Social Data Science

Cecilia Linn Hansen

18 aug 2016

First draft of structure

1. **Intro**

i) Problem
   - What model? Prediction or causality?

ii) Motivation

iii) Ethics
   - No *robots.txt* file. Neither at http://www.ku.dk/ or http://karakterstatistik.stads.ku.dk/
   - NOTE: Students are not mentioned by name, but teachers are!

2. **Data**

i) Collecting (too good to be true)
   - Scraping process (Selenium, Stack overflow?)

ii) Cleaning
   - Limitation (from 32,000 obs to ??)

iii) Final data (short)
   - Descriptive statistics

3. **Analysis (Model)**

i) Focus on problem

ii) Analytic statistics (graphs, tables, etc.)

4. **Discussion and conclusion**

i) Discussion: Different views and further investigation

ii) Conclusion: FOCUS ON PROBLEM

Potential problems

"Can the type of assessment explain outcome of exam (grades)?"

- Our hope: No, outcome should be explained by ability

- But what if there is a significant explanation?

"Is there a difference between ordinary exam outcomes and reexamination outcomes?"

- Focus should be on the mandatory courses.

"Is there a difference in outcomes if the lecturer is new versus experienced in teaching the class"

–> Discussion on Friday: We will focus on the first problem :-) Make the analysis primarily on economic courses and compare our results with political science (statskundskab) if we have room and time.

Introduction

Ethics

When working with public assessable data, an important issue to consider is the privacy of the individuals involved. Since this is a rather new field of data science, there are few formalized standards for privacy boundaries. For some poeple it can be of personal interest not to have their private information floating around in the World Wide Web, but since we live in a digital society it is the case that personal information is available. Working with social science data it is our responsibility not to cross the *** line but still gain as much information as possible.

Our project is based on sensitive data regarding peoples grades, where each transcript is only distributed to the individual student. The policy of the University of Copenhagen is to keep the grade information for each student private, and it is up to the student to decide whether to share the information or not. Some students might find it intimidating to let other people know their results, which is why we need to keep this in mind when analyzing our data.

One way to consider privacy issues is to look at the official robots.txt file of a web page. This file describes what is allowed and not allowed when scraping their page, but it is often only used by the larger web companies like e.g. http://www.google.com or http://wwww.amazon.com. Since we are working with data from the University of Copenhagen, which does not provide a robots-file, we have the possibility to use all available data. That does not mean that we did not face situations where we needed to consider privacy issues.

Looking at the grade distribution data by itself there were no problems regarding private information, since neither professors nor students are displayed by ID. Furhtermore, courses with less than three students do not display grade distributions, thus it is not possible to recognize individuals and no privacy boundaries are crossed. However, merging with the course description data led to the exposal of the names of the course responsibe for each course.