Red Wine Quality

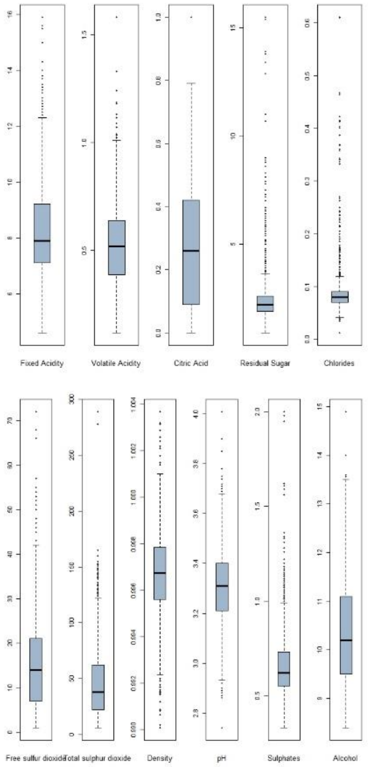
Summary: Red wine is a type of wine made from dark-colored (black) grape varieties. Wine industry shows a recent growth spurt as social drinking is on the rise. The price of wine depends on a rather abstract concept of wine appreciation by wine tasters, opinion among whom may have a high degree of variability. 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. 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 categorical variable with possible ranking from 1 (worst) to 10 (best). Each variety of wine is tasted by three independent tasters and the final rank assigned is the median rank given by the tasters. The main purpose of this analysis is to determine which physiochemical properties makes a 'good' wine. To predict good wine a new binary and categorical variable was created where **quality >=7** based on various given parameters alcohol, sulphates, chlorides, pH etc.

The target audience for this project is for wine producers to increase sales in countries like China where wine market is growing. The Chinese like to drink wine. China's millionaires, of which there are beau coup, don't think twice about dropping $300 on a bottle of wine. And, since China's economy looks like it's emerging from its recent doldrums, wine consumption is on the upswing. Not only is per capita consumption rising, but the way the Chinese drink wine is changing.

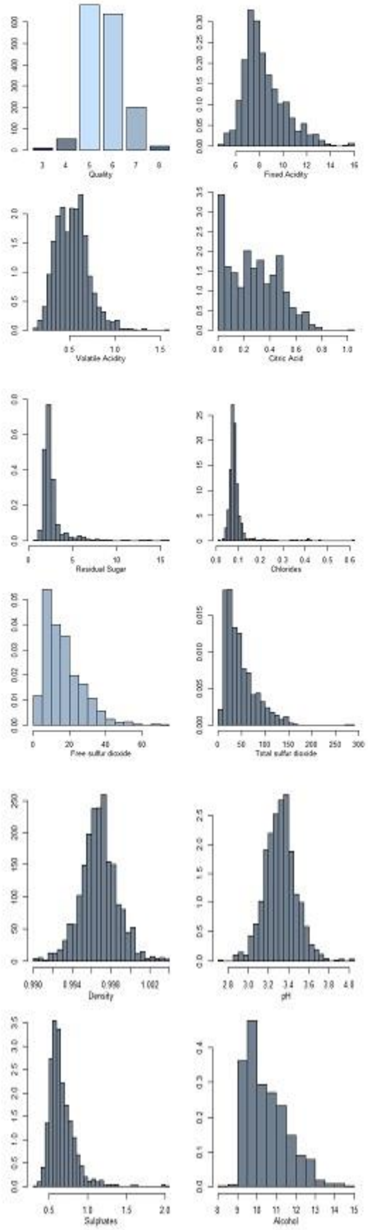
Exploratory Data Analysis: Objective of the analysis is Prediction of Quality ranking from the chemical properties of the wines. A predictive model developed on this data is expected to provide guidance to vineyards regarding quality and price expected on their produce without heavy reliance on volatility of wine tasters. All variables are summarized and univariate analysis with plots. All variables have outliers.

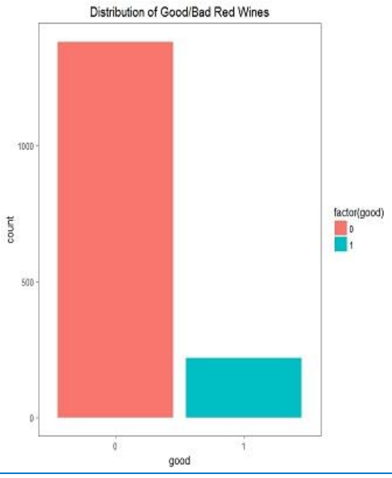
- Quality has most values concentrated in the categories 5, 6 and 7. Only a small proportion is in the categories [3, 4] and [8, 9] and none in the categories [1, 2] and 10.
- Fixed acidity, volatile acidity and citric acid have outliers. If those outliers are eliminated distribution of the variables may be taken to be symmetric.
- Residual sugar has a positively skewed distribution; even after eliminating the outlier's distribution will remain skewed.
- Mostly outliers are on the larger side. Alcohol has an irregular shaped distribution, but it does not have pronounced outliers.
- Some of the variables, e.g. free sulphur dioxide, density, have a few outliers but these are very different from the rest.

Boxplots for each of the variables as another indicator of spread.

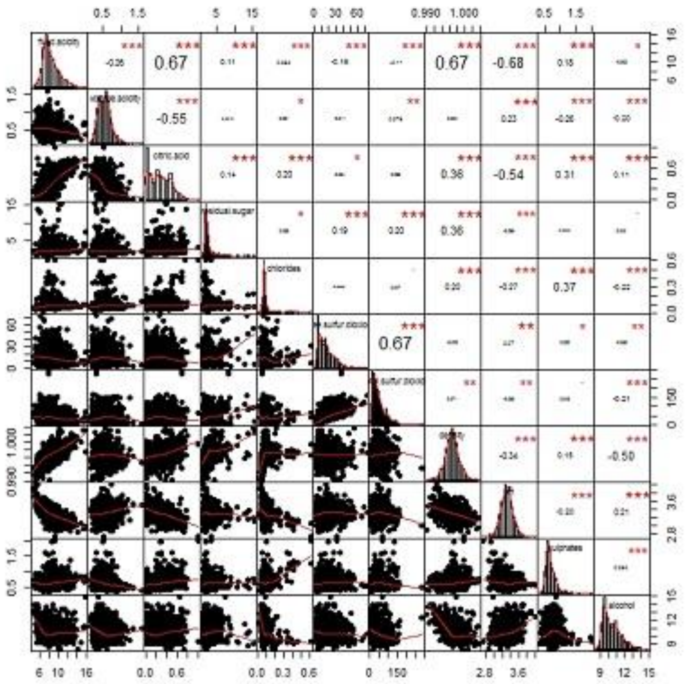


Histograms *to show the distribution of the variable values.*

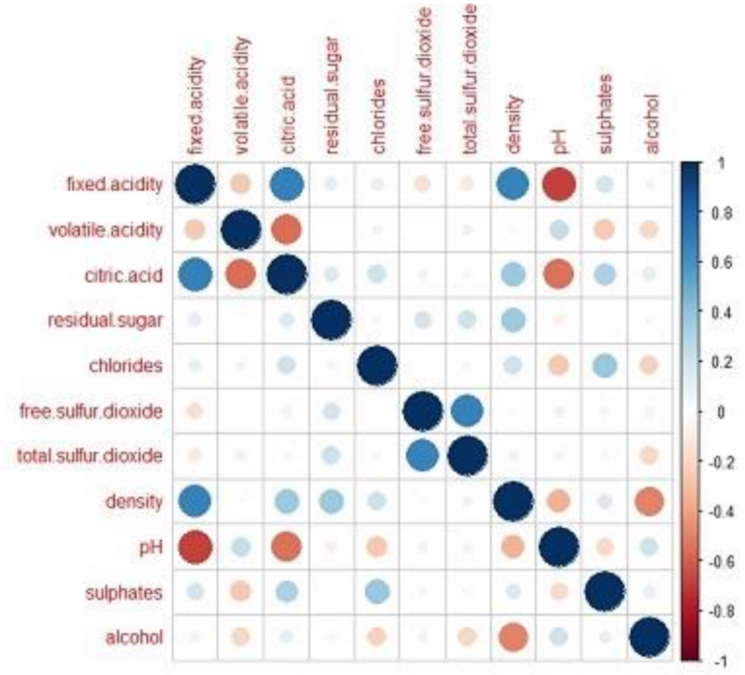




Correlation matrix chart



Visualizing the correlation matrix



Correlation: Highly correlated features provide redundant information. By eliminating highly correlated features we can avoid a predictive bias for the information contained in these features. Pearson's correlation and Spearman rank correlations were performed, and values looks close, hence only Pearson's correlation is considered. Correlations between all features are calculated and visualized as in figure. Correlations (≥ 60% in absolute value) are identified and highlighted in diagram. Since linear relationship among variables ≤ 70%, which doesn't establish any strong relationships, all variables were accounted for modeling process.

Mining techniques: For Prediction of Quality ranking from the chemical properties of the wines, three data mining techniques has been employed in this project. The first method is *Binomial Logistic Regression* – predicts the probability that an observation falls into one of two categories of a dichotomous dependent variable based on one or more independent variables that can be either continuous or categorical. The second method used is *Decision Tree* - a method of supervised learning that can be used for either classification, or regression problems. They're represented by diagrams, where the end of each line is called a leaf, or a terminal node. And these represent the predictions. The places where one-line splits into two are called the internal nodes. And these divide up the predictor space. The third method used is *Random Forest* - is a ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. The way random forests work is that at each split that we generate in a tree, we only consider a randomly chosen subset of the possible predictor variables.

Why selected data mining techniques are appropriate: A frequent problem in data mining is to avoid predictors that do not contribute significantly to model prediction. As part of binomial linear regression predictor thinning was done by stepwise selection methods. Such methods are aimed to thin the original number of predictors to a manageable size. We get a set of most useful predictors for the response, which is particularly helpful in achieving an accurate estimation with collinear or correlated predictors. In case of tree based method, our goal is to find the variable / split that best separates the response variable, which yields the lowest Residual Sum of Squares (RSS). Decision trees are very similar, so their predictions will be highly correlated. In a random forest, the one variable simply wouldn't be an option at some of the splits. So, some of the trees would have to have their first split based on some other variable. This leads to greater differences between the trees and lower correlation between their predictions. In this dataset alcohol is not only an option of split but some of the tree split based on sulphates and volatile acidity also. As part of random forest predictor thinning was done by importance method to rank predictor variables for model creation.

Model Selection: For the model selection, cross validation was employed, where for every run, different variables were included in model and accuracy was computed. Performed an outer level of cross-validation around inner cross-validation (the "double" cross-validation) to properly assess the model-fitting process. Then performed (one-level) cross-validation to select the best model, out of all considered methods. Lastly took the best model, and fit it on the full data set (all data) and get the final fitted model.

Conclusion: Comparing the accuracy of the above three data mining technique employed, the Random Forest model is found to be the one with least error and chosen to be the best technique for predicting the Quality ranking from the chemical properties of the wines.

References:

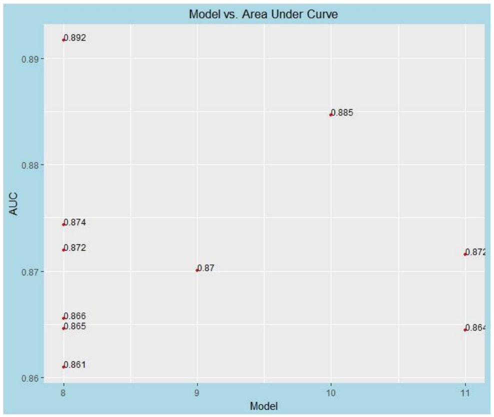
Source Dataset:
https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009

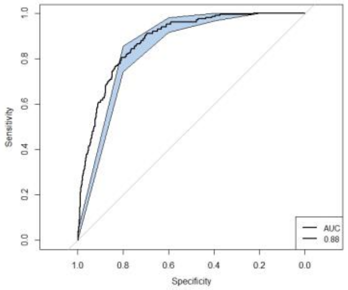
Input variables (based on physicochemical tests):
1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol
Output variable (based on sensory data)
12 - Quality(score between 0 and 1)

Binomial Logistic Regression

Model	AUC
8	0.860971
11	0.864443
8	0.864546
8	0.865556
9	0.870071
11	0.871547
8	0.871953
8	0.874346
10	0.884666
8	0.891674

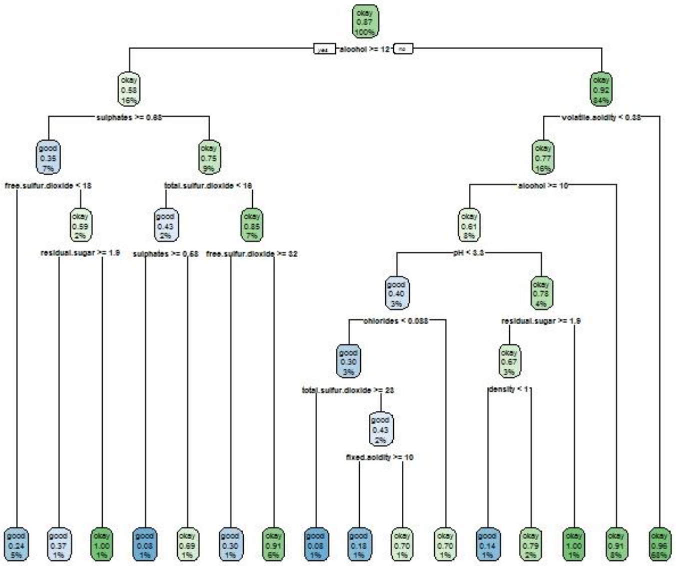
Area Under Curve based on "double" CV

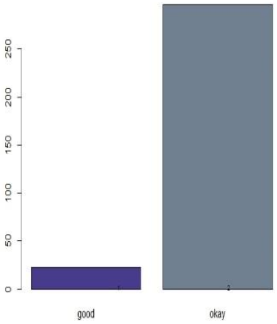




Accuracy is measured by the area under the ROC curve. An area of 1 represents a perfect test; an area of .5 represents a worthless test. As per rough guide for classifying the accuracy of a diagnostic test, Model 8 and Model 10 shows good accuracy. However best model was chosen based on highest accuracy which is Model 8. Predictor variables for model 8 are: alcohol + volatile. acidity + sulphates + total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar + density

Decision Tree





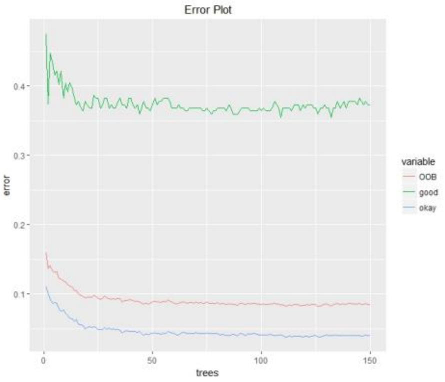
Use 10-fold cross-validation on the train ing data to choose the number of leaves that minimizes the classification error r ate and find out the optimal numbers of leaves. Create a pruned tree with 5 leav es. Error rate of the pruned tree on the validation set: 11%, Accuracy:88%

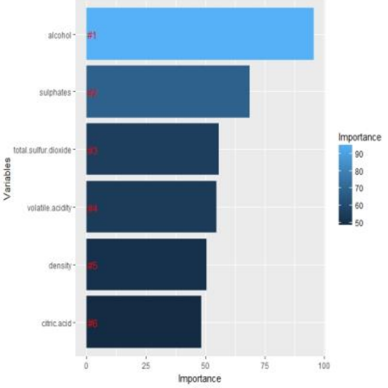
Misclassification error rate:10%, Accuracy:86%

Random Forest

Model	Accuracy
10	0.88125
9	0.9
11	0.9
10	0.9
5	0.9020833
5	0.903125
6	0.9044643
7	0.90625
11	0.9083333
6	0.910569

Model Accuracy based on "double" CV





Based on "double" cross-validation Model 6 and Model 11 shows good accuracy rate. However best model was chosen based on highest accuracy which is Model 6. Predictor variables for model 8 are: alcohol + sulphates + volatile.acidity + density + citric.acid + total.sulfur.dioxide.