Statistical Learning - Assignment 2

The data for this part of the assignment are from a large epidemiological study on the course of depressive and anxiety disorders among adults living in the Netherlands. The data are from a subsample of 1500 subjects, who at the start of the study were suffering from an anxiety and/or depressive disorder, and were diagnosed as such. Twelve months after the start of the study, the severity of each subject's depressive symptoms was assessed again (variable `dep_sev_fu`; short for depression severity at follow-up). The goal of the analyses is to predict the severity of depressive symptoms after twelve months, using the characteristics assessed at the start of the study.

The dataset is available on Blackboard as 'MHpredict.csv'. It contains 20 potential predictor variables, which were assessed at the start of the study, and are described in Appendix I. You can read it into R as follows:

```
MH_dat <- read.table("MHpredict.csv", sep = ",", header = TRUE,
                     stringsAsFactors = TRUE)
```

For each of the questions below, word limits are provided. Note that these are maxima, so it is not required to write this many words. A perfect answer can certainly be given with less words.

1. Select three supervised learning methods from those that were discussed in weeks 9 through 11 for analyzing this dataset. Justify why you would select each of these methods for this specific prediction problem. (Use max. 200-250 words per method.)

Now apply the three methods you selected to the dataset. Beforehand, randomly split the dataset into a training (n=1000) and test (n=500) dataset. Use your student number to set the seed of the random number generator.

2. Motivate your choice of the main model-fitting parameters. Thus, make a well-informed choice for a fixed value of each parameter and/or use cross-validation to set their values. Your answer should reflect understanding of what each parameter does. (Use max. 200-250 words per method.)

3. Provide an interpretation of each of the resulting models:

- Describe which variables are most important in determining the value of the outcome variable, and which measure(s) you used to determine their relative importance.
- Describe the effect of the most important variables (e.g., describe the shape and direction of the effect on the outcome and/or provide and discuss plots of the variables' effects) for each method.

(Use max. 150-200 words per method.)

4. Assess and compare the predictive accuracy of each of the models using the test set. Which model predicts best? Bonus: Using a suitable approach, compute confidence intervals for the (pairwise differences in) predictive performance (not taught during the lectures). (Use max. 100 words, max. 200 words including bonus.)

5. Based on 3 and 4: Provide a short overall conclusion regarding which predictors are related to the outcome. (Use max. 100 words.)

6. A psychologist has seen David Edgar Pression for an intake today. The psychologist wonders whether they should refer David to an intensive depression treatment program. The results of David's intake assessment are provided on BlackBoard, in the file 'Patient.csv'.

   You can read it into R as follows:

   ```
   pat_dat <- read.table("Patient.csv", sep = ",", header = TRUE,
                         stringsAsFactors = TRUE)
   ```

   Hint: Check whether factor variables are correctly coded. To assign a variable in `pat_dat` the same factor levels as a variable in `MH_dat` (use of function `levels` has a different effect):

   ```
   pat_dat$fact <- factor(pat_dat$fact, levels = levels(MH_data$fact))
   ```

   The psychologist asks you to provide them with an estimate of the severity of David's depressive symptoms in 12 months. Patients with predicted depressive symptom severity equal to or greater than 17 are referred to the intensive treatment program. What is your estimate? Should David be referred to the intensive treatment program? Bonus: Using a suitable approach, quantify the uncertainty of your estimate (not specifically taught during the lectures). (Use max. 100 words, max. 200 words including bonus.)

Guidelines for the report:

- Produce two separate documents: A textual report that answers all questions (e.g., a .docx or .pdf document) as well as a code file (e.g., .R, .Rmd, .py file).
- Do not refer to the script in your report. Your answers must be self-contained.
- Your report should be aimed at a broad audience of researchers who might be interested in these analyses. Assume they have some knowledge of statistics. Write in full sentences. Do not use code language (e.g., variable names like "dep_sev_fu" are somewhat arbitrary and have no meaning outside of the code, so need to be explained and/or referred to in a different manner).
- In the report, clearly divide the answers to each of the numbered assignments above (e.g., use section numbering or numbered headings).
- It is good practice to use graphs. Make sure it is clear what is on the x- and y-axes. E.g., this might require you to carefully rescale sizes of labels or plotting symbols, and/or to explain what axis labels mean, or what is the scale of the variable on the axes.
- In the script, clearly divide the code into one section for each numbered question. Make sure that your code is readable (meaningful variable names, comments etc.).
- Upload both the report and the script via Brightspace.
- The deadline can be found on Brightspace.

Good luck!

Appendix I

Variable name	**Explanation / values**
`disType`	Type of disorder (depressive disorder, anxiety disorder, or comorbid disorder (i.e., having a diagnosis for both types of disorder))
`Sexe`	Male or female
`Age`	Age in years
`Aedu`	Years of education completed
`IDS`	Testscore on the Inventory of Depressive Symptomatology
`BAI`	Testscore on Beck's Anxiety Inventory
`FQ`	Total score on the Fear Questionnaire
`LCImax`	Percentage of time in which symptoms of anxiety and/or depressive disorders were present during the past four years
`Pedigree`	Presence of a first-degree relative with an anxiety and/or depressive disorder
`Alcohol`	Alcohol disorder diagnosis
`bTypeDep`	Subtype of depression
`bSocPhob`	Diagnosis of social phobia
`bGAD`	Diagnosis of generalized anxiety disorder
`bPanic`	Diagnosis of panic disorder
`bAgo`	Diagnosis of agoraphobia
`AO`	Age at onset of the disorder
`RemDis`	Whether the anxiety and/or depressive disorder is currently in remission
`Sample`	Whether subject is a patient in specialized mental health care, a patient in primary care, or not currently receiving (mental) healthcare
`ADuse`	Whether subject uses anti-depressant medication
`PsychTreat`	Whether subject receives psychological treatment for the disorder(s)