NHH

SCHOOL EXAMINATION

BAN404

Spring, 2024

Date: 27.05.2024

Time: 9.00-15.00

Number of hours: 6

An invigilator can contact course responsible by phone: 47078891

SUPPORT MATERIALS PERMITTED DURING THE EXAMINATION:

Calculator

All written support material permitted (category III)

Number of pages, including front page: 5

Number of attachments: 1, The file “airline.csv”.

Instructions and recommendations

General recommendations: Be pragmatic if you run into difficulties. You need to find models that can answer the questions asked. Don´t let the perfect model stand in the way for a good one! Also, plan your work so that you do not use all time to solve one task.

R coding can be done in different ways, e.g. using base R or the tidyverse approach. Given a correct and clearly written result there is no preference for any particular approach.

Throughout the exam you will work with training and test data and you will have to make appropriate choices on when to use one or the other. In cases when you think there are more than one valid choice on working with the training, test or the full dataset, explain your choice!

Tasks

1. (50 points) This task is about some of the methodological topics treated in the course.

   (a) (10 points) Explain what the following R-function is doing.

```r
f <- function(x0,x,y,K=3)
{
  d <- abs(x-x0)
  o <- order(d)[1:K]
  xl <- x[o]
  yl <- y[o]
  xydata <- data.frame(xl=xl,yl=yl)
  reg <- lm(yl~xl,data=xydata)
  ypred <- predict(reg,newdata=data.frame(xl=x0,yl=1))
  return(ypred)
}
```

   (b) (10 points) Consider the following small dataset

x	1	2	3	4	5	6	7	8	9	10
y	5.26	9.13	11.17	15.64	25.32	25.55	41.39	48.17	58.65	68.24

   Use leave-one-out cross-validation to determine the optimal K in the function `f` for this dataset.

   (c) (10 points) Plot y against x and add a line with the predictions based on the optimal K.

   (d) (10 points) In the function `f`, there is only one predictor. One way to allow for more than one predictor is to compute `d` in the `f`-function in (a) differently. Explain how such a modification can be done.

   (e) (10 points) Another way to allow for several predictors is to use backfitting. Explain how you would implement backfitting for this method (you do not need to do it).

2. (50 points) In this task you shall analyze the dataset in the file `airline.csv`. This data is downloaded from kaggle.com, see Jana (n.d.). You can read it into R with the `read.csv` function (or `read_csv` if you prefer the tidyverse approach).

```r
airline <- read.csv("airline.csv")
```

   The dataset contain answers from a survey of customers of an airline. The customers are anonymous to the airline. You will help the airline with two tasks. First, you shall analyze **why some of the customers are dissatisfied**. Secondly, you shall help the airline to **predict if a new customer will be dissatisfied on the next flight**. The airline will use your prediction model to identify customers likely to be dissatisfied and

put in place measures to reduce this likelihood. The variables in the data set are given in the table below. For the detailed customer satisfaction questions the answers is a number between 0 and 5 meaning where 0 = no answer, 1 = very unhappy,. . . , 5 = very happy.

Variable name	**Explanation**
satisfaction	Overall satisfaction of customer
Age	Age of customer
Arrival_Delay_in_Minutes	Delay of arrival
Baggage_handling	Customer's satisfaction of bagage handling
Checkin_service	Customer's satisfaction of checkin service
Class	Passenger class
Cleanliness	Customer's satisfaction of checkin service
Customer_Type	Loyal or disloyal customer
DA_time_convenient	Customer's view on departure/arrival time
Departure_Delay_in_Minutes	Departure delay
Ease_of_Online_booking	Customer's satisfaction of online booking
Flight_Distance	Distance of flight
Food_and_drink	Customer's satisfaction of foods and drinks
Gate_location	Customer's satisfaction of gate location
Gender	Gender
Inflight_entertainment	Customer's satisfaction of inflight entertainment
Inflight_wifi_service	Customer's satisfaction of inflight entertainment
Leg_room_service	Customer's satisfaction of leg room
On_board_service	Customer's satisfaction of onboard service
Online_boarding	Customer's satisfaction of online check-in
Online_support	Customer's satisfaction of online support
Seat_comfort	Customer's satisfaction of seat comfort
Type_of_Travel	Customer's type of travel, e.g. business

(a) (3 points) If necessary, recode categorical variables as factors. Motivate your choices in words.

(b) (3 points) In tasks (a)-(g) the methods and models are used to answer the question "Why are some customers dissatisfied?". With this in mind, motivate your choice to evaluate the models on all/training/test data.

(c) (3 points) Use descriptive statistics to find variables associated with `satisfaction`. First, explain which type of descriptive statistics (types of tables and figures) that you are using. Give examples of R code but do not show all code and output if you are doing the same thing many times. Summarize your results as text, mentioning the relevant numbers.

(d) (5 points) Based on the results in the previous tasks, formulate a logistic regression model and use it to evaluate which variables are associated with `satisfaction`. Evaluate the model with one or more appropriate measures of model fit.

(e) (5 points) Fit a classification tree, plot it and interpret it. Evaluate the model fit.

(f) (5 points) Can you improve the predictions by using bagging or a random forest? Compare the model fit with the models in the previous tasks. Compute a variable

importance measure and interpret it.

(g) (4 points) Based on the data analysis so far, what is your answer to the airline's question "Why are some customers dissatisfied?"

(h) (4 points) You should now, and for the rest of this task, shift perspective. The airline also would like to know **how they can reduce the dissatisfaction of those customers**. Since the survey data is anonymous, a prediction model for who will potentially be unhappy about their flight must therefore account for the fact that all variables will not be available for a customer who have bought a ticket. State and explain your assumptions about which of the variables are observed for a customer who has bought a ticket.

(i) (2 points) In tasks (h)-(l) the methods and models are used to answer the question "How can we predict the likelihood of a new customer to be dissatisfied?". With this in mind, motivate your choice to evaluate the predictions on all/training/test data.

(j) (5 points) Formulate a logistic model which can be used to predict the probability of a customer, who has already bought a ticket, will be unsatisfied if nothing is done to stop it. Evaluate the predictions in an appropriate way.

(k) (7 points) Use a random forest to predict whether a customer will be dissatisfied and evaluate the predictions. Compare with the predictions from logistic regression.

(l) (4 points) Consider all variables in the original dataset used in (a)-(g). Are there some of the information that is realistic to collect in the future which can help to improve the predictions. Assume that survey questions about customer's satisfaction of different aspects of the airline's service must be collected anonymously.

References

Jana, Sayantan. n.d. "Airlines Customer Satisfaction. Data from Kaggle." https://www.kaggle.com/datasets/sjleshrac/airlines-customer-satisfaction.