HOD 3200: Final Project Guidelines

Mark Chin

2023-08-22

Final Project Overview

For the final project for HOD 3200, you will be working in pairs (or, in rare cases, groups of 3) on a project that showcases the skills learned in this class. The primary aim of this project for you and your teammates will be to answer a question that relates to prediction. This is necessarily vague based on the broad range of material covered in the class, and also to allow for you to pursue projects that cover topics that you are most excited to dig into.

One of the most complicated tasks in completing the final project for this class will be to find data that is ready to be analyzed. Much of the data in the world that you access requires extensive processing (some of which you will learn how to do in this class) before analyses can occur. To minimize the time you spend on processing data and to maximize the time you spend on conducting interesting analyses for the project and identifying how to best display results from these analyses, I am *strongly* encouraging you to use "pre-packaged" data–or data that requires minimal additional processing–for your projects. For example, data from the following sources are in good shape and, for the most part, ready to be analyzed:

1. U.S. Election Data from MIT's Election Lab
2. U.S. Demographic Data from the U.S. Census and Amerian Community Surveys aggregated to different geographies (e.g., states, counties, census tracts) from IPUMS NHGIS
3. Data on U.S. K-12 Public School Academic Performance and Demographics from Stanford's Educational Opportunity Project
4. Decades of Movie Data Scraped from IMDb
5. Extensive Performance Data for NFL Players and Teams Scraped from the Web using the nflreadR Package

From these data sources you may, for example, choose to complete a project using skills learned in this class that answers the following prediction questions:

1. Which U.S. counties should political parties target based on their likelihood of flipping partisan alignment?
2. Which U.S. census tracts are predicted to experience the greatest influx of college-educated 24-35 year olds in the next 5 years?
3. What district and geographic characteristics are most important for predicting inequality in test score performance by student race/ethnicity?
4. How does performance in the first four weeks of the NFL season predict performance later on?

You are of course welcome to find other data sources than those listed above. There is a treasure trove of publicly available information covering sports, education and inequality, and politics. Websites like FiveThirtyhEight, Kaggle, and the UC Irvine Machine Learning Repository also may have data that is interesting and ready to be analyzed. However, heed the recommendation that you may want to maximize the time you spend on analyzing data, and not processing it.

Final Project Rubric

There are five major areas to the final project:

1. Data Analysis
2. Graphical Presentation
3. Written Description
4. Organization and Clarity
5. Coding

Data Analysis

A strong (A) data analysis will use at least two of the four algorithms (conditional means, regression, logistic regression, k-means clustering) that we discussed in class. The data analysis will include a measure of model fit and will describe which characeristics are closely related to the outcome. The analysis will include cross-validation, which will be correctly exectued and described.

An acceptable (B) data analysis will include two algorithms, but there may be some mistakes or inaccuracies in how the results are presented. A cross-validation will be included, but may not be correctly done.

A weak (C) data analysis will not use two algorithms, or will use them inappropriately. It will either not include a measure of model fit, or will misuse a measure of model fit. It will incorrectly describe relationships with predictors or not describe them at all. It will not include cross-validation or the cross-validation will be done incorrectly.

Graphical Presentation

A strong (A) final project will include nicely labeled, easy to understand graphics that describe exactly what is happening with the patterns in the data. The graphics will be complex, showing lots of numbers. The response could include (but doesn't have to include) interactive graphics.

An acceptable (B) final project will include graphics, but these may not be easy to read or may not be sufficiently detailed.

A weak (C) final project will include graphics that are poorly labeled and don't make much sense.

Written Description

A strong (A) final project will include a 1500-2000 word description that is easily understandable by an interested layperson. Assume that your audience is your boss– not me. It will be much easier to write this if you have a perspective.

An acceptable (B) final project will be written pretty well, but technical details may be poorly described or not described at all, and sentences will be hard to follow.

A weak (C) final project will be poorly written, with many mistakes regarding both the analysis and good writing practices.

Organization, Clarity, Formatting

A strong (A) paper will have a `.Rmd` file that generates a very nicely formatted document, suitable for professional presentation. What kind of report would you want to give to a supervisor? That's what I want back from you. The orgnaization should be very clear and easy to understand.

An acceptable (B) paper will have some formatting problems and may not look very nice.

A weak (C) paper will include code chunks, poor formatting, and will just be messy. ## Coding

A strong (A) paper will have code that can generate results from the raw data in an easy to understand way. The code will be commented and will run on my computer without me having to tweak it in any way. (Easy test is to knit the document, with all related files in same directory)

An acceptable (B) paper will have code that is relatively clear, but has some problems, and may not be commented in a way that makes sense.

A weak (C) paper will have code that is messy, hard to understand and not commented. It will not run on my computer, and cannot be easily debugged.