Term Project guidelines

You are encouraged to work in **groups of three to five** for the term project to analyze and develop a machine learning solution using (preferably) business-related data. Individual projects or groups of 2 are not allowed, given the size of the class. A group of 6 is possible but you will need to justify it and choose an adequately complex project.

**Grading**: This project is a critical part of the course, and a significant factor in determining your grade. By default, all team members will receive the same score for their project. Typically, all project teams go for this default option which is of course the preferred way. However, if a team or team member feels that this is unfair, perhaps due to HIGHLY imbalanced contributions, then every team member needs to provide feedback on the contribution of each of the other team members via email to the instructor by the last day of class. After that I will need to have a meeting with all the members together to mediate.

**Dates**:

1. **Project outline due Oct 25**. 2-3 pages describing the problem, data available, some possible approaches you will consider to address the problem, and a short list of references. (need not be fully flushed out, more for a sanity check).
2. **In-class presentation** of project results, Late Nov/early Dec, approx 15-20 mins per group.
3. **Blog due by midnight, Dec 10th**, via Canvas. One submission per group. You are also asked to submit supplementary materials (code, referenced papers) via a pointer to the appropriate URL/dropbox/github/.. location(s). Blogs are usually posted in a publicly accessible location, however if you don’t want to make the blog public, that is OK with me so long as I have access and thus can read it.

# Project presentation schedule

Project groups, title and schedule will be available on Canvas when ready. Guidelines for your in-class presentation and for the content of report and the criteria for its evaluation are uploaded into Modules 🡪 Projects

# Project topics

The project should be centered around some problem with associated data sets that you can mine to provide useful and actionable answers. At the least, this should be an exercise in analyzing a reasonably large dataset. In the process, if you invent new techniques/algorithms or processes, or make inferences that are useful and not done before, of course that is an added bonus, though this is not common. Two types of projects are suggested below.

## Type I: Based on a Competition or other Real-World Large Datasets

### Data Mining Competitions

There have been several data mining competitions such those hosted by Kaggle (www.kaggle.com). For several of these competitions, as well as those from KDD cup (<http://www.kdnuggets.com/competitions/kddcup>), the data is still available and you can also find papers on how others have fared on these data sets. There are also several other ongoing competitions (e.g. see <http://www.kdnuggets.com/competitions>).

Warning: these can be quite addictive, but also quite fun and a learning experience, specially if it is an on-going competition.

### Other Public Domain Datasets

There is an astonishing amount and variety of public domain datasets on the web. KDNuggets (<http://www.kdnuggets.com/datasets/index.html>) provides a long list. Google now provides [*https://toolbox.google.com/datasetsearch*](https://toolbox.google.com/datasetsearch) to help looking for datasets.

You could be even selective on the topic, for example, if you google “multilabel classification dataset”, the first hit is a bunch of datasets associated with the software Mulan.

Microsoft has made available a variety of datasets at <http://research.microsoft.com/en-us/projects/data-science-initiative/default.aspx>  
They periodically hold competitions as well.

The US Government's Open Data policy has also resulted in a treasure trove of data. See (<http://www.data.gov>)

## Type II: Based on Type of Analysis or Application Domain

You can formulate and address a suitable predictive modeling problem based on data from industry or government. It will be your job to acquire and manage the data. **The project should be doable within a couple of months, but also non-trivial: at the very least it should involve a large (say "rows" times "columns" > 1 million) data set**. Remember that your class presentation is public, however your class report is not, and I (and the TA) can sign NDAs if need be in order to work with you on such a project and to evaluate it. You can choose any topic you want. For example, you could look at healthcare data, or data related to recommendation systems. Some pointers to these two example topics are given below:

## Type III: Hot Topics ( needing a mix of theoretical/algorithmic and experimental work, several papers and code repos available to get started!)

# Probably the hottest new topic other than “MLOps” (deep learning is old hat!) is building responsible and ethical ML systems. There are multiple aspects to such systems, including software processes, security, privacy, etc, but on the algorithmic/statistical side the three key issues are (a) building models that are explainable (aka XAI or explainable AI), (b) building fair or unbiased models and (c) robustness, i.e. building systems that are more tolerant to malicious attacks, are less prone to unexpected behavior, etc. A good place to start looking at these issues is information and code from IBM: <https://www.research.ibm.com/artificial-intelligence/trusted-ai/> which includes pointers to the [The Adversarial Robustness Toolbox (ART)](https://adversarial-robustness-toolbox.readthedocs.io/en/latest/) and the [AI Fairness 360 Open Source Toolkit](http://aif360.mybluemix.net/). Also see the [LinkedIn solution announced Aug 2020](https://engineering.linkedin.com/blog/2020/lift-addressing-bias-in-large-scale-ai-applications), specially for testing and scalability aspects. Google and Microsoft are also introducing tools in this space, so are startups such as Fiddler.AI

There are also workshops and tutorials on these topics, and even some conferences like [FAccT](https://facctconference.org/2021/) and [AIES](https://www.aies-conference.com/2021/).