BA476: 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 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 2-3 slides and last up to 5 minutes.

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.
- Specify you outcome variable, and the predictors you will use to predict the outcome.
- Describe your dataset: how many rows, how many columns, what types of variables are included? You can use the descriptive analytics lab as you basis and tweak it to suit your needs.
- Visualize some interesting relationships between a few predictors and the outcome to understand the relationships between them.

You should avoid:

- Time series prediction/forecasting: avoid datasets where the main predictor is the outcome variable at a prior time (e.g. stock prices where a key predictor of price today is price yesterday).
- Classification problems: make sure you outcome is numeric and continuous. The exception is problems where the outcome is a 0-1 indicator. You can treat these as regression problems.
- Small datasets: too few rows (say fewer than 100), or too few columns (say fewer than 10).
- Datasets with a lot of missing values.

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 be 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 10 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 did you learn that can be put to practice?

You will need to submit two things before your presentation:

- Your slide deck
- An Python notebook that shows your work. This document will likely contain more analyses than your slide deck.

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