Data Analytics - Fall 2022
Assessment Info

Adalbert Wilhelm

Time and Format

The final assessment in this course is done by a project report. The corresponding project is a data analysis performed for a given dataset.

- The project and the report are done to be individually.
- The data set and your task(s) will be communicated on December 6, 2022 by 17:00 at the latest.
- Your report is due on December 09, 2022 by 23:59 at the latest.
- Submission is via moodle on the course page: `https://elearning.jacobs-university.de/course/view.php?id=5057`
- Please submit as a single archived file.

Goal and Tasks

The goal of the project is to perform a reasonable analysis of the data given. Each student will get a (slightly) different data set and a goal for model building. So, this will be a supervised learning problem, most likely with a regression and a classification component.

The specific analysis is up to you. The evaluation will not be based on how well your model(s) perform, but on the way you approach the task and put the different analysis steps and options together.

Structure of the Report

The report is supposed to consist of two parts:

- Executive Summary
- Detailed Report

It is ok to generate a notebook and have code and text mixed. But it is also fine to separate the report text and put the code in an appendix. Please note that grading will predominantly be based on the analysis steps you perform, how you interpret the intermediate results, how you align your modeling strategy to these results, and how well you explain what you are doing. The code will be looked at to see that you really have performed what you describe, to sort out any technical/computational issues, and to evaluate the correctness of your figures.

Executive Summary

The Executive Summary is supposed to be at most a 1-page rather non-technical description of the project addressing the following points:

- Task/Goals
- Data background (if information is available)
- Approach/methods used
- Results

Detailed report

The expected length of the detailed report is about 20 pages (depending on number of graphics included, formatting, etc.). A typical structure could look like this:

1. Introduction
2. Background of the Data (if information is available)
3. Data Preprocessing
4. Data Exploration
5. Data Analysis
6. Results and Conclusions
7. References (Bibliography)

In particular steps 3 to 5 are highly interrelated and can be organized in different ways. It is however important to keep a good structure.

Hints

For preparing for the project I strongly recommend that you construct a rather generic data pipeline that

1. reads the data
2. does a simple data cleaning check (e.g. any odd variable types, missing values)
3. splits the data into training and test
4. does a very simple data preprocessing step (e.g. remove all cases with missings)
5. does a simple data exploration (e.g. histograms/bar charts for all features)
6. trains some standard model (e.g. linear model for regression, logistic regression for classification) on the training data
7. evaluates your model on the test data

Feel free to take some inspiration from this blog entry: `https://towardsdatascience.com/beginner-guide-to-build-compare-and-evaluate-machine-learning-models-in-under-10-minutes-19a6`

Once you have a working pipeline, think about strategies to improve on the first, simple model, e.g.

- by further inspecting the raw data and transforming features correspondingly
- by using feature selection techniques
- by applying cross-validation to optimize hyper-parameters in your model(s)
- by applying other models (which alternatives do you know)

Please make sure that you allocate a sufficient amount of time for writing the report. It is not sufficient if you do the analyses, you have to talk about them, describe what you are doing, why you are doing it, and what the command outputs tell you. Python and R can do a lot of things, but **you** shall be assessed on **your** understanding, knowledge and skills in data analytics, not the developers of R and scikit.learn.