Stat 4630 Project: Final Report Guidelines

due Wednesday Dec 7

Each group will submit a final report based on the work done on this project for the semester.

- In Section 1, your group will provide a non-technical overview of the work done.
- Sections 2, 3, and 4 should mostly be taken from work done in previous Milestones, with edits made based on my comments from Milestones 3 and 4.
- Section 5 consists of reflection.

The final report should address the items listed in Sections 1 to 5:

1 Executive Summary

In this section, address the following:

a. Both questions of interest your group is seeking to answer.

b. Motivation for both questions of interest.

c. How does the analysis your group carried out in previous milestones answer these two questions of interest?

This section should be written with **no technical jargon**, and can be read and understood by any person with a college degree but has not taken any statistics classes. Think of how newspaper articles summarize results from a statistical study. As an example, look at this article from the New York Times (paragraphs 10 and 11) (`https://www.nytimes.com/2018/08/13/well/an-underappreciated-key-to-college-success-sleep.html`). This section should not be more than 2 pages.

2 Data Description

Provide the following information of your data set:

a. Who or what are the data about.

b. The source of the data set.

c. Descriptions of the variables the reader will encounter later in the report. Clearly state which variables were the response variables. If you had to create your own variables, please clearly state that these variables were created and not part of the original data set, and describe how these were created.

3 Regression Question

In this section, provide the following:

3.1 Exploratory Data analysis

Present graphical summaries that you thought were interesting and comment on how these summaries address your **regression** question. You do not have to present every graphical summary explored. Be sure to provide relevant R output to refer to, and not just state the results without any reference to output.

3.2 Shrinkage Methods

In this subsection, provide the following:

a. A table that compares the test MSEs from ridge, lasso, and OLS regression.

3.3 Regression Trees

Present a tree built by recursive binary splitting or pruning. Include the following:

a. Reason(s) why you chose to present the tree built with recursive binary splitting or pruning.

b. The output from the summary() function on the tree.

c. State how many terminal nodes your tree has.

d. The graphical output of the regression tree.

e. Output from the `importance()` and/or the `varImpPlot()` functions with random forests.

f. A table that compares the test MSEs from recursive binary splitting or pruning, and random forests.

3.4 Summary of Findings

Summarize your findings from subsections 3.1 to 3.3. Include the following:

a. A comparison of the test MSE with linear regression, ridge regression, lasso regression, regression tree (built with recursive binary splitting or pruning), and random forests, and comment on these values.

b. A discussion on how the findings from subsections 3.1 to 3.3 answer your group's question of interest. Comment on similarities and / or differences from the various methods.

c. Commentary on which method was best in terms of answering your question of interest, and implications. If various methods were better at addressing different aspects of your question of interest, provide a discussion about this.

4 Classification Question

In this section, provide the following:

4.1 Exploratory Data analysis

Present graphical summaries that you thought were interesting and comment on how these summaries address your **classification** question. You do not have to present every graphical summary explored. Be sure to provide relevant R output to refer to, and not just state the results without any reference to output.

4.2 Logistic Regression Model or Linear Discriminant Analysis

In this subsection, present the logistic regression model that your group is most satisfied with. If the assumptions for carrying out linear discriminant analysis (LDA) were met, you may instead present the LDA model your group used. Include the following:

a. For logistic regression: the `summary()` output from the `glm()` function;

   OR (if LDA assumptions were met): the output from the `lda()` function.

b. A confusion matrix for the test data and report the error rate (based on threshold of 0.5). If necessary, also report the false positive and false negative rate, and discuss if the threshold should be adjusted (and if so, provide a confusion matrix, FPR and FNR, with the new threshold).

4.3 Classification Trees

Present a tree built by recursive binary splitting or pruning. Include the following:

a. Reason(s) why you chose to present the tree built with recursive binary splitting or pruning.

b. The output from the summary() function on the tree.

c. State how many terminal nodes your tree has.

d. The graphical output of the regression tree.

e. Output from the `importance()` and/or the `varImpPlot()` functions with random forests.

f. A confusion matrix for the test data and report the error rate (based on threshold of 0.5) for recursive binary splitting or pruning, and for random forests. If necessary, also report the false positive and false negative rate, and discuss if the threshold should be adjusted (and if so, provide a confusion matrix, FPR and FNR, with the new threshold).

4.4 Summary of Findings

Summarize your findings from subsections 4.1 to 4.3. Include the following:

a. A comparison of the test error rates with logistic regression (or LDA), classification tree (recursive binary splitting or pruning), and random forests.

b. If necessary, compare the false positive rates and false negative rates from logistic regression (or LDA), classification tree (recursive binary splitting or pruning), and random forests.

c. Commentary on whether the thresholds should be adjusted (and if so, compare the error rates, FPRs, and FNRs, with the new threshold).

d. A discussion on how the findings from subsections 4.1 to 4.3 answer your group's question of interest. Comment on similarities and / or differences from the various methods.

e. Commentary on whether your logistic regression model (or LDA), classification tree, or random forest was better in terms of answering your question of interest, and implications. If various methods were better at addressing different aspects of your question of interest, provide a discussion about this.

5 Challenges and Further Work

Address any challenges your group faced, including work done in data wrangling. Also, if your group had more time to work on this project, what else would you have considered doing?

6 Grading Guidelines

Your report will be graded A, B, C, D, or F and then converted to a 0-100 scale.

- A (90 to 100): the elements listed in Sections 1 to 5, as well as my comments from Milestones 3 and 4, are fully addressed and addressed well.

- B (80 to 89): a few elements listed in Sections 1 to 5, as well as my comments from Milestones 3 and 4, are missing or a few are not addressed well.

- C (70 to 79): some elements listed in Sections 1 to 5, as well as my comments from Milestones 3 and 4, are missing or some are not addressed well.

- D (60 to 69): a lot of elements listed in Sections 1 to 5, as well as my comments from Milestones 3 and 4, are missing or a lot are not addressed well.

- F (below 60): elements listed in in Sections 1 to 5, as well as my comments from Milestones 3 and 4, are generally missing or not addressed well.

7 Additional Grading Guidelines

Your report should adhere to the following elements. Not following these will result in deduction of points (up to 5 points for each missing element).

- One member of the group will upload the report (.pdf or .html file), and attach the dataset as a .csv file. If the dataset comes from an R package, be sure to specify which package it is from in the report (no need to attach dataset).

- Include the names of the group members and group number in the heading of your report.

- Have sections that are clearly labeled.

- Aim for no more than 30 pages. If you go over this limit a bit, that is fine.

- Do not use appendices as a way to work around the page limit. Anything that belongs in the main body of the report should be in the main body and not be tucked away in an appendix. I will not read anything in the appendix.

- Use proper paragraphing, sentences, grammar, spelling, etc. Avoid using long paragraphs. Using tables or bullet points can be helpful in summarizing key pieces of information.

- The text in your document should be readable after printing out on letter-sized paper.

- Include relevant R output. Output is relevant if you use the information from the output and and is referred to in your report. I should know what output to look at when you are referring to it in the report.

- Your report does not to include the R code.

- Be written for the appropriate audience: general audience for Section 1, and your classmates in STAT 4630 for the other sections.

8 Submission

One member of your group will upload your group's report via Assignments in Collab by Wednesday, Dec 7. A 10 point penalty will be assessed for every 12 hour period that is past this deadline.

9 Group Evaluation (10 points)

A group evaluation is due with this milestone. Please complete the Project Group Evaluation Questions (Final Report and Presentation) via Test & Quizzes, by Dec 8. Please note that your comments will be forwarded to the relevant group members (with your name removed). You will be evaluating yourself and your group members on the following:

- being on time for group meetings
- participation in discussions
- listening to others
- completing their share of the tasks

Your score will be based on the consistency and average reviews you receive from your group members. Not turning in this group evaluation will result in a 5 point deduction.