric)
3. ESTC (factor)
4. NIVELEST (factor)
5. SIT_LAB (factor)
6. TIP_JOR (factor)
7. ACTIV (factor)
8. OCUPACION1 (factor)
9. OCUPACION2 (factor)
10. ING_HOG (factor)
11. TIP_H (factor)
12. TOT_MH (numeric)
13. TOTAL10_15 (numeric)
14. TOT_MEN16 (numeric)

15. TOT_MAY74 (numeric)
16. TOT_16_64_T (numeric)
17. TOT_16_64_NT (numeric)
18. TOT_16_24_E (numeric)
19. PROD12 (factor)

(1 point)

b) Replace spaces in the dataset with NA. **(1 point)**

c) Remove the existing missing values from the dataset. **(1 point)**

d) Adapt the variable types to correspond to the types mentioned in the list above. **(1 point)**

e) Recode the values of variable PROD12:
- Set values 1 and 2 to Buy, representing individuals who have purchased food online.
- Set value 6 to Notbuy, representing individuals who have *not* purchased food online.

Hint: data_Sel$PROD12 <- ifelse(data_Sel$PROD12 == "1" | data_Sel$PROD12 == "2", "Buy", "Notbuy")

After recoding, make sure PROD12 is a factor, and show explicitly its distribution.
(1 point)

f) Split the data into training and test set using a random stratified sampling based on PROD12. Proceed by creating a blueprint that prepares the data for analysis in the following order:
- Centre and scale all numeric features.
- Handle infrequent categories in categorical variables.
- Transform all the categorical variables into dummy variables.
- Eliminate all features with near-zero variance.
- Prepare and bake both training and test data using the blueprint created before.
- Report explicitly the size of the new datasets.

(5 points)

g) Using the package *caret* in R, apply k-fold cross-validation to train a set of classification models (minimum three) to predict the likelihood of an individual purchasing food online. **(5 points)**

h) Discuss the output of each model, focusing on their strengths and weaknesses in the context of the problem. **(5 points)**

i) Compare the models based on their performance metrics, and make a few recommendations on how the models should be improved without changing the predictors. **(2,5 points)**

j) Based on the model insights, advise the company on which segments of the population they should target in their next marketing campaign. **(5 points)**

k) Save the predicted probability of each model as a new variable. Combine them into a new variable representing the average predicted probability across the respective models. Evaluate the performance of this stacked model, comparing it to the individual models. Reflect on whether using a stacked model improves performance, and under which circumstances stacking might be beneficial or detrimental. **(7,5 points)**

Problem 2 (BV) – 35 points

The goal of this task is to predict the value of purchases (variable VCOMPRAS_cont) based on socio-demographic and economic factors. Proceed with the following:

a) Load the dataset into R. Select columns 1 to 12, and column 70 (VCOMPRAS_cont). During this loading, save blank (" ") as NA variable. **(1 point)**

b) Remove any rows with missing (NA) values. Then, remove any rows with entries "6" in column 10 (ING_HOG variable). (**1 point**)

c) Convert all variables excluding columns 2 (EDAD), 12 (TOT_MH) and 70 (VCOMPRAS_cont) to factors. **(1 point)**

d) Provide summary statistics for each of these factor variables, and remove factor levels that have less than 20 entries. Which variables are modified during this stage? What is the final dimension of the dataset? **(2 point)**

e) Using ggplot2 library, perform two tasks. First, plot the histogram of the VCOMPRAS_cont variable. Second, plot the box plot of VCOMPRAS_cont by the ING_HOG variable. Discuss the plots. **(3 points)**

f) Set your seed to 25. Randomly select 80% of the units into the training set and the remaining units into the test set. Do this using sample() function and without stratification. Define a matrix of dimension 5x2 called test.error with NA variables, which has to be used in the below parts to store test errors of five models across two metrics. (**2 points**)

g) The task is to predict the VCOMPRAS_cont variable. Run an intercept only model on the training set. Then, store its test errors with respect to the root mean squared error (RMSE) and mean absolute error (MAE) metrics. What are the values of the test errors? **(4 points)**

h) To predict the VCOMPRAS_cont variable, run a linear regression with all remaining variables as predictors on the training set. Then, store its test errors with respect to RMSE and MAE metrics. What are the values of test errors? Which variables are statistically significant? **(5 points)**

i) To predict the VCOMPRAS_cont variable, run lasso and ridge regression with all remaining variables as predictors on the training set and with 10-fold cross validation. Then, store their test errors with respect to RMSE and MAE metrics. What are the values of the test errors? What are the tuned lambda values (lambda.min) for both models? Which variables are selected by lasso? Discuss these results. **(8 points)**

j) The next task is to use three predictors, namely linear regression, lasso and ridge regression, from parts h) and i) and form a median-based model averaging predictor. To do this, run a for loop across the length of the test set to form this median-based predictor. Then, store its test errors with respect to RMSE and MAE metrics. What are the values of the test errors? **(5 points)**

k) Compare the RMSE and MAE test errors of the five models from parts g), h), i) and j) by extracting errors stored in the test.error matrix. Discuss your results. Based on the material we covered in the lectures, do you find these results expected or surprising? Provide arguments. **(3 points)**