Data Sources for Project

It is absolutely ESSENTIAL for the grade of your project that you find a good dataset & problem. The next most important thing is that you a) start early and b) talk to me often. This project COULD make your career! I have had students getting jobs based on what they did in this course. Read the below document VERY CAREFULLY.

Finding a good project:

1. It has to be a good problem with some ACTION somewhere
2. The scenario has to make sense
3. Needs to be technically feasible within your group (you are welcome to simplify)
4. You need to be able to get the data (you need the data NOW)
5. The data has to satisfy certain criteria: enough features, enough examples
6. Identify something you can actually evaluate. This typically means that you can simulate your ACTION on predictions you make on a subset of the data and where you can argue how much better the overall results are over some ‘dumb’ strategy that is not utilizing the predictions.

This document gives you some tips on finding a suitable dataset for your project. Typically projects using data from one of the team members tend to do better

Sources of data:

* YOURS – you work already with data – great – bring it in!
* Data Repositories such as
  + <http://www.kdnuggets.com/datasets/index.html>
  + <http://www.sigkdd.org/kddcup/index.php>
  + http://smartdatacollective.com/bernardmarr/235366/big-data-20-free-big-data-sources-everyone-should-know
  + http://kdd.ics.uci.edu/
* NYC has a great selection of really relevant data.
  + data.cityofnewyork.us
* DonorsChoose <http://data.donorschoose.org/open-data/overview/>
  + local NYC non profit is waiting for your help to get material to teachers in the area
  + This is REAL – they may actually use what you build!
* Kaggle http://www.kaggle.com/competitions
* Lending Club
  + <https://www.lendingclub.com/info/download-data.action>
* Kickstarter
  + https://sites.google.com/site/machinelearningkickstarter/dataset-and-learning-methods
* Yelp
  + https://www.yelp.com/dataset\_challenge
* Miscellaneous ML Collection (YUGE!!!)
  + https://blog.bigml.com/list-of-public-data-sources-fit-for-machine-learning/

Other tips:

You are at liberty to make up certain details of the task. For instance, you can pretend not to have data that you have or that the solution does not already exist.

The technical details are important but a great project originates in a good business case. See some of the examples I posted. Can you make the results of your analysis ACTIONABLE. So deriving ‘insights about consumer behavior that can be used by the strategy group to better target’ is NOT ACTIONABLE. Using the predictions of your model to allocate marketing spend proportional to the predicted likelihood of changing carrier IS ACTIONABLE. You need to think about how to evaluate the ACTION, not just the model. This is one of the hardest parts and requires some very deep thinking. I am looking for creativity here. Obviously you do not know what will happen without a full deployment but we would like to see a proposal on how it should be evaluated.