Final Project: Part 2

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

Fall 2021

The purpose of this project is to put into practice what you have learned in this course through the design and implementation of a typical data science workflow, including data cleaning, visualization, exploratory data analysis, feature selection, and modeling.

This is Part 2 of the project, where you will continue where you left off in Part 1. Using the information you gathered from EDA in Part 1, along with feedback from your TA, you will answer some guided questions on modeling different aspects of the data. After that, you will be able to use modeling to answer questions you may have come up with in the first part of the project.

Datasets

Project Guidelines

This part of the project involves carrying through the following steps.

1. **Part 1 Peer Evals** You will be evaluating your group (including yourself) on your contributions for Part 1 of the project. Please be honest about your group's contributions. Completing these evaluations will be a part of your project grade.

2. **Modeling**

    - Guided questions
    - Open-ended questions

3. **Model assessment.** Evaluate the performance of your model in 3) using five-fold cross-validation of the learning set to estimate the misclassification error rate, i.e., perform all the tasks in 3) (including feature selection) on the training sets and compute misclassification rates on the validation sets.

4. **Part 2 Peer Evals** You will be evaluating your group (including yourself) on your contributions for Part 2 of the project. Please be honest about your group's contributions. Completing these evaluations will be a part of your project grade.

Timeline

Date (by EOD at 11:59pm)	Event / Deliverable	Relevant Links
11/24	Part 2 Released	
12/2	Part 1 Peer Evals Due	https://forms.gle/jDWd23qivV6S2RXr8
12/13	Final Deliverable Due	
12/15	Part 2 Peer Evals Due	

Report Format and Submission

1. **Code.** Use the provided starter notebooks to complete the guided modeling aspect of the project. Use your submitted notebook from Part 1 to complete the open ended modeling. It may be useful to make a copy of your Part 1 notebook to track what was completed in which part.

   (a) Guided Modeling (provided notebook)

   (b) Open Ended Modeling (Part 1 notebook)

   Note: We will run the notebooks when grading, so please account for that.

2. **Open Ended Modeling Report.** This typed portion of the notebook should summarize your workflow and what you have learned. You should discuss the modeling you completed, along with what problems you solved with the modeling you completed.

   - The report should contain between 1500 and 3000 words
   - Describe the problem you addressed with modeling.
   - Analyze and evaluate your model, including visualizations showcasing your analysis. Ensure you're using the correct mechanism for determining the success of the model.
   - Describe the types of models you produced. Carefully describe the methods used and why they are appropriate for the question to be answered.
   - Show some steps that you have taken to improve your model.
   - Describe further work that could continue the work completed so far in the project.

Grading.
Part 2 of the project will be graded based on your modeling code and writeups.

Grading Breakdown

- **Part 1**: 50%
- **Part 2**: 50%

Project Component	**%**
Guided Modeling	20
Open Ended Modeling	30

Team work.
You must complete the project together with your assigned group. You will be graded equally.