Group Project Guidelines

STATS 415

Fall 2023

This project allows you to apply the statistical learning techniques covered in class to analyze real datasets. The project has two parts. The first part is open-ended investigation: you will pose and answer two questions about the National Health and Nutrition Examiniation Survey (NHANES). The second part is a Kaggle competition: you will test your ability to build powerful predictive models. You will submit final reports detailing your analysis.

All together, the project comprises four assignments. It is worth 15% of your final semester-end grade. That 15% is divided up into 100 points.

- Write a one-page proposal for the open-ended portion of the project (10 points).
- Write a final report summarizing the findings of your open-ended exploration (50 points).
- Enter a submission to the Kaggle competition (10 points).
- Write a final report summarizing how your group tackled the Kaggle parts of the project (30 points).

There are also three key dates you'll want to keep in mind.

- Nov 12, 8:59pm: proposal for the open-ended portion is due. No more than one page, submitted by each group via Canvas.
- November 12, 9pm: the Kaggle competition begins.
- December 3rd, 8:59pm: final project reports are due and Kaggle competition ends.

Open-ended investigation

The open-ended investigation seeks to uncover Truth About Reality by using data from the National Health and Nutrition Examination Survey (NHANES). You can download any data from the CDC's website. Your primary source of the data should be the section called "Continuous NHANES." You can use the R package "haven" to read in the XPT files. There are many datasets on that website, and you can decide which you want to use. You can also use combinations of datasets. The open-ended part of the group project has two assignments: a proposal and a final report.

Proposal

By Nov 12 at 8:59pm, each team submits one copy of the proposal, via Canvas. It must meet the following requirements.

- Not exceed one page.
- Include team member's names.
- Describe which NHANES datasets you will use.
- Specify two questions you intend to explore.

- Specify the response variables you will focus on.
- Specify how you intend to use to investigate the two questions. Your approach must involve at least three of the following seven tools: (1) KNN, (2) LDA/QDA, (3) lasso/ridge penalties, (4) curvature penalties, (5) svm, (6) trees, (6) bootstrap, and (7) cross-validation. Note that you must use three tools overall, not three tools per question (e.g., you could use two tools for the first question and one tool for the second question).

The proposal counts for 10 points, and will be graded on two counts. First: does it meets the requirements above (5 points)? Second, do the questions/methods make sense (5 points)?

Report

By Dec 3rd at 8:59pm, each team submits a report on your open-ended exploration. Each team will also submit the code used to make the report. Unlike homework, the pdf for the report should NOT include R code or raw outputs from R. Any output should be included in the form of tables and figures. Models should be written as equations, e.g., $y = \beta_0 + \beta_1 x$. The final report does not need to exactly match the proposal you submit on Nov 12. The report should be less than 8 pages and include the following sections.

1. Introduction. What questions are you answering?
2. Data. What datasets are you using? This should include some visualizations of the data.
3. Methods. What analyses are you using to answer them?
4. Results. What are the results of the analyses? This should include some visualizations of the results.
5. Conclusions. How do the results of the analyses shed light on the questions? In what way might the analyses be misleading?
6. Contributions. What did each team member contribute to the project?
7. Reproducibility. If we wanted to run the same analysis as you, what should we do?

This report counts for 50 points, and it will be graded according to the following criteria.

- Are the questions good? Do they make sense for the data, can they be answered with the data at hand, are they something people might care about in real life? (5 points)
- Are the datasets clearly described? Are there useful visualizations to help orient the reader? (5 points)
- Does the approach involve at least three of the seven tools, used correctly? (10 points)
- Do you explain *why* the three tools are well-suited for answering your question? (5 points)
- Do you help the reader understand the results of the approach using tables or plots? (5 points)
- Do you give answers to the questions you have posed? Do you justify why your answers follow from the results of the analyses you performed? Do you mention at least one reason that your answers might be wrong (i.e., at least one way that your statistical analyses could be misleading)? (5 points)

- Is the report professional? Is it easy to read? Does it have all the required sections? Are the font-sizes legible? Is it written for the appropriate target audience (i.e., your GSIs)? You do not need to give detailed explanations of material from this class; the GSIs already know that. Is it less than 8 pages? Does it include all team member names? (10 points)
- Is the report reproducible? Does the appendix and attached code include everything necessary to reproduce the plots/tables in your report? (5 points)

Kaggle competition

This Kaggle competition has two parts: a training dataset (including many inputs and responses) and testing dataset (including inputs and responses). You will get to see all of the inputs in the training set, all of the responses in the training set, and all of the inputs in the testing dataset. You will not be able to see all of the responses for the testing dataset. A Kaggle "entry" is simply a list of guesses about the responses in the testing dataset. You will have an opportunity to submit up to five entries per day. Every time you submit an entry, the system will tell you how good your guesses were (measured in variance explained) on a small subset of the testing data. Your final performance, however, will be judged by the quality of your guesses on the entire testing dataset.

The Kaggle competition part of the group project has two assignments: a Kaggle competition entry and a report.

Entry

Teams submit an entry to the Kaggle competition by December 3rd at 8:59pm. This submission is worth 10 points. Your submission will attempt to make predictions about new held-out test points. You will be limited to five Kaggle submissions per day. More details will be posted on November 12th, but you can begin to familiarize yourself with Kaggle competitions now by visiting the website.

The Kaggle entry is graded by the test set R^2 (i.e., proportion of variance explained). Nine of the points will be awarded based on absolute R^2: you will get 5 out of 9 points if your R^2 is at least 0.5 and you will get 9 out of 9 points if your R^2 is at least 0.75. The remaining point is based on your ranking relative to your classmates: if you are in the 90th percentile, for example, you will get .9 points.

Report

Teams submit a report by December 3rd at 8:59pm. This report is worth 30 points. It should be a pdf file. No code is necessary. It must meet the following requirements.

- It should indicate the name of the account that was used to submit your Kaggle entry.
- It should not exceed 3 pages. Much less may suffice, depending upon the simplicity of your method. Simple can be good.
- It should include all team member names.
- It should describe the method you used.
- It should describe how you decided to use that method.
- It should describe what each team member contributed to the project.

- It should should include no R code at all and no reference to R packages. Pseudocode is permitted.
- It should abide by the following rules for describing your method.
  - If you wish to refer to estimators or methods covered by the textbook or the lecture materials, you should do so without explaining exactly how they work (of course you do have to explain exactly how you used them).
  - If you wish to refer to an algorithm that was developed by other people and not covered by the class, you must cite an article (ideally peer-reviewed) that describes the algorithm. You do not need to explain exactly how it works (of course you do have to explain exactly how you used it), but you must provide a few sentences explaining the general idea. Note: to obtain a perfect score on this report it is not necessary to use any methods going beyond the course material, but you may if you wish.
  - If you wish to refer to an algorithm that your group developed, you must explain the algorithm with sufficient detail that one of the GSIs could re-implement your method.

  Note that achieving good R^2 performance on this data challenge does not require any fancy new algorithms or techniques beyond what waas covered in this course. For example, my submission was made by applying a method taught in the course to a refeaturized version of the data.

It will be graded according the following criteria:

- Does it fulfill all of the requirements above? (7 points)
- Is it professional? Nothing hand-written. Should be organized into proper paragraphs and include section headings if there are more than five paragraphs total. No plots or formulas are strictly necessary for this report, but if they are included they should be legible. (8 points)
- Does it clearly describe the method you used? Is it clear enough that one of the GSIs could re-implement your method by studying your report and reading all of the reference you cite? (15 points)

Academic integrity

Along with the information about academic integrity that can be found in the syllabus, a few things worth mentioning about the group procjet in particular.

- In working together with the other members in your group, you should strive to find a way of working together that feels fair to all of you. If this is proving very difficult, and you think I might be able to help moderate, don't hesitate to reach out.
- Any cross-group conversations about the project must occur on the Kaggle discussion board or the canvas discussion board, so that all members of the class can get the info.
- Any cross-group conversations about the project should be limited to one of three topics.
  1. High-level discussion of ideas
  2. Nitty gritty details about getting the kaggle platform to work (e.g., how to construct the csv files in the right format, how to make a kaggle, etc)

3. Nitty gritty details about how to parse the NHAMES datasets, figure out the meanings of the variables, and things of that nature