Final Project Proposal: Statistical Learning (UNL STAT 983) Imbalanced Classification and Prediction of Wine Quality

by Sarah Aurit, Alison Kleffner, Emily Robinson

March 21, 2021

1 Introduction

Marketing and industry success is often affected by the quality of their products as determined by a quality certification in order to promote the sales of their products. 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, lab 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 to automate the quality assessment process 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 address the challenges of classification with imbalanced data and compare the accuracy of three classification techniques.

2 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). This poses a challenge of working with imbalanced classes as there are many more normal wines than low or high quality wines (Figure 1). Below, we propose resampling techniques to improve the accuracy of classification.

3 Proposed Methodology

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. The first one is random forests. This is because in prior research this technique seemed to work well just using the white wine data, so 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.

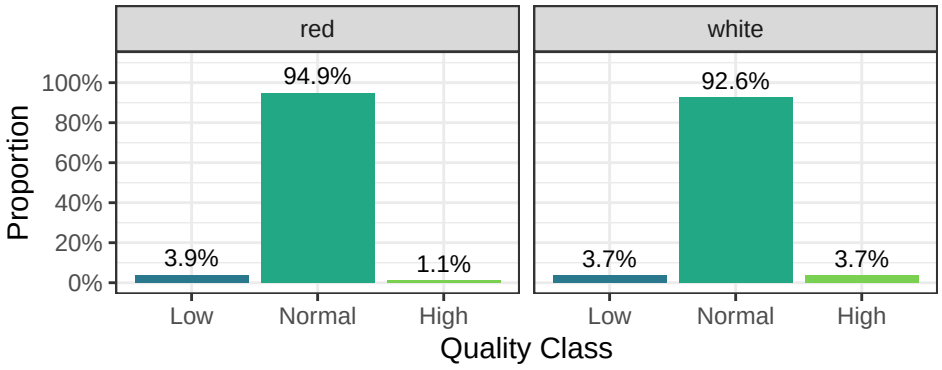

Figure 1: Wine quality class imbalance.

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. So we are also interested in looking at different resampling techniques in order to determine if resampling is necessary here, and if so what kind of resampling method is best. We plan on running each of the classfication methods to find predictions using each of the resampling techniques in order to compare them. So we are first going to look the results of the classfication 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 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 classfication method, which will be determined using MSE.

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.