CptS 475/575: Data Science
Fall 2023

Assignment 1: Create Data Science Profile of Yourself and Reflect on an Article on Data Science

Release Date: 08-25-2023. **Due by: 09-01-2023, 11:59pm.**

This assignment consists of two tasks. The first task is about a data science profile of yourself and it builds upon the discussion we had in class on 08/25/2023. The second task asks for your reflection on an article that is one of the required further readings for the lecture we had on 08/23/2023.

Task 1 (50 points total)

As I explained in class, the purpose of this task is to create a visual data science profile of yourself. Specifically, you will create two instances of profiles. The first will show the way you see yourself now. The second will show how you would like to see yourself by the end of the course.

The profile is simple. On the horizontal-axis you will have seven "areas of skills" that could generally be regarded important to Data Science:
1) Computer Science. 2) Math. 3) Statistics. 4) Machine Learning. 5) Domain expertise. 6) Data visualization. 7) Communication and presentation skills.

On the vertical-axis you will have a relative scale (think percentage) of your skill level in each of these areas. The area in which you have the strongest skill will be close to 100, and the area in which you think you have very little skill would be close to zero.

As an example, see the slide in the lecture slides of Aug 25 (posted on Canvas under the module titled What is Data Science in the file What is DataScience-partII) that shows the data science profile of the author of the book "Doing Data Science". As a context, the author, Rachel Schutt, has a PhD in Stat and has held several senior and executive-level Data Science positions in industry.

Your task is to create your own profile – two to be exact, one showing current and the other projected. You are still a student and you may not feel you have a lot of skill in some of these areas. Allow yourself a generous interpretation of skill level and keep in mind that this is on a *relative* scale. Also, keep in mind that it is perfectly okay to have zero skill level in some of these areas. For example, if you are a computer science major, it is natural that "Domain expertise" would be the area in which you have the lowest skill level among the seven, and it is okay for it to be close to zero. That said, if you have had an internship some place or an interest/hobby that you think has helped you acquire some expertise in an area, you could take that into account in deciding the level of your "Domain expertise". In any event, make sure to mention what the domain is if you indicate your domain expertise to be non-zero.

You can use any tool (Excel, R, Python, etc) you wish to make the plots.

Here are a few associated presentation considerations and discussion points you are asked to address as part of this task.

1.a. (40 points) The areas in the horizontal axis could be ordered in a number of different ways. What ordering in your opinion would be most effective (and aestetically pleasing) and why? Create your profile in the order you chose.

1.b. (10 points) Is there a skill (bucket) you think should be added to this data science profile? A skill you think should be removed? Specify and justify briefly.

Task 2 (50 points total)

As you recall, we briefly discussed the article "Data Science and Prediction" by Vasant Dhar in class in connection with the topic "what is data science?" A link to a copy of the article is posted on Canvas. Read the article and briefly answer the following questions.

2.a. (15 points) The author identifies a few ways in which data science differs from statistics. What are those ways?

2.b. (25 points) In the section of the article headed "Knowledge Discovery" (pages 70 to 72 of the article), the author makes a distinction between domains in terms of the predictive power of their theories (models). Secifically, the author points out that models in the physical sciences are generally expected to be "complete", whereas in the social sciences they are generally "incomplete". The author discusses ways in which "big data" could potentially put domains on both ends of this spectrum on firmer grounds in terms of theory development. Give a brief summary of the ways the author identifies. Do you see any additional ways than what the author sees?
(If the discussion in this section of the article reasonated in some ways with your own research or work you do, feel free to incorporate that in your answer.)

2.c. (10 points) Imagine you were asked to write a "head-line" (as you see in newspapers) for this article, followed by two or three very telling summary sentences. What would your headline and the summary sentences be?

Weight:

Task 1 carries 50% – broken down as 40% for 1.a and 10% for 1.b
Task 2 carries 50% – broken down as 10% for 2.a, 15% for 2.b and 25% for 2.c.