EBC1045
Knowledge Discovery and Data Visualization
Take-home Exam

Roselinde Kessels and Nele Raskin

Monday 19 December 2022, 9h00 -
Thursday 22 December 2022, 24h00

Exam regulations: The exam is a take-home exam for which you will need to perform data analysis in R and write an R Markdown document to report on your analysis process and findings. You will need to convert the R Markdown document into either PDF, Word or HTML format. Both the R Markdown document and the converted file format (i.e. the eventual report) need to be handed in through Canvas, provided via "Course Modules" > "Take-home exam December 2022". Name your files starting with "Your name + Name tutor + Tutorial group number" according to your tutorial group 1 through 6. The report should be based on the standard Markdown template and consist of maximum 25 pages in the A4 format.
You are allowed to consult external resources, but your document should be written independently by you. We therefore request you to fill out and upload the form "Declaration of Originality" together with your individually produced files.

Deadline: The exam instructions and all files needed for the data analysis will be made available via Canvas on Monday 19 December 2022, 9h00. Your R Markdown file, report, and Declaration of Originality should be handed in via Canvas before Thursday 22 December 2022, 24h00.

Data: You will work on the 'Housing' data set that lists 506 housing values in towns or suburbs of Boston. The data are provided in the file *housing.csv*. All variables are listed in order and are explained in the file *housing.names*. The target variable or variable of interest is the continuous variable "MEDV" or median value of owner-occupied homes expressed in $1000s.

Questions

1. Generate a random subset of 480 observations (i.e. housing values) from the 'Housing' data set in *housing.csv*. Set the seed to a value of choice before generating the subset. (0.5 points out of 12)

2. Summarize, visualize and explore your subset of housing values. Apply different exploratory data analysis techniques that you have learned during the course to discover interesting or meaningful relationships between the variables in the data. Make sure to apply appropriate preprocessing techniques. You may consider binning the target variable. You should only focus on those variables and relationships which you think are relevant for managerial or research purposes. Comment, interpret and discuss your findings carefully and in detail. You can use all the material covered in Chapters 1 till 3 of the book "Discovering Knowledge in Data". (5.5 points out of 12)
   Here is the detailed score sheet to get you started:

   (a) Perform a numerical method to identify outliers and visualize in case of extreme outliers (1 point)

   (b) Report ONLY interesting or important relationships (especially relating to the target variable) (1.5 points)

   (c) Provide ONLY useful visualizations with correct graph labels (especially relating to the target variable) (1 point)

   (d) In the case of numeric variables, perform a correlation analysis (0.5 points)

   (e) In the case of categorical variables, provide frequency tables (0.5 points)

   (f) Interpret and discuss all your results (1 point)

3. Construct and compare at least two conceptually different data mining models for the target variable based on your subset of housing values. Partition your subset randomly into a training set and a test set. Validate the partition for the most relevant variables by performing hypothesis tests. Build the data mining models on the training set and evaluate their prediction performance on the test set. Use one or more evaluation metrics that are appropriate to compare the performance of the models. Interpret, discuss and summarize your findings carefully and in detail. Present your best model and discuss the predictions it generates on the test set. You can use all the material covered in Chapters 4 till 8 of the book "Discovering Knowledge in Data". (6 points out of 12)
   Here is the detailed score sheet to get you started:

   (a) Partition the data into a training and test set (0.5 points)

   (b) Validate the partition or check its consistency by performing hypothesis tests (1 point)

   (c) Include only relevant variables in the model building process (0.5 points)

(d) Normalize the data when required and motivate your choice (0.5 points)

(e) Construct at least two different types of data mining models for the target variable (1.5 points)

(f) Compare and discuss the performance of the models using one or more evaluation metrics (1 point)

(g) Interpret, discuss and summarize all your results (1 point)

Please note the following:

- All tables and graphs in your report should be discussed in at least one or two sentences.
- If the converted file format (i.e. the eventual report) in PDF, Word or HTML from the R Markdown document is missing in your submission, you lose 2 points.
- Late submissions are not accepted.

You have reached the end of this exam. Good luck!