Stat 6021: Project 2 Proposal

Due November 4

The dataset chosen is from UCI Machine Learning Repository (wine). The data set is meant for binary class classification to predict whether the quality of wine is good or bad. Our client’s main concern is that they can enjoy good quality wine. There are a total twelve of attributes, consisting of 1 categorical and eleven numerical types and are named as follows:

1 - fixed acidity (numerical)  
2 - volatile acidity (numerical)  
3 - citric acid (numerical)  
4 - residual sugar (numerical)  
5 – chlorides (numerical)  
6 - free sulfur dioxide (numerical)  
7 - total sulfur dioxide (numerical)  
8 – density (numerical)  
9 – pH (numerical)  
10 – sulphates (numerical)  
11 – alcohol (numerical)  
Output variable (based on sensory data):  
12 - quality (score between 0 and 10) - **$quality$ >= 6.5 => "good" , $quality$ < 6.5 => "bad"**

1. The data needs to be processed by data cleaning, data outlying check and handling of missing data. Adopt null/hypothesis tests to identify statistical significant predictors.
2. Start the model building process. The first model considers is the one which evaluates based on the client’s main concerns (our client wants to taste the best wine and less predictors). The relevant R output and graphical summaries will be offered.
3. The initial model considered: since our client’s primary concern is to ensure the model will predict whether the wine is high quality or not. The objective is to apply Logistic Regression on the data to split test and train sets.  Receiver Operating Characteristic (ROC) curve and the Area under the ROC Curve (AUC) is used to assess how well our logistic regression model performs in classifying outcomes. Apply confusion matrix to adjust threshold to see how the True positive and false negative varied.
4. Summary of findings: our recommended model(s) including suggestions for fine-tuning the model(s) and further development.
5. Summary of findings for the client: addressing client's main concerns and why we are recommending a model(s).