Analysis of Data Science Pedagogy at UC Berkeley

Honors Thesis Project Prospectus

Christopher Pyles

September 18, 2019

This project will look into the success of the data science pedagogical practices at UC Berkeley, concerning both its mainstream course offerings (e.g. Data 8, Data 100) and its other programs through the Data Science Education Program (DSEP). DSEP offers connector courses and modules, which are meant to increase the awareness of the data science community at UC Berkeley and to provide students with a greater amount of data literacy in an increasingly data-centric world.

The goal of this project is to determine whether or not taking data science courses and/or taking a non-data science course in which there is a module are related to better student outcomes. That is, the goal is to find out whether the outreach being performed through data science courses and DSEP valuable to the UC Berkeley community at large. Given the scale of data science courses, it seems unrealistic to be able to poll every student, however the scale is also beneficial in that even sampling from it gives a sizeable sample from which to draw conclusions.

Because the foundational question of this project relates to outcomes for two different groups of people, it is likely that hypothesis testing will be a large part of the analysis for this project; specifically, A/B testing on the groups who are and are not exposed to data science curricula. Another goal for this project is to build a classification model that can be used on smaller data sets (e.g. data from pilot offerings of courses) to determine whether or not those courses will be beneficial. This model would be trained on the data that is collected for this project, and would involve a good deal of machine learning techniques (e.g. support vector machine, neural network, logistic regression).

Because this project will involve parsing student feedback, it is fair to say some rudimentary natural language processing will be involved. This includes a bag-of-words approach to parsing student reviews and sentiment analysis to help determine the overall attitude of students towards their interaction with data science or DSEP.

A good deal of data is already housed at the Division of Data Science and Information (DDSI), especially as it relates to the outreach performed by DSEP. There will be some data collection involved in this project to get data on the mainstream courses. This will involve going to lectures for Data 8, Data 100, and Data 102 to poll the students. Ideally, data from the course evaluations for these courses would also be used, but it is not yet known whether or not that data is accessible for projects such as this.

The process of EDA would involve quite a bit of looking at what kinds of features there are in the text of each student's answer and testing which kinds of sentiment analysis algorithms seem to have a relationship with the outcome that is being predicted. It would also involve creating cluster models to group responses based on their attributes, e.g. using Gaussian mixture models or hierarchical agglomeration.

The main challenge of the EDA process will be in figuring out how to synthesize one or a few rectangular data sets from a multitutde of different information of varying samples, granularity, and sources. For example, merging modules surveys and course evaluations will be very difficult, and the key to making these conditions comparable would be to find metrics that both data sources have in common and focus on those, or in transmuting differing sources into a common form.

This project's link to the data science major is clear, as it is a full-scale implementation of the data science lifecycle that is the focus of the DS undergraduate curriculum. It is also related in that the project focuses on the same ecosystem that contains the data science major, and which will touch it in its analytical scope. If the DDSI is viewed through the lens of a startup, which in some ways it is, then this project also relates to my domain emphais, Business & Industrial Analytics, as an analysis of the effectiveness of the Divion's key product: data science pedagogy. Many of the skills that I am using in this project common from the courses that I took for my domain emphasis, especially UGBA 147: Advanced Business Analytics.

It also links to my professional interests because of my involvement at the Division. My background in the Division has encouraged in me an interest in the budding data science pedagogical landscape, a place in which I hope to have a career after graduation. Working on analyzing the effectiveness of the program that we are creating at UC Berkeley will give me a better understanding of the good practices we are developing, and those practices which would be better left out of the program, an expertise I hope to take into the professional world.

The main link to Human Contexts and Ethics is in the fact that the data this project uses will be student data, and it is important to prevent the reidentification of individuals after anonymizing the data set. It is also important to think about the ways in which

the models used in this project are trained, to prevent them from learning human-induced biases. Introducing such biases could result in models that create wild inaccuracy in prediction and which produce results which, if acted upon, could be a detrimented to the students at UC Berkeley.