Data Science 311

Final Project (50 points)

Read all of the instructions. Late work will not be accepted.

Overview

In the final project, you will complete a project that puts together just about everything we will have covered in this course, covering the full extent of the data science pipeline from coming up with a question to presenting your results. The project will run for just over four weeks, with roughly one deliverable due each week.

- **Proposal / Data Acquisition**: Propose a topic / question; demonstrate that the data exists, and (unless your project includes a heavy data acquisition component) get the data you will need to answer your question.
- **Milestone 1:** Preliminary results. Demonstrate preliminary results on your data, sufficient to demonstrate that your question is answerable. Implement one or more baselines for your prediction task. If your project needs to pivot in any way, you will include an explanation and updated roadmap.
- **Milestone 2:** Full Analysis. Complete a thorough analysis, including any exploratory analysis and predictions.
- **Final Submission:** Complete polished analysis and visualizations. Provide tuned prediction results, and a thorough discussion of your findings.

Topics

The topic for the project is left open-ended so you can explore something of interest to you. The requirements are:

- The project should be substantial – larger than any of the labs we have done so far.
- The project should involve some exploratory analysis component, which must yield some interesting insights; if you find a negative result (no interesting trends in the data), you will need to pivot to new data or a new topic.
- The project must involve a machine learning component; this may be classification, regression, or clustering. The machine learning component should be
  - Well-motivated: there should be some good reason to want to predict the thing you are proposing to predict.
  - Successful: this might involve discovery of notable structure using unsupervised learning, or prediction results that do better than one or more sensible baselines. Again, if you are unable to get reasonably good results on your prediction task, you will need to switch to a different task.

A Note on Flexibility

It is pretty easy to scope a project incorrectly or go down a path where good data is not available, so we may need to iterate a little on the proposals before settling on a topic. It is also possible that you will encounter unexpected road blocks; in this case, well-justified pivots will be allowed at each milestone, provided that you end up with a quality project at the end. Pivots because you did not start early enough and ran out of time are not considered well-justified.

Tasks and Deadlines

Proposal (10 pts), due Wednesday, 11/2 at 10pm

Write a proposal document for your final project. One person from each group will submit a single proposal as a PDF by the deadline above. Make it as short or long as needed (ideally no longer than needed), but I expect these to be about 1-2 pages. The document should include the following sections:

- A list of group members
- The goal of your analysis:
  - Your motivating research question. What are you looking to learn from this project?
  - Describe the data you are working with. By the time you submit the proposal, you should be very sure the data exists, and ideally have acquired it and made sure it is readable and contains what you think it does. If your project involves a substantial scraping component, you should have a proof-of-concept scraper running, though you need not have all the data collected yet.
  - At a high level, describe how you plan to use your data to answer your question. Be sure to talk specifically about an exploratory component and a machine learning component.
- Milestone 1 deliverable: describe what you plan to have done at the first milestone deadline. This should tell me how the Milestone 1 guideline listed in the Overview section above applies to your project.
- Milestone 2 deliverable: describe what you plan to have done at the second milestone deadline. This should tell me how the Milestone 2 guideline listed in the Overview section above applies to your project.
- Roadmap: do your best to break the project into subtasks that will take one group member no more than a week to accomplish. For each task, give a tentative allocation of which group member(s) will accomplish it and when it will be done.

Milestone 1 Report (5 pts), due Saturday, 11/12 at 10pm

Submit a milestone report describing your progress, and supporting notebooks, code, etc. Submit a PDF file with the following information:

1. A list of group members

2. If any feedback was given on the proposal regarding missing details, address that here.
3. A description of the status of each of the tasks requested for Milestone 1. This will include:
   - The data is acquired and cleaned.
   - The bulk of the exploratory analysis is done and presentable, though it need not be highly polished yet.
   - The evaluation environment is set up for the prediction task. This means that:
     - Data splits are created; make sure that the splits are repeatable by setting the seed of the random number generator, e.g. by passing a `random_state` to `train_test_split`.
     - Appropriate evaluation metrics have been chosen and implemented.
     - Sensible baselines have been implemented, and their performance has been assessed using the evaluation metrics.
   - Any other deliverables mentioned in your proposal or requested in my feedback.
4. If any of the above goals were not met, explain why and detail your plan for completing them. If any changes in scope, goals, or roadmap are necessary, explain why and what your updated plan is.

Alongside your report, you should submit a zip file containing artifacts showing your progress. This will likely take the form of ipynb files, but if something else makes sense, go for it. I expect to see at least evidence of exploratory analysis results and baseline/evaluation setup. Any curated datasets you have created for the project should be saved in csv or other appropriate format with a corresponding Readme.txt. You do not need to submit the raw uncurated data, but ***you should submit the notebooks saved with up-to-date outputs that show us what we would see if we did run them.***

Milestone 2 Report (5 pts), due Saturday, 11/19 at 10pm

Submit a milestone report describing your progress, and supporting notebooks, code, etc. Submit a PDF file with the following information:

1. A list of group members
2. A description of the status of each of the tasks requested for Milestone 2. This will include:
   - Exploratory analysis is completed.
   - You have achieved the best performance you can on your prediction task. This likely will involve preprocessing, model selection, and hyperparameter tuning; you should try multiple different classification or regression models provided by sklearn and do at least some hyperparameter tuning to get the best possible performance you can on each. Submit a notebook showing this analysis and discussing different avenues you tried.
   - Any other deliverables mentioned in your proposal or requested in my feedback.

3. If any of the above goals were not met, explain why and detail your plan for completing them. If any changes in scope, goals, or roadmap are necessary, explain why and what your updated plan is.

Alongside your report, you should submit a zip file containing artifacts showing your progress. This will likely take the form of ipynb files, but if something else makes sense, go for it. This should include (some of this may not have changed since MS1 if it was satisfactory at that point, but you should include it anyway):

- Data collection, cleaning, and preprocessing where applicable
- CSV file/s (or other applicable format) with curated dataset/s.
- Readme.txt describing curated dataset/s.
- Exploratory analysis results
- Predictions, including baselines and model selection work.

You do not need to submit the raw data necessary to run the notebooks, but you should submit any curated data along with the notebooks with up-to-date outputs that show us what we would see if we did run them.

Final Report (20 pts), due Sunday, 12/4 at 10am

Notebook Blog Post

This is a writeup of your project for a general audience. It will be a notebook (titled `blog.ipynb`), but you should think of it as a blog post. It should talk not only about your results, but also provide background on what you set out to do, why is it interesting and worth reading about, the size, source, etc. of the data you used. It should walk the reader through your most interesting findings and provide discussion and interpretation of what the results mean and their implications. The blog post should address both the exploratory analysis you did and the predictions you made, though it can focus more on one or the other if one of them turned out to be more interesting. The blog post should include at least some Lab 2-quality visualizations to support the exposition; that is, the visualizations should be highly polished and adhering to the principles of visualization aesthetics, though you don't need to explain your designs here as you did in Lab 2.

Supporting Notebooks

Your blog notebook should also describe one or more additional notebooks (in .ipynb format) containing the details of your analysis. These notebooks should be correct, clear, and thorough according to the guidelines used in several labs this quarter. The audience here is an interested reader of your blog post who wants to dive into the details of your analysis, and potentially reproduce it – as such, the notebooks together should contain everything needed to reproduce your results.

Submission

Submit a single zip file to Canvas containing:

- The `blog.ipynb` file containing your blog post.
- The supporting notebooks (in .ipynb format) with your full analysis.

Submissions will be posted on the course webpage and linked from a final project showcase page for posterity.

Final Presentations (10 pts), Week 11

We will use Monday, Wednesday and Friday of Week 11 give brief (10 minute) presentations of the final projects. These are informal presentations that will give you a chance to see the fun and interesting results found by other groups. You may make separate presentation slides, or just show and talk about your blog post, but make sure that you're keeping it short and talking about the highlights of what was interesting about your findings. If you have slides or any other visuals to show, you will need to send them to me the night before the presentation so we can present from a single computer. I will announce the presentation schedule when it gets closer to Week 10.

Acknowledgments

This assignment is based upon Scott Wehrwein's final project assignment from Fall 2021.