PQHS 471 (2018) Final

Due by the **midnight of Friday, May 4, 2018**.

Goal: Evaluate the performance of machine learning and data mining methods using the Caravan Insurance data. The data have been used in a data mining competition. Our purpose is a little different. Details are below.

Data: Data and basic description: http://kdd.ics.uci.edu/databases/tic/tic.html . The data dictionary contains all the variables names. For your convience I created a file for all the names so that you can read them in and assign them as the variable names for the data. There are 86 variables. The last variable is the outcome.

These are the filed needed:

- ticdata2000.txt: Training data (86 columns)
- ticeval2000.txt: Test data (85 columns)
- tictgts2000.txt: Outcomes for the test data
- varnames: 86 variable names (on course website)

For more description and links about this dataset, go to the following page and search "Caravan": http://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html

What to do:

1. (80 points) Supervised learning: Apply random forest, boosting, and support vector machine to the data. For each method, build the best model (as best as you can) to predict if a person would buy a caravan insurance policy. Then compare the performance of the 3 best models. Try to add neural network to the list of methods.

2. (20 points) Unsupervised learning: Apply K-means, hierarchical clustering to cluster the data, and MDS to visualize the data. This part is more open because there is no definite answer. Use only the variables 6-41. Can you cluster your training data into subsets? Which method performs better for the data?

What to turn in: (electronic copies through email)

1. Your answers in either PDF or Word, with tables or figures if necessary. There is no need to make the tables or figures in publication quality, but they should have enough labels so that I can follow.

2. R code, with comments to tell me what a line or section is about. Before turning in your code, start a new session and run through your code to make sure there are no error messages.