Principles of Data Science – DS UA 112

Capstone project

Assessing Professor Effectiveness (APE)

The purpose of this capstone project is to tie everything we learned in this class together. This might be challenging in the short term, but is consistently rated by students as being extremely valuable and useful in the long run. Please read these instructions carefully.

It is widely recognized that academia is in severe crisis and dire need for fundamental reform. Armed with your domain knowledge as a student and your technical skills as a data scientist, you heed the call of duty and come to the rescue in support of this noble cause. In this capstone project, you will focus on the assessment of professors from a sufficiently large dataset and the insights you can glean from that.

As a data source for this undertaking, the professor has scraped the website ratemyprofessor.com ("RMP", from now on). Students enjoy RMP as a source of information to inform their course choices and also contribute ratings to RMP to help out their fellow students by giving back to the community. Whereas overall response rates are low and potentially affected by response bias (the students who do provide a rating might do so in a complimentary or retaliatory fashion). Research shows that this data source is not entirely invalid, as the correlation between RMP ratings and student evaluations of teaching obtained at the end of a class are on the order of 0.7. In contrast to student evaluations of teaching, RMP data is considerably larger and publicly available, which is why this is the dataset we use for this project.

As this is not a data engineering class, almost all the necessary data munging and scaffolding (e.g. scraping the actual site, converting the html to data, collating the information from individual ratings, anonymization, etc.) has already been done by the professor (you are welcome). The remaining pre-processing steps are data science relevant (e.g. how to identify and handle missing data) and are thus left for you to implement.

Mission command preamble: As usual, we won't tell you *how* to do something. That is up to you and allows you to showcase your creative problem-solving skills. However, we will pose the questions that you should answer by interrogating the data. We might also give hints.

Format: The project consist of your answers to 10 (equally-weighed, grade-wise) questions. Each answer *must* include some **text** (describing both what you *did* and what you *found,* i.e. the answer to the question), a **figure** that illustrates the findings and some **numbers** (e.g. test statistics, confidence intervals, p-values or the like). Please save it as a pdf document. This document should be 4-6 pages long (arbitrary font size and margins). About ½ a page/question is reasonable. In addition, open your document with a brief statement as to how you handled preprocessing (e.g. data cleaning), as this will apply to all answers. Make sure to include your name.

Deliverables: Upload two files to the Brightspace portal by the due date in the sittyba:
*A pdf (the "project report") that contains your answers to the questions, as well as an introductory paragraph about preprocessing, how you seeded the RNG, etc.
*A .py file with the code that performed the data analysis and created the figures.

Academic integrity: You are expected to do this project by yourself, individually, so that we are able to determine a grade for you personally. So make sure this works reflects your intellectual contribution – not that of third parties. You can use generative AI like chatGPT to aid you in this task. There are enough degrees of freedom (e.g. how to clean the data, what variables to compare, aesthetic choices in the figures, etc.) that no two reports will be alike. We'll be on the lookout for suspicious similarities, so please refrain from collaborating.
To prevent cheating (please don't do this – it is easily detected), it is very important that you – at the beginning of the code file – seed the random number generator with your N-number. That way, the correct answers will be keyed to your own solution (as this matters, e.g. for the specific train/test split or bootstrapping). As N-numbers are unique, this will also protect your work from plagiarism.
Failure to seed the RNG in this way will also result in the loss of grade points.

We do wish you all the best in executing on these instructions. We aimed at an optimal balance between specificity and implementation leeway, while still allowing us to grade the projects in a **fast**, **fair** and faithful (=consistent and accurate) manner (FFF).
Everything we ask for should be doable from what was covered in this course.
If you take this project seriously and do a quality job, you can easily use it as an item in your DS portfolio. Former students told us that they secured internships and even jobs by well executed capstone projects that impressed recruiters and interviewers.

Considerations: *There is some missing data, you'll have to handle it somehow.
*Note that the *average* rating is more meaningful if it is based on more ratings, as discussed in class. You have to handle this somehow. Either you can accept all data (which will likely yield extreme average ratings of 1 or 5, based on a single rating), set a threshold (only accept data with more than k ratings as valid) or weigh the average somehow. This is a judgment call. Argue how you make it.
*As we are concerned about false positives, as you have sufficient power and as we have to correct for multiple comparisons, consider an alpha level of 0.005 as the threshold for statistical significance for all of your results (Button et al., 2018).

Description of dataset: The datafile rmpCapstoneNum.csv contains 89893 records. Each of these records (rows) corresponds to information about one professor.

The columns represent the following information, in order:
1: Average Rating (the arithmetic mean of all individual quality ratings of this professor)
2: Average Difficulty (the arithmetic mean of all individual difficulty ratings of this professor)
3: Number of ratings (simply the total number of ratings these averages are based on)
4: Received a "pepper"? (Boolean - was this professor judged as "hot" by the students?)
5: The proportion of students that said they would take the class again
6: The number of ratings coming from online classes
7: Male gender (Boolean – 1: determined with high confidence that professor is male)
8: Female (Boolean – 1: determined with high confidence that professor is female)

There is a second datafile rmpCapstoneQual.csv that has the same number of 89893 records in the same order, but only 3 columns containing qualitative information:
1: Major/Field
2: University
3: US State (2 letter abbreviation)

With this dataset in hand, we would like you to answer the following questions:

1. Activists have asserted that there is a strong gender bias in student evaluations of professors, with male professors enjoying a boost in rating from this bias. While this has been celebrated by ideologues, skeptics have pointed out that this research is of technically poor quality, either due to a low sample size – as small as n = 1 (Mitchell & Martin, 2018), failure to control for confounders such as teaching experience (Centra & Gaubatz, 2000) or obvious p-hacking (MacNell et al., 2015). We would like you to answer the question whether there is evidence of a pro-male gender bias in this dataset.
Hint: A significance test is probably required.

2. Is there an effect of experience on the quality of teaching? You can operationalize quality with average rating and use the number of ratings as an imperfect – but available – proxy for experience. Again, a significance test is probably a good idea.

3. What is the relationship between average rating and average difficulty?

4. Do professors who teach a lot of classes in the online modality receive higher or lower ratings than those who don't? Hint: A significance test might be a good idea, but you need to think of a creative but suitable way to split the data.

5. What is the relationship between the average rating and the proportion of people who would take the class the professor teaches again?

6. Do professors who are "hot" receive higher ratings than those who are not? Again, a significance test is indicated.

7. Build a regression model predicting average rating from difficulty (only). Make sure to include the R^2 and RMSE of this model.

8. Build a regression model predicting average rating from all available factors. Make sure to include the R^2 and RMSE of this model. Comment on how this model compares to the "difficulty only" model and on individual betas. Hint: Make sure to address collinearity concerns.

9. Build a classification model that predicts whether a professor receives a "pepper" from average rating only. Make sure to include quality metrics such as AU(RO)C and also address class imbalances.

10. Build a classification model that predicts whether a professor receives a "pepper" from all available factors. Comment on how this model compares to the "average rating only" model. Make sure to include quality metrics such as AU(RO)C and also address class imbalances.

Extra credit: Tell us something interesting about this dataset that is not trivial and not already part of an answer (implied or explicitly) to these enumerated questions [Suggestion: Do something with the qualitative data, e.g. major, university or state by linking the two data files]

References:

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... & Johnson, V. E. (2018). Redefine statistical significance. Nature human behaviour, 2(1), 6-10.

Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching?. The journal of higher education, 71(1), 17-33.

MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a name: Exposing gender bias in student ratings of teaching. Innovative Higher Education, 40(4), 291-303.

Mitchell, K. M., & Martin, J. (2018). Gender bias in student evaluations. PS: Political Science & Politics, 51(3), 648-652.