Instruction

Student number	
Full Name	

SAVE:

Save this document (File > Save As) with LASTNAME_1234567.pdf as filename in which "LASTNAME" is replaced by your last name and "1234567" is replaced by your student number (=7 digits, not your VU-net ID!) on T:\Exam on the computer.
Save your progression regularly by clicking the "Save"-button.
After finishing the exam submit this document digitally and follow the instructions at the end of this file.

Question 1: Deodorant sales forecasting (13 credits)

This is a manual exercise for which you need to specify the formula and detail all calculation steps.

The number of aggregated Kruidvats' sales for deodorant are:

Day	Week 1	Week 2	Week 3	Week 4
Sunday	50	57	61	66
Monday	60	66	73	75
Tuesday	70	73	78	83
Wednesday	65	67	77	79
Thursday	60	66	69	77
Friday	55	59	66	73
Saturday	51	59	61	66

a) Which time series components does this series have? Which smoothing method would you suggest? (4 credits)

b) What would the trailing moving average (W=7) forecast be for week 5? (2 credits)

c) Calculate the first 3 values with simple exponential smoothing (3 credits)
$L_t = \alpha\, y_t + (1-\alpha) L_{t-1}$, $\alpha = 0.2$. Use for initialization $L_0 = y_1$.

d) Calculate the first 2 values of L_t and T_t with Holts method (double exponential smoothing) and produce F_3. (4 credits)

L
$(¿¿ t-1 + T_{t-1})$, $\alpha = 0.2$
$L_t = \alpha\, y_t + (1-\alpha) ¿$
L
$(¿¿ t - L_{t-1}) + (1-\beta) T_{t-1}$, $\beta = 0.15$
$T_t = \beta ¿$
$F_{t+1} = L_t + T_t$

Use for initialization $L_0 = y_1$, and $T_0 = y_2 - y_1$

Question 2: Classification with R (12 credits)

The data set "Boston" is part of the MASS library and it contains data about the house values in suburbs of Boston. Use R studio to answer these questions below and copy your R code to the answer box below each question.

a) Create a dummy variable crime01 that is 1 if a given suburb has a crime rate above 0.2 and 0 otherwise. (2 credits)

b) Split the data set randomly in an equally sized training and test set. Use a seed value of 99. (2 credits)

c) Perform logistic regression on crime01, use all relevant variables. What is the test error rate? (4 credits)

d) Predict crime01 with KNN (K=10). Use the following variables zn, indus, chas, nox, rm, age, dis, rad, tax, ptratio, black, lstat, medv. (4 credits)

Question 3: Subset selection and cross validation (14 credits)

The data set “Hitters” is part of the ISLR library and it contains the Major League baseball data from the 1986 and 1987 seasons. Use R studio to answer these questions below and copy your R code to the answer box below each question.

a) Perform linear regression on Salary using all relevant variables. Which variables are significant? (2 credits)

b) Find the ``best model” (use adjusted r-squared) with backward selection. How many and which variables does this model contain? (4 credits)

c) Repeat b) but with best subset selection. (2 credits)

d) Explain the differences between the models in questions b) and c). Briefly discuss the differences between backwards and forward selection. (2 credits)

e) Cross-validation could give better estimations. Repeat a) with the validation set approach. Split the training and test set randomly in two equal sized sets using a seed value of 33. Report the test MSE. (4 credits)

Question 4: Australian sparkling wine forecasting (16 credits)

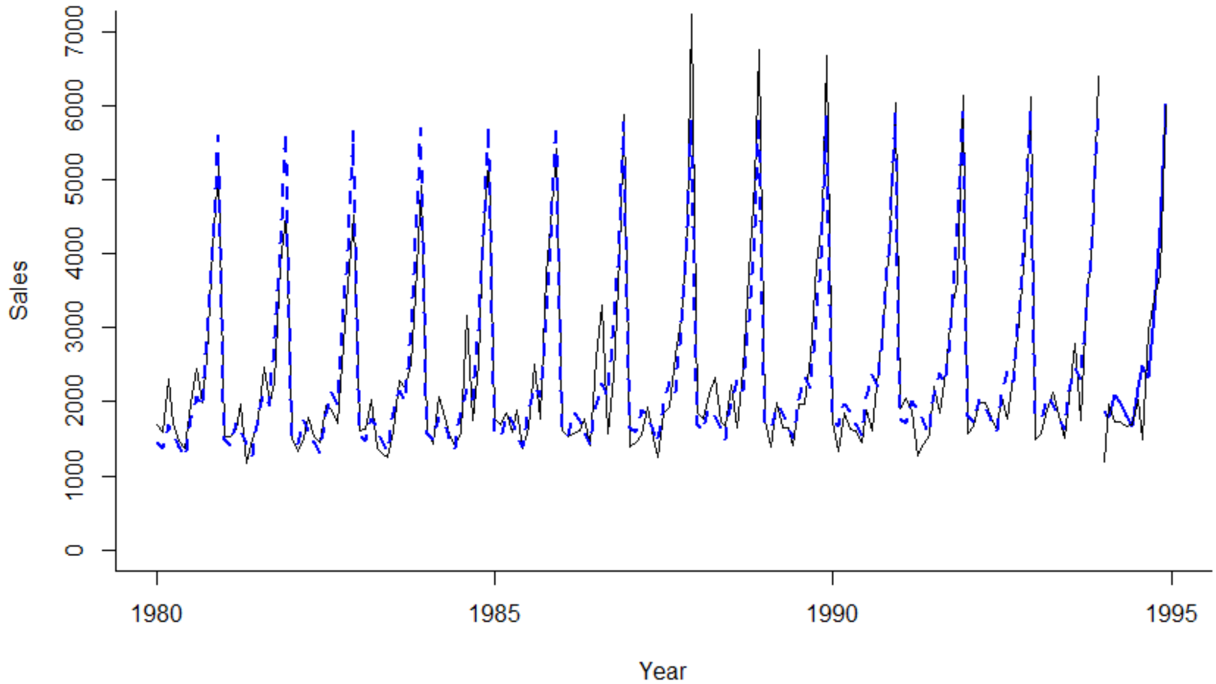
The file AustralianWines.xls contains data on the monthly sales for wine. The goal of this question is to find a model that minimizes the MAPE. Use R studio to answer the questions below and copy your R code to the answer box below each question.

a) Import AustralianWines.xls, make a time series object for the sparkling wine sales and split the data into a training and a validation set. Use the data up to December 1993 as training set and the remaining data as validation set. (3 credits)

b) Make a seasonal naive forecast for the validation period. What is the value of the MAPE? (2 credits)

c) Make a forecast with help of a regression model with a linear trend and seasonality. What is the MAPE? (3 credits)

d) Produce the following plot with the regression model from question c) (3 credits):



Report the R code to create the graph above. If you were unable to answer question d) , show your plotting skills by plotting train.ts, valid.ts and the following forecasts: 1866, 1789, 2096, 1992, 1814, 1665, 2180, 2485, 2335, 3220, 4348, 6022.

e) Plot the autocorrelation and interpret the results. Give 2 options to improve our results. (3 credits)

f) Make a forecast with help of a multiplicative regression model with a trend and seasonality. What is the MAPE? (2 credits)

Question 5: Logistic regression (15 credits)

This is a manual exercise for which you need to specify the formula and detail all calculation steps.

This question is about the credit card default data set.
Remember that the multiple logistic function is defined as follows:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−3.5041	0.0707	−49.55	<0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

Figure 1: Logistic regression results for credit card default data

a) Calculate the probabilities that a student and a non-student default based on the logistic regression result from Figure 1. Are students more likely to default? (3 credits)

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
student[Yes]	−0.6468	0.2362	−2.74	0.0062

Figure 2: Multiple logistic regression results for credit card default data.

b) Given the results of Figure 2 and a balance of 2000 euro's calculate the probabilities of default for students and non-students. Who is more likely to default students or non-students? (3 credits)

c) Explain the difference between answer a) and answer b). You may (also) use Figure 3 for your answer. (3 credits)

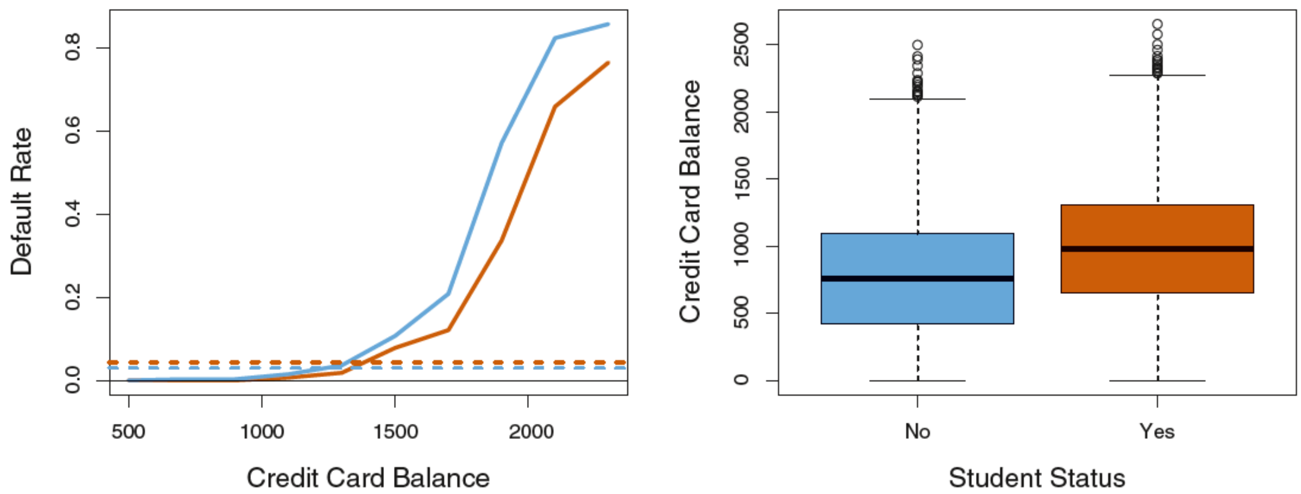
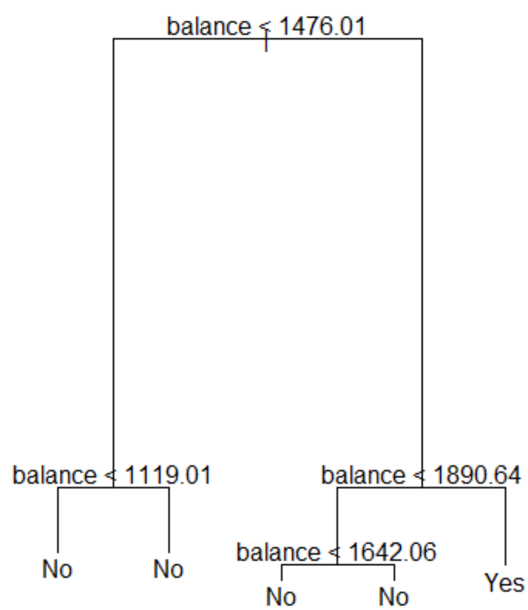


Figure 3: Default rate and credit card balance.

d) What is the maximum balance rate that a student can have such that his/her default probability is 0.2 or lower. (4 credits)

e) Assume that we use a decision tree instead of logistic regression and we get the following tree:



Briefly explain this tree, does it make sense to use balance multiple times? Why would we use a branch resulting in two different leaves with the same decision? (2 credits)

SUBMIT your exam

Click on the icon “Submit Exam” on the desktop. The website that is opened allows you to upload your document.

Log in to the website with your VUet ID and password.

Follow the instructions on the website. Click the “Choose File”-button and select the file you want to upload.

Click the button “Turn in exam” (Tentamen inleveren).

END