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. If your notebook will crash on the autograder, comment out that code for your autograder submission; otherwise, all of your autograded parts will fail.

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 bullet points below, preceded by dashes, are the rubric items we will use for grading the required portions of your project.** We will expand the rubric to reward extra effort with extra credit, but the expanded rubric won't affect the curve or the required portions. These rubric items may be split up, to maximize partial credit we can give.

   - The report should roughly contain between 1500 and 3000 words. Fewer than 1500 words is highly discouraged.
   - **(Optional) Open-Ended EDA**: Re-include your open-ended EDA visualizations *if you intend to clobber your design document score.* See the clobber section below for details.
     - ⋆Plot 1 Title, Axis Labels, Takeaway.
       - Dataframe screenshots do not count.
       - Takeaway should help the reader understand what is interesting in the plot. What is the insight that this plot provides?
     - ⋆Plot 2 Title, Axis Labels, Takeaway. Same requires as Plot 1.
     - ⋆Performs EDA that covers **two** of the following categories:
       - Repeated guided EDA analysis on new data (pre, post, pre-post difference)
       - Used a different temporal split of data (e.g., weekend vs. weekday)
       - Used a different grouped locational split of data or max/min/upperbound/lowerbound time, destination count, number rides, etc.
       - Had additional insight from a plot with new split of data.
     - ⋆Includes open-ended questions for further open-ended EDA
   - **Problem**: Describe the hypothesis you addressed. Your hypothesis is judged on testability:
     - ⋆Is it explicitly explained how the hypothesis will be confirmed or rejected?
       - Not needed if hypothesis obvious, like ”we expect A > B”.
       - Needed if hypothesis vague. e.g., ”X and Y are correlated”. To what significance? Positively? Or negatively correlated? No correlation is exactly 0 with real-world data.
       - Explicitly saying ”null hypothesis” and ”alternate hypothesis” is acceptable, as implying you will use an A/B test.
     - ⋆Can you confirm or reject this hypothesis in principle, assuming you have unlimited access to all data? (This rubric item is designed to catch hypotheses that *could* technically be answerable. If your hypothesis satisfies the next rubric item, it will satisfy this one too.)
     - ⋆Can you confirm or reject this hypothesis with existing datasets? Or is it reasonable to expect the dataset exists? Item not awarded if project needs inaccessible ground truth (e.g., Uber prices for all trips) unless design document mentions clever proxy (e.g., scraping uber-estimates.com)

  - *You will miss these points if your document does not state how hypothesis will be confirmed or rejected. In all of the following, how do you gauge success or failure? Examples:*
    - *We believe we can/can't predict X. (Possible fix: believe can predict to Y accuracy)*
    - *We believe X and Y are correlated. (Possible fix: believe are positively or negatively correlated)*
    - *X is explained by Y. (Possible fix: check positive or negative correlation)*
  - ⋆Considers a "creative" data source or feature in the hypothesis. If your hypothesis relates two variables X and Y, then at *minimum*, consider making X or Y a feature *computed* on the original features. (Examples: distance between census tracts, weekday vs. weekend, estimated uber ride cost or driver payout).
    - *You will only get half credit if both X and Y are features in the original dataset (e.g., speed and latitude). Includes basic arithmetic operations like raising to a power, average, difference, variance, max or min.*
    - *You will get no credit if your hypothesis is not testable.*
    - Any feature that involves an external data source is generally a good idea. Examples:
      - X or Y is a feature involving general regions, like city boundaries, nature boundaries, city districts (e.g. Finance District). Includes bridges, since there are only 3. Awarded this rubric item if unsure how you would compute the attribute (e.g., commercial vs. residential)
      - X or Y is a granular feature from another *existing project's dataset. e.g., COVID or AQI
      - X or Y is a granular feature from another *existing dataset. Could include population attributes like income, density etc. OR geographic attributes like elevation
      - X or Y is a high density point of interest from another existing dataset. Could include: type of location like sports venue, number of businesses, tourist attraction etc. OR traffic-related attribute like car accidents, highway
      - X or Y is another method of transportation like BART or bikeshare
- **Modeling**: Describe the types of models you produced. Carefully describe the methods used and why they are appropriate.
  - ⋆Mentions model to train (e.g., linear regression, multiple linear regression, decision tree)
  - ⋆Describes inputs to the model.
    - Requires first rubric item (needs a model)
  - ⋆Mentions output
    - Requires first rubric item (needs a model)
    - For supervised method, mentions variable to predict.
    - For unsupervised method like clustering, mentions possible clustering interpretation
  - *There was formerly an "open-ended questions" rubric item here. That was moved to "(Optional) Open-Ended EDA" above.*
  - Includes an explanation of the model choice, potentially related to EDA or insight.
    - For example, "We noticed a linear trend between X and Y. Given these two variables are used as inputs, we use linear regression to model this relationship".
    - The reason could be to use linear regression as a baseline for example. It could also be rooted in your experimentation. For example, you tried linear regression as a baseline and it underfit horribly. Then, you might use feature engineering to combat that underfitting.
  - Includes an explanation of the input features, potentially related to EDA or insight.
- **Model Evaluation and Analysis**: Analyze and evaluate your model, including visualizations showcasing your analysis. Ensure you're using the correct mechanism for determining the success of the model.

  - Includes evaluation result for model.
  - Evaluation method is appropriate for the task and model.
    - For example, don't use MSE on a classification model.
  - Includes an explanation of whether the model result is "good" or "bad".
    - For example, "We trained a binary classifier, and our resulting accuracy is 80%, significantly higher than a baseline random guesser which would attain 50%".
    - Only exception is if you argue this is a baseline (e.g., The current model uses only 2 features and is a simple linear regression model. We call this naive approach our baseline).
    - Points will not be awarded if there is no explanation, or if the explanation is insufficient (e.g., "MSE 0.20 is clearly great". Why? What if your inputs are all between 0 and 0.5? Then the MSE is relatively large, which makes 0.20 MSE pretty bad)
  - Plot 1:
    - Must be related to model to earn credit.
    - Title
    - Axis Labels
    - Key/Legend if applicable (For example, multiple colors or lines the plot mean this is necessary. You could replace this with a caption.)
    - Takeaway: Should help the reader understand what is interesting in the plot. What is the insight that this plot provides? Does it show that your model predictions are effectively useless? (e.g., predictions are all negative, which would be useless if you're predicting traffic speeds)
    - Connection to project: Should help the reader understand *why* this plot is being included. Does the plot show a difficult case that the model learns successfully? Does it motivate a further model improvement? (e.g., predictions are clearly poor in a specific region of the graph)
  - Plot 2. Same requirements as Plot 1.

- **Model Improvement**: Show some steps that you have taken to improve your model.
  - "Improvement" 1:
    - Problem: Identify what was wrong with the original model. Maybe it made an intolerable, critical mistake that needs fixing. Maybe it was overfitting. This must be justified, either in words or with plots.
      - Incorrect justifications won't get credit. For example, observing 0% training accuracy and declaring the model is overfitting. Mistakes won't be as obvious as this, so be careful.
    - Solution: Identify a solution, and again provide justification for why this solution is the right thing to do. Provide an intuition based on your knowledge in the course. For example, "We applied a LASSO loss term to exploit the sparsity in the problem".
    - Result: Report the result of your experiments. "Explain" the result: If it worked, then simply say your intuition was correct. If not, provide an educated guess for why. The guess must be substantiated with valid intuition or better yet, evidence. The most interesting explanations are ones where you observed a tradeoff or realized some side effect. For example, you fixed the critical mistake but overall performance is lower. Then, explain why this tradeoff is worthwhile. Or maybe you noticed better validation accuracy *also* came with more overfitting.
  - "Improvement" 2: Same as above.

- **Future Work**: Describe further work that could continue the work completed so far in the project.

    - Describe a direction for future work.
        - Relate the direction to your current work.
        - Explain what future work would explore. Can be a trend you noticed, a flaw you failed to address etc.
        - Explain why this new direction would be interesting. For example, it may boost model accuracy for change points, helping us better predict how traffic will change with future lockdowns.

Grading.

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

Design Document Clobber Policy

The above rubric items that are preceded by a red asterisk, like ⋆, overlap with the design document rubric items. To effect a clobber, we will:

1. Compute your score for all the red-star rubric items. This gives us a percentage. We will call that percentage A.
2. We will take the corresponding rubric items from your original design document grade, and compute a percentage B.
3. We use `max(A, B)` when computing your score for the required portion of the design document.
4. The clobber does not apply to the extra credit portions of the design document. The curve for the design document will likewise only include the required portions of the design document (the rubric items included above).

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.