BA810: Team Project[1]

Instructor: Nachiketa Sahoo

1 Team project

For your team project you must formulate and solve a prediction problem using the techniques discussed in the course. You can work in teams of up to four students. The project will proceed in two phases.

1.1 Phase I (5 of 20 points)

During Phase I find a dataset and formulate an interesting prediction problem. It is 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 a two-page proposal (see syllabus for due date).

You should perform the following tasks and answer the following questions in your proposal:

- 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 Python, for example by showing a couple of interesting figures motivating your project and showing summary statistics.
- What are the anticipated results? Describe the type of analysis you’ll do and metrics, types of results you’d produce. You won’t be able to say which method will work best, but you can discuss, as currently planned, that you’ll present comparison based on certain metrics, perhaps across certain types of preprocessing, or regularization, or parameter search methods.
- What are the potentials implications of your results? How can they be used in practice? Why is the project worth undertaking?

Please submit only one copy of the proposal per team.

1.2 Phase II (15 of 20 points)

During Phase II you will execute your idea. 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 (for example, if you decide to use a completely different dataset) consult with me.

You should try to apply most ML methods that are taught in class. 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 7 minutes followed by 2 minutes for question and answers. 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 your data – where it came from, what it contains
- Present some interesting descriptive analyses (plots/tables) that motivates your exercise
- Present your main results

[1] Acknowledgement: This project assignment is based upon the structure offered by Giorgos Zervas.

- Which methods worked best for your problem?
- What were the challenges you faced? Tell us about the biggest challenge you faced and how you overcame it (or, tried but did 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* (by 9am on the day of) your presentation:

- Your slide deck
- The link to the Colab notebook that shows your work — place it on the last slide of the deck. Make sure to switch on sharing in this Colab notebook so that we can access. 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 and the results that you obtained that were reported in your presentation.

You do not need to submit your dataset. **Please submit only one copy of the slides per team.**
We'll conduct the presentation in the team number order.

2 Dataset pointers

Here are some pointers to useful data sources, but feel free to use any data source you like (with permission—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
- https://learn.microsoft.com/en-us/azure/open-datasets/dataset-catalog
- https://datasetsearch.research.google.com/

There are many datasets (often from UCI repository) which have 100 to 1000 records. Many such datasets are often too clean and considered "toy" datasets. Aim for somewhat larger datasets (10k to 100k records) with 10-20 non-anonymous columns. If it's closer to business applications (purchase, churn, demand prediction, etc.) that's a plus, but can be something else if you are really interested in it. Business relevant project is more likely to help you in the future.