Final Project Tips and Suggestions

Stats 415

March 18, 2022

Project Overview

Project Guidelines and Timetable

The project guidelines can be found on Canvas

A reminder on due dates

- ~~Form groups: Feb. 25, 2022~~
- Proposal: **Mar. 20, 2022, 8:59 pm** (this Sunday)
- Final project reports: Apr. 17, 2022, 8:59 pm

Project Outline

Your report should have roughly six sections (the order can vary)

1. Introduction
2. Open-ended problem 1 (regression)
3. Open-ended problem 2 (classification)
4. Kaggle write-up
5. Conclusions
6. Contributions of each team member (one paragraph)

The report should be **no longer than 10 pages** including figures and tables (but not counting any appendix/supplementary material)

Reproducibility

Regardless of how you write up your report, all code used for the project should be submitted to Canvas (.R or .Rmd files) with a short appendix (not counted towards the 10 pages) detailing how we should run your code to reproduce the analysis

Proposal Requirements

Each team submits one copy of their project proposal via Canvas.

- The proposal must include the team members' names, the questions you intend to answer with the data, and the methods you intend to use to answer the questions.
- Include a rough timeline for project completion and specify the proposed responsibilities of each team member.
- Join the Kaggle competition, make one Kaggle submission (you can submit random numbers), and report the name of the account (as it appears on the leaderboard) who will be submitting entries for the competition.

There is a strict **1 page maximum** for the proposal.

Proposal grade will mostly reflect completion/following all instructions.

Kaggle Competition

Loading the data

- To load the data, use the read_xpt function from the haven library
- The best way to create the training/testing sets is to
  1. load in all the data sets from the same year
  2. left join those on SEQN
  3. stack each year's large data frame using bind_rows (creating the "massive data frame")
- Finally, left join the massive data frame onto train.csv which should give you 8921 rows and 149 columns (after eliminating any column with NA values)
- You can eliminate the columns specified in the appendix of the final project guidelines

Understanding the data

- Outcomes (y) are synthetic - they were generated from a subset of size $k \ll p$ variables.
- y is a nonlinear function of the k variables with interactions

Strategies for model selection

Start early (you are limited to 5 Kaggle submissions per day)

Ideas:

- Use CV to tune on training set
- Use Lasso to select variables to use in other, more flexible, models
- Use principal components as input to other regression models (including non-linear and tree-based models)

Submitting to Kaggle

- Use the `write_csv()` function to save your predictions to a .csv file
- Your submission file should have two columns: `SEQN` and `y`
- Make sure your kaggle team name/username is clearly marked on the proposal and on the final report

Scoring submissions

- Your submission will be scored on the squared correlation (r-squared) between your predictions and the true value of y in the test set.
- There is a public leaderboard (computed on 30% of the test data) which you can use to estimate your ranking.
- Your final class ranking will be determined by the private leaderboard (computed on the remaining 70% of the test data)
- The public leaderboard can be viewed here.

Write-up

You should use *approximately* 1 page (not a strict limit) to talk about your Kaggle submission

Things to discuss:

- Models you tried
- How you did hyperparameter tuning (if any)
- How you did variable selection (if any)

Open Ended

General strategies

- Ask questions that have scientific interest
- You don't have to only use the data sets that were a part of the Kaggle competition (or the same years)
- Focus on inference, rather than prediction
- You can search online for papers that use the NHANES data to get ideas of questions to ask
- Questions for regression and classification should have different response variables

Exploratory data analysis

- Should be meaningful, not just a box-checking exercise (don't only show paired scatterplots/basic graphs with no comments)
- Show trends that you will later elaborate on with your models
- Put effort into making your graphics visually appealing

How to write/organize your thoughts

- Make sure the question you are asking is *clear, concise, and well-motivated*
- Give an answer to the question (or state why you do not have enough evidence to properly answer the question) and explain how the data support your conclusions.
- Focus more on the conclusions than on describing the statistical methods in detail (e.g., you do not have to explain how LDA works)

How to talk about statistical methods

- Motivate the methods you are using (e.g., I am using ridge regression because $n \ll p$)
- Talk about how the method works to the extent that it relates to the problem you are solving (don't explain how KNN works just to fill space)
- Highlight strengths and limitations of the method

Presenting data

DO NOT INCLUDE CODE OR RAW R OUTPUT

- Use tables and graphs (no raw output - R packages `stargazer` and `xtable` are useful if you're writing the report in LaTeX)
- Make figures visually appealing (use consistent color scheme, should be easy to read, etc.)

Collaboration

Strategies for effective collaboration

- Start early
- Create a schedule
- Organize your code in a github repository (or, at the very least, in a shared Google Drive/Dropbox folder)

Writing the report

You can use RMarkdown, Microsoft Word/Google docs, Overleaf, or any other word processing software to write your report

Even though different people will write different parts of the report, the whole report should read as one continuous piece of writing

Grading

Though you will detail the responsibilities of each group member at the end of the report, all members should proofread/endorse all sections of the project

All group members will receive the same grade for the project (unless there are extenuating circumstances)

Sample Report

Example of applied statistics report

A sample applied statistics report can be found here (relevant for open-ended portion).

Similarities:

- Exploratory data analysis section focuses on trends that are later supported by model output
- All data/output is formatted in tables or graphs

Caveats:

- You do not need to cite references in your report
- You should not use the same question/data as this report
- This report is only focused on a regression problem, you'll also need to have a classification problem
- The report covered material from different classes
- This was from a timed exam