Final Project Report: Statistical Learning (UNL STAT 983)

by Alison Kleffner, Sarah Aurit, Emily Robinson

April 11, 2021

1 Introduction

Successful marking campaigns and productive selling strategies are directly linked to communication about key indicators of quality; hence, objective measurements of quality are essential. Within the wine industry, there are two types of quality assessment: physiochemical and sensory tests. Sensory tests require a human expert to assess the quality of wine based on visual, taste, and smell [Hu et al., 2016]. Hiring human experts to conduct sensory tests can take time and be expensive [Gupta, 2018]. In addition, taste is the least understood of all human senses [Cortez et al., 2009]. Unlike sensory tests, laboratory tests for measuring the physiochemical characteristics of wine such as acidity and alcohol content do not require a human expert. The relationship between physiochemcial and sensory analysis is not well understood. Recently, research in the food industry has utilized statistical learning techniques to evaluate widely available characteristics of wine. This type of evaluation allows the automation of quality assessment processes by minimizing the need of human experts [Gupta, 2018]. These techniques also have the advantage of identifying important the phsiochemical characteristics that have an impact on the quality of wine as determined by a sensory test. In this report, we will investigate the accuracy of three classification techniques, and also address the challenges of classification with imbalanced data.

We will be working with the "Wine Quality Data Set" found on the UCI Data Repository at `https://archive.ics.uci.edu/ml/datasets/Wine+Quality`. [UCI, Cortez et al., 2009]. This data set consists of white (4898 samples) and red (1599 samples) Portuguese "Vinho Verde" wine samples. Each wine sample was processed to obtain measures for 11 physiochemical variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. A sensory test conducted by at least three sommeliers was used to determine a quality rating on an 11 point scale from 0 - very bad to 10 - very excellent. Following Hu et al. [2016], we separated the wine samples in to 3 classes: Low Quality (≤ 4), Normal ($5-7$), and High Quality (≥ 8). Basic descriptives are provided below; the Mann-Whitney and chi-square tests were used for comparisons of continuous and discrete variables, respectively.

The quality categorization poses a challenge of working with imbalanced classes as there are many more normal wines than low or high quality wines (Figure 1).

In this paper we are going to be looking at two main research questions. The first one focuses on developing a model that would work well to classify wines into three categories, which are: poor quality, normal quality and high quality. As stated in the problem description, it is desirable to classify wines using physicochemical properties since this does not involve human bias that would come into play with human tasters. We plan on using three different classification techniques. First, we will utilize random forests given prior investigation of the white wine data showed this technique seemed to work well; we would like to test this method using both white and red wine. Our second method is XGBoost, which is a gradient boosting framework that can be done in R. Previous papers on this data set have used different versions of gradient boosting like adaptive boost. Our final proposed method is KNN, as this seems to be a method that has not yet been used in any sort of wine classification.

Additionally, as described above, an issue with our data is the imbalanced nature of the data set, with most of the wines being classified into the normal quality category. We are also interested in looking at different resampling techniques in order to determine if resampling is necessary here, and to identify which method of resampling method is best. We plan on running each of the classification methods to find predictions in conjunction with each of the resampling techniques for the sake of comparison. Initially, we will first look at the results of the classification method predictions if no resampling is done. Since the data is very imbalanced, we would expect this to probably not do well, but it gives a baseline. The second resampling method we will test is SMOTE, which is an algorithm where

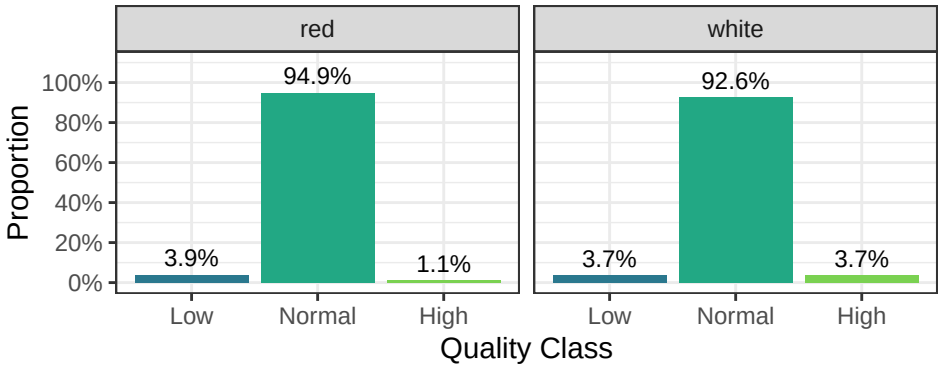

Figure 1: Wine quality class imbalance.

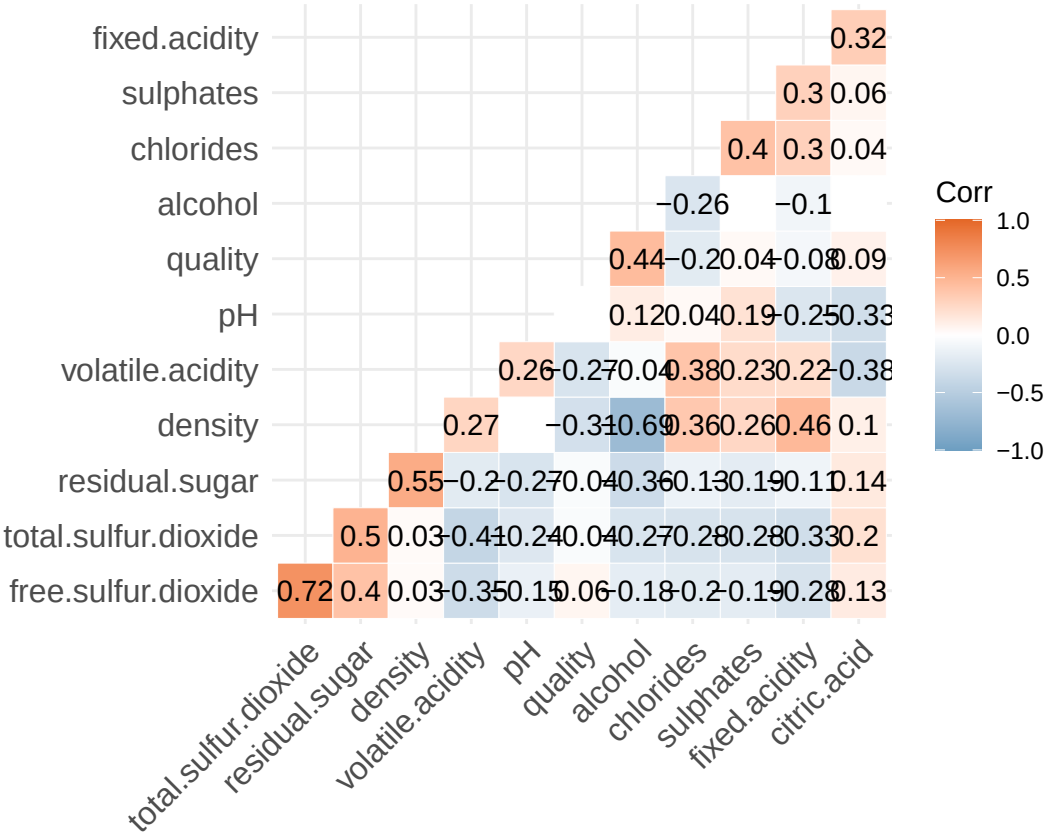

Figure 2: Correlation Plot

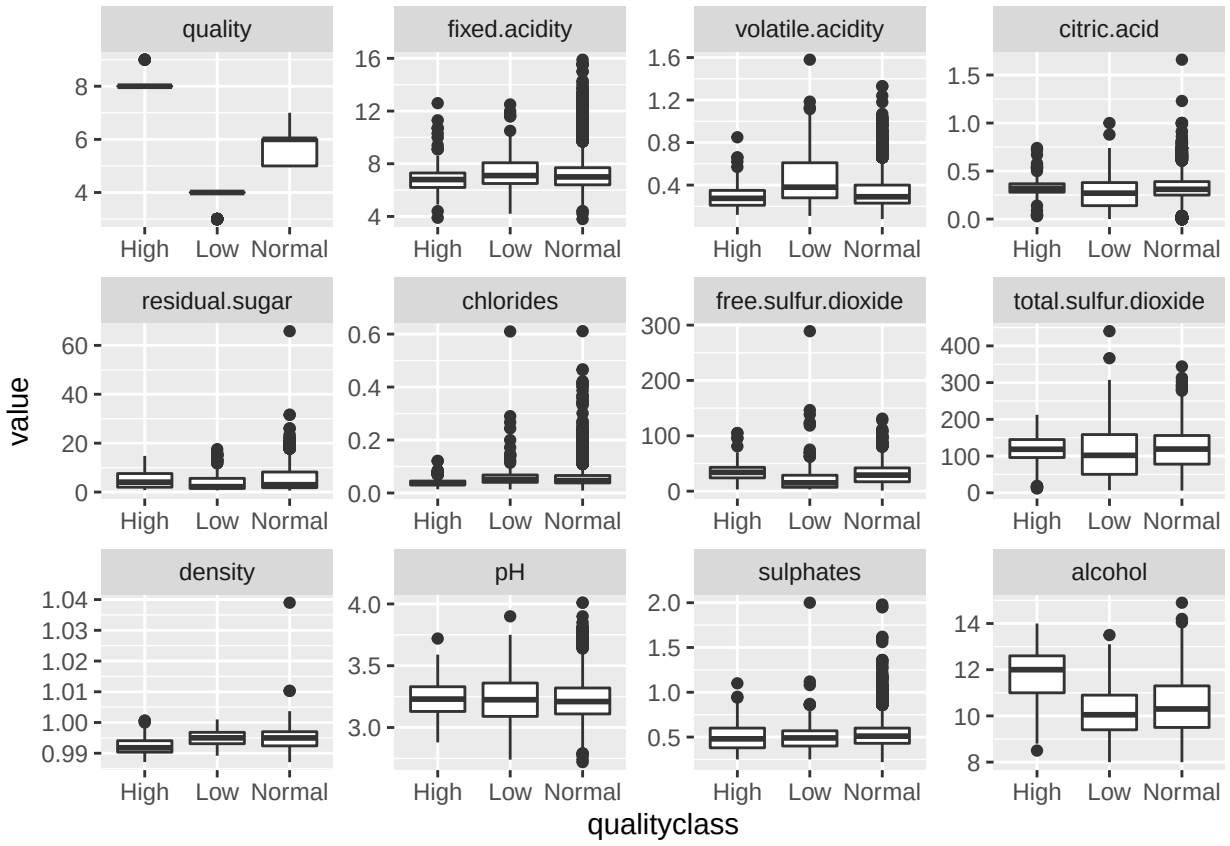

Figure 3: Distribution of the Predictor Variables

the minority class (in this case low and high quality), will be oversampled. This method runs the risk of overfitting the mode. The final resampling method that we will look at is random under sampler, where majority class data are randomly removed from the data set. This risks losing valuable information in our model. We are looking for the best prediction model using the combination of resampling and classification method, which will be determined using MSE.

2 Methods

2.1 Resampling

Before we are able to consider the different classification methods, the imbalanced issue in our data must be examined. This is because typical classifier algorithms assume a relatively balanced distribution, so with imbalanced data they tend to be biased towards the majority class [Yanminsun, 2011]. Classification rules for the minority classes tend to be undiscovered or ignored, so the minority class is misclassified more often than the majority class. If we were to look at our data set from a binary standpoint where "Low" and "High" were classified as rare, and "Normal" was classified as not rare, the majority class has 6053 observations, and the minority class has 444 observations, which is a ratio of about 13:1. So due to this imbalanced nature in the data, classifying algorithms will be biased towards classifying wines as Normal, and causing poor predictions for the Low and High classes. When creating the resampled data sets, the binary classification of "rare" and "not rare" is used due to the resampling algorithms needing a binary class.

A method on how to deal with the imbalanced issue in the data is through resampling the original data set either by oversampling the majority class, or undersampling the minority class. The goal with these methods is to create a data set that has close to a balanced class distribution, so the classifying algorithms will have better predictions, so it will predict to the minority class more accurately [Chawla, 2013]. The oversampling method involves replicating the minority class and creating synthetic data, or creating new instances using heuristics. This method tends to have an overfitting problem, where it'll predict the minority class more than it should. The undersampling method removes instances from the majority class randomly or using some heuristics. Since data is being removed, potentially valuable information is not considered [Hu et al, 2016]. To see how valuable resampling is in our data set, we ran the classifying algorithms with just using the original imbalanced data set, randomly undersampling method, and SMOTE, which is an oversampling method.

The undersampling method that we considered is random undersampling. In this method, instances of the majority class are discarded at random until reaches balanced with the minority class [Chawla, 2013]. For example, say there are 1000 observations in the majority class, and 100 in the minority class, observations in the majority class will be randomly discarded until this class is also 100. The benefit of this method is that since the data frame is being reduced, it is less costly. However, potentially useful information is discarded, which may make the decision boundary between the majority and minority class less clear, causing a decrease in prediction performance [Chawla, 2013].

SMOTE, which stands for Synthetic Minority Over-Sampling Technique, is an oversampling technique which uses interpolation of the minority class to create synthetic data. The process begins by finding the k nearest neighbors of each observation of the minority class based on some distance measure. Then a point between the minority class observation and one of its nearest neighbors is randomly picked by first finding the difference between the observation and its nearest neighbor. This difference is then multiplied by a random number between 0 and 1. This is then added to the observation, which becomes the new synthetic data point that is then added to the data set [Chawla, 2002]. In Hu et al, the used k=5, so that is what we used to create our resampled data set [2016]. Oversampling tend to have an overfitting problem since now the minority class extends into the majority space, however this generally poses less of an issue with SMOTE [Luego, 2010].

2.2 Classification

3 Results

4 Conclusion

References

Uci machine learning repository: Wine quality data set. URL `https://archive.ics.uci.edu/ml/datasets/Wine+Quality`.

P. Cortez, J. Teixeira, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Using data mining for wine quality assessment. In *International Conference on Discovery Science*, pages 66–79. Springer, 2009.

Y. Gupta. Selection of important features and predicting wine quality using machine learning techniques. *Procedia Computer Science*, 125:305–312, 2018.

G. Hu, T. Xi, F. Mohammed, and H. Miao. Classification of wine quality with imbalanced data. In *2016 IEEE International Conference on Industrial Technology (ICIT)*, pages 1712–1217. IEEE, 2016.

A Code

Need to work on "Environment shaded undefined."