Data Science and Machine Learning in Python

Assignment 7 – Own project

Stephan Weyers

Fachhochschule Dortmund
University of Applied Sciences and Arts

Assignment 7

Provided files

- W07_Results.xlsx
- W07_Peer review.xlsx
- W07_Task.pdf (this document)

Instructions

- Work together with the team mates who have been assigned to you (during lecture or via email).
- Read the task description on the following pages
- The end product should be one ipynb-file with your Python code and one pdf-presentation. Submit these files in ILIAS (one version per team, only one submission required. If two or more team members upload the presentation, the most recent version counts). Do not put your names in the presentation to enable anonymous peer feedback
- Evaluate the contributions of your own team members in the Excel file
- Submit the populated xlsx file in ILIAS (one version for each individual student)
- After all teams submitted their results, each individual students has to review 2 other team distributions, so that each team gets at least 6 student feedbacks
- By default the grading will be the median assessment of the 6 peer reviews weighted by the contributions of the team mates. Stephan Weyers will (selectively) check the grading. In addition, teams can ask for revision, if they are not satisfied with their grades.

Due date

- July 12th (23:59 German time) for submitting your solution
- On July 14th the files will be shared for peer review
- July 26th (23:59 German time) for providing your peer reviews

Own project task description (1/2)

Do not view the steps below as an entirely linear process. Data science is the creation of a hypothesis based on exploration and testing of that hypothesis is through analysis. You may need to go through many of these steps multiple times before you arrive at meaningful hypothesis or conclusions.

Step 1: Choose a dataset or datasets

Based on your interest, identify a dataset which you want to examine. You can use data that we already looked at in the course or completely new data. In the “01 Ressources” folder of the ILIAS course there are links to publicly available data sources, but you can also use any other source or even create your own dataset. Feel free to use a combination of datasets which you can merge in some meaningful way.

Step 2: Explore the dataset(s)

Explore what is present in the data and how the data is organized. If using multiple datasets, you’ll need to determine what common features allow you to merge the datasets. You’ll want to answer the following questions: Are there quality issues in the dataset (noisy, missing data, etc.)? What will you need to do to clean and/or transform the raw data for analysis?

Step 3: Identify research question(s)

Now that you have a better understanding of the data, you will want to form a research question which is interesting to you. The research question should be broad enough to be of interest to a reader but narrow enough that the question can be answered with the data. Some examples:

- Too Narrow: What is the GDP of the U.S. for 2011? This is just asking for a fact or a single data point.
- Too Broad: What is the primary reason for global poverty? This could be a Ph.D. thesis and would still be way too broad.
- Good (example 1): Do science fiction movies tend to be rated more highly than other movie genres? You can pull out the ratings by genre and see how they stack up to one another. You could also see if the distributions of ratings within genre are comparable across genres (maybe science fiction movies tend to be either highly or poorly rated, with little in between).
- Good (example 2): Can you use sentiment analysis on comments about movies on Twitter to predict its box office earnings? If you have, or can obtain, tweets which refer to a variety of movies and you have their box office earnings, this is a question which you can potentially answer well.

Source: Adapted from UCSanDiegoX DSE200x – Python for Data Science (https://www.edx.org/micromasters/uc-san-diegox-data-science)

Own project task description (2/2)

Step 4: Describe your dataset(s) and your research question(s)

Step 5: Use appropriate methods to explore your data

Based on your research question, you can now explore your data to answer the question. Use appropriate methods to answer that question. For example, if you are looking for a relationship between two items (say, CO2 emissions and GDP), you may wish to use scatter plots and statistical correlation. If you are trying to predict an outcome based on input data, you'll need to identify the appropriate methods from machine learning. Be sure to document, in your notebook, what you are exploring and why.

Step 6: Present your findings

What did you learn from the data and how do your findings help answer your research question? Use visualizations to present these findings. Try to keep your entire presentation at around 10 total slides.

Keep the presentation formal and coherent as a self-contained entity. Anyone reading your presentation should have a full understanding of the question, approach and the results.

Be professional. You should be comfortable giving this presentation to the general public, your boss, or your academic advisor.

Write for a diverse audience including:

- General public: Reads only the title and abstract looking for high-level point mainly for conversational purposes.
- A company CEO: Reads introduction, research question, findings and conclusions looking for business value
- An academic advisor (or company CTO): Reads the full presentation AND your Jupyter Notebook paying particular attention to technical coherence, academic value, and technical data science strengths.

Presentations should be for the above-mentioned three audiences. Think of the diversity of the audience. The whole point is to tell a story - so you should motivate a reader to care based on the question you are exploring, answer that question in a clear and concise manner, provide an honest appraisal of your results, and give the reader valuable insights. Use charts whenever possible. Avoid slides with a lot of text and bullet points - break the slide into multiple slides when this happens. Be concise!

Source: Adapted from UCSanDiegoX DSE200x – Python for Data Science (https://www.edx.org/micromasters/uc-san-diegox-data-science)