Predictive Modeling - Final Take Home Exam

Jun.Prof. Dr. Timo Dimitriadis

Heidelberg University, Summer Term 2021

Exam rules

1. The solution has to be handed in until **August, 5th, 23:59:59pm**. Late submission will be graded as failed.
2. Please submit your solutions through the "Exam Submission" tool on the Moodle platform.
3. You have to submit an electronically written document as a PDF file (preferably by using LaTeX) consisting of your solution in the form of tables, figures, explanations and interpretations together with the associated .R file containing your code, where the code has to be explained in comments (in the .R file). You can also use R Markdown and submit both, the source .Rmd file and the resulting PDF.
4. The grading will be based on the solutions, explanations, and interpretations in the written document, and on the code together with its explaining comments in the .R file. It is recommended that you add comments to all (non-trivial) lines of code, explaining in this way, line by line, what your code is doing.
5. The solutions and the code have to be prepared by everyone themselves. Any form of collaboration is forbidden!
6. The exam consists of three individual problems. All three have to be solved.

Problem 1 (40 Points)

Consider the data set `heart` (available on Moodle), which contains a retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa. The binary response variable `y` contains whether a person has a coronary heart disease (yes=1, no=0), which shall be predicted based no the remaining variables in the data set. These covariates are given by:

- "sbp": systolic blood pressure of the patient,
- "tobacco": cumulative tobacco (in kg),
- "ldl": low density lipoprotein cholesterol,
- "adiposity": an (unknown) numeric measure of obesity,
- "famhist": family history of heart disease (Present=1, Absent=0),
- "typea": type-A coronary prone personality behaviour as measured by a self-administered Bortner Short Rating Scale (possible total scores can range from 12 to 84),
- "obesity": a numeric measure of obesity,
- "alcohol": current alcohol consumption,
- "age": age.

For this, perform the following tasks:

1. Split the data in a training and test data set of equal size.
2. Use three different *reasonable* prediction methods for the probability of $y = 1$. Shortly explain why you think these three methods are suitable for prediction in this example.
3. Evaluate the predictions using the methods (for this type of forecasts) introduced in the lectures.
4. Explain and discuss your findings.

Problem 2 (40 Points)

Consider the data set `insurance.csv` (available on Moodle), that contains a sample of health insurance holders and their associated insurance claims (continuous variable `charges`) together with the explanatory variables:

- "age": age of the policy holder,
- "sex": sex of the policy holder,
- "bmi": body mass index of the policy holder,
- "children": number of children of the policy holder,
- "smoker": binary variables indicating whether the policy holder smokes,
- "region": categorical variable where the policy holder lives.

Perform the following tasks:

1. Split the data in a training and test data set of equal size.
2. Use three different *reasonable* prediction methods for the *mean* of the variable "charges". Explain why you think these three methods are suitable for prediction in this example.
3. Evaluate the predictions using the methods (for this type of forecasts) introduced in the lectures.
4. Explain and discuss your findings.

Problem 3 (20 Points)

Repeat the prediction (and evaluation) exercise of Problem 2, but for predictions for the 10%-*quantile* of the variable `charges`. It remains to consider two competing prediction methods here. Further discuss why such predictions might be of interest.