STATISTICS 652: Statistical Learning and Prediction

November 18, 2020

PROJECT 3

Due Dates: Individual presentations delivered on Zoom at agreed-upon times between Wednesday, Dec 9–Friday, Dec 11, times to be determined.

ASSIGNMENT

As graduate students, likely have interests in areas besides Statistics. This project is an opportunity to practice your new skills on a regression or classification prediction problem that may be of interest to you. It may be your own research of something related to your research or from some other topic that you are interested in. You need to identify the goals of the research and how the data relate to the research (there should be a regression or classification prediction problem). Then you will do a thorough analysis, similar to first two projects, with the goal of providing a process by which you or the data's owners could perform the prediction task. Finally, you will provide some measure of prediction error to accompany the process. (Since there is no "right" answer here, I do not need to collect predicted values.)

For the analysis portion, I expect you to do a thorough job of model development and comparison, including some assessment of variable importance, You may use any techniques learned in class. You may also use techniques from outside of class for this project, provided you explain them in your talk and are prepared to answer questions about them. You should be able to explain them to me at the same level that I have been explaining things to you. Specifically, you may not use an R function for analysis if you don't understand what its analysis does. I would much rather you stick to class elements and really do a nice job than watch a 5-minute YouTube video or read a blog post and not really understand what you are doing. That's kind of a rule for life, too.

You are welcome to use domain knowledge! If you know something about the data already, explain this and how it can help you. But it is also OK to know very little about your data, or to know something about your data with no clue how it can help—this is pretty common! No need to mention it if you are not using it.

DELIVERABLES

You will prepare and deliver a presentation of **no more than 8 minutes** describing the problem, the data, the process, the results, and the measure of error. The talk will be delivered over Zoom to me privately. Explain your problem and solution clearly, and emphasize important or potentially useful findings. Provide evidence to support your decisions. You can include a few simple bits of `R` code (e.g. whatever produces the final results) and highlight the results briefly. You won't have time to do much more than that. You may/should reference, quote, and borrow from my class notes as much as you need to (cite the lecture number, if you do this!).

The presentation should clearly explain *in your own words*[1]

1. Why you chose the data
2. What the data are about
3. An outline of the analysis process
4. Description of all the relevant steps
5. Graphics/tables or whatever you need to explain what decisions you made
6. How you did your final predictions
    (a) I don't need to see a long list of predicted values unless there is something very interesting you want to talk about.
7. A measure of uncertainty on the predictions, and how it compares to
    (a) Confusion matrix and misclassification rate
    (b) Root mean squared prediction error and comparison to sample variance of Y
8. Any relevant summary or conclusions.

After the report, I may ask questions about anything that I didn't understand or that I might have approached differently. Questions will be focused on the report and what you did. For example, if you are doing classification, I won't ask random questions about regression or details that are not relevant to your data or analysis. And I am not planning to deliberately find questions designed to trip you up. However, I may ask questions about any of the techniques used or about why you chose *not* to use certain techniques. So be sure you know what you are doing and why.

[1]Please be aware of the definition of *plagiarism*, which includes quoting or minimally rewriting material without citation. Also, please look up and understand what "patchwriting" is. Don't do this. Practically speaking, I firmly believe that you don't truly understand something unless you can "put the book away" and explain it in your own words. If you have to recite something from the source, be sure to say that it comes from there. However, since I want you to show me what *you* understand, if you use too much text from an outside source, or simply rearrange words from another source, I will assume that you do so because you don't understand your subject well enough to explain it yourself. This will result in a very bad grade. *I will check!* **You are warned!**

GRADES

Your grade will be based on the quality and coherence of your presentation and partly on your ability to answer follow-up questions. My presentation rubric includes marks for

- clarity of presentation (can I reconstruct your analysis in my head based on what you tell me?),
- quality and thoroughness of analysis,
- quality of slides
  - use bullets in a readable font on screen
  - don't read paragraphs of text!
- and timing of the report (not too short or too long).

Also, a *small* portion of the marks may be associated with the "degree of difficulty" associated with the analysis. If you choose a data set with $p = 2$, you will not score as well as if you select something with a little more complexity to it. But don't choose something that is beyond your ability to handle. You won't be able to show me much you have learned if you choosing $n = 10^{10}$ or $p > n$ and discover too late that you are limited by time to using only linear regression. Feel free to run your data size by me if you are worried that it is too simple. You are left to judge for yourself if you have something too large. In particular, it may make sense to subsample a set that has millions many observations in it, just to make analysis feasible in R. You should explain if you do this.

POLICY

1. Since analyses will fall into two general categories, classification and regression, students will be doing a similar kinds of analyses on their projects. Therefore, you should treat this as being like a substitute for a final exam and not talk among yourselves about analysis ideas.
2. I will make myself available for questions by appointment, rather than re-purpose the QA sessions for individual meetings. (I still need them for the 452 students.) I can try to answer any questions you may have, but I will likely treat it like the other projects and try to get you to answer your own questions.

FINAL COMMENTS

I hope this is a useful experience for you. I hope that many of you can find a data set that you *want* to learn more about. After this is over and marks are returned, I would gladly hear comments about whether this was interesting and useful, or what would have made this a better experience for you.