CptS 575: Data Science

Fall 2018

Assignment 1

Submitted To

Assefaw Gebremedhin

September 1,2018

Submitted By

Sukhjinder Singh

Task 1:

1.a. Data Science profile based on “areas of skills”

According to my opinion, to become a domain expert we should have basics strong knowledge related to the domain. Computer science is the generalization of all the fields, this fields let us to know what's the most interesting field we want to work. My specialization is data science and I want to become an expert in the data science, I should know the mathematics, statistics and data visualization for the datasets to analysis and visualize in perspective sense. Now, how we use analysis the dataset that depend on the algorithms. Machine learning provides different types of algorithm to solve the problem related to data analytics. For example, Web page classification: various spam and junk pages, like soft404, parked domain, etc. It runs inside Bing index generation pipeline processing billions of pages. The one of most important field in every field is communication and presentation skills. According to me, I gave the first preference to communication and presentation skills, computer science which help you to decide your interest in the field. Mathematics, statistics, data visualization, and machine learning is the basics component to become domain experts. First Histogram shows currently data science profile and skills set currently I think I have. I have looking for number of factors or skills I want to improve in this class. Second Histogram shows projected data science profile.

Figure 1: Currently Data Science Profile

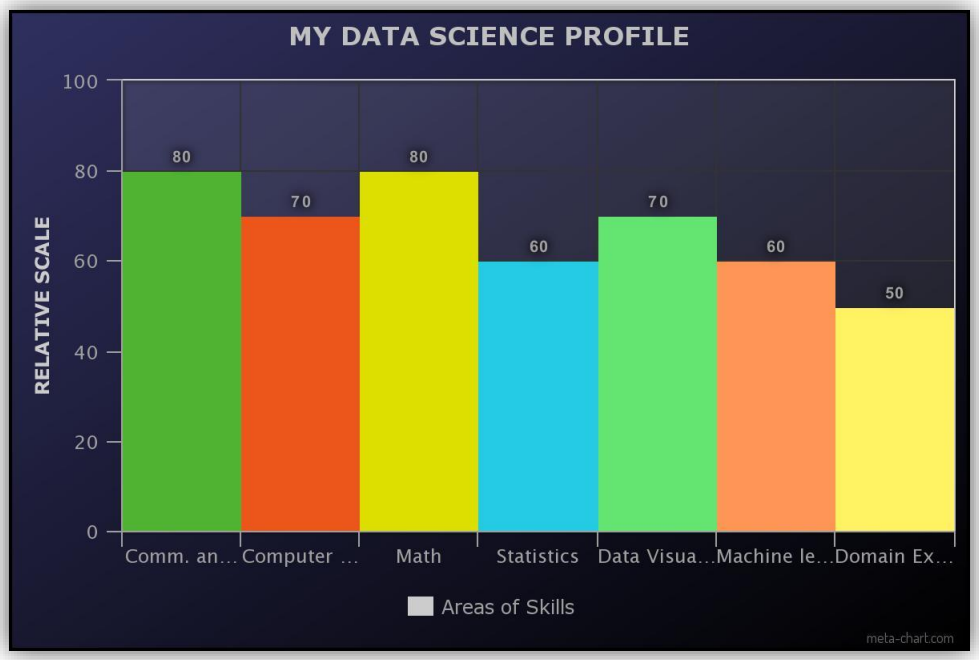
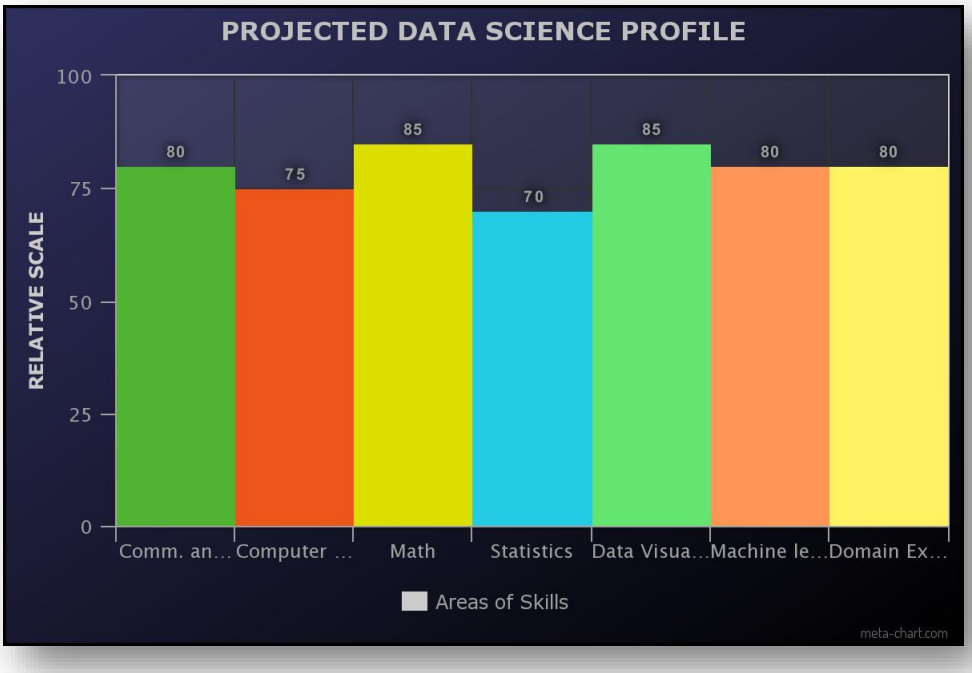


Figure 2. Projected Data Science Profile



1.b.

I think the most fundamentals of a data scientist's skill set- the job of data scientist is much more applied than that of a traditional statistician. I think **programming** is important in multiple ways, including three ways below:

- We can create different type of **tools to do better in computer science**. This includes everything from building systems that different community can use to visualize data, create frameworks to automatically analyse experiments, and managing the data pipeline.
- The ability to analyse **large datasets**. The datasets we get to work with in industry are huge-we easily get data that reaches millions of rows and many more.
- We should be able to program **augments** your ability to do statistics. If we have a bunch of statistics knowledges but no ways to implement it, our statistics knowledge become much less useful.

The normal software engineering training here will help us to develop programming skills.

Task 2.

2.a.

Data Science can be defined in two different terms which is Data and Science. Data Science imply a focus around data whereas Statistics, is a systematic study about the organization, properties and analysis of data and their role in inference, including our confidence in such inference. The author identifies few ways to differentiate between data science and statistics.

- The raw material, the "data" part of Data Science, is increasingly heterogenous and unstructured- text, images, and videos, often emanating from network with complex relationships among its entities. In the paper, author shows the volume of structured and unstructured data between 2018 and 2015, projecting a difference of 200 PB in 205 compared to a difference of 50 PB in 2012 which is four times.
- Analysis of the homogenous data which required integration, interpretation and sense making, increasingly based on tools from linguistics, sociology, and other disciplines. Moreover, the proliferation of mark-up languages, tags, etc. are designed to let computer interact with the data automatically, making them active agents in the process of sense making. In contrast to early markups, such as HTML that were displaying information for human

consumptions, most of the data now being generated by computers is for the consumption by the other computers.

- Computers are increasingly doing the background work for each other. This allows decision making to scale: it is becoming increasingly common for the computers to be the decision maker, unaided by humans. The authors show the shift form human towards computer as decision makers raise a multitude of issues ranging from the cost of incorrect decision to ethical and privacy issues.

2.b.

According to the author, big data allows us to significantly reduce **the misspecification of the model error** that means a linear model that attempts to fit a nonlinear phenomenon will generate an error simply because the linear model imposes an inappropriate bias on the problem and eliminate **simple estimates error** which become reasonable proxies for the population. For example, large amount of data allows us to consider richer models than linear or logistics regression simply because there is a lot more data to test such models and compute reliable error bounds. In the last, Big data eliminates **randomness** from the model.

2.c.

Headline - "Data Science is Changing and Data Scientists will Need to Change Too"

Summary

The field of Data Science is in a transitional mode in terms of how the latest data technologies are being used to solve business problems for a strategic advantage. Soon, Deep changes are underway in how data science is practiced and successfully deployed to solve business problems and create strategic advantage. These same changes point to major changes in how data scientists will do their work. The current market trends in Business Analytics indicate that the platform strategy will soon shift from being a "one-stop, general purpose" platform to a domain-specific solution geared to industry sectors such as ecommerce, finance, HR, manufacturing and so on.