Final Project

Data Science 602: Data Analysis and Machine Learning

Spring 2022

1 Project Objective

Using a data set of your choice, identify a business question and apply data analysis and machine learning to address that question. Projects will require the following elements:

- Define a business problem or question (which could be any real-world academic, business, or scientific, or similar inquiry)
- Identify data sets that are responsive to the business problem
- Analyze the data
- Pre-process the data to form a data set suitable for machine learning
- Identify modeling approaches
- Identify a final model
- Use the outcome of the final model to answer the business problem

2 Deliverables

2.1 Project Proposal

A project proposal will be required mid-semester (see deadlines in Section 7). The proposal will indicate:

1. The business problem or question to be answered
2. The data sources that will be used to answer the question
3. The type of modeling approach (e.g., classification, regression, or unsupervised)
4. The project approach if not CRISP-DM
5. Significant assumptions or constraints
6. Any requested deviations from project requirements

2.2 Final Project

Projects will consist of three parts:

1. A self-contained Jupyter notebook of all code used to generate the dataset. (You may break the problem into multiple notebooks if you wish)
2. An executive summary document (2-3 pages) that summarizes the business problem, approach, and conclusions
3. An oral presentation delivered to the class the week of December 5.

3 Requirements

Below are high-level requirements for the project. Not every requirement may be applicable to every problem, but please ask for guidance in advance to request that a requirement be modified or waived.

1. You should follow the CRISP-DM methodology for this project; or identify an alternative Data Science Trajectory that is responsive to answering the identified business question.
2. Real data is preferred. For applications where true data is not readily available (e.g., fraud), synthetic data may be used.
3. The total number of observations in the original dataset should be at least 50,000.
4. The data analysis should be novel (e.g., an attempt to apply machine learning to a new problem).
5. The approach should identify candidate models from at least 3 modeling families.
6. You will need to evaluate the final model against test data to show predictive validity.
7. The executive summary report should use effective visualizations to show outcomes. You may target the summary to a business or academic audience, as you prefer. The summary should introduce the problem, explain the methodology, and summarize key conclusions.
8. Curated data (e.g., from Kaggle) is not permitted unless there are no other available alternatives. Problems from Kaggle competitions, or equivalent, cannot be used for the final project.

4 Team projects

Group collaborations of up to 2 students are allowed. Group projects are expected to tackle more difficult problems than individual work. Additional presentation time will be allotted to group projects, and both members of the group are expected to present.

5 Presentation

Presentations will be conducted in the final class. You may assume that the audience is technical, but does not have background in the subject matter on which you are presenting. The expected length of the presentation will be announced, and depend on the final class size and number of group presentations. To ensure all project teams have time to speak, please ensure your presentation does not go over the allotted time.

6 Grading

Projects that do not meet the project requirements may not receive credit. For projects that meet the project requirements, grading will be based on the following factors:

- Completion of the project proposal (10%)
- Complexity of the business question being answered (10%)
- Thoroughness of data analysis and pre-processing (25%)
- Execution of model selection (25%)
- Evaluation of selected model (10%)
- Cleanliness and readability of submitted code (5%)
- Readability and effectiveness of the executive summary (10%)

- Effectiveness of oral presentation (5%)

Final success of the project in achieving predictive validity will not be a factor in grading, i.e., if after identifying a sound modeling approach and technical design, the model does not show predictive validity, the negative outcome will not in itself be penalized in grading.

7 Schedule

Date	Target
March 31	Project Proposal
May 12	Project Presentation
May 19	Project Due