Data200 Final Project

Due April 8th

Description

Purpose: Apply the concepts learned in the second half of this course to an interesting data set. Choose your own path and demonstrate your mastery of the learning objectives!

You're encouraged to use the same data as the midterm project, but you're welcome to find a new one. The same collection of dataset sources is described at the end of this document. Make sure the data you choose has plenty of modelling opportunites! The data cannot be data that has already been used in this course. The data must be available publicly.

Your submission should be a single pdf (exported from a Jupyter notebook) uploaded to GradeScope. Groups must be two to four people (inclusive). Absolutely no groups larger than 4, and groups of 1 are similarly discouraged. Groups may be the same as the midterm project, or you may wish to change it up.

The next page includes a rubric for how the project will be graded. The list below is a much more concise description of what I want.

- A clear **research question** that can be answered by the data available to you.
- An **exploratory data analysis**, including **cleaning** and **visualization** that are relevant to the research question. Make sure everything is relevant to the goal of the project! No extraneous plots!
- A **regression task** that addresses your research question.
- A **classification task** that addresses your research question.
  - Do **not** just bin a continuous variable into categories!
- Proper use of **bootstrapping** to estimate the sampling distribution of a parameter of interest to address an **inference** task.
- Proper use of **cross validation** to choose hyperparameters to ensure out-of-sample predictive accuracy in a **prediction** task.

  - Note: You may have up to 4 different tasks: 2 if you combine, say, the regression and inference and also combine the classification and prediction, 4 if you do regression, classification, inference (possibly just estimating a mean or means of different groups), and prediction (which may be slightly different from your stated research question) as separate tasks.

Rubric

- Cohesion

  1. Submission consists of disconnected applications of course material.
  2. It is clear why each aspect of the submission is present, but there is no clear direction.
  3. The submission has a clear overarching idea.
  4. All aspects of the submission work together to answer a single, clear research question.

- Content Descriptions

  1. Notebook contains a lot of irrelevant information or is clearly missing information.
  2. Notebook would benefit from editing or editorializing, such as removing/combining sections or adding new sections.
  3. Notebook is a reasonable length; all information is present and nothing extra is present.
  4. Notebook contains exactly the information needed, presented concisely and completely.

- Readability

  1. Most of the information is contained within the code cells and their output.
  2. It is required to read the code cells and output in order to understand what was done, or there is a mismatch between the claimed procedure and the one in the code.
  3. Document can be read without the code cells and output, but some important information about how the model was fit is constrained to the cells.
  4. The document can be read and fully understood - including understanding of all models used - without reading the code cells. (Do not hide code cells - we need to be able to verify this.)

- Code Cells

  1. Notebook is either mostly code or mostly text, including irrelevant output.
     - "Irrelevant output" refers to any outputs from code that is never mentioned in the text. Adding superfluous references to the output does not make it relevant.

2. Notebook includes reasonable amounts of code and text, but includes irrelevant code/output.
3. Notebook is well structured, with code used in appropriate places.
4. Notebook is well structured, with clean, readable code and clear, concise interpretations of the code.

- Code Output
  1. Submission includes at least one output that takes up more than half of a page.
     - You get a 1 in this category if this happens even once. I'm being more strict about this!
     - This strictness does not apply to plots.
  2. Output includes information that is not relevant to the research question.
     - Again, I'm being strict about this. Don't just print out `head(df)`, if showing the data is relevant then choose which columns are relevant and only show those!
  3. Output is all relevant and concise.
  4. Only relevant code is shown, it is formatted nicely, and it does not disrupt the flow of the document.

- Regression Usefulness
  1. Regression not used.
  2. Only a basic linear model is used.
  3. The regression model and data used makes sense in the context of the research question. Multiple regression models were compared.
  4. Regression is used correctly and the results clearly inform the research question. Multiple regression models were compared.

- Regression Application
  1. Regression wasn't used.
  2. Regression was used, but some parts are incorrect.
  3. Regression was used correctly and diagnostic checks were present.
     - Diagnostic checks: residual plots, summary statistics.
  4. Regression was used in a clever or intelligent way. Diagnostic checks were done correctly.

- Classification Usefulness
  1. Classification not used.
  2. Only a basic logistic regression model is used.
  3. The classification model and data used makes sense in the context of the research question. Multiple classification models were compared.
  4. Classification is used correctly and the results clearly inform the research question. Multiple classification models were compared.

- Classification Application
    1. Classification wasn't used.
    2. Classification was used, but some parts are incorrect.
    3. Classification was used correctly and diagnostic checks were present.
        - Diagnostic checks: confusion matrices
    4. Classification was used in a clever or intelligent way. Diagnostic checks were done correctly.
- Inference
    1. Inference not discussed.
    2. Inference task was tacked on, unrelated to the research question.
    3. Inference task is relevant to the research question.
    4. The results of the analysis clearly say something meaningful about the population of interest. Population of interest is clearly defined.
- Prediction
    1. Prediction not discussed.
    2. Prediction task was tacked on, unrelated to the research question.
    3. Prediction task is relevant to the research question.
    4. The predictions from the final model could be used to make important decisions.
- Conclusion
    1. There is no conclusion section; or the conclusion section does not actually contain a conclusion.
    2. Conclusions either do not adress the research question, or are not supported by the models fit.
    3. Conclusion adequately addresses the research question.
    4. Conclusion ties together all aspects of the project.
- Benefit-of-the-Doubt Marks
    - Each project will be given 3 marks by default.
    - Marks can be removed for major errors not covered by this rubic, such as not including names or submitting an assignment with errors (or no outputs).
- Available Bonus Marks
    - (2 marks) Notebook is hosted publicly on GitHub, GitLab, Bitbucket, etc.
        - Be careful not to include personal information - such as student numbers - in the version hosted publicly. Your GradeScope submission should have this information, but not the version on GitHub.
    - Up to 2 marks can be added for extra creativity in the modelling process.

Example Project Ideas and Data Sources

- The Hansard is the collection of all debates in the Canadian Parliament.
  - Regression: Are longer statements more emotionally charged? Use the VADER lexicon to find sentiment, then compare sentiment to number of words.
  - Classification: What words might indicate party membership? (Note that this can be an inference task!)
  - Data can be downloaded here (warning: 850MB download). The page also has extra data on parties and politicians - sounds like a good chance to show off your `join`ing skills!
- Search for the transcripts of your favourite show, find out if things change over the course of the series!
  - For example: Avatar: The Last Airbender has a nicely formatted dataset on Kaggle. Note that there are some "stage directions" inside square brackets, e.g., `[Happily Surprised.] Sokka, look!`, which should be regex'd out. There are also datasets of user ratings that could be joined, and you might suspect certain words from certain characters correlate with epsiode rating.
  - Alternatively, you could use the VADER Lexicon to find the sentiment of each character.
- Analyse the StackOverflow Developer Survey for trends in the technology that you care about.
  - Filter by people who have used Python, R, and/or Julia via regex (and/or find jobs related to data science), visualize and analyze the responses to other questions.