Project Final Instructions

Data Science in Practice, Fall 2019, CS4973/CS6463

1 Summary

1.1 Downloaded Report Package

The PR03.zip file contains three items:

- **PR03-ProjectFinalInstructions.pdf**
- **PR03-ProjectFinalReport.ipynb**
  Use this Jupyter notebook template to submit your project report. See further instructions below.
- **PR03-ProjectFinalCode.ipynb**
  Use this Jupyter notebook template to submit your project code. See further instructions below.

1.2 Due Dates

The final project report is due before midnight on **Monday, 2 December 2019**.

I will do one early submission review of the Project Final Report notebook

- If you submit your PR03-ProjectFinalReport.ipynb by midnight on **Monday, 25 November 2019,** I will review and return them with suggestions by the end of the day on Tuesday, 26 November 2019.

1.3 Grading

The final project is worth 18 points, broken down as follows:

- Project Report Notebook – **5 points**
- Final presentations – **4 points**
  CS6463 students must present their project and be present at both sessions
  CS4973 students must be present at both presentation sessions
- Project Code Notebook – **9 points**

1.4 Presentations

Graduate students must present their project to the class. There are 10 graduate students that will be presenting. Presentations will be on Tuesday, 3 December and on Tuesday 10 December.

- Each presenter will be allotted 10 minutes for their presentation and 5 minutes for questions.
- A schedule will be emailed to everyone designating the random order chosen for the presentations.
- Undergraduate students must be present to get credit for the presentations.

2 PR03-PROJECTFINALREPORT.IPYNB

2.1 INSTRUCTIONS

The project final report notebook will be manually graded, but must still be processed through the automated grading system. The template notebook has very limited content, consisting of read only markdown sections that contain a title for the content that you must add in subsequent cells. The following sections describe the expected content and the point values assigned to each section.

Try to structure your project notebook to tell a story. Your work will be judged on quality, not quantity. Remember that you are demonstrating the principles that you have been studying all semester. Be clear and concise. Please review your work prior to submitting your notebook. If graphics are not visible after the notebook is run through the grader, you will lose points.

The majority of the content of the project report must be in markdown cells.

- All tables must be in markdown cells. If you want to show the output from a Pandas dataframe rather than crafting a markdown table, consider taking a screen shot, saving the file, then including it as an image.
- Static graphics must be embedded in a markdown cell as images.
- You may use code cells if necessary to demonstrate interactive graphics, but any data required to run the graphic must be loaded from a preprocessed, saved file.
- If you include interactive graphics, you MUST also include one static image of the graphic in a markdown cell along with a description of how the interaction works and what insights the user is expected to gain from the graphic.

The points assigned to each section are detailed below. Each section will be graded primarily on content, but appearance matters, and points may be lost for incorrect formatting.

2.2 NOTEBOOK SECTIONS

This notebook already contains the following required level 2 section headings. You may include additional level 3 headings as needed to tell your story.

2.2.1 Project Final Report (0.1 points)

This section must contain only 1 markdown cell with the following information. The index must be a hyperlinked list of sections.

```
Title:
Student Id:
Course: (either CS6463 or CS4973)
Index:
```

2.2.2 Research Objective (0.1 points)

This section must contain only 1 markdown cell. Please review your research objective and update it as necessary to match your final project accomplishments

2.2.3 Data Description (0.8)

This section may contain as many markdown segments as needed. You may include one code section if needed for an interactive graphic demonstration.

The intent of this section is to acquaint your reader with the data that you used. Where did it come from? How much data is there? What are the important fields?

2.2.4 Data Exploration (1.1)

This section may contain as many markdown segments as needed. You may include one code section if needed for an interactive graphic demonstration.

The intent of this section is to describe to your reader a discovery process that helps the reader to understand the data. This should include items like statistical summaries, summary tables, and appropriate graphs. What problems did you encounter and how did you overcome them? What kinds of data aggregation were necessary, especially if more than one data set was involved? What did you do to clean, scale, or otherwise prepare your data? If fields required special handling, describe that. For example, if sales prices were collected over an extended period, were they corrected for inflation? If you needed to aggregate data expressed as percent values, what basis did you use for correcting potential size differences (for example, how would you aggregate two tax revenue streams expressed as percent of gross revenue? One group had 10% tax on $1000, the other had 2% tax on $100,000. The aggregate tax rate is certainly NOT 6%, but rather 2.1% !)

2.2.5 Modeling (1.4)

This section may contain as many markdown segments as needed. You may include one code section if needed for an interactive graphic demonstration.

Briefly describe the type of modeling that you employed (classification, regression, unsupervised clustering). What attributes did you use and how did you select them? What were you trying to predict? What types of parameters were required for your models? Did you perform any searches through the parameter space? What were the results of your modeling?

This section should contain at least one graphic (preferably more) to demonstrate the effectiveness of your modeling. This could include a confusion matrix, ROC curve, prediction plots, etc.

2.2.6 Summary (1.4)

This section should only contain markdown cells. In this final section, summarize what you learned about your dataset. This must include at least one presentation quality graphic that adheres to the principles of Tufte and Cairo that we studied.

2.2.7 References (0.1)

Cite all data sources used in this section. Consider referencing the licensing associated with the data source. If you used other notebooks or published code, you must cite them here. All references MUST be formatted properly using numbered paragraphs.

2.2.8 Presentation (4.0)

This section is for crediting either giving a presentation or attending the presentations. Do not put any cells below this one.

3 PR03-PROJECTFINALCODE.IPYNB

3.1 INSTRUCTIONS

The project final code notebook will be manually graded, but must still be processed through the automated grading system. The template notebook has very limited content, consisting of read only markdown sections that contain a title for the content that you must add using subsequent cells. The following sections describe the expected content and the point values assigned to each section.

The purpose of this notebook is to provide the code that you developed during your project. The code will be reviewed for structure, commenting, and organization.

IMPORTANT: It is not intended to execute the code in the auto-grader. The code may be executed and evaluated separately.

- Do NOT include your data with your project submission.
- Your code must access your data using the same environment variable approach that we have been using in class. For example:

```
dataroot = os.environ['DATASETS_ROOT']
sDataDir = os.path.join(dataroot, 'Kaggle/amazon-fine-food-reviews')
```

- You must include a markdown section with instructions on how to obtain the data in order to reproduce your results.

The majority of the cells in this notebook should be code cells.

- You may include markdown cells as necessary for extended comments, but embedding graphics is not necessary.
- Most of your code should be in functions that compile but do not execute, or in separate .py files.
- For each section, include ONE code cell that executes your functions to perform the tasks. THIS CODE SHOULD BE COMMENTED OUT.
- The code to produce your graphics must be included. Again, use functions for the majority of your work and comment out the few lines that execute the functions.

The points assigned to each section are detailed below.

3.2 NOTEBOOK CONTENTS

3.2.1 Project Final Code (0.1)

This section must contain only 1 markdown cell with the following information. The index must be a hyperlinked list of sections.

```
Title:
Student Id:
Course: (either CS6463 or CS4973)
Index:
```

3.2.2 Obtaining the Data (0.1)

This is a markdown section.

- Describe how to obtain the dataset(s).

- If you wrote code to download or access the data, you may include a code cell with the code. HOWEVER, you MUST comment out the code. Otherwise, the auto-grader will attempt to execute the code.

3.2.3 Setup (0.5)

This must be a single code cell similar to the ones on the homework assignments. If you moved some of your code to separate .py files, you should preload those here.

3.2.4 Data Preparation (1.1)

One or more code cells.

3.2.5 Data Exploration (1.2)

One or more code cells.

3.2.6 Modeling (1.4)

One or more code cells.

3.2.7 Presentation Graphic (1.4)

One or more code cells.

3.2.8 Project approach and overall execution (1.2)

These last three sections are for grading. Do not put any cells below this one.

3.2.9 Code Structure and Organization (1.0)

3.2.10 Code Commenting (1.0)

4 ASSIGNMENT SUBMISSION

4.1 CHECKLIST

Before you submit your notebooks for grading, you must do the following or you risk losing points.

1. Remove all comments that are inserted by the auto-grader release process. These include:

```
# YOUR CODE HERE
raise NotImplementedError()
# YOUR ANSWER HERE
```

2. Restart the kernel (in the menubar, select Kernel→Restart)
3. Run all cells (in the menubar, select Cell→Run All).
   For the report notebook, the ONLY thing that should execute are any interactive graphics.
   For the code notebook, NOTHING should execute.
4. Review the results to ensure that all graphics and output are as you expect.

4.2 SUBMITTING YOUR PROJECT

All submissions must follow the normal procedure with all artifacts placed in the root directory of a zip archive named **PR03.zip**.

The submission **MUST** include:

- PR03-ProjectFinalCode.ipynb
- PR03-ProjectFinalReport.ipynb

The submission **MAY** include:

- Supporting code (*.py) files that are loaded from your code notebook
- Small data files required to support any animations in your report notebook

The submission **MUST NOT** include:

- Large data files
- Any subdirectories (e.g. .ipynb_checkpoints, __pycache__, _tmp)