AARHUS BSS

DEPARTMENT OF ECONOMICS AND BUSINESS ECONOMICS
AARHUS UNIVERSITY

EXAM ASSIGNMENT

Study Programme and level	**MSc Business Intelligence + elective**						
Term	Winter 20-21 ordinary						
Course name and exam code(s)	Machine Learning for Business Intelligence 1						460202E004
Exam form and duration	WOA, changed to WHAI due to Covid restrictions						4 hours
Date and time	12 December, 2020						09.00 – 13.00
Supplementary material/aids	All	X	Specified		No		
Other relevant information	See below and page 2-3.						
Hand-in of hand-written material allowed	Yes		No	X	Comments:		
Anonymous exam?	Yes	X	No		Comments: Please do not write your name or student ID number anywhere.		
Number of pages (incl. front page)	6						

How to hand in your exam paper

Start preparing the hand in well in advance of the exam deadline. Your exam paper must be handed in as **one PDF file** in WISEflow. The maximum permitted file size is **200 MB**.

Additional material/appendices may (if permitted) be uploaded in other file formats. The total maximum permitted file size is 5 GB.

If you experience problems uploading and handing in your exam paper in WISEflow, you can send the paper to the following email address: bss.exam@au.dk. You need to ask for permission to hand your paper for final assessment in WISEflow. Use the formula “Exemption” under “Applications to Study Councils” in the Student Self-Service. You need to apply as soon as possible after sending your paper to the email address.

If you need technical assistance during the exam, you can contact **BSS IT-support, phone: 8715 0933.** Contact the invigilator if the exam is on-site.

Be aware that exam papers are as a rule only permitted for final assessment if handed in in the right format/size and within the exam deadline.

PLEASE NOTE

This exam has been changed to online format. Please note that the rules regarding plagiarism or the use of unauthorized auxiliary material and communication still apply.

If some suspicion is raised upon submitted solutions, it will be reported and led to the Board of Studies of Economics and Management.

QUESTIONS DURING THE FIRST HOUR OF THE EXAM

If you have any *clarification* questions to the exam assignment within the first hour of the exam, please email them to Phillip Heiler (**for problem 1**) at pheiler@econ.au.dk or Ana Alina Tudoran (**for problem 2**) at anat@econ.au.dk. Do not expect instant answers. Answers to questions of a general concern will be posted to Blackboard.

Practical information:

This is an open book exam. You are allowed to use any course material provided throughout the lectures. Unless explicitly stated otherwise, you ARE allowed to use external packages to solve the exam.

To submit your exam, you must upload a .ZIP or similar archive file containing all the source code files necessary to reproduce your results. In addition, the 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.

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 regards to each question.

Communication during the exam is strictly forbidden. The exam has to be done individually by yourself, strictly without consulting anyone. If some suspicion is raised upon submitted solutions, it has to be reported to the Board of Studies.

This exam contains 2 set of questions. There are 70 points on this exam. The exam will be graded on the Danish seven-point scale. You have 4 hours to upload your solutions.

Good luck!

Data Description

For all problems, you work with modified audit fraud data from file *audit2.csv.* It is a cross-section containing 12 variables for 1550 observations about exhaustive non-confidential data in the year 2015 to 2016 of firms collected from the Auditor Office of India to build a predictor for classifying suspicious firms. In particular, one would like to detect audit fraud (”Risk”). Many potential risk factors are examined from various areas like past records of audit office, audit-paras, environmental conditions reports, firm reputation summary, on-going issues report, profit-value records, loss-value records, follow-up reports etc. After in-depth interview with the auditors, important risk factors are evaluated and scored. It contains the following variables:

Variable	Type	Description
”Sector_score”	continuous	Historical risk score value of the target-unit
”PARA_A”	continuous	Discrepancy found in the planned-expenditure of inspection and summary report A (measured in 10 Mio Rupees)
”Score_A”	categorical (0.2,0.4,0.6)	Risk score value of the target-unit from summary report A
”PARA_B”	continuous	Discrepancy found in the unplanned-expenditure of inspection and summary report B (measured in 10 Mio Rupees)
”TOTAL”	continuous	Total amount of discrepancy found in other reports (measured in 10 Mio Rupees)
”numbers”	categorical (5.0, 5.5, 6.0, 6.5, 9.0)	Historical discrepancy score
”Score_B.1”	categorical (0.2,0.4,0.6)	Risk score value of the target-unit from summary report B.1
”Money_Value”	continuous	Amount of money involved in misstatements in the past audits (measured in 10 Mio Rupees)
”Score_MV”	categorical (0.2,0.4,0.6)	Risk score value of the target-unit from money involved in misstatements
”District_Loss”	categorical (2,4,6)	Historical loss of a district in the last 10 years
”History”	categorical (0-9)	Average historical loss suffered by firm in the last 10 years
”Risk”	binary (0,1)	Risk Class assigned to an audit-case. (Target Feature)

Problem 1:

40 P

In this problem, you are supposed to predict observations Risk class.

1. Load *audit2.csv*. Standardize all predictors. Split the data set using the first 1162 observations for training and save the remaining observations in a test data set. How large is the share of Risk assigned observations in each data set? 3 P

2. Estimate a model for predicting Risk using logistic regression and all available predictors. What is the training and test set accuracy? Interpret your result. 5 P

3. Estimate a model for predicting Risk using logistic regression and all available predictors plus their two-way interactions. Calculate the in-sample error (log-likelihood function) and the AIC and BIC for this model and the model from 2). Which model is suggested by each of the three criteria? Interpret your results **thoroughly**. 9 P

4. Use a model averaging method suitable for logistic regression models using all available predictors for predicting risk. Which model receives the majority of the weight? Is this the true model? 8 P

5. Extend the predictor set further by adding all squared predictors and two-way interactions. Use a regularization approach for logistic regression that yields sparse solutions to estimate a model for predicting Risk. Propose a method to select your tuning parameter. Predict the Risk class using this approach. What is the test and training accuracy? **Briefly** interpret your results. 9 P

6. Use a super learner to combine at least two reasonable classification methods to predict the Risk class. Which method receives the most weight? Calculate the test error of the super learner. **Hint:** *For binomial likelihood methods you can specify the options:* method = "method.NNloglik" *and* family = binomial *in SuperLearner().* 6 P

Problem 2:

30 P

1. Load *audit2.csv*. Discretize all the continuous variables (High/Low) using their median as a cutoff. **Note:** *If you are not able to do this task, consider working only with the discrete variables in the dataset.* 5 P

2. Split the data using the 1162 observation for training and the rest of the observations for testing. Consider the task of classifying the firms as risky or non-risky using a Naive Bayes Classifier. How do you interpret a-priori and conditional probabilities in the output? Evaluate the performance of the model using Type-I error, Type II error, Sensitivity, Specificity, Accuracy, and AUC for suspicious firm classification. 10 P

3. Discuss to what extent the Naive Bayes outperforms alternative models (e.g. from problem 1) in terms of their accuracy in predicting the risk-class, and theoretically what are the advantages of using a Naive Bayes model? 5 P

4. Estimate the probability of fraud for a firm with a high risk score value of the target-unit from summary report A and a high risk score value of the target-unit from summary report B.1. Using the Naive Bayes model with all the predictors (point 2.), evaluate which combination of predictors and their corresponding values is associated with the highest probability of fraud. For these inferences, use the train dataset. 10 P