IE 583 Homework 1

*This homework is due on **Tuesday February 6th**, by the end of the day (midnight).*

Data

This assignment uses the attached dataset of credit risk, called "credit-g". Background information on the data can be found by following the link below:

- https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)

This is a popular dataset in the data mining world, so a Google search will reveal many results – of decisively mixed quality. Its fine to look at those, just remember that they may be wrong!

Assignment

Use R and the caret package to build a good classification model to predict credit risk.

Try all three of the algorithms discussed so far, that is:

- Decision trees (use "J48" in caret)
- Naïve Bayes (use "nb" in caret)
- kNN (use "knn" in caret)

As you start experimenting with caret you will see that it has many more sophisticated learning algorithms available. It's tempting to try them but do **not** use those yet! This assignment is not about finding the best learning algorithm to be used with this data. We'll get there soon enough.

What you **should do** is the following. For each of the three simple methods you should determine how to tune and test the method:

- **Tuning**. Experiment with the parameters of the induction algorithm and determine a good setting for the parameters. Can you determine the best parameters for each learning algorithm when applied to this dataset? Make sure you vary all available parameters.
- **Testing**. Experiment with different methods for estimating the error. Specifically try both using an independent test dataset and cross-validation. For cross-validation try at least three different numbers for the number of folds. Can you determine the best method for estimating the error for each learning algorithm when applied to this dataset?

Report

Your report should include an explanation of why you believe your parameter values are good and your assessment of which of the three models should be used. Also you should explain which method for estimating the error is best and why. Make sure that you pay close attention to the type of errors made by your model when making the comparison. Note that it is not necessary to turn in all of your outputs – restrict it to those that add to the discussion. There is no fixed page number for your report.