Data Journal 1

Ludek Cizinsky, luci@itu.dk

September 2020

I decided to focus my first journal on the following question: "What are the key challenges data scientists have to face when dealing with survey data?" I observed that except from technical challenges, there are also many social as well as ethical challenges when evaluating survey data. Since these challenges were also discussed in an Introductory of the book Weapons of Math destruction from Cathy O'Neil, I will try to put them in a context with observations from the author of the book who actually saw the challenges from both academic as well as business perspective. According to one of the O'Neil's observations, the importance of understanding how data about ourselves are being evaluated is rapidly increasing since our lives are more and more influenced by the patterns derived from the data.

I would like to start with pointing out a few interesting observations from the survey data, which was sent out to all first-year data science students. First, only half of the students is using bike in order to get to the school. I expected the number to be at least around 75 %, but when I looked on the survey from Cycling-embasy.dk conducted in a year 2017 focused on Copenhagen area, it also stated that around half of the respondents is using a bike to get to work or school, thus the result of the internal student survey might reflect reality accurately.

Second, there were only two people who voted they knew a lot about Data Science before coming to study at ITU and almost half of the respondents did barely know anything. There are two things that caught my attention. Since I was one of the persons who voted for the highest grade of knowing about the field of Data Science prior coming to study it, I would like to know whether my claim was correct or compare to other people knowledge not. Additionally, I find surprising that many people chose to study Data Science without knowing a lot about it. My assumption for this behavior is that Data Science has been many times discussed in the news, articles and other media, thus leading many people to a conclusion that this field will definitely secure them (at least) a good job for the next decade or so.

Finally, I would like to discuss what were the challenges I encountered when evaluating survey data and how these challenges are related to the observations from the Cathy O'Neil's book. The privilege to design an algorithm and decide about its logic is only in hands of a few human beings. This of course raises a big social responsibility. If I were to imagine myself in this situation, it is very

hard to make important conclusions without including my own experience as an input. A great example of this behavior are my observations of people's prior knowledge of Data Science field. Moreover, Cathy O'Neil in her book argues that we often tend to look on the bright outputs of the given algorithm and to ignore the wrong outputs which as consequence leads to big social-economic imbalance. In other words, the algorithms tend to favor those in a good financial situation and put down those whose financial situation and background is not that bright. Therefore, my conclusion for this journal is a question: "How to be socially responsible when designing algorithms that may have an impact on millions of people? Moreover, is it even possible to find a perfect solution?" I hope I will be able to answers these questions in the next journals.