INFO 7390 Midterm Case studies

- Answer all questions
- Submit an executive report(in MS Word) with your detailed analysis, explanation and interpretation of your analysis
- Deadline: Midnight – 3/18/2015
- You should include
    - One report summarizing all problems in WORD format
    - Share your data files and code through googledrive with analyticsneu@gmail.com
    - Slide deck for all the problems
- On 19th, each team will be asked to present one problem randomly for 5 minutes. Be prepared to present any of the problems

West Roxbury dataset (WestRoxbury.xlsx) (20 points)

You are given a dataset and are expected to build a model to compute the expected value of a home.

- Perform exploratory data analysis using TABLEAU
- Build prediction models for expected value of home using Regression, CART and Random forests
- Discuss the model and evaluate the model's performance.
- Which model would you use and why?
- Write an executive report to your management team on your model and recommendations

Mortgage Defaults (MortgageDefaulters.xlsx) (20 points)

Review MortgageDefaulters.xlsx. You are given a dataset of mortgage characteristics and whether a loan defaulted or not (See OUTCOME).

- Perform exploratory data analysis using TABLEAU
- Build classification models using Logistic regression, CART and Random forests
- Discuss the model and evaluate the model's performance.
- Which model would you use and why?
- Write an executive report to your management team on your model and recommendations

Detecting Spam (spambase.xlsx) (20 points)

Detecting Spam E-mail (from the UCI Machine Learning Repository). A team at Hewlett-Packard collected data on a large number of e-mail messages from their postmaster and personal e-mail for the purpose of finding a classifier that can separate e-mail messages that are spam versus nonspam (a.k.a. "ham"). The spam concept is diverse: It includes advertisements for products or websites, "make money fast" schemes, chain letters, pornography, and so on. The definition used here is "unsolicited commercial e-mail." The file Spambase.xls contains information on 4601 e-mail messages, among which 1813 are tagged "spam." The predictors include 57 attributes, most of them are the average number of times a certain word (e.g., mail, George) or symbol (e.g., #, !) appears in the e-mail. A few predictors are related to the number and length of capitalized words.

- Perform exploratory data analysis using TABLEAU

- Partition the data into training and validation sets; then perform a Logistic regression, cart and Random forests on the predictors.
- If we are interested mainly in detecting spam messages, is this model useful? Use the confusion matrix, lift chart, and decile chart for the validation set for the evaluation.
- Discuss the model and evaluate the model's performance.
- Write an executive report to your management team on your model and recommendations

Blog feedback problem (40 points)

https://archive.ics.uci.edu/ml/datasets/BlogFeedback

Data Set Information:

This data originates from blog posts. The raw HTML-documents of the blog posts were crawled and processed. The prediction task associated with the data is the prediction of the number of comments in the upcoming 24 hours. In order to simulate this situation, we choose a base time (in the past) and select the blog posts that were published at most 72 hours before the selected base date/time. Then, we calculate all the features of the selected blog posts from the information that was available at the base time, therefore each instance corresponds to a blog post. The target is the number of comments that the blog post received in the next 24 hours relative to the base time.

In the train data, the base times were in the years 2010 and 2011. In the test data the base times were in February and March 2012. This simulates the real-world situation in which training data from the past is available to predict events in the future.

The train data was generated from different base times that may temporally overlap. Therefore, if you simply split the train into disjoint partitions, the underlying time intervals may overlap. Therefore, you should use the provided, temporally disjoint train and test splits in order to ensure that the evaluation is fair.

- Build prediction models for expected value of home using Regression, CART and Random forests
- Discuss the model and evaluate the model's performance.
- Which model would you use and why?
- Write an executive report to your management team on your model and recommendations
- See attached paper for more information
- Feel free to use R, Python, Wrangler, XLMiner etc for this exercise.