Eco 5385 Prof. Tom Fomby
Data Mining Techniques for Economists Spring 2010

SUPER PROBLEM

The **Super Problem** represents 10% of your grade. You are to work on the problem **in teams**. See the teams and their memberships below. The first person on the team is designated as the team captain and is responsible for getting the team together to start work on the project. I will be e-mailing the team captain the e-mail addresses of the team members so that a meeting can be coordinated. This project is due on Monday, April 5 at the time of our class.

The Super Problem is in the form of a **Prediction Case Study**. The data to be analyzed is an artificial data set with 1,000 cases, one target variable labeled Y, and 10 input variables labeled X1 – X10. The data is contained in the EXCEL spreadsheet "Superprob.xlsx" that is available on the class website. Using XLMiner and the random number seed 12345, you should partition the data into 50% training data, 30% validation data, and 20% test data. Each team is responsible for writing a paper having a main body of no longer than 10 pages. (Appendices containing tables and graphs can be added to the end of the paper but keep the appendices to no more than 5 pages.) In essence the paper is to describe the building and validation of **3 prediction models** for Y using some or all of the 10 inputs X1 – X10. The three prediction models are to be based on **best subset multiple linear regression, neural networks, and regression trees**.

In terms of evaluation, each group will be judged on the quality of the exposition of the paper (50%) and the accuracy of the best model (possibly an ensemble) in the test data (50%). So a team may have the most accurate predictive model of all of the groups **but if the exposition in the paper is poor** the grade could be a **C** (70%) on the project. So don't take the writing of the paper lightly!

Each team should write up its paper with the following outline in mind:

Outline of Paper:

I. Description of the Data and How the Data is Partitioned into Training, Validation, and Test data sets
II. Training and Selection of Model I and Its Performance in the Validation data set
III. Training and Selection of Model II and Its Performance in the Validation data set
IV. Training and Selection of Model III and Its Performance in the Validation data set
V. Build an Ensemble Model using the Validation data set scores of the 3 models
VI. Apply the individual models I, II, and III as well as the ensemble model to the Test data set
VII. Conclusion: Which model has the greatest predictive accuracy in the **Test data set**? Model I, II, III or the ensemble model?

As supporting documentation, each team should provide 4 separate XLMiner spreadsheets – one for each of the prediction models, and one for the ensemble model building and validation exercise. If any team has any questions, the team captain should e-mail me at tfomby@smu.edu.

TEAM I

Alekseenko, Tatiana*
Baldwin, Charles
Bonner, Jason
Brister, Michael
Carter, Clayton

TEAM II

Carter, Jason*
Catalani III, Frank
Chen, Josephine
Crooks, Zachene
Crum, Christina

TEAM III

Gutierrez, Arturo*
Hamilton, John
Holmes, Amanda
Jensen, Astrid
Li, Jing

TEAM IV

Li, Xian*
London, Eric
Lopez, Julio
Mack, Adrienne
Malherbe, Julia

TEAM V

Marcelius Jr., Georges*
Nasir, Sumaiya
Odom III, Elzie
Ramachandran, Meenakshi
Reedy, Jamie

TEAM VI

Shi, Rui*
Tran, Man
Wilkerson, Bryan
Wong, Graham
Wood, Diann
Zorsky, Joshua

* Captain of team