Introduction

This is a brief analysis of Red Wine quality data. The accompanying code can easily be altered to analyze the White wine data. The approach taken was to:

1. Load the data
2. Perform EDA on the data set to address any obvious data issues
3. Perform a correlation analysis and heatmap
4. Perform a principal components analysis to view the multidimensional data in a 2-D plane
5. Build a Linear Regression classifier
6. Evaluate the performance of the model

Background

1. The data load was a simple process given the direct URL to the dataset. Thinking forward, this approach of loading data could be supported easily in a “production environment” given a frequency of data refreshing. Python also has the ability to pull from API and database infrastructures, thereby facilitating any data loading routines. This data is very text book and provides very good guidelines for “clean” data, i.e. lack of missing data and all data types are continuous, although the outcome of “quality” may be viewed as categorical.

Approach

1. The red wine dataset is composed of 1599 observations and 12 variables. Exploratory data analysis was performed to identify any initial properties of the data, the breakdown of Wine Quality scores is shown below:



We can see that the Quality score is heavily centered, suggesting that differentiating between 5 and 6 is not obvious. Next, we examined all possible X-Y plots to examine any direct correlations with the Quality score:



There are many strong correlations among the Data inputs, for example we see that Density is highly correlated with Residual Sugar:



It is worth noting that the most correlated variables with quality are “Alcohol” and “Volatile Acidity”:

|  |  |
| --- | --- |
| **Variable** | **Correlation W Quality** |
| ﻿fixed Acidity | 0.124 |
| volatile Acidity | -0.391 |
| citric Acid | 0.226 |
| residual Sugar | 0.014 |
| chlorides | -0.129 |
| free Sulfur Dioxide | -0.051 |
| totalSulfur Dioxide | -0.185 |
| density | -0.175 |
| pH | -0.058 |
| sulphates | 0.251 |
| alcohol | 0.476 |

A heatmap demonstrates the strength of all relationships amongst one another, we see the strong correlations of Volatile Acidity and Alcohol with Quality.



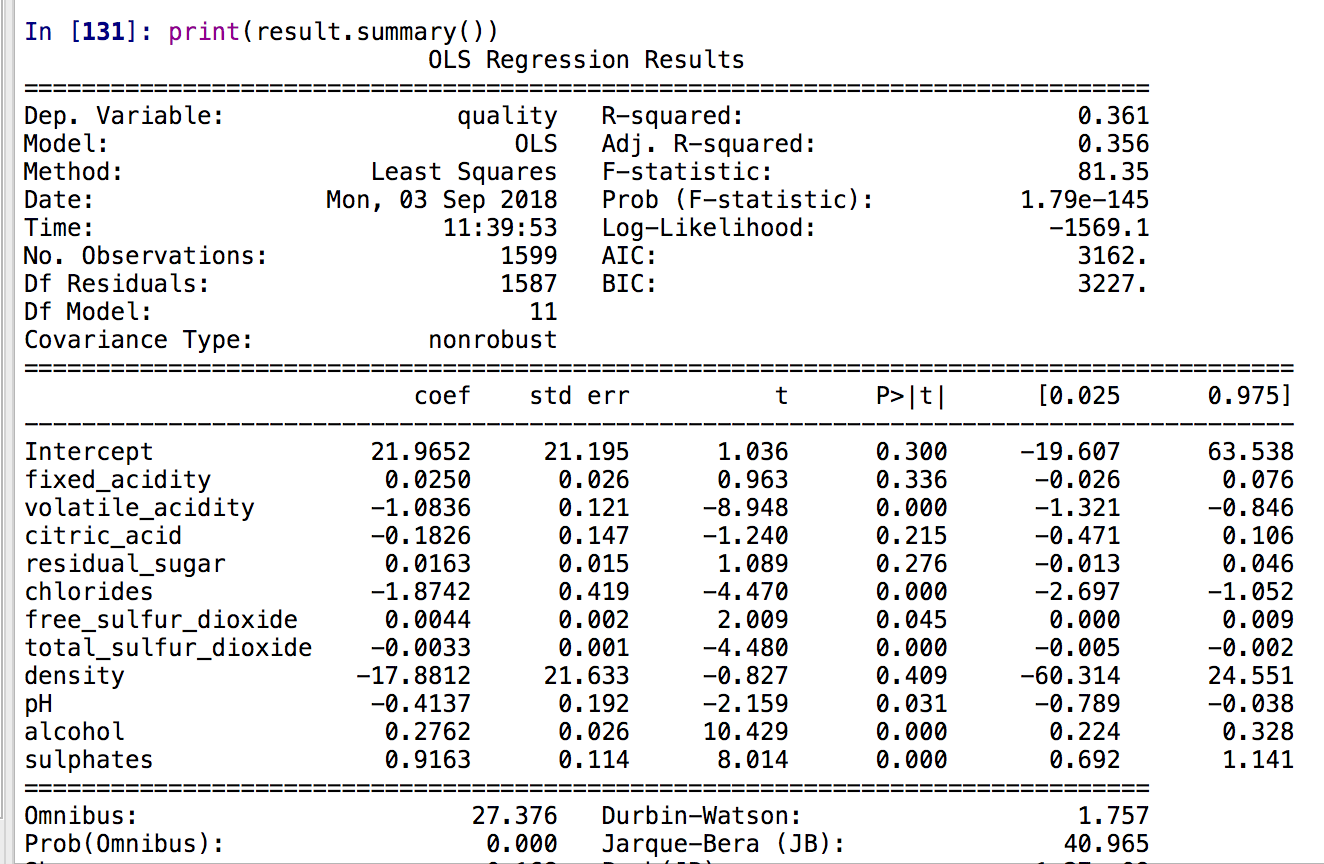


We performed a PCA analysis to reduce the dimensions of the input variables from 11 to 2, while retaining the maximum information from the inputs. It is worth noting that grouping and similarities of the wines with quality scores 5 and 6 (blue and yellow respectively). The PCA analysis summarize all inputs and applies differentiation to the quality scores, this analysis suggests there will be a challenge differentiating between 5 and 6.

1. Training Data:

We built a simple linear regression model to test that theory by partitioning the dataset into 2 sets with the training data composed of 75% of the initial dataset and the remaining 25% utilized for validation. We should note that the resulting model will likely be impacted by overfitting and collinearity, given proper time constraints we can apply a subset selection method to address these issues.

The results of the Training data set are below:



Amongst our variable we can see that Volatile Acidity, Total Sulfur, Alcohol and Sulfates are the most statistically significant as suggested by P-value. Next, we look at the difference between the projected values and actual values:



For this exercise we rounded the predicted value to compare with actual value, we see that the model correctly classified over 600 of the 1119 observations. This observation can be measure by the Mean Squared Error which is 0.427. In further models we wish to decrease this number which implies improved model accuracy

1. Test Data:

We validated the results utilizing the remaining 25% of the data, below are the results of utilizing the model to score the remaining data:



We see the model again correctly classifies over 50% of the data with a model MSE of 0.399. Both the training and test model achieved similar MSE’s of 0.4 suggesting the model is relatively robust despite the multicollinearity concerns.

Conclusion

1. The customer should experience a lift in properly classifying quality of wine, however as with most models there is room for improvement, namely alternative statistical methods such as a Decision Tree or Support Vector Machine may offer some improvement over this “base” model. It is worth noting that with any statistical model, that model will have limitations given the input data and the behavior of the outcome, “quality” in this case. Perhaps revisiting the definition of quality and applying stricter scores will help the overall progress and performance of statistical modeling. In addition, any additional input data that the end-user feels is predictive should be considered as well.

Not Surprisingly we see that Volatile Acidity and Alcohol content appear to be strong indicators of Wine Quality, as suggested by the correlation analysis and the regression analysis. In further iterations it is advised to reduce the variables to avoid overfitting the model, in fact reducing the model to 3 variables produced a similar MSE of 0.444 in the training data set suggesting there is definitely room for improvement in the statistical approach taken.