Final Project: Data Science Workflow

Data 100/200A: Principles and Techniques of Data Science

Due date: Wednesday, May 13th 11:59PM

Project Guidelines

The purpose of this project is to carry through a data science workflow and put into practice what you have learned in this course in a more open-ended setting than the assignments.

To make the project more self-contained, you are provided with four datasets. For undergraduate students, you must pick from one of these three topics; *for graduate students, you can pick from one of the provided datasets or you can choose your own dataset with approval from the GSIs. Please find details of the approval process on Piazza.*

In this project, you should carry through the following steps:

1. Perform exploratory data analysis (EDA) and include in your report at least two data visualizations.
2. Describe any data cleaning or transformations that you perform and why they are motivated by your EDA.
3. Apply relevant inference or prediction methods (e.g., linear regression, logistic regression, or classification and regression trees), including, if appropriate, feature engineering and regularization.
4. Use cross-validation or test data as appropriate for model selection and evaluation. Make sure to carefully describe the methods you are using and why they are appropriate for the question to be answered.
5. Summarize and interpret your results (including visualization).
6. Provide an evaluation of your approach and discuss any limitations of the methods you used.
7. Describe any surprising discoveries that you made and future work.

The analysis must involve at least one of the inference or prediction methods presented in this course.

Datasets

We provide the following three topics and a few potential directions you can explore for each of the topics: A brief description of each is included, as well as some guiding questions. For undergraduate students, you must use at least one dataset provided for the topic of your choice, but you are welcome to add other datasets to help with you analysis.

Covid-19

The coronavirus datasets contain different information about COVID-19. They include information about daily cases and fatalities, as well as hospital level data. We provided multiple datasets for this topic and you are required to use at least one of them if you choose to work on this topic.

The sources of the datasets come from:

- https://github.com/Yu-Group/covid19-severity-prediction/tree/master/data
- https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data

The links of the four .csv files provided for this topic are in the README file under the covid19 data folder.

Below are some potential directions for this dataset:

- What factors can be beneficial in predicting how fast coronavirus spreads in a region?
- Suppose you wanted to model the number of coronavirus cases in a region. What kind of model and features would you use? How might historical trends inform our predictions on the number of coronavirus cases in a given day?

Basketball

The basketball datasets include college player's statistics, and box score information for NBA players over the last 7 years. For these datasets, we have included descriptions of them in the pdf files in the data folder.

Below are some potential directions for this dataset:

- Which college produces the best rebounders, 3 point shooters, scorers, defenders? Which colleges produce the longest tenured NBA players?
- Can you predict how good a player will be based on their college statistics? Is there more of a correlation of one and done players or four year players?
- Since certain statistics exist solely from 2012-2018 while others exist from 1960 on wards, how are you going to account for that? What bias do you introduce to your statistics?

Contraceptive

The contraceptive dataset was conducted on a set of married women (who were not pregnant at the time) and contains information about their background (e.g. age, number of children, education) and the type of contraceptive they use.

You can find more description about the dataset here:

- https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice

Below are some potential directions for this dataset:

- What are the factors that best determine the type of contraceptive used? What are the factors that best determine the number of children a woman has, based on her other information? What could be good models for these tasks?
- What other data-sets could you bring in to add to this analysis? What confounding factors do you anticipate affect this dataset? Cost? Insurance?

Team work

You may complete the project together with up to two other classmates. Students on the same team will receive the same score. Each team will only have to submit one report described in the next section.

Report Format and Submission

The project submission should include the following components, to be submitted on Gradescope as a zip file: Jupyter Notebook, and a CSV file of the data if you choose to use your own data (*for graduate students*).

1. **Jupyter Notebook.** A single notebook that contains all the code run for the project. The code should be clear, easy-to-read and well-documented. In addition, the notebook should explain in text what you can conclude from the output of your code cells.

2. **Project narrative.** This typed PDF document should summarize your workflow and what you have learned. It should include a title, list authors, abstract, introduction, description of data, description of methods, summary of results, and discussion. Make sure to number figures and tables and include informative captions. Specifically, you should address the following in the narrative.

   Note: There is a page limit of 6 pages, excluding figures and tables.

   - What type of question are you trying to answer with the data.
   - (*Graduate students only)* Description of the data if you choose your own dataset and provide a link to the dataset you chose.
   - Perform exploratory data analysis (EDA) and provide data visualizations.
   - Describe any data cleaning or transformations that you perform and why they are motivated by your EDA.
   - Carefully describe the methods you are using and why they are appropriate for the question to be answered.
   - Create a complete statement of the model and assumptions they are using for inference
   - Summarize and interpret your results (including visualization).
   - Address the following seven specific questions.
     - (i) What were two or three of the most interesting features you came across for your particular question?
     - (ii) Describe one feature you thought would be useful, but turned out to be ineffective.
     - (iii) What challenges did you find with your data? Where did you get stuck?
     - (iv) What are some limitations of the analysis that you did? What assumptions did you make that could prove to be incorrect?
     - (v) What ethical dilemmas did you face with this data?
     - (vi) What additional data, if available, would strengthen your analysis, or allow you test some other hypothesises?
     - (vii) What ethical concerns might you encounter in studying this problem? How might you address those concerns?
   - Provide an evaluation of your approach and discuss any limitations of the methods you used.
   - Describe any surprising discoveries that you made and future work.

3. **Dataset.** *(For graduate students)* If you choose to use your own dataset, please submit the data source files. Please make sure that you name these files and put them in the right folders so that we can run your code on loading the dataset without error.

4. **Video.** *(For graduate students)* Graduate students are required to make a video explaining their work and link this video in the PDF document. Please limit the length of the video to 2 minutes.

Grading

You will be graded based on the Jupyter notebook and the narrative pdf you submitted.