ST563 Final Project

Mana Azizsoltani

09 March, 2021

Contents

Introduction

There has been a push towards the implementation of different analytics solutions within the agriculture industry, especially in the realms of crop health and yield. These sort of analyses can help farmers to reduce waste and improve profits. In particular, the viticulture industry stands to benefit from such methods, being that the production of wine is multifaceted. Winemakers can work to optimize revenue by analyzing their input and output (crops and wines). Specifically, costs may be cut out if machine learning methods were able to predict the quality of wine based on different metrics, since wine producers would no longer have to hire expert wine tasters to determine wine quality.

In this study, our objective is to try and predict the quality of wine using a variety of statistical learning techniques. We will look at this problem from a regression standpoint as well as a classification standpoint. Using regression, we will try and predict the exact rating (0 to 10) of a particular wine. On the other hand, when we look at classification, we will try to classify a particular wine as either good or bad, taking wines rated from 0-5 as low quality wines and wines rated from 6-10 as high quality wines. We will be assessing the prediction accuracy of the following methods:

Table 1: ML Methods

Classification	Regression
Classification Tree	Regression Tree
Random Forest	Boosted Tree
Support Vector Machine (SVM)	Multiple Linear Regression (MLR)
KNN Classification	KNN Regression
Logistic Regression	LASSO Regression

Required Libraries

To run the code for the project, the following libraries are required:

- `caret`: to do the heavy lifting of training and tuning the models
- `tidyverse`: for all the data reading and wrangling
- `knitr`: for rmarkdown table outputs
- `rmarkdown`: for output documents
- `corrplot`: for correlation structure visualization
- `leaps`: for subset selection
- `rpart`: fitting the basic classification tree
- `rpart.plot`: visualization of the classification tree
- `randomForest`: fitting random forest models
- `kernlab`: fitting the support vector machine
- `class`: fitting the k-nearest neighbor model

Data

The data set used is the `Wine Quality` dataset from the UCI Machine Learning Repository. The data set is composed of 1600 observations of different variants of the Portuguese "Vinho Verde" red wine. For each wine, various features of the wine were measured, including a measurement of quality, given by an expert wine taster. Our objective it to use the different features to predict the quality of the wine.

Variable Descriptions

Each entry in the dataset represents the different metrics of the following attributes of a single type of wine.

- **fixed acidity**: the quantity of fixed acids found in the wines. The predominant fixed acids found in wines are tartaric, malic, citric, and succinic, all of which come from the grapes with the exception of the succinic acid, which comes from the yeast in fermentation.
- **volatile acidity**: the steam distillable acids present in wine, primarily acetic acid but also lactic, formic, butyric, and propionic acids. These acids generally come from the fermentation process.
- **citric acid**: added to the wine as a natural preservative or for acidity and tartness.
- **residual sugar**: the quantity of sugar left in the wine after the fermentation process.
- **chlorides**: the amount of salt in a wine.
- **free sulfur dioxide**: the amount of SO_2 that is not bound to other molecules.
- **total sulfur dioxide**: total amount of SO_2 in the wine. Sulfur Dioxide is used throughout all stages of the winemaking process to prevent oxidation and bacteria growth.
- **density**: density of the wine.
- **pH**: pH of the wine.
- **sulphates**: quantity of sulphates in the wine. Sulphates are used as a preservative.
- **alcohol**: alcohol content by volume.
- **quality**: score of quality of the wine given by expert tasters (score between 0 and 10).

Methods

Data Processing & Partitioning

Fortunately, the `wine` data set that we worked with was very tidy; there was no missing values. We didn't standardize the data initially, but when training the support vector machine as well as the k-nearest neighbors models we centered and scaled the data.

As mentioned in the introduction, our focus is on the classification accuracy of machine learning methods. In particular, we will be using traditional cross-validation methods to assess the accuracy, sensitivity, and specificity of each of the models. We split the data into a training and test data set in order to later evaluate the model's prediction accuracy. For this project we used a 75/25 split, training the data on the 75% and testing the trained models on the withheld 25%. We will then repeat this process over 5 folds of the data, averaging the results.

The tree-based, LASSO, and K-nearest neighbors models require parameters to be tuned (more on those later). Since we used the `caret` package in R to fit all of our models, we used the tunes of the parameters that were deemed "best" by the `train()` function.

Variable Selection