Stat 6021: Project 2

Overview

As you have progressed through the course, this may a good time to review the learning outcomes for STAT 6021 as laid out in the syllabus. Upon successful completion of this course, you will be able to:

1. Statistical Methods

   (a) Use regression analysis to answer questions of interest in a wide variety of application environments.

   (b) Determine the most appropriate regression model for a given data set.

   (c) Identify the assumptions and limitations of a given regression model.

   (d) Diagnose and remedy common problems with the regression model found in real data.

2. Computing

   (a) Work with various data structures and primitive data types.

   (b) Process R dataframes into the forms necessary for subsequent analysis including subsetting by rows, columns, condition, changing column names, removing missing values, combining dataframes with vectors.

   (c) Use the appropriate numerical and graphical summaries based on the question of interest and type of data.

   (d) Use the statistical software R for regression analysis.

3. Communication and Professionalism

   (a) State appropriate context-specific conclusions from an analysis.

   (b) Present and discuss orally and in writing, statistical ideas, methods, and results to lay and professional audiences.

   (c) Work in teams to demonstrate the skills of a professional statistician in organizing and managing projects.

4. Statistical Theory and Mathematics

(a) Describe the mathematical framework of regression models.

(b) Describe the importance of assessing the assumptions and limitations for a given regression model.

Project 2 will incorporate all of these skills to answer a data analysis question. You will be working in assigned groups of 3 to 4 students. A lot of work in data science is completed in a group setting, so this project will provide some insight into how to create a successful group project. Also, you have seen that there may be more than one way to approach a data analysis question, so feel free to explore various ideas brought up by each group member and learn from each other!

This project will take place in the last few weeks of the semester with various components due at different dates. The list of deliverables is shown below:

- Part 1: Group Expectations Agreement (due April 11)
- Part 2: Proposal (due April 11)
- Part 3: Typed Report (due May 9)
- Part 4: Video Presentation (due May 9)
- Part 5: Feedback on Classmates' Presentation (due May 10)
- Part 6: Self- and Peer-Evaluation on Project (due May 10)

The subject of the project will not be assigned; instead it will be chosen collaboratively by the members of each group. Each group will complete a written report and oral presentation. Each project should feature the following:

- Clear central analytic goal(s) and/or question(s) to answer – the more interesting, the better.
- Identifiable data that can be obtained in support of the project.
- Appropriate use of various methods learned in this course.
- Clear communication of results to a variety of audiences.

More information on how to complete each part and how each part will be evaluated is provided in the sections below.

Part 1: Group Expectations Agreement

As mentioned, this project will be completed in randomly assigned groups of 3-4 students. A lot of work in data science is completed in groups, so knowing how to successfully complete group work will be beneficial! You have also seen that there may be more than one way to approach a data analysis question, so feel free to explore various ideas brought up by each group member and learn from each other!

For this part of your project, your group will submit a Group Expectations Agreement. Samples are provided on Collab as suggestions. Your group will create a document, put your names, and list the expectations you agree as a group to adopt for this project. Given the online nature of this course, you group should address at least the following aspects:

- Mode of communication between group members.
- How to schedule meetings.
- How to conduct meetings.
- How to handle disagreements.

The expectations are for your use and benefit; they will not be graded or commented on unless you specifically ask for comments. Note that the list should be fairly thorough without being unrealistic. For example, "We will solve every problem completely" or "We will get a perfect score for the project" or "We will never miss a meeting" are unrealistic, but "We will be sure to understand our tasks prior to our next meeting" and "We will try to be as flexible as possible in scheduling meetings" are realistic.

Group work is not always easy; group members sometimes cannot prepare for or attend all meetings due to other responsibilities, and conflicts may arise from different skill levels and work ethics. When groups work and communicate well, the benefits more than compensate for the difficulties. One way to improve the chances a group works well is to agree beforehand on what everyone on the group expects from everyone else. Reaching this understanding is the goal of the Group Expectations Agreement.

Submission

- Please submit your Group Expectations Agreement (.pdf file) via Assignments.
- Every student will submit a document.
- Documents within a group should be identical.
- Submission indicates agreement with the document.

Failure to submit a Group Expectations Agreement will result in getting a 0 for Project 2.

Part 2: Proposal

Each group will submit a project proposal for approval by the instructor, to ensure that an appropriate data set is being used and an appropriate amount of analysis will be done to produce a successful 4 week project. Each proposal should provide:

- The associated data set for the project should be provided, in one of the following ways: 1) providing the names of the R package and the R dataframe; 2) a link to the data set; 3) a file containing the data, as an Excel spreadsheet, a .csv file, or a .txt file. Please note that a list of some publicly available sources of data sets is provided on Collab.

- Project objectives/goals. What questions is the group trying to answer, as well as potential practical implications (the more interesting and/or practical, the better) of the results. Your project should involve both linear regression and logistic regression, so clearly state the response variables involved.

- Some data visualizations and commentary related to the project objectives/goals. At least one visualization related to linear regression, and at least one visualization related to logistic regression.

The proposal should be no longer than 6 pages. The proposal will be graded on a pass / fail basis. The proposal will be evaluated on the following:

- The appropriateness of the data set.
  - The instructor should be able to access the data set easily.
  - The data set should be formatted in a manner where each row represents an observation, and each column represents a variable.
  - Please note that regression methods assume the observations are independent. One way of assessing whether your observations are independent is to ask if the order of the rows in your dataframe matters. If you can scramble the order of the rows without affecting any structure in your data, then your observations are likely to be independent. If scrambling the rows upends the structure of the dataframe, then your observations are not independent, and regression methods are not meant to handle dependent data.
  - For the categorical response variable, be sure it is binary. We have not covered enough material to tackle categorical response variables with more than 2 classes.

- The objectives are appropriate for a project that spans 4 weeks. The instructor will assess if you have at least one question involving linear regression and at least one question involving logistic regression.

Feedback will be provided for group proposals that are rejected. Groups will have the option of submitting one revised proposal (but given the relatively short time you have to complete the project, this should be avoided).

Submission

Please submit your group's proposal (.pdf file) via Assignments (1 upload per group).

Part 3: Report (100 points)

Your group is to type up a report for this project. One member of your group is to upload the report and the R script via Assignments on Collab. The report is to include the following:

1. An executive summary that describes the high-level results of the analysis. This executive summary should be written in a way that can be understood by a wide variety of readers, including readers with no background in statistics. A way to think about this is how newspaper articles report results from various studies, so avoid technical jargon. This section should be no more than 2 pages.

2. A detailed description for a professional audience. The audience for this section is another classmate your client may hire to review your report. This section should include:

    - Detailed description of the data including challenges, data cleaning before, during, or after analysis
    - Exploratory data analysis and what you learned
    - Clear reasons given for model(s) considered
    - Appropriate model diagnostics provided and checked
    - Attempts to improve the model, as well as reasons for decisions made on how to improve the model.
    - Relevant R output and graphical summaries

3. Discussion of analysis should always be done contextually.

Report Guidelines

- 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.
- The report should contain correct grammar, clear explanations, and professional presentation.
- I should be able to repeat your analysis without looking at your R code.
- Your report does not need to include any R code. Relevant output from R (e.g. graphs, results from hypothesis tests, etc) should be included if the output is referenced to in the report.
- The text in your document should be readable after printing out. Difficult to read documents will be penalized.

Submission

Please submit your group's report (.pdf file) and a separate R script of your code via Assignments (1 upload per group).

Part 4: Presentation (100 points)

Each group will give a presentation that is no more than 15 minutes long. This presentation should be designed to be understandable by anyone familiar with the course material, but who has not read the project report. Each group is free to organize who talks about which topics during the presentation. Not everyone needs to talk, but all group members should be active contributors to the presentation materials. The presentation is to include the following:

- What are the goals/questions of this project?
- What are the results of the analysis? (include any recommendations)
- What is the nature of the data used?
- Does your project answer interesting and/or important questions? Explain.
- Use of graphical summaries to explore the data.
- A description of the model building process. This part will be much more technical. The previously listed items should be aimed for a non technical audience.

Presentation Guidelines

- The presentation should use PowerPoint or something similar as a visual.
- Each slide should be clear and easy to read.
- The presenter should be clear, with good pace and logical flow.
- R output should be clearly labeled.
- The recording should be hosted on zoom (preferred). Other popular video hosting sites such as youtube and vimeo are acceptable.

Submission

Please submit your group's presentation (link to where the recording is hosted) via Assignments (1 upload per group).

Part 5: Feedback on Classmates' Presentation (20 points)

After you have completed your project, you will need to review another presentation in Collab:

- View the presentation you have been assigned to review.
- Provide anonymous, constructive, and concise feedback.
- Your feedback will be evaluated on the helpfulness and conciseness of the feedback you provide.
- Half the points will be awarded for turning the feedback in on time.

Submission

Each student will submit the feedback via Assignments.

Part 6: Self- and Peer-Evaluation of Group Participation in Project 2 (20 points)

- You will anonymously evaluate each group member's contributions to the project.
- Complete this by giving an honest of your own performance and that of your group members in project 2.
- Half the points will be awarded for turning the evaluation in on time.

Submission

Each student will submit the group participation evaluation via Test & Quizzes.