CptS 575 Data Science: Assignment 1

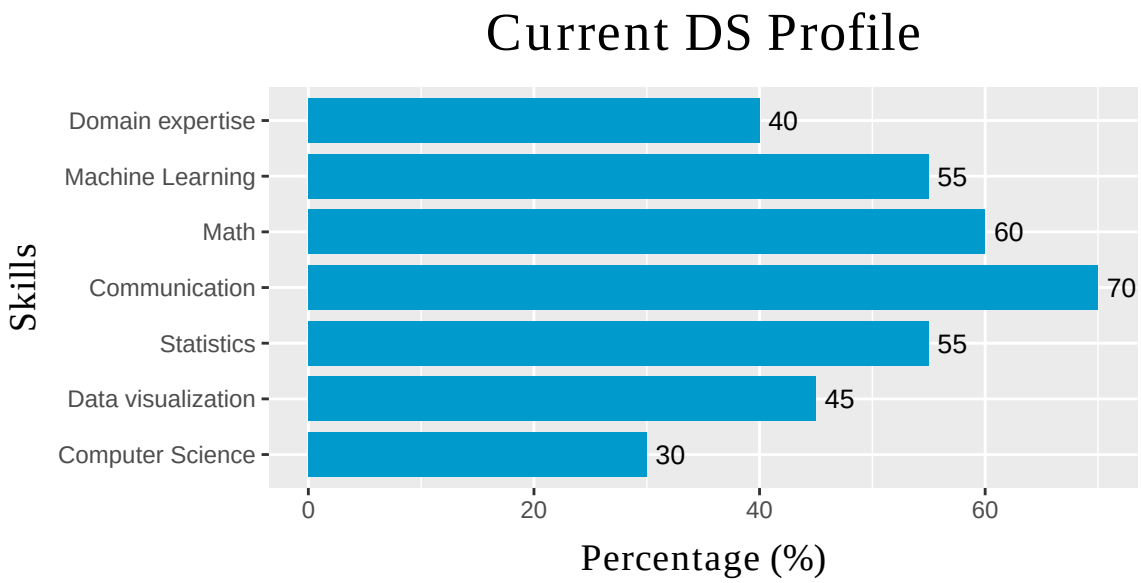
Md Muhtasim Billah

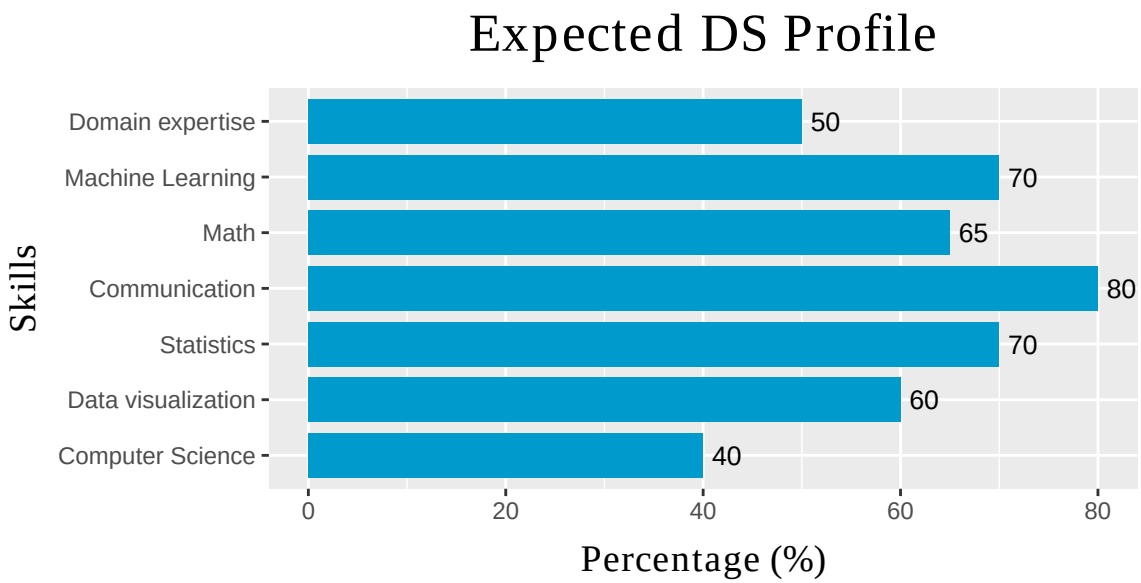
9/2/2020

Task 1 : Create Data Science Profile of Yourself

Answer:

My current visual data science profile (at the beginning of this class) and the expected profile (at the end of this class) is shown below. Here, as "Domain Expertise", computational modeling has been considered which is the area of research for my PhD.





1.a.

The data science skills (y-axis) could be ordered in number of ways acccording to the relative skill percentage (x-axis) such as in an ascending order, a descending order or putting the skills with the highest relative percentage in the middle. For visualization, I have chosen the third option to create my profiles for both aesthetical and effectiveness reasons.

In terms of aesthetics, I believe that the barplot looks like a histogram of a normally distributed dataset in the shape of a bell curve which is pleasing to eye. It also seems more effective to me because it shows the relative difference among the two adjacent skills and thus easier to point out the skills which require to be worked with as well as the amount of works required for them.

1.b.

Though not within the scope of this class, since this is a booming era for big data, I think the skills regarding database management systems (DBMS) can add value to someone's profile and might be added in this chart. I wouldn't remove any of the remaining skills since all of them possess significant importance though not equal in their extent.

Task 2: Reflection on the article "Data Science and Prediction" by Vasant Dhar

Answer:

2.a.

The author identifies a few ways in which data science differs from statistics. Those ways are as follows:

i) While data is the raw material of both statistics and DS, it vividly varies in nature. DS deals with the data which are extremely heterogenous and unstructured (text, music, video) and it is ever increasing at an exponential rate.

ii) Outgrowing the scope of statistics, DS requires integration of tools and techniques from several other scientific fields such as computer science, econometrics and sociology etc.

iii) Big data, integrating with computer science, is allowing computers to interpret the data automatically which is making them take decisions on their own intuitively. DS has paved this way for automated decision making.

iv) Traditional statistical methods and databases are inadequate for knowledge discovery since they are only useful for well-formulated query of the user. On the other hand DS, taking advantage of big data, can discover actionable insights about data patterns even when the user is in lack of a well-formulated query.

v) Traditionally, a theory originates in the human mind based on prior theory. Then data is gathered to demonstrate the validity of the theory. DS, equipped with Machine learning, has turned this process around. Given a large amount of data, the computer can provide interesting insights if the right question is asked.

2.b.

The ways that the author identified for bridging the gap between physical science and social science in terms of theory development are noted below.

i) Big data can help develop more accurate theories reducing the prediction error of the model for social sciences.

ii) Two common sources of error - misspecification of the model and a small sample size - can be overcome using big data, since the model has to make lesser assumption about its parameters (reducing the bias) and a huge sample size better represents the true population.

iii) The third source of error, the randomness, can also be tackled to a limited extent by conductng inexpensive large-scale randomized experiments on social behavior. Internet and big data provide a basis for testing them.

iv) Though I am from mechanical engineering background (so, "hard science"), my PhD research is on developing and working on a computational model of a biophysical system. I have to run extensive simulation to study the paramter effects on drug delivery which takes weeks. While there is no solid theory in place, only experimental and computational investigations for measuring importance of these parameters, with a significantly large dataset on the paramters of this biological pricess can go a long way to minimize the use of resources for carrying out the same research.

2.c.

My headline with a few summary sentences for this article will be as follows.

Data Science : The True Art of Clairvoyance

This article presents an educative disection of the intertwined aspects of statistics and data science as well as the implications and promises offered by big data and machine learning in the upcoming decades. The skillset spanning over several disciplines which are required to thrive in the industry as data scientist and the challages posed at the management levels have been discussed. How the emerging technologies are influencing the age-old theory development and knowledge discovery process by elevated predicition capabilities are entertained as well.