Rubric for 432 Project A Portfolios and Presentations

Thomas E. Love

2022-03-07 15:26:49

Contents

What is the Purpose of this Document?

There are two purposes.

1. Provide the teaching assistants with detailed instructions on how to evaluate each part of the portfolio and what checks are needed for the presentation videos.
2. Provide the students in 432 with a clearer understanding of what we were looking for in the portfolios, in particular.

Initial Check of Canvas Submissions

At the start, **Dr. Love** will check quickly to see that each of the following elements of the proposal are submitted properly through Canvas, including

- the raw data (for projects that need to submit raw data),
- the cleaned .Rds,
- the complete .Rmd file
- the .HTML result knitted from the .Rmd
- the slides for the presentation
- the presentation video
- as well as a one-page note from the non-reporting partner if the proposal comes from a team.

Check of the Presentation Video

Dr. Love noted the length of the presentation video (it is not supposed to exceed 5 minutes), and clicked into it in a couple of random spots to ensure that the video is clear and the audio can be heard. If there's a problem, they will contact the student to resubmit.

Dr. Love will watch and evaluate each video in its entirety. The TAs are welcome to watch the videos and provide comments if they like, but my guess is that the portfolio work set out in the remainder of this document will be more than enough work already.

The 50 Things That the TAs (and Dr. Love) will Look for in the Portfolios

In addition to the TA review specified below, Dr. Love will also look for some of these things, and Dr. Love will also look for other things required in the Instructions but not listed below as part of his review of the portfolios.

The TAs will store their comments on each of these 50 elements for each project in a Google Sheet provided on the Shared Drive between the TAs and Dr. Love.

Overall Checks

As you walk through the submission, please check that:

1. (TEL checked) the HTML result uses `code_folding = show` and that this functionality is working
2. (TEL checked) the HTML result uses a dynamic or floating table of contents that works and includes numbered sections
3. (TEL checked) the R Markdown document does not use `echo = FALSE`, so that all code in the entire document can be seen or hidden at the whim of the reader, **and**
4. **TAs**: any graphs or tables are completely legible in the HTML, and not (for instance) outside the size of the page, **and**
5. (TEL checked) the student has used the template provided, or something equivalent that maintains nearly the same headings in the Table of Contents to facilitate finding the materials so you have no difficulties using the HTML file.

Section-by-Section Checks

The sections that must exist in the document and the things that should be checked in each section are listed below.

R Packages and Setup

6. (TEL checked) Please check that the tidyverse is loaded last.
7. (TEL checked) Please provide a count of the number of unique packages loaded in this section. No packages should be loaded here which are not used in the project.
8. (TEL checked) Please check that no packages are loaded more than once.
9. (TEL checked) Please check that no packages that are part of the core tidyverse (`dplyr`, `ggplot2`, `tidyr`, `readr`, `purrr`, `tibble`, `stringr` or `forcats`) are loaded on their own.
10. (TEL checked) Please check that the function `library()` does not appear after this section.

Data Source

11. **TAs**: Just check that this exists.

The Subjects

12. **TAs**: Just check that this exists.

Loading and Tidying the Data

13. **TAs**: Just check that this exists.

The Tidy Tibble

14. (TEL checked) Please check that this includes a subsection where the tidy tibble is saved using the same file name as is presented in the submitted .Rds file.
15. (TEL checked) Please ensure (from its printout in the HTML) that the tibble is in fact a tibble, and not just a data frame.

The Code Book

16. **TAs**: Please ensure that there is a table defining all of the variables included in the listing of the tidy tibble.
17. (TEL checked) Please ensure that there is also a numeric description of all variables listed in the code book using `describe()` from the `Hmisc` package.

Linear Regression Plans

18. **TAs**: Please ensure they have clearly stated a research question here that applies to Models A and B below, and be sure that question is actually stated as a question, ending in a question mark.
19. **TAs**: Please ensure that the outcome and predictors specified here are part of the tidy tibble and codebook, and also match exactly those used in their Model A later.

Logistic Regression Plans

20. **TAs**: Please ensure they have clearly stated a research question here that applies to Models Y and Z below, and be sure that question is actually stated as a question, ending in a question mark.
21. **TAs**: Please ensure that the outcome and predictors specified here are part of the tidy tibbble and codebook, and also match exactly those used in their Model Y later.

Linear Regression Modeling (All checked by TAs)

22. Please ensure that they begin with an appropriate discussion of missingness, including how much they have and how they have decided to deal with it for Models A and B.
23. Please ensure that both a Model A and a Model B are presented,
24. If they have transformed their outcome, we want to note what transformation they used: we want them to restrict themselves to the square, square root, logarithm (any kind) or the inverse, so if they used something else, we want to note that. If their outcome includes values that are not positive, then they would likely have added something to each one before applying a transformation, in which case we want to note that.
25. Note the number of predictors in their model A. Please ensure that Model A contains at least four predictors, and that at least one is quantitative and at least one is categorical (it's supposed to be multi-categorical but if they had to collapse the categories they may have wound up with only binary predictors.)
26. Note the number of non-linear terms added to model A in their model B. Please ensure that Model B adds at least one and no more than three non-linear terms to Model A.
27. Please ensure that the selection of Model B non-linear terms matches the logic they provide (usually a Spearman rho-squared plot.)

28. Please ensure that if they fit both an interaction term and a spline to the same quantitative variable, they use `%ia%` rather than `*` to fit the interaction.
29. Please ensure that any splines or polynomial terms in Model B are applied only to quantitative predictors.
30. Please ensure that they don't have any explosions in coefficient sizes or standard errors in either Model A or Model B by looking at the tidied coefficients, or the `plot(summary())` results, or a nomogram.
31. Please ensure that they've incorporated a validation assessment of prediction quality (which should include R^2 at a minimum, but can also include other things like root mean squared error or mean absolute error) to make a clear decision between Model A and Model B as their final model.
32. Please ensure that they've provided and interpreted a residuals vs. fitted plot (at least - they are welcome to show all four plots, but only residuals vs. fitted is mandatory) for the final model they choose (Model A or Model B).
33. Please ensure they've provided a nomogram for the final model they choose (Model A or Model B) that looks nice in the HTML output. If they used a transformation for the outcome variable, then the nomogram should include the back-transformation.

Logistic Regression Modeling (All checked by TAs)

34. Please ensure that they begin with an appropriate discussion of missingness, including how much they have and how they have decided to deal with it for Models Y and Z.
35. Please ensure that both a Model Y and a Model Z are presented here.
36. Note the number of predictors in their model Y. Please ensure that Model Y contains at least four predictors, and that at least one is quantitative and at least one is categorical (it's supposed to be multi-categorical but if they had to collapse the categories they may have wound up with only binary predictors.)
37. Note the number of non-linear terms added to model Y in their model Z. Please ensure that Model Z adds at least one and no more than three non-linear terms to Model Y.
38. Please ensure that the selection of Model Z non-linear terms matches the logic they provide (usually a Spearman rho-squared plot.)
39. Please ensure that if they fit both an interaction term and a spline to the same quantitative predictor, they use `%ia%` rather than `*` to fit the interaction.
40. Please ensure that any splines or polynomial terms are applied only to quantitative predictors.
41. Please ensure that they have clearly and consistently specified the direction of the outcome variable (whether they are using a 1/0 variable for the outcome or Yes/No or something else) and that they've done so correctly, so that if they say they are predicting Pr(outcome = Yes) they really are. One way to assess this is to look at whether area under the ROC curve that is plotted for their final selected model matches the value of the C statistic in sample for that same model.
42. Please ensure that they don't have any explosions in coefficient sizes or standard errors in either model by looking at the tidied coefficients, or the `plot(summary())` results or the nomogram.
43. Please ensure that they've incorporated a validation assessment of prediction quality (likely the C statistic but Nagelkerke R^2 is also a good option) to help make a clear decision between Model Y and Model Z as their final model.
44. Please ensure they've provided a nomogram for the final model they choose (Model Y or Model Z) that looks nice in the HTML output. The nomogram should show the appropriate probability for their outcome in addition to the linear prediction.

Discussion (All checked by TAs)

45. Please ensure that they have clearly restated their research questions in this section and that they mirror what they presented earlier in the linear and logistic regression plans.
46. Please ensure that they have answered each of their research questions in this section using results from their selected models.

47. Please ensure that they have also provided a discussion of the Project 1 process that addresses at least two of the following four questions, and noting which of these (a, b, c, d) they address.
    a. What was substantially harder or easier than you expected, and why?
    b. What do you wish you'd known at the start of this process that you know now, and why?
    c. What was the most confusing part of doing the project, and how did you get past it?
    d. What was the most useful thing you learned while doing the project, and why?

Affirmation

48. (TEL checked) Just check that this exists and that it says what it needs to say.

References

49. (TEL checked) Just check that this exists. They should provide, at minimum, a reference for the data set.

Session Information

50. (TEL checked) Just check that this exists and note the version of R they used - ideally 4.1.2, and whether they used a Windows, Macintosh or a Linux setup.

Thank you!