Final assignment/Exam

GRA4157

7. November 2022

Write a data science report on a subject of choice

In the final assignment, you should write a data science report. The report can be based on the data and results you have presented as part of the visualization and machine learning projects.

Practicalities

The page limit is 8 pages (pdf, including figures and references), but you may add code snippets as .py-files as attachments or in an appendix. The deadline for submission is **9. December 12:00 (noon)**. The group size is 1–3 persons. For a group, the expectations are slightly higher than for an individual.

Structure

The data science report needs to consist of at least the following sections: 1) Introduction, 2) Methods, 3) Results, 4) Conclusions (or Summary). You may rename the sections (e.g. "background" instead of "introduction" or "methodology" instead of "methods") if you prefer other headings. Additional subheadings under each heading is encouraged to structure the report. The sections need to include the following:

Introduction

Here, you should introduce the topic to a general reader. First, introduce and explain why the topic is important. Then see if you can find any sources on previous work or reports on the subject, and explain what has been done. Continue with what has *not* been done, and then write what is the goal of the current report. Tip: Very often, most of the stuff you plan to do has already been done (in some form). Therefore you may be very specific in the "this has not been done"-part of your introduction. If your data set is on house-pricing from SSB you can write something like "However, assessing whether the housing-prices found at SSB.no could be predicted based on all other features in the dataset has not yet been done. Therefore, the goal of this report is to...". You need to describe both the rationale for gathering the data, and also why you think machine learning may be useful on the data set you chose (or for analyses on similar types of data in the future).

Methods

The first thing to do here is to describe the data set. You need to provide the source (i.e. where you found the data) and the format of the data. It is also useful to discuss the features of the data. You can include some simple visualizations of the data, but save the detailed analysis and statistics for the Results-section. When including visualizations it could be wise to refer to an attachment (.py-file) that can reproduce the figure shown in the report, e.g. something like ”The map was created with ipyleaflet (see myplot.py)” or ”The bar chart was created with matplotlib and seaborn (see myplot.py)”. The point of the method section is to provide the reader with detailed information, but also ensure that the results-section can be read without too many interruptions on details from the methodology.

Results

This should be the main part of the report, and contain most of the plots and graphs. Remember that categorizing data and visualize it on a map or with statistics is a result in itself. For each statistic or quantitative measure you present, describe briefly the process on how you found it. As the data should be very rich, you need to explain why you decided to focus on the part you did. Why did you focus more on some features than others? The results section should be mainly observatory, and should not jump too much into conclusions. For the presentation of machine learning results you should clearly state which machine learning algorithm and model parameters you used. You need to discuss these choices in light of (at least one) other algorithm(s) and parameters. It is thus wise to test a few different algorithms and parameters on your data before deciding which one to use. You have to include both a training and a test set of data, and include model accuracy in your report.

Conclusions

In this section you first summarize your findings in a few sentences. Then you need to discuss why your findings are important. You can relate your findings to how a business can improve, how the population can gain general knowledge about a subject, or how politicians need to change policies. On these matters, you are wise to be relatively subtle in your recommendations if you are not 100% certain. Moreover, you should discuss whether the accuracy of you model is sufficient to provide recommendations, as well as the practical implications when your model fails. Try to rely on what the data tells you. It is also useful to write a few sentences about any limitations of your analysis.

Grading

Grading (from 0 to 100) is based on the following criteria:

- Is the topic well introduced? Does it provide the right amount of context and facts?

- Is the dataset adequately described and discussed? You will not be evaluated on the data set you chose as long it is large enough for machine learning.

- Does the result section provide detailed explanations and statistics about the data sets? Are the visualizations appealing and explanatory?

- Is the machine learning pipeline well explained? Was other types of machine learning models tested? For instance different algorithms, or even different problems (e.g. classification, clustering and regression on the same data set?).

- Is there a thorough evaluation of the performance of the machine learning model(s) of choice? Is model accuracy explained and sensitivity to model parameters tested?

- Did the conclusions provide a concise summary of the findings? Do the conclusions link back to the problems and goals addressed in the introduction and are the conclusions supported by the data?