CptS 575 Data Science
Washington State University

Assignment - 1

Contents

1 Task 1

1.1 1.a.

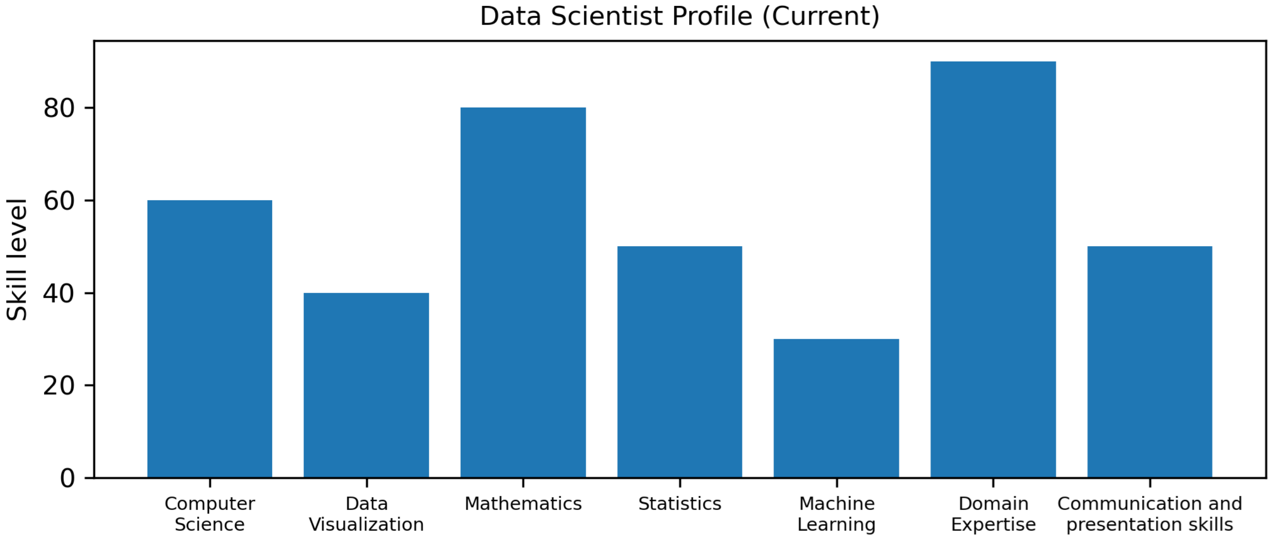

Figure 1: Data Scientist Profile (Current)

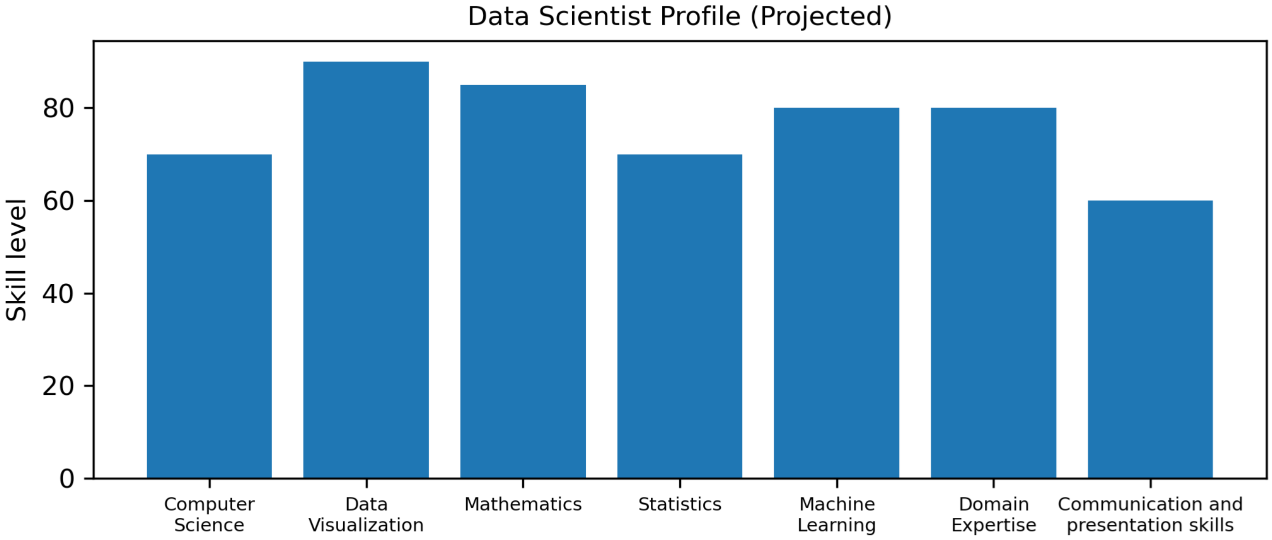

Figure 2: Data Scientist Profile (Projected)

The order of the skills which I have came up with is from a point of view of a recruiter. As a Data scientist, it is good to see the computer science and data visualization capabilities first which are important in day to day tasks of a data scientist. Later the knowledge in the math and stat can be analyzed as it provides analytical foundation required for data scientists. Domain expertise comes later which explains about the speciality of that particular candidate. Finally, communication skills can be represented last to access it more easily from last column.

1.2 1.b.

I would like to add one more bucket named ”Project Experience”. Even though you have enough theoretical skills, having a hands on experience is good for analysing a data scientist profile. This will give an info that I have worked in some projects and comfortable with the topics and ready to take on a project. Also I think no buckets to be removed since each represent critical aspects of a data scientist.

2 Task 2

2.1 2.a. Difference between data science and statistics

In the article, Vasant Dhar outlines several key distinctions between data science and traditional statistics:

1. **Structure of Data:** Traditional statistics mainly deals with structured data whereas Data science handles a wider range of data types, including unstructured data such as text, images, and videos.
2. **Scale of Data:** Compared to statistical methods, Data science is able to handle massive datasets, often referred as big data.
3. **Emphasis on Prediction:** Statistics mainly revolves around hypothesis testing and inference while Data science has a huge emphasis on predictive models.
4. **Interdisciplinary Approach:** Statistics is majorly mathematical and theory based. However Data science is interdisciplinary which integrates methods from computer science, machine learning and other fields.
5. **Automated Decision-Making:** Data science involves automated processes where computers not only analyze data but also make decisions based on that analysis, which is less common in traditional statistical practices.

2.2 2.b. Summary

In the ”Knowledge Discovery” section, Vasant Dhar distinguishes between domains based on the completeness of their models:

- **Physical Sciences:** Models in the physical sciences are typically considered complete, providing accurate predictions because they are based on well-understood causal relationships. Big data enhances these models by offering more data points, leading to more precise predictions and validating the underlying theories.

- **Social Sciences:** In contrast, models in the social sciences are often incomplete and lack the predictive accuracy of physical science models. However, big data can contribute to developing more accurate predictive models, even when the underlying causal mechanisms are not fully understood. These predictive models can serve as a foundation for future theory development by uncovering patterns that were previously unnoticed.
- **Additional Perspective:** Beyond what Dhar discusses, big data could also enable *real-time feedback loops* in theory development, particularly within the social sciences. This would allow theories to be continuously updated and refined based on new data, leading to more dynamic and adaptive models.

2.3 2.c. Headline and summary

Headline: **Beyond Statistics: The Predictive Power of Data Science**

The article explains how data science differs from traditional statistics by emphasizing predictive power and capability of handling big data. With the help of big data, data science is able to contribute to theory development across various domains, from physical to social sciences.