Term Project guidelines

You are encouraged to work in **groups of three to five** for the term project to analyze and develop a predictive modeling 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.

This project is a critical part of the course, and a significant factor in determining your grade. Teams are required to hand in a brief report and prepare a short class presentation of their work. By default, all team members will receive the same score for their project. Typically all project teams go for the 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 16**. 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. On Nov 1, each group will make a 3-4 min **in-class presentation** (4 slides max) on what they plan to do.
3. **In-class presentation** of project results, Late Nov, approx 15-20 mins per group.
4. **Written project** **report due by midnight, Dec 14th**, via Canvas. One submission per group. Your report should be 20-30 pages (1.5 spacing) including figures, tables and/or references in the form of a single pdf file. You may want to refer to the [guidelines](http://cseweb.ucsd.edu/users/elkan/250B/writing.pdf) posted by Professor Elkan at University of California, San Diego for writing your paper. If you want to submit supplementary materials (code, referenced papers) make a folder and give me access via a pointer to the URL/dropbox/github/.. location.

# 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.

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.

Yahoo has complex web-related data, typically quite challenging: <https://webscope.sandbox.yahoo.com/>The US Government's Open Data policy has also resulted in a treasure trove of data. See (<http://www.data.gov>)

Austin’s own <https://data.world/> has many datasets and tools, but some of these datasets are very simple.

## 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:

### Two recent hot topics: needing a mix of theoretical/algorithmic and experimental work

### (i) explainable machine learning, or more broadly, explainable AI, e.g. see <http://home.earthlink.net/~dwaha/research/meetings/ijcai17-xai/>

### (ii) fairness and trust in ML: <http://www.fatml.org/>

### Data Mining for Health Care

[List of some Health Care Data Sets](http://www.ideal.ece.utexas.edu/courses/ee380l/Health-Care-Data-Sets.xls). CMS, for example has recently released data about Medicare Provider Utilization and Payments, http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/ which has lots of possibilities. An amazing large and rich ICU dataset has been released recently <http://eicu-crd.mit.edu/about/eicu/> (I’ll need to write a note on your behalf for permission, but it is routine).

### "Affinity" Data Sets

[List of some "Affinity" Data Sets](http://www.ideal.ece.utexas.edu/courses/ee380l/Affinity-Data-Sets.xls). These problems involve finding the affinities among two (or more) sets of entities, such as users and movies, users and web pages/advertisements, etc. Clearly they include recommendation or ranking problems. Often "side information" in the form of additional attributes of these entities (e.g. demographic information for the users, a social network etc.) is also available to improve predictions. Another associated problem is learning to rank, for which the LETOR datasets/benchmarks are public-domain.