Lab 1 - PECARN TBI Data, Stat 215A, Fall 2024

September 6, 2024

Submission

Push a folder called `lab1` to your `stat-215-a GitHub` repository by 23:59 on Friday, September 20nd. I will run a script that will pull from each of your GitHub repositories promptly at midnight so take care not to be late as late labs will not be accepted.

Follow the general lab instructions in stat-215-a-gsi/disc/week1/lab-instructions.pdf for more details. Please do not try to make your lab fit the requirements at the last minute!

I have provided a template (both `.tex` and `.ipynb`) as a guideline for the writeup. You will note that it essentially outlines an example of **PCS documentation** for the first few steps of the DSLC. Since the template is intended to make grading easier, please do not deviate from it significantly without good reason. Please restrict your writeup to twelve pages, including figures and bibliography. This is a strict limit: I will crop anything that appears beyond the twelfth page. In your `lab1` folder, please provide everything (except the data) that is needed for someone else to be able to compile your report and receive the exact same pdf. Your peers will be reviewing your code and attempting to recompile your report (they will manually add the data folder).

Special coding instructions for lab1: You must have a `clean.py` Python file in your `code` folder. This file should contain a function called `clean_data` that takes in the raw DataFrame (i.e. immediately after loading with Pandas) and possibly additional parameters, and returns the cleaned data as a DataFrame. I am requiring this for lab1 so that you have to get familiar with defining Python functions in `.py` files if you are not already. If you are used to working in Python notebooks, I'd recommend you start by defining the function in the notebook, then eventually move it over to `clean.py`.

Additional Remarks on Grading

Due to the importance of good communication, readability and grammar of your writeup will also be part of your grade. Moreover, we emphasize that the domain problem and context matters greatly in practice. While we do not expect you to be an expert on clinical decision rules a priori, we do expect you to learn a little about the area through reading Kuppermann et al. and to incorporate some bits of this domain information throughout your report. Ideally, in every (sub)section, you should try to ground your discussions of your findings/analyses in the domain context. A great report will also tell a story, where the writing flows from one section to the next and each plot has a reason for being included.

Here is an itemized overview of what you will be graded on for this lab:

- Readability and grammar.
- Readability of code (+ comments). This includes style (see the general lab instructions)
- Reproducibility of report. I should know how to run your code to reproduce your results (see the general lab instructions)
- Data cleaning (description and validity). Describe any problems/inconsistencies you see with the data, how you cleaned the data, and why you cleaned the data in that way.
- Three findings (creativeness, interestingness, and quality of figure). Add titles, axis and legend titles, choose appropriate color schemes, adjust sizes of figures.
- Figures that are not for the findings (relevance and quality).
- Overall quality and level of detail of report.

- Incorporate domain information (from the paper) and place your analysis in the domain context.

Academic honesty

Academic honesty statement

I ask you to draft a personal academic integrity pledge, addressed to Bin, that you will include with all of your assignments throughout the semester. This should be a short statement, in your own words, that the work in this report is your own and that all sources you used are properly cited, including your classmates. Please answer the following question: Why is academic research honesty necessary? If you feel it is not, make a clear argument why not.

Collaboration

You are welcome to discuss **ideas** with me or other students, but your report must be written up and completed individually. Do not share code or copy/paste any part of the writeup. If you discuss with other students, you must acknowledge these students in your lab report.

LLM usage

You are not allowed to use any sort of LLM (ChatGPT, GitHub Copilot, etc.) to help with completing this lab. You cannot ask any lab-related questions to an LLM, or use one to help code, write the report, get ideas, write emails to me, and so on. This policy will be loosened in future labs.

Background

In this lab, you will perform an extensive exploratory data analysis and preprocessing for a data set from clinical medicine. The Pediatric Emergency Care Applied Research Network (PECARN) performs research into acute injuries and illnesses among children in a wide range of demographics and institutions. The data set that we study in this lab is related detecting the risk of traumatic brain injuries (TBI) in patients younger than 18. TBI is a leading cause of death of in patients younger than 18, and often the patients with high risk must be identified urgently in order to prescribe the appropriate interventions. A computed tomography (CT) scan is the standard for detecting the presence of traumatic brain injuries. The benefit of a CT scan must be weighed against the risk that ionizing radiation from the scan will cause further health problems in the patient. Due to this trade off, it is of interest to develop algorithmic risk scores in order to screen patients based on their clinical characteristics and avoid unnecessary CT scans. In this lab, you will clean and perform exploratory data analysis on the PECARN TBI data. In your final project, you will develop and stress test clinical decision rules derived from this data.

The data for this lab is taken from Kuppermann et al., which can be found in the `lab1` folder of the `stat-215-a-gsi` GitHub repo. You should read this paper before doing the lab and understand the source of the data.

Data

The raw data can be found in the `lab1/data` folder of the `stat-215-a-gsi` GitHub repo. The files of interest are TBI PUD 10-08-2013.csv (the raw data containing patient level features) and TBI PUD Documentation 10-08-2013.xlsx (the documentation sheet describing each feature). The goal of this task is to simulate receiving data in a collaboration. Your first goal is to explore the data on your own. Try to understand how variables behave, and what their relationships are. This also involves carefully cleaning the data set. Do not take data consistency or correctness for granted. The following is a suggestion on how you might proceed.

Your first task will be to check the data quality and explicitly address the issues we discussed in class, such as the data collection method and data entry issues (e.g. missing values, errors in data, etc). Be sure to

discuss all inconsistencies, problems, and oddities that you find with the data.

Bearing the data quality in mind, your second task will be data cleaning. See `https://vdsbook.com/04-data_cleaning#sec-examine-clean` for guidelines on how to clean data. This data set is quite raw—it contains some outliers, inconsistencies, and lots of missing values. Note that many missing values are coded as a number in the data, so you cannot assume that only NA values in the data frame represent missingness.

These codes are listed for each variable in the TBI PUD Documentation 10-08-2013.xlsx file. You will need to make your own choices on how to most appropriately clean the data. Record in your report the steps you take to clean the data, and when necessary, explain why you cleaned the data in that way. Here, you may choose to bring in domain knowledge (from Kuppermann et al.) or common knowledge to support your data cleaning decisions.

Next, think of some questions you would like to ask of the data and use plots to answer them graphically. Try to show what interesting findings can be gained from the data. You may show general patterns or anecdotal events. Again record in your report your process—include plots you make. Don't be afraid to try methods that are new to you and be critical of your own graphics.

The following sections give guiding questions to think about while performing your data cleaning and EDA.

Domain problem to solve

How can we best vet and/or improve the clinical decision rule for traumatic brain injuries? Most importantly, the clinical decision rule should be highly predictive and minimize the amount of missed diagnoses (i.e. have a very high sensitivity). It should also be easy-to-use, using variables that clinicians can readily have access to when making their decisions. Finally, the interpretability of the rule helps to check whether its predictions will make sense for new patients and makes it easier to apply in new settings.

Data Collection

What are the most relevant data to collect to answer the domain problem?

Ideas from experimental design (a subfield of statistics) and active learning (a subfield of machine learning) are useful here. The above question is good to ask even if the data has already been collected because understanding the ideal data collection process might reveal shortcomings of the actual data collection process and shed light on analysis steps to follow.

The questions below are useful to ask: How were the data collected? At what locations? Over what time period? Who collected them? What instruments were used? Have the operators and instruments changed over the period? Try to imagine yourself at the data collection site physically.

Meaning

What does each variable mean in the data? What does it measure? Does it measure what it is supposed to measure? How could things go wrong? What statistical assumptions is one making by assuming things didnât go wrong? (Knowing the data collection process helps here.) Meaning of each variable—imagine being there at the ER and giving a Glasgow coma score, for example, and also a couple of variables. What could cause different values written down.

How were the data cleaned? By whom?

Relevance

Can the data collected answer the substantive question(s) in whole or in part? If not, what other data should one collect? The points made in (2.3) are pertinent here.

Comparability

Are the data units comparable or normalized so that they can be treated as if they were exchangeable? Or are apples and oranges being combined? Are the data units independent? Are any two columns of data duplicates of the same variable?

Presenting findings

Choose three of your interesting findings and produce a publication quality graphic for each along with a short caption of what each shows. This is where I expect to see very polished graphics—refer to the lab session materials on data visualization. Think carefully about use of color, labeling, shading, transparency, etc. This is your chance to do something innovative. If you are feeling bored or ambitious consider doing something dynamic or interactive (show a static version in the pdf) and provide either an additional .html document with the interactive graphic or a web link to where the interactive graphic is hosted.

Reality check

At this point, you've done some extensive data cleaning and made many judgment calls along the way. Using common sense, check your cleaned data against some external reality. Do your decisions during the data preprocessing align with the domain problem you are aiming to solve? This could be from your prior understanding of the world or you could cite an external source of information. Either way, be sure to clearly state your assumptions. Does your cleaned data pass the reality check or are there issues? Discuss.

Stability check

Next, select one of your judgment calls in your data cleaning or presentation of findings and perturb it somehow. By this, I mean to modify the judgment call in some way that seems reasonable to you. Clearly explain the call and your reasoning behind it, and explain the change you intend to make. Choose one of your findings from above. How does this change affect your finding? Create a before-and-after comparison visualization to show what – if anything – changes in the presentation of the finding and discuss.

Discussion

Did the data size restrict you in any way? Discuss some challenges that you faced as a result of the data size. Recall the three realms of data science: data/reality, algorithms/models, and future data/reality. Where do the different parts of this lab fit into those three realms? Do you think there is a one-to-one correspondence of the data and reality? What about reality and data visualization?