MA321 Group Coursework

Due in: 23th March 2022

Submission of the project report: submit a copy via FASER.
The **same** report has to be submitted by **all** group members.

All members of each group should participate in the editing and writing of the submitted version of the project report. The allocation of the marksbetween the group members will be based on the written statement listing the contribution of each member of the group, which has to be included in the project report.

Students are encouraged to equal contributions within groups.

Task: Suppose that you work as a data analyst at an estate agency. They are interested inusing machine learning to understand the housing market. With the help of R, each groupwill analyse the dataset "house-data.csv" from the point of view of classification, prediction and subsequently validation, as below. You should use R in order to conduct your statistical analysis. **You should include the R code as part of an Appendix of your report which should run without errors.** When answering the questions you should explain the method used and justify your answers. For each task, you are expected to comment on your findings.

1. Provide numerical and graphical summaries of the data set and make any initial comments that you deem appropriate. **[20 marks]**

2. Divide houses based on their overall condition (OverallCond) as follows:

   - Poor if the overall condition is between 1 to 3.
   - Average if the overall condition is between 4 and 6.
   - Good if the overall condition is between 7 and 10.

   (a) Fit a logistic regression model which predicts the overall condition (OverallCond) of a house. **[10 marks]**

   (b) Carry out a similar study using a different classification method you learned in MA321 to classify the house condition. **[10 marks]**

3. Predicting house prices:

   (a) Employ two methods you learned in MA321 in order to predict house prices. Justify your choice of model and comment on the results you obtained. **[10 marks]**

(b) Use two re-sampling methods of your choice to estimate the test error associated with fitting the above regressors in 3(a) on a set of observations. Comment on the results you obtained. **[10 marks]**

4. Using this data set, "house-data.csv", what else would be interesting in investigating? Identifya research question and employ a clustering analysis methodology you have learned in MA321 to answer this question. **[20 marks]**

General rules and hints:

- Follow the guideline: 'Writing Reports: a brief guide'.
- Plan and structure your work. Structure your report, for example: Page 1: cover page (title, your name, date,...). Page 2: abstract, contents and word count. Pages 3-7: introduction; preliminary analysis; analysis; discussion; conclusion; references. Page 8-10: appendix: R-code with explanations, etc..
- Use R. Put all R code, which was necessary for your report in an appendix and explain your R code (add comments within the R code). Do not include R code of an analysis which is not used for your report. Make sure, that YOU wrote the R code (the use of some R code, without citing the source, can be viewed as plagiarism).
- Use an appropriate word processor (MS Word, Open office,... or type setter (Lyx, Latex,...).
- UG: The report can have a length of 1600 to 2400 words (without cover page and appendix). Not more than 8 pages without counting the cover page and the appendix. More than 2400 words or more than 8 pages (without counting the cover page and the appendix) will reduce the marking.
- PG: The report can have a length of 2400 to 3600 words (without cover page and appendix). Not more than 12 pages without counting the cover page and the appendix. More than 3600 words or more than 12 pages (without counting the cover page and the appendix) will reduce the marking.
- Use point size 12, Times New Roman; line spacing 1.5.
- UG: Do not use more than 6 figures and 4 tables within the main text. You may include further figures and tables into the appendix, if necessary.
- PG: Do not use more than 7 figures and 5 tables within the main text. You may include further figures and tables into the appendix, if necessary.
- In addition your report should include a clear account of any assumptions made in the analysis of the data.

Marking: Marks for individual students will be based on the mark for their group's project, on the written statement listing the contribution of each member of the group, which has to be included in the project report. The markers reserve the right to make inquiries about the contributions of the members of the group if they feel they need to. If a member of the group contributes less than other members of the group, the markers will reduce the individual mark. If a member of the group contributes not at all, the individual mark will be zero.

Additionally, marks will be awarded as follows:

Report guidelines (in total 20):

- 0 of 20: group did not follow the guide lines.
- 10 of 20: group followed the guide lines; but understanding of specific parts of the guidelines/report structure is weak; e.g. no table legends, citation style inappropriate, etc.
- 20 of 20: group followed the guide lines.

Tasks 1, 2(a), 2(b), 3(a), 3(b) and 4 each (in total 80):

- 0: is missing or makes no sense.
- 50% of full mark: group describes a main analysis, which was suggested in the lectures and classes; tables and/or figures should support the results. The discussion and/or conclusion summarises the data analysis and result of the study.
- Full mark: group describes a main analysis, which includes justification of assumptions, provides further tables or figures which support the argument of the report. The discussion effectively communicates the results to the reader.