**Modeling wine quality**

**By - Darpan Shah**

**BIA 652**

**Guided By -**

**Professor** [**Feng Mai**](https://sit.instructure.com/courses/38429/users/20030)

**INTRODUCTION**

According to *Business Wire* U.S. is **#1** importer of Vinho Verde wines not only in volume but value also, very slight difference with Germany for the top spot. Once viewed as a luxury good, nowadays wine is increasingly enjoyed by a wide range of consumers. Portugal is a top ten wine exporting country with 3.17% of the market share in 2018. Exports of its vinho verde wine have increased by 36% from 2007 to 2018.

Vinho Verde refers to Portuguese wine that originated in the historic Minho province in the far north of the country. The modern-day ‘Vinho Verde’ region, originally designed in 1908, includes the old Minho province plus adjacent areas to the south.

Vinho Verde is not a grape variety, it is a system of protected designation of origin for wines, cheeses, butters and other agricultural products from Portugal. The name means “green wine”, with wine being released three to six months after the grapes are harvested. To support its growth, the wine industry is investing in new technologies for both wine making and selling processes.

Certification prevents the illegal adulteration of wines (to safeguard human health) and assures quality for the wine market. Quality evaluation is often part of the certification process and can be used to improve wine making (by identifying the most influential factors) and to stratify wines such as premium brands (useful for setting prices).

**PROBLEM STATEMENT**

In this paper, we present a case study for modeling wine quality based on analytical data that are easily available at the wine certification step. Building such model is valuable not only for certification entities but also wine producers and even consumers. It can be used to support the oenologist wine evaluations, potentially improving the precision and speed of their decisions.

Moreover, measuring the impact of the physicochemical tests in the final wine quality is useful for improving the production process. Furthermore, it can help in target marketing i.e. by applying similar techniques to model the consumer’s preferences of niche and/or profitable markets.

The main highlight of the analysis:

* Physicochemical information about the wine is easily available before the production start, thus with the help of this analysis we would be able to see the impact of each component on final wine quality at the start of the process whether the wine is going to be good or bad.
* The variable selection is based on sensitivity analysis, which is a computationally efficient method that measures input relevance and guides the variable selection process.

**MODEL APPROACH**

**Regression Analysis:**

It is a statistical method that helps us to analyze and understand the relationship between two or more variables of interest. The process that is adopted to perform regression analysis helps to understand which factors are important, which factors can be ignored and how they are influencing each other.

**Types of Regression:**

There are different types of regression analysis like linear regression, polynomial regression, logistic regression, discriminant analysis. Each with their own assumptions and methodology. For this dataset I have tried different regression approaches like Lasso, Ridge and Logistic Regression. And ultimately found the logistic regression best fits the dataset.

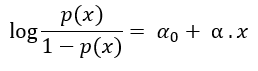
**Logistic Regression:**

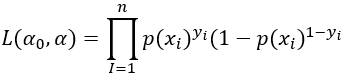
A statistical model typically used to model a binary dependent variable with the help of logistic function. It establishes a relationship between dependent and independent variables. Another name for the logistic function is a sigmoid function and is given by:

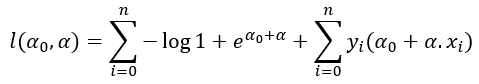
This function assists the logistic regression model to squeeze the values from (-k, k) to (0, 1). Logistic regression is majorly used for binary classification tasks; however, it can be used for multiclass classification.

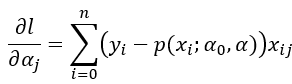
The reason behind this model is that just like Linear Regression, logistic regression starts from a linear equation. However, this equation consists of log-odds which is further passed through a sigmoid function which squeezes the output of the linear equation to a probability between 0 and 1. And, we can decide a decision boundary and use this probability to conduct classification task.

**Math behind Logistic Regression:**

So it all start with a linear function p(x) and then using log function with p(x) we are able to bound this function to 0 to 1. So the function will be like

Since Logistic regression predicts probabilities, we can fit it using likelihood. Now the likelihood can be written as:

Further, after putting the value of p(x):

In order to increase the probability of occurring we can use Maximum Likelihood function and differentiating the equation with respect to different parameter and setting it to zero.

= 0

**Assumptions:**

* The dependent variable is categorical. Dichotomous for binary logistic regression and multi-label for multi-class classification
* Attributes and log odds i.e. log (p / 1-p) should be linearly related to the independent variables
* Attributes are independent of each other (low or no multicollinearity)
* In binary logistic regression class of interest is coded with 1 and other class 0.

**MODEL ACCURACY**

**Mean Squared Error**:

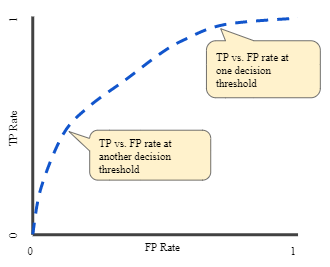
The mean squared error tells you how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line (these distances are the “errors”) and squaring them. The squaring is necessary to remove any negative signs.

**Confusion Matrix**:

It is a table that is often used to describe the performance of a classification model on a set of data for which the true values are known. This consist of four different parts:

* True Positive (TP) - These are cases in which we predicted yes, and that’s actually yes.
* True Negative (TN) - We predicted no, and it’s actually no.
* False Positive (FP) - We predicted yes, but they actually was no.
* False Negative (FN) - We predicted no, but actually it is yes.

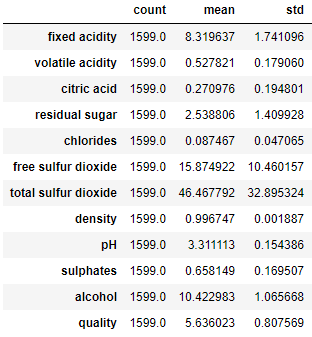
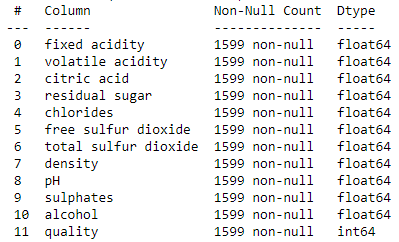
**ROC Curve and AUC:**

An ROC Curve (receiver operating characteristic curve) is a graph showing the performance of a classification model with the help of True Positive Rate and False Positive Rate at different classification thresholds. AUC stands for Area under the ROC curve measures the entire two-dimensional area underneath the entire ROC curve.

**EXPLORATORY DATA ANALYSIS**

**Data**

The dataset consist 1599 distinct rows each representing different test values with 12 variables in which “**Quality**” is the dependent variable (y) and rest are the independent variables (x).

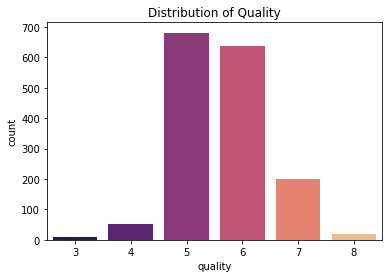
Let’s see some Summary Statistics:

* We can see that there are no null values in the dataset and the data type for all variables other than the dependent variable (quality) are float64 whereas data type for quality is int64.

Dependent Variable:

This is a variable which is used as a resultant variable in the model. This is predicted using all the independent variables. In our case dependent variable is “**Quality**” which has six distinct values going from 3 to 8, where 8 being the highest quality and 3 be the worse quality.

Now let’s check the distribution of the quality variable in our dataset:

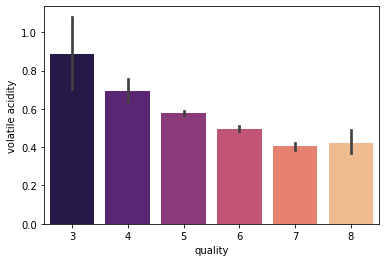
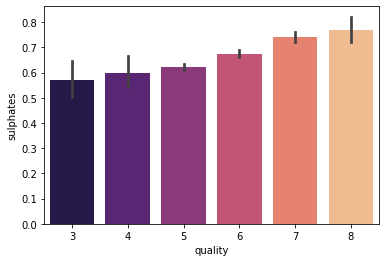
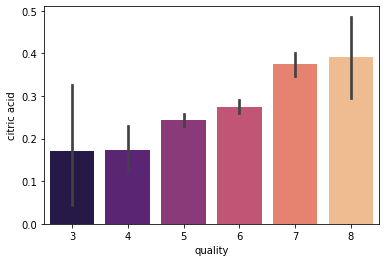


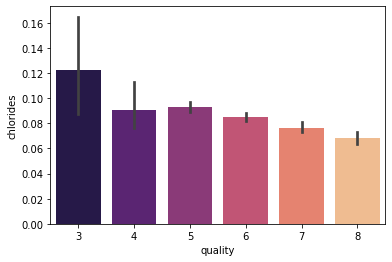
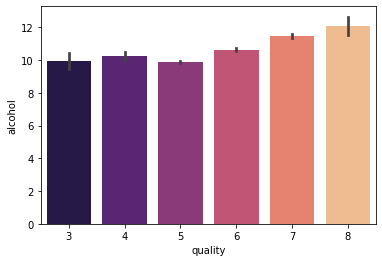
We could see that most of the data is concentrated between 5 and 7 quality. Hence later on we converted this variable into binary.

Independent Variable:

1. Fixed acidity (tartaric acid - g/dm3)
   * most acid involved with wine which are fixed or nonvolatile (do not evaporate readily)
2. Volatile acidity (acetic acid - g/dm3)
   * the amount of acetic acid in wine, which at too high level lead to an unpleasant, vinegar taste
3. Citric acid (g/dm3)
   * found in small quantities, citric acid can add ‘freshness’ and flavor to wines
4. Residual sugar (g/dm3)
   * the amount of sugar remaining after fermentation stops, it’s rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet
5. Chlorides (sodium chloride - g/dm3)
   * the amount of salt in the wine
6. Free sulfur dioxide (mg/dm3)
   * the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine
7. Total sulfur dioxide (mg/dm3)
   * amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
8. Density (g/dm3)
   * the density of water is close to that of water depending on the percent alcohol and sugar content
9. pH
   * describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
10. Sulphates (potassium sulphate - g/dm3)
    * a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant
11. Alcohol (% by volume)
    * The percent alcohol content of the wine

**Visualization:**

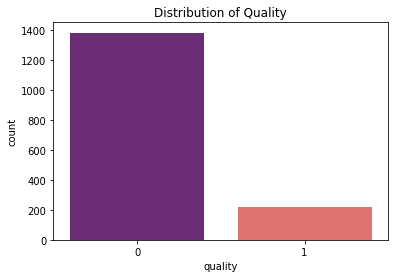
****Below we have independent variable named **Volatile Acidity**, **Critic acid**, **Chlorides**, **Sulphates** and **Alcohol,** that has a relation with the dependent variable **Quality.**

****

**PRE-PROCESSING**

**Transforming Variable:**

* According to exploratory data analysis we can observe that not all the categories of dependent variable (‘quality’) has sufficient amount of data. Thus it would lead to unnecessarily complex analysis.
* Thus it is suitable to change the variable to binary which is if the quality is greater than 6 it will be considered good otherwise bad.
* Hence the final distribution of the dependent variable (‘quality’) is:



Bad

Good

1,382

217

**Correlation Matrix:**

It is defined as the covariance between two variables divided by the product of the standard deviations of the two variables.

The value of lies between -1 and +1. Values near +1 indicate strong positive relation and -1 represents strong negative relation.

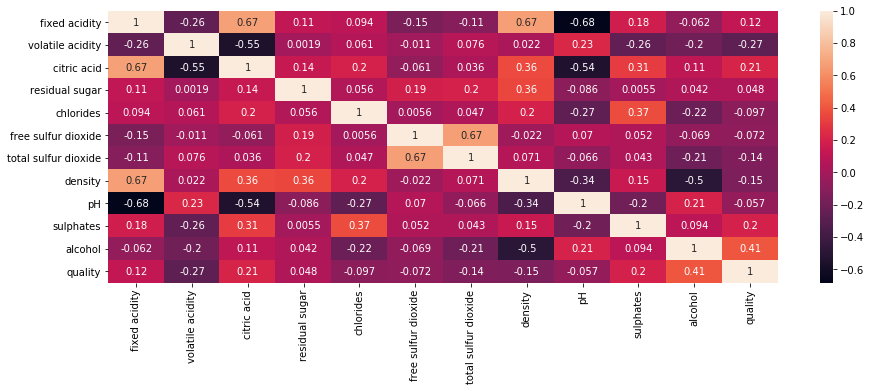
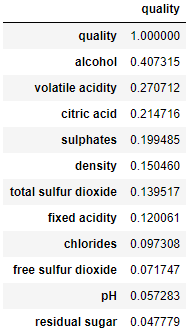
****

Figure - Correlation Matrix

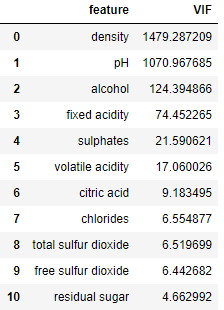
* From the above table we will find values which are greater than ±0.5 to remove variables which are highly correlated as it will generate multicollinearity.
* So we could see following pairs like (pH-fixed acidity), (critic acid-fixed acidity), (critic acid- pH), (total sulfur dioxide – free sulfur dioxide), (fixed acidity - density) have high correlation among each other hence must be treated or removed.

Correlation with Dependent Variable (‘quality’):

* From the table to the right we can observe the correlation strength of each variable with the dependent variable (‘quality’).
* Like alcohol has a correlation coefficient of 0.4 with quality so they are highly correlated.
* A high correlation will help the model to predict the correct value for the dependent variable (‘quality’).
* Thus we will keep all the variable which display a high correlation coefficient with the dependent variable.

**Variance Inflation Factor:**

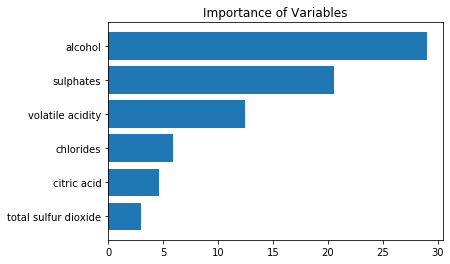
A variance inflation factor is basically a tool to help identify the degree of multicollinearity. Multicollinearity exists when there is a linear relationship, or correlation, between one or more of the independent variables or inputs.

VIF scores for our dataset variables:

* From the table to the right we can easily observe that variables like **density, pH, fixed acidity** must not to taken into consideration while building model as they have very high VIF score and thus would lead to multicollinearity.

**Variable Selection:**

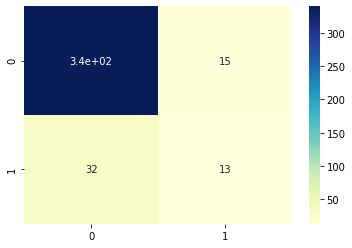
* Ultimately we have reached at the point where we could decide on which variables needs to be selected for model building.
* So using the above pre-processing techniques we conclude the list to be



* + Alcohol
  + Sulphates
  + Volatile Acidity
  + Chlorides
  + Citric Acid
  + Total Sulfur dioxide

**INFERENCE**

* The Mean Squared error is 0.1175 which is small and considered to be good, leads to a score of 0.88. Thus using the model we are 88% sure to predict the quality of the wine with the following parameters.
* To check the bracket of score I have used cross-validation matrix on score and got the bracket of 0.84 and 0.90. That shows our score is in the good range.
* Equation model is reached with the help of coefficient for each variable and using this equation we can grade any wine to know its quality.

**Confusion Matrix:**

This represents-

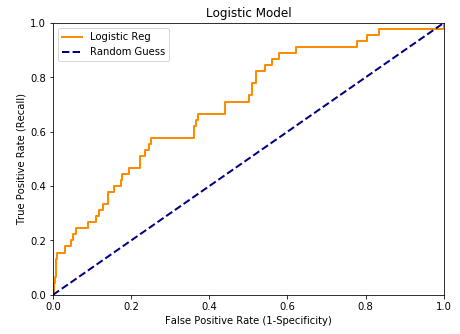
* True Negative – 340
* False Positive – 15
* False Negative – 32
* True positive – 13

By seeing this we can say that we have correctly predicted 353 (340+13) and 47 (15+32) wrongly predicted. Thus the accuracy score is 0.88 which is the score for the model.

**Classification Report:**

Using the classification report we have got more detailed analysis:

* Precision (91%), this means out of all values of quality we were able to predict 91% times correctly.
* Recall (96%), this means 96% of the actual values are predicted correctly.
* F1 score is 94% which is the average of precision and recall.

**ROC Curve and AUC:**

* Using True Positive and False positive rate we have the ROC curve on different classification threshold.
* AUC is 0.7 for this model that means overall we are able to classify differentiate the quality very well.

**RECOMMENDATION**

* Finally we are ready with the equation using the variables through which we can easily predict the quality of wine.
* This model can be made more useful if we are able to add the cost of false positive and true negative predictions. So we are able to optimize the cost and improve the values of one among them which is less costly using different classification thresholds.
* This model is specific to a single type of red wine, thus to make it more universally applicable it needs to be tested on different type of red wines produced in different region.
* Model is solely dependent on the features of wine but if we are able to include human preferences according to their test, it would have been more realistic and acceptable.

**REFERENCES**

[1] A. Asuncion and D. Newman. UCI Machine Learning Repository, University of

California, Irvine, http://www.ics.uci.edu/∼mlearn/MLRepository.html, 2007

[2] J. Bi and K. Bennett. Regression Error Characteristic curves. In Proceedings of 20th Int. Conf. on Machine Learning (ICML), Washington DC, USA, 2003.

[3] Fahrmeir, L., Kneib, T., Lang, S. and Marx, B., 2007. *Regression*. Springer-Verlag Berlin Heidelberg.

[4] Kleinbaum, David G., et al. *Logistic regression*. New York: Springer-Verlag, 2002.

[5] Ferri, C., Hernández-Orallo, J. and Modroiu, R., 2009. An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, *30*(1), pp.27-38.

[6] Lewis, K.P., 2004. How important is the statistical approach for analyzing categorical data? A critique using artificial nests. *Oikos*, *104*(2), pp.305-315.