Estelle Hooper (ehh52) and Eva Ruse (ewr43)

Dr. Yang Ning

STSCI 4740

8 December 2021

Wine Quality Prediction Report

**I. Introduction**

Using machine learning methods, we sought to discover what characteristics of wine determine a high-quality or low-quality wine, such as acidity, sweetness, and alcohol content. The dataset we used only contains red and white wine variants from the Portuguese wine Vinho Verde. In this dataset, the various wine variants are given a score from 1-10. While this prediction task could either be regression or classification, we chose to treat this as a classification problem so that we could apply numerous classification methods learned in class. In addition to selecting the best model for predicting wine quality, we also wanted to discover which features are the best predictors for quality, and if they differ based on white or red wine.

**II. Exploratory Data Analysis:**

Before modeling the wine dataset, we explored the data to understand general relationships between predictors and other information to indicate what methods should be used. The dataset had 6497 observations and 13 features: type (white or red wine) fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, alcohol, and quality scores. The dataset did not require much cleaning because there were no NA (missing) values. However, we did convert our outcome column (quality, originally a rating of 1-10) to a binary column to expand the breadth of models we could create. If the quality was over 5, the quality=1 (high quality), and if the quality was 5 and below, the quality=0 (low quality). Additionally, we converted this binary quality column to a factor so that the models would treat the outcome as categorical, not numeric.

|  |
| --- |
| **Fig. 1** Heatmap from the wine dataset. |

We created a correlation matrix to determine correlated variables because predictor variables should be independent of each other. As seen in the heatmap in Fig. 1, alcohol and density are highly correlated (~ 0.7) and density was dropped from the dataset to prevent multicollinearity effects on our models’ prediction accuracy. Lastly, we split the dataset into red and white wine to discover differences in the best predictors for each type of wine.

**III. Data Analysis and Model Selection**

We tested several models to best predict the quality of wine: logistic regression, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), Naive Bayes (NB), and decision trees.

1. **Logistic Regression**

We used best subset selection to select our features for the logistic regression model because the algorithm must consider every single possible combination of predictor variables and choose one according to some selection criteria (cP, AIC, BIC, adjusted , etc.). In order to do subset selection for a logistic model (opposed to a subset selection for a linear model as we discussed in class) we used the package “bestglm.”

When we conducted subset selection for both the white and red wines datasets, we compared the logistic models that bestglm chose based on different selection criteria: AIC, BIC, and cross-validation (CV).

|  |  |  |
| --- | --- | --- |
|  | **Fig. 2** White wine best subset selection |  |

As seen in Fig 2, the three selection criteria selected a different number of variables for white wines. The model based on AIC used 7 variables: fixed.acidity, volatile.acidity, residual.sugar, free.sulfur.dioxide, total.sulfur.dioxide, sulphates, and alcohol as its predictors. The model based on BIC used 6 variables: fixed.acidity, volatile.acidity, residual.sugar, free.sulfur.dioxide, sulphates, and alcohol. The model based on CV used 6 variables: volatile.acidity, fixed.acidity, chlorides, free.sulfur.dioxide, sulphates, and alcohol. We calculated the test error rate, accuracy score, recall score, precision score, and area-under-curve (AUC) for each of these logistic models to compare to the rest of our models later.

|  |  |  |
| --- | --- | --- |
|  | **Fig. 3** Red wine best subset selection |  |

Similarly, as seen in Fig 3, the three selection criteria selected a different number of variables for red wines The model based on AIC used 8 variables: fixed.acidity, volatile.acidity, citric.acid, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, sulphates, and alcohol as its predictors. The model based on BIC used 6 variables: volatile.acidity, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, sulphates, and alcohol as its features. The model based on CV used 6 variables: volatile.acidity, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, sulphates, and alcohol as predictors.

It is not surprising that BIC chose the model with the least number of predictors for the white and red wine datasets. BIC has a higher penalization for models with a higher number of predictors, therefore causing models chosen based on BIC to be parsimonious.

1. **Linear Discriminant Analysis/Quadratic Discriminant Analysis/Naive Bayes**

Generative classification models are other algorithms that perform well for classification problems. We chose to explore linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and naive Bayes (NB) as potential models for the wines dataset. We fit these three models to our white wines and red wines dataset and computed the test error rate for these models to compare later.

1. **Decision Trees**

Since our methods thus far resulted in test error rates > 0.20 (Fig. 6, Fig. 7), we decided to implement decision trees as a final classification method. We implemented a basic tree for the white and red wines dataset, and then used random forests, bagging, and boosting to see if we could improve our models. We recorded the test error rate to compare these models with the rest in the following section.

**IV. Interpretation**

1. **Best Predictors/Variables**

While we did perform feature selection using subset selection for our logistic models, these models did not perform well in comparison to other models. Since random forests had the lowest test error rates, the variables these models chose are the best predictors. According to the Variable Importance plots derived from random forest, the best predictor for both white and red wine was alcohol. However, the subsequent good predictors are not the same for white and red wine. As seen in Fig 4, the top 3 best predictors for white wine (in order) besides alcohol are volatile acidity, free sulfur dioxide, and residual sugar (according to the mean decrease in accuracy their absence from the model causes). On the other hand, seen in Fig 5, the top 3 best predictors for red wine (in order) are sulphates, volatile acidity, and total.sulfur.dioxide (according to the mean decrease in accuracy their absence from the model causes). Of course, since these plots were only derived from the random forest output, they do not represent all the models used, but can still give insight on which variables are the most important in predicting quality. The plots below also show that features hold a different influence on the models depending on whether it is a model for white or red wine.

|  |
| --- |
| **Fig. 4** Increase in MSE of predictions (left) and Mean decrease in node purity (right): white wine |
|  |
| **Fig. 5** Increase in MSE of predictions (left) and Mean decrease in node purity (right): white wine |

1. **Best and Worst Models**

As seen in Fig. 6 and Fig. 7, the best models for both white wine and red wine were random forest at the error rates of 0.174 and 0.208, respectively. The worst model for white wine was Naive Bayes at the error rate of 0.302. Contrastingly, the worst model for red wine was the pruned tree at the error rate of 0.308. In general, the rank of error rates for white and red wine was consistent except for AIC and BIC. For white wine, BIC performed slightly worse than AIC, but for red wine, BIC performed much worse than AIC, despite both methods being logistic regression. For white wine, pruned trees largely outperformed QDA, but for red wine, QDA largely outperformed pruned trees.

The fact that random forests appeared to be the best models for predicting wine quality for white and red wine suggests that there may be a non-linear and complex relationship between the features (for each the white and red wine model) and the outcome-response variable quality.

|  |
| --- |
|  |
| **Fig. 6** Error Rate Plot: white wine |
|  |
| **Fig. 7** Error Rate Plot: red wine |
|  |

**V. Conclusion**

By treating the prediction task of wine quality as a classification problem, we were able to observe key differences in what makes high-quality white wine and red wine, as least in the Portuguese wine variant Vinho Verde. Since we only looked at this particular wine variant, it is entirely possible that the best predictors observed for this wine variant would differ if we ran the same models for a dataset of different wine variants. Additionally, our best model, random forest, may not be the same for other wine variants.

Additionally, though we were able to determine that certain characteristics of wine are more important than others, it is important to remember that the quality of wine itself is determined by human preference and wine experts. While providing statistical models based on machine learning algorithms may be beneficial in improving quality for wine production, it is only a general resource to assist wine industries in formulating the flavor of wines.

**VI. Appendix**

Link to [R-Code](https://drive.google.com/drive/folders/1zZYt_AURvbErLEQSGj6PJuVoBhmS9kc_?usp=sharing)