Math 505 Final Project

You may work in groups of up to three people, and each group should turn in a single, professional quality report detailing your findings. At the end of your report, provide a breakdown of what each group member did. Each member is expected to contribute significantly to the project.

1. The goal of this project is to build a statistical model for predicting some variable Y from one or more variables $X_1, \ldots, X_p$ based on real data. It is preferred that you have multiple predictor variables X_j.

2. Your report should start with an overview of the real world aspects of the problem, including a description of the variables. Why are you interested in predicting Y, and why is it reasonable to expect that such predictions can be made from the X_j's?

3. Your goal is to build the best model possible, and you should assess each model with appropriate diagnostics. You should do your best to correct any deficiencies before settling on a final model. For instance, if you build a linear model, the error term assumptions, functional form, and overall goodness of fit should be assessed, and any problems should be addressed, possibly through transformations or adding additional terms.

4. It is preferred, though not required, that you build multiple models. If Y is quantitative, you could try a linear model and a nonparametric model, based on a technique such as LOWESS. If Y is binary, you could use logistic regression and discriminant analysis. Feel free to use methods from outside of the course, such as data mining techniques, or to create your own model from scratch, using a "brute force" or "common sense" approach.

5. The adequacy of each model should be assessed with cross-validation, using mean square prediction error if Y is quantitative and classification accuracy percentage if Y is categorical. If multiple models were built, they should then be compared on this basis and on the basis of the diagnostics.

6. Your report should end with a realistic assessment of the practical value of the models obtained. Is your model capable of making predictions that are accurate enough to be of value? How does your model compare to models constructed by others for the same problem? This ties in nicely with cross-validation, and these issues can be combined into one section.