Final Project

------Wine Classification with a Neural Network

Introduction

Our group focus on the wine quality datasets (Cortez et al., 2009), which include the red and white vinho verde wine samples from the north of Portugal. We are curious about what factors make two type of wine different and decide to classify the types of wine by analyzing physicochemical variables.

Table 1.1-Summary of Classifier

Type	Number	Relabel
Red	1599	1
White	4898	2

Before we analyzing the data, we combined the red wine and white wine datasets with the `rbind()` function in R and deleted the non-using column (quality). We then relabeled the type of the wine with `Red` as 1 and `White` as 2. From *Table 1.1* we can see there are 1599 samples of red wine and 4898 samples of white wine.

We then summarized our predictors. *Table 1.2* describes the summary statistics together with the units of variables. Specifically, there are three different types of acidity. Most acids involved in wine is fixed acidity, which does not evaporate readily. Volatile acidity is the amount of acetic acid in wine. Its high levels can lead to an unpleasant, vinegar taste. Citric acid is found in small quantities, but it can add ‘freshness’ and flavor to wines.

Besides acidity, residual sugar is the amount of sugar remaining after fermentation stops. Chlorides describe the amount of salt in the wine. Free sulfur dioxide represents the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion, while total sulfur dioxide is the amount of free and bound forms of S02. Sulphates is a wine additive which can contribute to sulfur dioxide gas (S02) levels, acting as an antimicrobial and antioxidant.

Table 1.2-Summary of Predictors

	Min.	1st Qu	Median	Mean	3rd Qu	Max.	Units
fixed acidity	3.8	6.4	7	7.215307	7.7	15.9	g(tartaric acid)/L
volatile acidity	0.08	0.23	0.29	0.339666	0.4	1.58	g(acetic acid)/L
citric acid	0	0.25	0.31	0.318633	0.39	1.66	g/L
residual sugar	0.6	1.8	3	5.443235	8.1	65.8	g/L
chlorides	0.009	0.038	0.047	0.056034	0.065	0.611	g/L
free sulfur dioxide	1	17	29	30.52532	41	289	mg/L
total sulfur dioxide	6	77	118	115.7446	156	440	mg/L
density	0.98711	0.99234	0.99489	0.994697	0.99699	1.03898	g/mL
pH	2.72	3.11	3.21	3.218501	3.32	4.01	none
sulphates	0.22	0.43	0.51	0.531268	0.6	2	g/L
alcohol	8	9.5	10.3	10.4918	11.3	14.9	percent

After conducting the basic summary of the predictors, we employed several methods to investigate the difference between two types of wine. We first created box plots for each predictor, and then focused on the ones that we were interested in. Later we searched for the correlation among variables and figure out the reason behind them. Also, we utilized the PCA and K-means clustering models on our combined datasets. Unfortunately, the results show a very high False-Positive Rate. You can read more detailed of the preliminary analysis research in the results and discussion part of this report. Therefore, in order to improve performance, we tried Neural Network in our analysis and received satisfactory results.

Exploratory Data Analysis

Figure 2.1-Alcohol

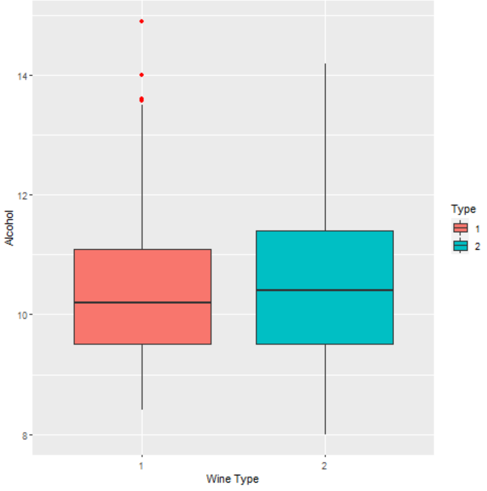


Figure 2.2-PH

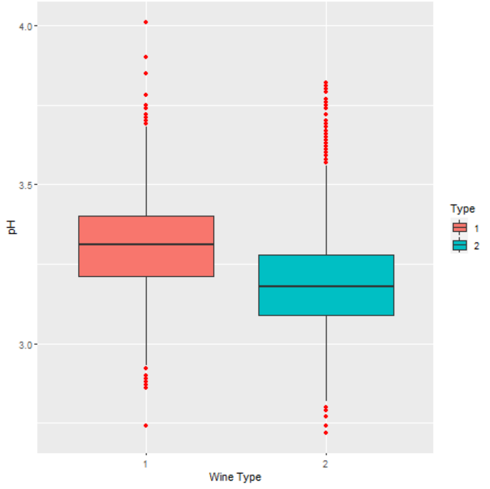


Figure 2.3-Residual Sugar

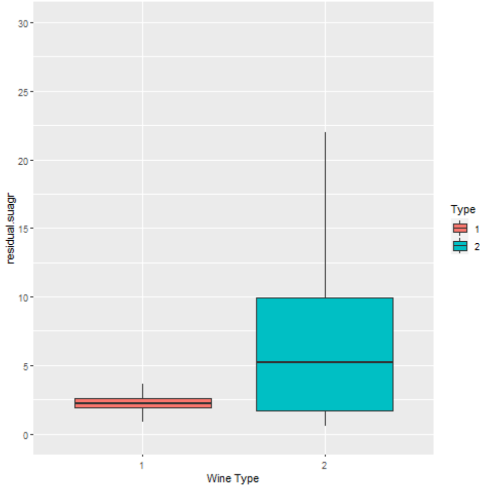
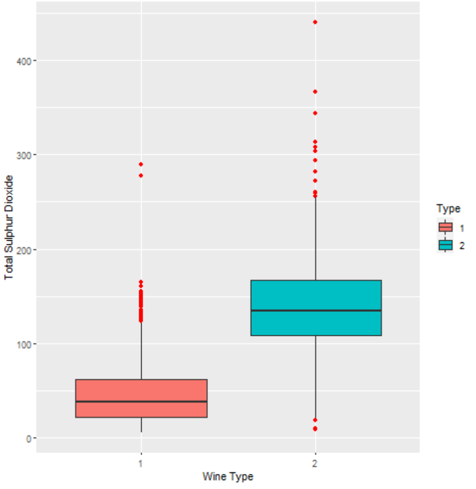


Figure 2.4-Total Sulphur Dioxide



In our preliminary analysis, we created box plots for the variables we thought would have different means/medians and spreads for Red and White wine. The results are shown in the figures above. We can see that the alcohol concentration is roughly the same across both wines. However, pH, residual sugar and Total SO_2 concentrations differ in spread and measures of central tendencies.

On further investigating why this might be, we found some pretty interesting results. Before we discuss them, we think it is important to know what these terms exactly mean.

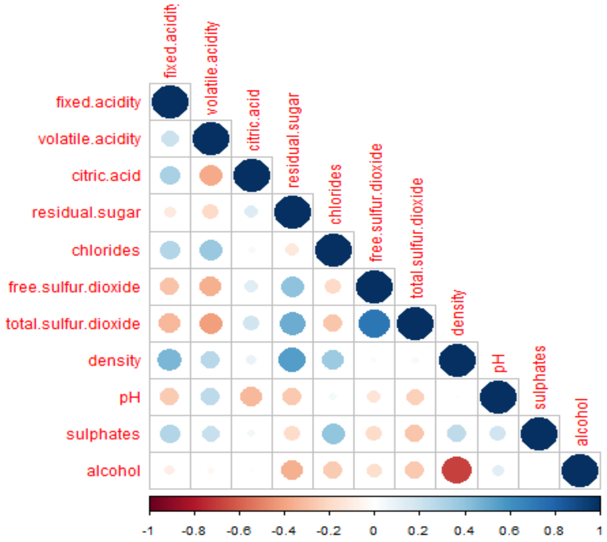
pH: It is a measure of expressing the acidity or alkalinity of a solution. It is measured on a log scale where the value of 7 is neutral and anything less than 7 is acidic and more than 7 is alkaline.

SO_2: Sulfur dioxide is added as preservative and prevents wine from turning into vinegar. It is also a natural by-product of yeast fermentation.

Residual Sugar: It refers to the natural grape sugars leftover after fermentation ceases.

Bringing your attention back to the results, we found that, in general, white wine has a lower pH and is more acidic than red wine. Furthermore, White wine has more residual sugar on average as they are usually fermented for a shorter period of time. The increased acidity helps balance the sweetness and gives the wine a fresher and crisper taste. If you are a wine connoisseur, you are probably familiar with the notion that red wines cause headaches because of their higher SO_2/sulfite levels. This is not true. As we can see from our data, The SO_2 levels for red wines is significantly lower than that of white. This is because red wines naturally contain tannin which is a stabilizing agent and serves the same functions as SO_2/sulfites. Thus, less SO_2 is needed to protect the wine during winemaking and maturation. Given these differences, we decided to make the goal of our project to classify a given sample of wine to determine if it is Red or White.

Figure 2.5-Correlation graph of predictors



Next, we investigated whether there was any correlation between the variables we were about to use as predictors. The `corrplot()` function in R helped us generate this graphic for our task. In the figure we can see that density and alcohol are highly correlated (negatively). There is a very simple reason for this.

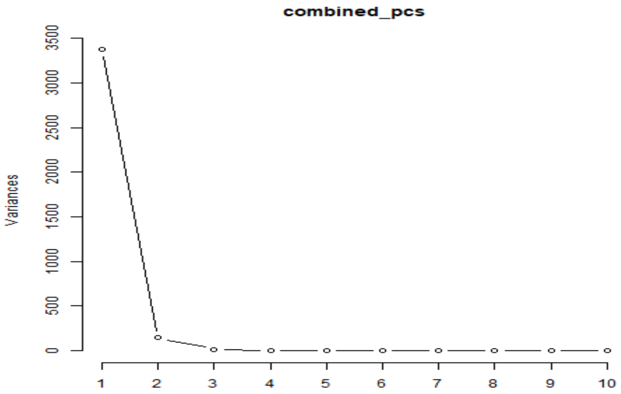
Density = concentrations of (Water + Sugar + Alcohol).

Water concentration = 100%(total concentration) – alcohol (concentration) - (sugar concentration)

The density of alcohol (ethanol) is 0.80 g/ml at 20C and that of water is 1 gm/CC. Therefore, more the alcohol concentration, the less the water concentration and thus, the density is closer to 0.8 g/ml. As a result, lower the alcohol concentration, higher the density (closer to 1g/ml).

PCA

Figure 2.6-Combinded PCs



Then, we decided to make use of the Principal Component Analysis to visualize whether there was some separation of the 2 classes. From the scree plot we can see that the first 7 PCs pretty much explain all of the variation in the data (99.999% of all the variation). So we decided to use the projection of our data on the first 7 dimensions. From the plots that follow, especially on the plots of PC1 vs PC6 and PC1 vs PC4, we can see that there is quite a bit of separation of the data points belonging to the 2 classes. This led us to run a K-means clustering algorithm on the PCs and see what the results were.

Figure 2.7-PC1 vs. PC6

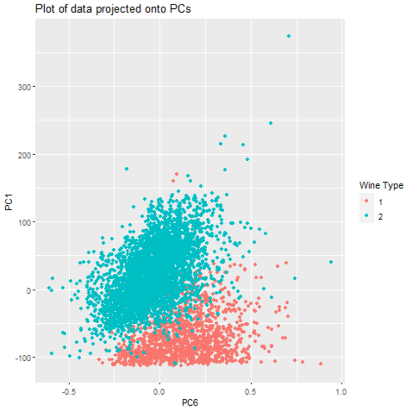


Figure 2.8-PC1 vs. PC6

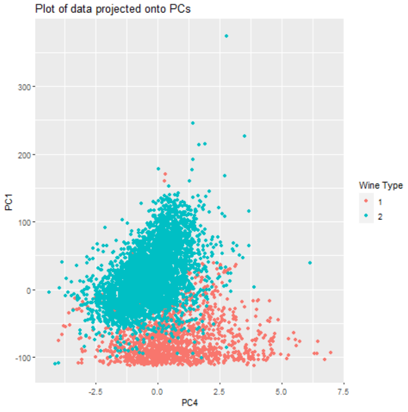
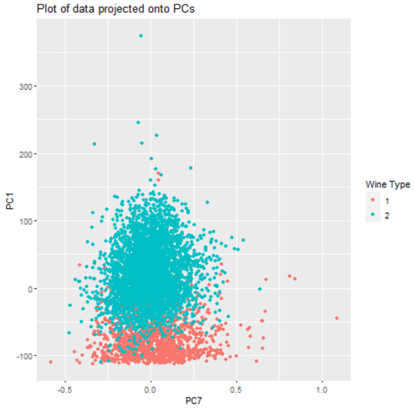
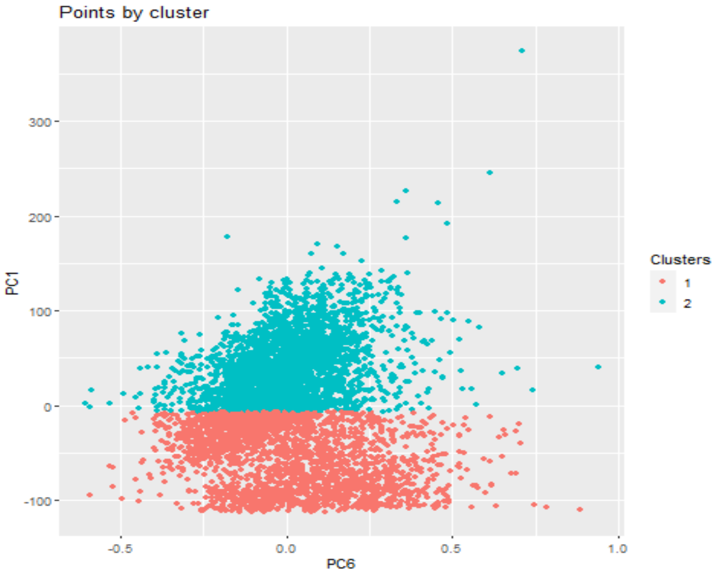


Figure 2.9-PC1 vs. PC6



Clustering

Figure 2.10-Points by Cluster



The figures here show you the results of running the K-means clustering algorithm on the first 7 PCs. It is important to compare and contrast this plot and with the previous ones with correct labels. The K-means pretty much divided the data horizontally.

Table 2.1-The confusion matrix: (1-Red wine; 2-White wine)

	Predicted		
Actual		**1**	**2**
	1	1516	83
	2	1310	3588

We can see that K-means clustering has a very high False-Positive rate (Incorrectly classifies a large number of white wine samples as red). We need to do better! Thus, in our main analysis we decided to use a Neural Network to improve performance.

Learning Method

Definition of Neural Network

Neural Network is an algorithm that is powerful and widely used in the field of Deep Learning. It allows computers to learn from observations, recognize potential patterns, and perform clustering on unlabeled data. It can also be used for classification after being trained on labeled data, which satisfies our need for classifying the wine types in the data.

“A neural network is put together by hooking together many of our simple ‘neurons’, so that the output of a neuron can be the input of another.” (Andrew, 2011, p. 3)

Figure 3.1-Map of Neural Network

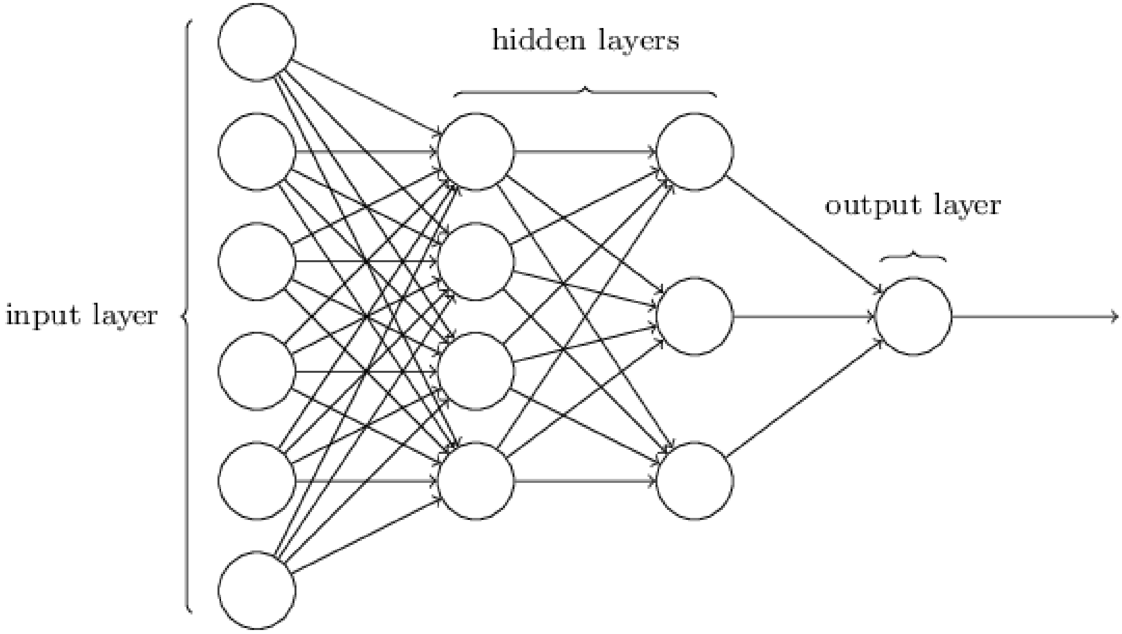


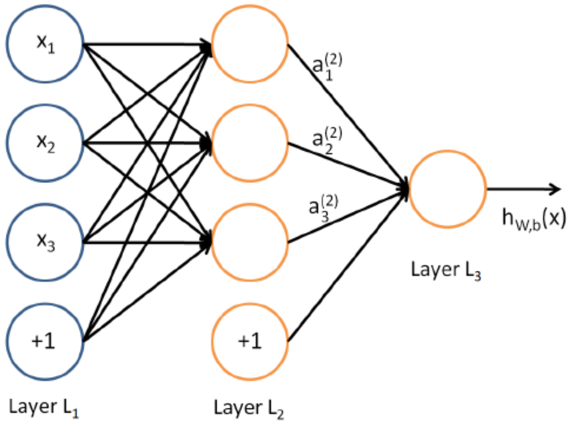
Figure 3.1 is a simple illustration of how Neural Network works. Each circle in the graph is a “neuron”. Neural Network is said to be inspired by the functioning of human brains. Human thought is the result of firing neurons inside human brains. Neurons help coordinate the functions of the brain by using electrical and chemical signals. Therefore, Neural Network is also called the Artificial Neural Network because it tries to mimic human brains.

Neural Network consists of several layers of nodes. The left-most layer is the input layer. The right-most layer is the output layer. The layers between are the hidden layers where most computations take place. From the perspective of this project, physicochemical data are taken in at the input layer. They are then processed at the hidden layers. Information on wine types is given at the output layer.

Forward Propagation

Figure 3.2 is a small Neural Network that can be seen as a part of the big Neural Network in *Figure 3.1*. It is an example of Forward Propagation because it does not have any directed loops or cycles. This small Neural Network has one input layer, one hidden layer, and one output layer. Nodes labeled x_i are inputs, which can be the observations from the data or outputs from neurons in previous layers.

Figure 3.2-Example of Forward Propagation



In this example, it has parameters $(W,b) = \left(W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}\right)$ where $W_{ij}^{(l)}$ denotes the weights associated with the connection between unit j in layer l, and unit i in layer $l+1$. The nodes labeled "+1" are the bias units denoted by b_i^l that associate with unit i in layer $l+1$. $a_i^{(l)}$ denotes the activation of unit i in layer l. $a_i^{(l+1)}$ denotes the output from the ith neuron in layer l. The computation at each neuron inside the hidden layers can be represented by

$$a_i^{l+1} = f(W_{i1}^{(l)} a_1^l + W_{i2}^{(l)} a_2^l + \cdots + b_i^{(l)}) \text{ (Andrew, 2011, p. 4)}$$

In this small neural network, $a_i^l = x_i$, f denotes the activation function of the neuron and $\sum_{j=1}^{n} W_{ij}^{(l)} a_j^l + b_i^{(l)}$ is the input function that takes in the activations on the links feeding into this node.

Backpropagation and Gradient Descent

Backpropagation is a crucial step in training the Neural Network. Briefly, the Neural Network is established by repeatedly performing Forward Propagation, Backpropagation and Gradient Descent. Such repetition serves to minimize the cost function of the Neural Network. In the example of Figure 2.2, its cost function is defined as:

$$J(W,b;x,y) = \frac{1}{2}\left\|h_{W,b}(x) - y\right\|^2 \text{ (Andrew, 2011, p. 6)}$$

If a training set of size m is given, the overall cost function of it is defined as

$$J(W,b) = \left[\frac{1}{m}\sum_{i=1}^{m} J(W,b;\, x^{(i)},\, y^{(i)})\right] + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}} (W_{ji}^{(l)})^2$$

(Andrew, 2011, p. 6)

The goal of training the Neural Network is to minimize $J(W,b)$ which is a function of W and b. When first performing a Forward Propagation on the Neural Network, parameters W and b are randomly initialized to be small values near zero, and then are updated by applying an optimization algorithm, such as Gradient Descent. In each iteration, Gradient Descent updates the parameters W and b and the process is defined as follows:

$$W_{ij}^{(l)} := W_{ij}^{(l)} - \alpha \frac{\partial}{\partial W_{ij}^{(l)}} J(W, b)$$

$$b_{i}^{(l)} := b_{i}^{(l)} - \alpha \frac{\partial}{\partial b_{i}^{(l)}} J(W, b)$$

(Andrew, 2011, p. 7)

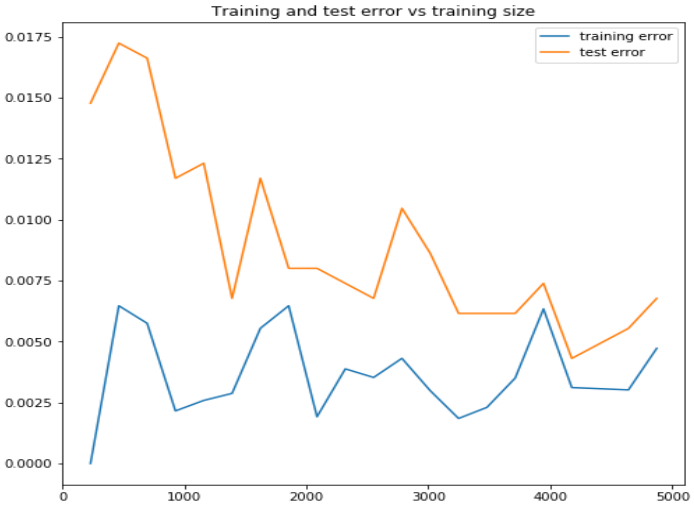
Backpropagation is the algorithm that serves to compute the partial derivatives used in the Gradient Descent algorithm. By repeatedly performing Forward Propagation, Backpropagation, and Gradient Descent, the parameters W, b are optimized to minimize $J(W, b)$, and the Neural Network can, therefore, be trained.

The Neural Network- Our Model

We used Python’s `sklearn` library to implement a shallow feed forward neural network. We split the data into a training and testing set to fit and evaluate the neural network. We also normalized the data before feeding it in to the network. The neural network had 2 hidden layers. The first one had 2 two hidden units and the second had 1 hidden unit. The hidden layer used the RELU activation function and output layer used the sigmoid activation so the output was the probability that a sample was a Red wine. We trained the neural network using the Binary cross entropy loss. The optimization method we used was ADAM which helps the gradient decent algorithm to converge faster and provides more accurate results than regular back-propagation.

We also plotted the train and test error vs the size of the training set. We can see that training error decreases as data size increases. However, the random spikes in errors show up because the neural networks are vulnerable to local minima.

Figure 3.3-Training and Test Error vs. Training Size

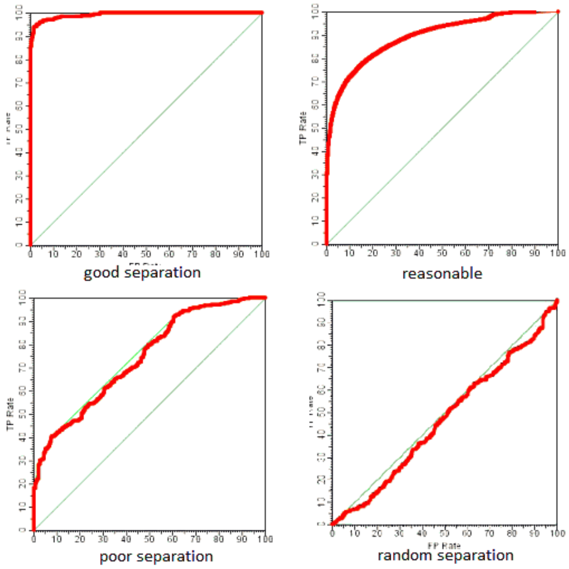


ROC Curve

A useful tool for measuring the success of our model is a receiver operating characteristic (ROC) curve. A ROC curve displays the performance of a classification model across all possible classification thresholds by graphing the true positive rate versus the false positive rate that results from every threshold that could be applied to the final output. In most cases, a threshold of 0.5 (after using the sigmoid function) is most appropriate, but there are some situations where other thresholds are of interest. In the medical world, for example, a negative threshold is often used when testing for diseases because doctors would rather trade false negatives (patients with the illness that go undiagnosed) for false positives (patients without the illness that are flagged).

The ROC curve takes all of these possibilities into account by plotting a point for how the model performs against each possible threshold that could be chosen. For example, if the threshold is 1, everything will be classified as a 0, resulting in the point (0, 0). On the other hand, if the threshold is 0, everything will be classified as a 1, resulting in the point (1, 1). The thresholds in between these values will plot points that create a curve from (0, 0) to (1, 1). Here are some examples of what a typical ROC curve looks like for a model with good, reasonable, poor, and random separation:

Figure 3.4-Models of Good, Reasonable, Poor and Random Separation



good separation

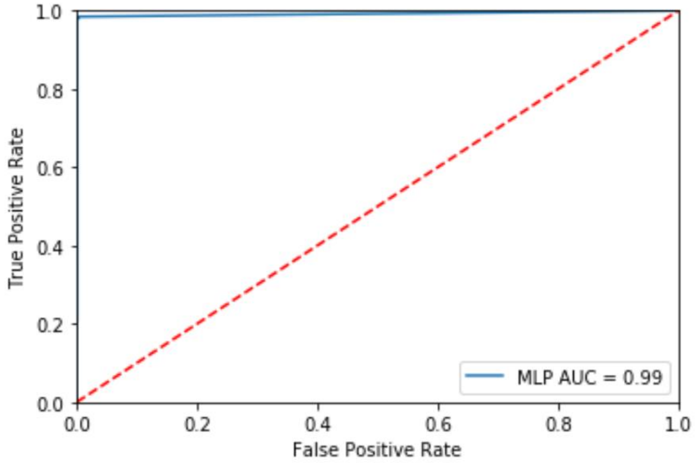
reasonable

poor separation

random separation

As demonstrated by these graphs, the more that the curve rests above the line TPR = FPR, the better the model is. For this reason, the area under the curve (AUC) is often used as a metric to judge model's accuracy because it is directly related to how well the model separates the data and does not depend on the chosen threshold. For example, an AUC of 0.5 means that the model separates the data randomly for all thresholds while an AUC of 1 means that the model separates the data perfectly for all thresholds. Below is an estimate of the ROC curve for our model:

Figure 3.5-ROC Curve of Our Model



The curve remains well above the TPR = FPR line and has an impressive AUC of 0.99. This means that this model performs well with all thresholds, so it would still be useful even if someone was only interested in identifying bottles that were white wine with 99% certainty, for example. In our case, we simply want to classify red and white wine as best as possible, so a threshold of 0.5 still makes the most sense.

Conclusion

Using this threshold, the model performs very well, as expected. It only misclassified 11 out of the 1,625 test cases for a 0.0066 overall error rate. By class, it had a 0.0153 error rate for red wine and a 0.0041 error rate for white wine. To answer our original question, this model has demonstrated that there are detectable differences between red and white wine in the predictors of interest. However, because of the complicated nature of a neural network, the model does not divulge exactly what those differences are. While the neural network does an exceptional job of identifying red versus white wine, a simpler method (such as logistic regression) may be more appropriate for explicitly identifying the main differences between the two with respect to the given predictors.

Citation

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Andrew Ng. (2011, January). CS294A Lecture notes

https://www.totalwine.com/wine-guide/wine-acidity-crispness

https://www.wineturtle.com/white-red-more-sugar/

https://www.chemwine.com/home/why-is-wine-density-important-1-sl6yl