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. Don’t do something impossible (predicting stock prices or the weather)
4. Needs to be technically feasible within your group (you are welcome to simplify)
5. You need to be able to get the data (you need the data NOW, latest in 2 weeks)
6. The data has to satisfy certain criteria: enough features, enough examples
7. 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

Some examples of projects that tend to be difficult to pull off in terms of making a good business case:

* Predict Movie Revenue
* Predict Flight delay
* Predict Yelp ratings
* Genetic Data

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 (Drew Conway) and I can connect you.
  + 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!
  + They have more internal data than you will see online
* Kaggle http://www.kaggle.com/competitions
  + Basketball
  + ASUS Malfunction
  + Loan Default
* Lending Club
  + https://www.lendingclub.com/info/download-data.action

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 evaluate the ACTION, not 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 – but please try to ‘simulate’ with test data what is most likely going to happen.