CPT_S 575(Fall 2020)

Assignment 1

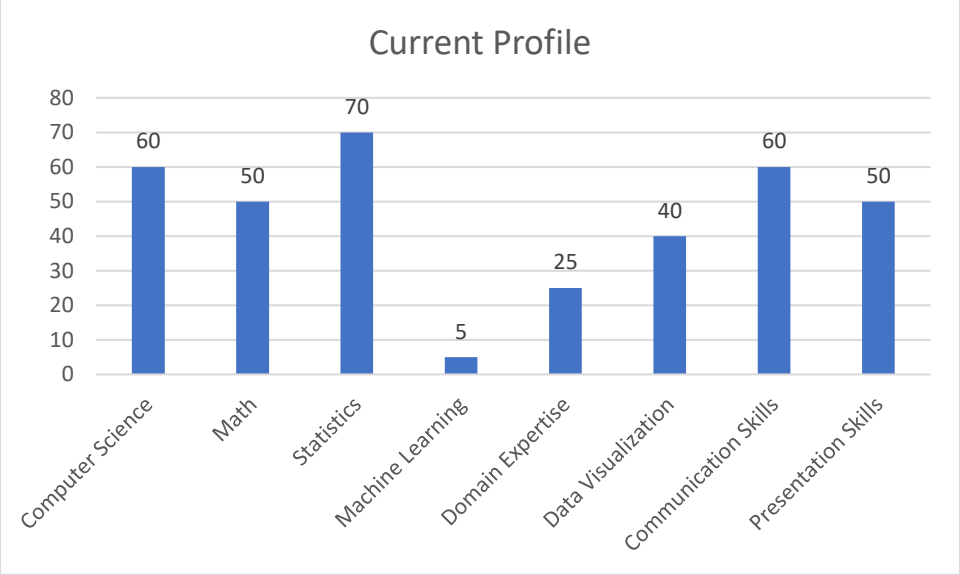
Tayyab Munir

11716089

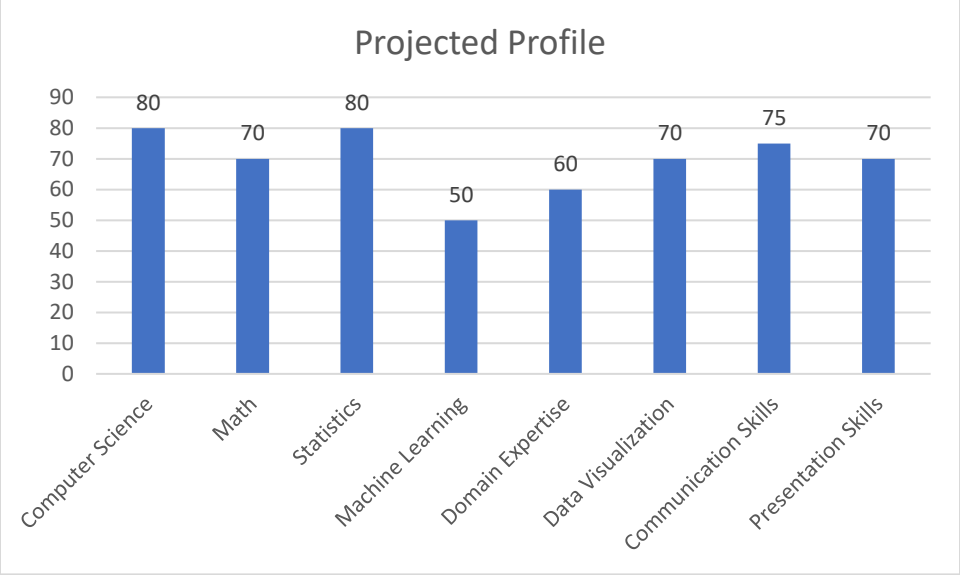
Date: 09/02/2020

Task 1

a.

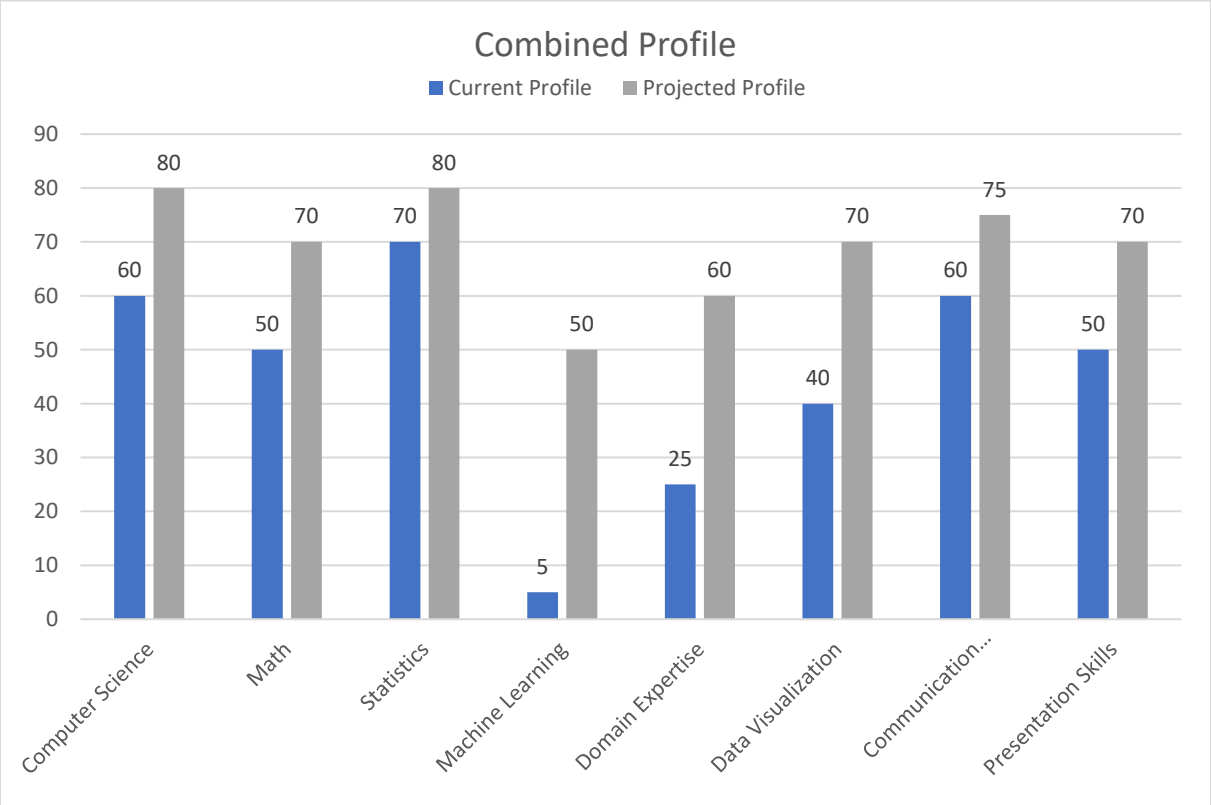


My current profile shows my Statistics skill to be the highest. My Computer science and communication skill is at 60%. Math is at 50% Machine Learning is at 5% as I have not taken any course of ML but I have a little idea about that. Domain expertise is at 25%. My Data Visualization is at 50% as I have done some projects related to Databases during my bachelors but that was 4 years ago now and I need to brush it up to improve. Since I moved to US for Graduate degree, my communication skills (60%) have improved as I am getting more exposure here and moreover my presentation skills (50%) as well.



My projected Skills graph shows that I will improve in almost every related area discussed in this assignment. My computer science skills, for sure, will increase as this semester goes and, in my

view, a 20% increase in this will make a huge impact for my future. My Math skills will also increase as I am simultaneously studying math related courses as well and it shows an increase of 20%. My statistics skills will increase from 70% to 80%. Moreover, I am also taking Machine Learning course as well which I am sure that it will help me with this course as well and my knowledge related to ML will increase from 5% to at least 50%. Domain expertise, again, will improve as I will be studying Data science, Big Data and ML tis semester. Data visualization skills will be brushed up this semester and I am very much excited for this to happen. Last but not the least My communication and Presentation skill will also improve as I will be communicating and working in teams with my classmates to work on different projects this semester.



b. 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.

I think critical thinking is a skill that should be added to this data science profile because the person working as a data scientist must have the ability to critically analyze and interpret of facts and data. Critical thinking is required to come up with a conclusion at the end of the process.

Task 2

a. The author identifies a few ways in which data science differs from statistics. What are those ways?

- Type of Data: In Statistics we primarily deal with structured data but in Data Science we work with increasing amounts of heterogeneous and unstructured data (text, images, video).
- Required Disciplines: To perform Statistical analysis we only require knowledge of Mathematics but when dealing with unstructured data we need to have a combination of different disciplines like Computer Science, Mathematics, Linguistics and Sociology amongst others to better analyze the data.
- Decision making: Decision making using Data Science is more efficient when compared to using Statistics since we have a large amount of data that is being generated by humans and computers and the computers are able to understand the data that is being created and take better decisions automatically.
- Knowledge Discovery: Statistics is not optimal for Knowledge Discovery since they provide data based on the given query while in Data Science, we are able to find patterns within the same data.
- Prediction Capability: One of the key factors that is given the most consideration in Data Science is the predictive accuracy of future observations based on a given set of data while Statistics mainly deals with just the analysis of the given data.

b. 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). Specifically, 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?

In theory development, whether we start with a pre-conceived idea of the theory or try to develop a theory from the results of the data, we use data science. In physical sciences, we proceed with data analysis after understanding the phenomena (relation) of what we are looking for therefore the explanatory and predictive models are the same. In social sciences we make assumptions about the data and then proceed with the analysis and based on the results of the analysis we extract causal models.

In both physical and social sciences, we encounter three kinds of errors: misspecification of a model, the samples used for estimating parameters and randomness. These errors have been solved using data science. In misspecification of a model which occurs because of choosing an incorrect model for a problem is rectified by using large amounts of data to train and test the model such that it makes fewer assumptions and has reliable error bounds. The samples used for estimating parameters causes the model to have a greater bias in its estimate when the sample is

small. This issue has been resolved by using high volumes of data such that the sample estimates to be a reasonable proxy data. The error due to randomness occurs because the data that we process is passive and performs predictions for the given information. It does not predict the output when one of the variables are changed during run time. In short randomness occurs because the data that we receive is not from a controlled environment.

Using data science, physical and social sciences are now able to analyze data at a granular level, draw conclusions from the results, find patterns and connections they would have never otherwise figured out as well predict the future. This leads to formulating theories from the data with less assumptions and bias.

Data science also aide's theory development through its powerful data visualization and communication tools. It helps produce a graphical representation of the analysis and the output of theory in a manner which can easily be understood by people from all levels of expertise. Its ability to convert the report on the theory that is created once into multiple formats without additional effort helps researchers to reduce the unnecessary time taken to generate the same report in different formats.

c. 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?

"Data Science: The key to thrive in the New Era of Data".

In a world where data is increasing exponentially by the millisecond, we have found an efficient way to manipulate this data and predict future outcomes using Data Science. The skills needed to perform data science and its applications to real world problems have been discussed.