STAT 101A Final Project Final Submission is on Friday December 15 @ 11:59 PM

Analysis of Wine Quality Data

In the second example of data mining for knowledge discovery we consider a set of observations on a number of red and white wine varieties involving their chemical properties and ranking score by tasters. Wine industry shows a recent growth spurt as social drinking is on the rise. The price of wine depends on a rather abstract concept of wine appreciation by wine tasters, opinion among whom may have a high degree of variability. Pricing of wine depends on such a volatile factor to some extent.

Another key factor in wine certification and quality assessment is physicochemical tests which are laboratory-based and takes into account factors like acidity, pH level, presence of sugar and other chemical properties. For the wine market, it would be of interest if human quality of tasting can be related to the chemical properties of wine so that certification and quality assessment and assurance process is more controlled.

Two datasets are available of which one dataset serves as your training data set and it contains 7000 observations on red and white wine and have 12 different predictors (including Wine Color) and the other serves as your testing data set and it contains 3000 observations on red and white wine and have the same 12 predictors but without the values of the response variable (QualityNew). All wines are produced in a particular area of Portugal. Data are collected on 12 different properties of the wines one of which is Quality, based on sensory data, and the rest are on chemical properties of the wines including density, acidity, alcohol content etc. All chemical properties of wines are continuous variables. Quality is a continuous variable with possible scoring from 1 (worst) to 10 (best). Each variety of wine is tasted by three independent tasters and the final score assigned is the median score given by the tasters.

The main task is to use the training data set to get yourself familiar with the data and create your "best" valid predictive model. Then use this model to predict the wine quality in your testing data and you will get a score on how well you have predicted the wine quality relative to their true value (only I have these values).

- You are allowed to use polynomials of your predictors.
- You are allowed to transform the variable and you are allowed to used any technique discussed in class.
- You are allowed to delete cases that you think are going to have a bad influence on your predictions.
- You are not allowed to use any libraries or R modeling functions not discussed in this class.
- You are restricted on using only the following MLR functions lm or lm.beta functions and none of the data mining packages or libraries which are beyond the scope of the class.

Objective of the Analysis

- Prediction of Quality score from the chemical properties of the wines
- A predictive model developed on this data is expected to provide guidance to vineyards regarding quality score and price expected on their produce without heavy reliance on volatility of wine tasters.

The following analytical approaches is taken:

Multiple regression: The response Quality is assumed to be a continuous variable and is predicted by the independent predictors, all of which are continuous expect for the Wine color as Categorical (White or Red Wine).

Observations regarding variables: All variables have outliers

Attribute Information:

For more information, read [Cortez et al., 2009].
Input variables (based on physicochemical tests):
1. Wine Color: R (Red) or W (White)
2 - fixed acidity
3 - volatile acidity
4 - citric acid
5 - residual sugar
6 - chlorides
7 - free sulfur dioxide
8 - total sulfur dioxide
9 - density
10 - pH
11 - sulphates
12 - alcohol

Output variable (based on sensory data):
13 - Quality (score between 0 and 10)

Submitting

- You will be asked to submit predictions in a particular format: a CSV file with exactly two columns. The first must be called ID and contain the original ID numbers (or Case Number) of the values to be predicted, and the second must have the same name as the response variable (to be given on Friday).
- Once submitted, your predictions are compared to the true values and you will be given a score based on your rank in your class and it is relative to a simple model done by the professor or your TA.
- Make sure that you keep your values of the predicted quality in the same units. So if you decided to transfer your response variable in your model, you need to transform the predicted values back to their original units. Otherwise your score is going to be negatively affected.
- Your true score is going to be posted on ccle (twice before your final submission) based on how well your predictive model performs.
- You are allowed to submit up to three models. One on week 9, one on week 10 and your final submission is due **Friday December 15 @ 11:59 PM**
- You are required to submit a written report (PDF format) that tells your story on how you came up with your final predictive model (no R codes, graphs are OK). Due on **Friday December 15 @ 11:59 PM. (Max is 5 pages)**
- You are required to submit a copy of your final model R codes (RDM format) that should run once I include the data file to check your final predictive model. Due on **Friday December 15 @ 11:59 PM**

Source:

Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez
A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal
@2009