PPOL 670 Data Science Class Project

Assigned March 2017

Taking a data science class does not make one a data scientist. It's just the beginning. For many teams in industry, the following points are commonly relied upon to assess one's standing:

1. It is expected that data scientists are able to construct a data actionable question and experiment design – if you don't know what a counterfactual is then don't progress to #2 until you know what it is;
2. The main languages are Python, R or another scripting language;
3. Try only to use a method or technique unless one is able to write the math and code from scratch (e.g. Do not use a random forest if you don't know how it works);
4. For a model to be pushed as a policy or product option the accuracy must be a minimum of 0.75. Coefficients alone are not enough;
5. Do not be overly confident about being a data scientist, especially if the first requirements are not met;
6. Accept technical feedback as that's the only way to become the best.

These are ideals, but attainable ideals. The objective of the class project is for you to develop a usable data science product so that you are able to signal some degree of control over the first four points. Working in teams of **1 to 4 people** (3 people is optimal), you will:

- Write a proposal for a project (for Homework #4 due **March 29th**)
- Execute upon the proposal (Final Deliverable due **May 4th)**.

As an added incentive to deliver a professional-grade project, *the best project will be submitted to the Innovation Policy team at Google[x] – and their recruiting team.*

Options

You will have the option of developing a project on one of two tracks: a data science project or a library.

#1: Process and analyze data using a data science technique.

Architect and implement a data science workflow, starting with data wrangling and then training a model for actionable insight.

Requirements

Produce a one-page research proposal, which outlines:

- For homework #4, produce an one- to two-page research proposal that outlines: (1) the goal, hypothesis and research questions, (2) data experiment design, (3) the data, (4) proposed processing methods, (5) proposed analytical methods, (6) ethical considerations of such a project. On the proposal, please indicate who are members of the team. Submit this as a PDF via blackboard. Get us excited and sold on your idea. We will return with comments and feedback.
- For the final deliverable, provide a well-documented .Rmd file and generate a stylish PDF report.

Potential Ideas

There are many public data sources such as Open Data Network – a large repository of city and state-level data, Kaggle Datasets among others. Below are a few example ideas:

- Unsupervised Learning: Determine what Presidents focused on as indicated in their public documents.
- Supervised learning: Combine Global Landslide Data with NOAA weather data to predict landslides
- Supervised learning: Forecast San Francisco's 311 case volumes

- Develop a classifier to predict which host from This American Life is likely to say certain things (supervised/unsupervised): https://github.com/SigmaMonstR/getThisAmericanLife
- Supervised learning: Estimate the likely salary of an employee based on the title

#2: Write a library to handle (Advanced)

Big picture: A library is comprised of classes of methods that are designed to handle very specific tasks. These methods are designed to handle tasks that do not currently exist in the R ecosystem. By taking on this task, you will be creating a series of functions that facilitate a specific use case.

Requirements

- For homework #4, produce a one-page proposal that describes the specific case that the new library supports. Provide a few high-level descriptions of types of methods your library will include.
- For the final deliverable:
- Develop a Github-hosted library of at least four methods. Usage of each method should be well-documented and preferrably part of the Georgetown McCourt Github group (speak with Jeff and Dan for details).
- Provide a well-documented .Rmd file that illustrates how to use the library, preferably using the same workflow but on two separate data sets (generalizability test)

Potential Ideas

- Develop a library to get access to different data sources: US-EU library
- Develop a library that eases interpretation of models for policy analysts
- Develop a library to process drug data for prediction purposes

Scoring

This project is 50% of your class grade and will be graded as follows:

Component	Description	Weight
Relevance	Your project should answer a problem. The size of the problem is irrelevant, but the solution must be relevant. For example, a prototype forecast algorithm reliant on a publicly, frequently updated data source could be pitched to the data producers as a way to apply the data. A library that processes and analyzes a certain type of data can be offered as a utility to other users of the data.	10 pts.
Code and Documentation Clarity	The extent to which your code can be read and understood is an asset getting support for your technical work. Functions should be annotated, data manipulation choices should be described, and any assumption should be documented.	15 pts.
Soundness	Your work should be technically sound and meet your stated objectives. For supervised learning, the statistical experiment design should minimize bias and selection. For unsupervised learning, the results should be stable. For libraries, there should not be any arbitrary decisions.	25 pts.

Note that objectives may change during the course of developing a project. It is acceptable for your project objective to change, but be sure to include a section in your write up or library documentation that describes the revised objective.