BA810: Team project

Instructor: Georgios Zervas
<zg@bu.edu>

Team project

For your team project you have to formulate and solve a prediction problem.

The project will proceed in two phases.

Phase I

During Phase I your goal is to find a dataset online, and formulate an interesting prediction (and, potentially, inference) problem around the dataset. I leave it up to you and your teammates to work on a dataset and a problem that excites you. However, you will have to "defend" your proposed project in class 5 (see syllabus). These presentations should contain 6-7 slides and last 7-8 minutes. We will allot 2 minutes for Q&A by the audience. Thus, each group will have about 10 minutes total.

During the presentation you should perform the following tasks and answer the following questions:

- Clearly state the problem you are solving.
- Explain what data sources you are going to use.
- Describe your dataset: how many rows, how many columns, what types of variables are included?
- Demonstrate that you can load the data in R: for example, you can do this by presenting a couple of interesting figures you produced that motivate your proposed problem.
- What are the anticipated results?
- What are the potentials implications of your results? How can they be used in practice? Why is the project worth undertaking?

This presentation is developmental and not evaluative; it does *not* count towards your final team grade.

Phase II

During Phase II you will execute your idea. I encourage you to start working on your project as soon as it is approved. It is OK for your idea to change as you work on it. But if it changes substantially, I ask that you consult with me (for example, if you decide to use a completely different dataset).

In general, I expect teams to try all ML methods that are taught in class, though there may exceptions. If you decide not to try a particular method explain why.

Phase II will conclude with a presentation of you results during the last week of class (see syllabus). Each presentation will take 15 minutes. You should clearly communicate your results. The presentation format is up to you but in the very least:

- State the problem
- Tell us who cares about this problem and Why
- Describe you data – where it came from, what it contains
- Present some interesting descriptive analyses (plots/tables) that inform the question your are answering
- Present your main results
- Which methods worked best for your particular problem?
- What were the challenges you faced? Tell us about the biggest challenge you faced and how you overcame it (or, not – that's fine too – not every problem has a solution.)
- Conclude – what have you learnt that can be put to practice?

You will need to submit two things before your presentation:

- Your slide deck
- An R markdown document (in PDF – or, you can export an HTML document and then print as PDF) that shows your work. This document will likely contain more analyses that your slide deck. It will also contain your code. At a minimum this document should show the code that you used to obtain the results in your presentation.

You do not need to submit your dataset.

Dataset pointers

Here are some pointers to useful data sources, but feel free to use any data source you like (as long as you are permitted – please no proprietary data.)

- `https://nycopendata.socrata.com/`
- `https://www.kaggle.com/datasets`
- `https://webscope.sandbox.yahoo.com/`
- `http://www.census.gov/data/developers/data-sets.html`
- `https://www.yelp.com/dataset_challenge`
- `http://datamarket.azure.com/browse/data`