BIOST 544 Final Project

Alison King, Katie McFarlane, and Thu Vu

12/11/22

Introduction

Problem of Interest

Vinho verde is a unique variant of wine from the northwestern region of Portugal that can vary in quality and taste. We are interested in understanding how the physicochemical attributes of vinho verde red wines are associated with a wine being perceived as good quality. Each red wine in our sample has been scored on a 10-point scale, with wines receiving a score of 7 or higher being classified as "good quality" wines.[2] There is potential that the physical and chemical properties of a wine, such as the acidity and chemical content, may have an effect on the perception of that wine's quality. We are interested in model feature selection to identify statistically significant predictors of wine quality as well as testing the model's ability to classify wines into good or poor quality based on their physicochemical characteristics.

Dataset Description

The dataset and further descriptions can be found at the UC Irvine Machine Learning Repository[3].

Variables of Interest

The output variable for our data is a quality score ranging from 0 (very poor) to 10 (excellent), which we have dichotomized into "good quality" (7 or higher) and "poor quality" (less than 7). The quality score for each wine is based on the sensory evaluation of three or more tasters using blind taste tests.[1] The potential physicochemical predictors in our dataset are measurements regularly taken during the wine certification process, and they are listed here:

1. Fixed acidity (g(tartaric acid)/dm^3)
2. Volatile acidity (g(acetic acid)/dm^3)
3. Citric acid (g/dm^3)
4. Residual sugar (g/dm^3)
5. Chlorides (g(sodium chloride)/dm^3)
6. Free sulfur dioxide (mg/dm^3)
7. Total sulfur dioxide (mg/dm^3)
8. Density (g/cm^3)
9. pH
10. Sulphates (g(potassium sulphate)/dm^3)
11. Alcohol (volume %)

All predictor variables are quantitative variables. In addition, all of these variables have a potentially significant relationship with wine quality, and as such all will be included in our full model when performing stepwise selection.

Missing Values

Our dataset includes 1599 wines, and there are no missing values in any of the variables or quality outcomes for any of them.

Statistical Methods

Descriptive Statistics

For each variable, we display the mean and standard deviation, stratified into poor quality and good quality wines, as well as for the entire dataset. The number of wines in each category is also displayed.

Primary Model

In order to select our model variables, we use multiple logistic regression and forward stepwise selection with AIC. Our initial model contains solely an intercept term, while the upper model includes all eleven variables outlined in the variables of interest section above as predictors. Both models use the dichotomized quality value as the outcome. This process will be used to determine the model features we want to use. We use a significance level of 0.05 to make this decision. We are also interested in evaluating the model's classification abilities, and as such we assign 70% of the data randomly to a training set in order to fit this model, leaving the remaining 30% for the testing set. We will then calculate the misclassification error rate on the test set.

When reporting model estimates for all models, robust standard errors are used and all confidence intervals are reported at the 95% level. Hypothesis tests are conducted using Wald tests.

Sensitivity Analysis

For comparison purposes, we use our training set to also fit a multiple logistic regression model and conduct forward stepwise selection with BIC, as using AIC can sometimes overfit the model. Similarly to our primary model, we use the same initial and upper models with the same outcome and predictor variables, and we again compute the misclassification error rate on the test set. In addition, we compare the AIC values for both the primary model and this second model.

Results

Descriptive Statistics

Descriptive statistics for the dataset can be found in Table 1. We note that roughly 13.6% of wines are considered to have good quality. While the mean and standard deviation of some variables appear to be consistent for both poor and good quality wines (such as chlorides, density, and pH level), there are notable differences in the mean values for many of the other variables when comparing between poor and good wines.

Primary Model

When using forward stepwise selection with AIC, the following nine predictors are included: alcohol, volatile acidity, sulphates, fixed acidity, chlorides, total sulfur dioxide, density, residual sugar, and free sulfur dioxide. Table 2 shows an output of the coefficient estimates, as well as robust standard errors, 95% confidence intervals, and p-values associated with all variables selected through forward stepwise selection using AIC. We find that all of the predictors in the model are significant at the 0.05-level except for the variable free sulfur dioxide. We return an AIC value of 585.12.

Sensitivity Analysis

When using forward stepwise selection with BIC, we return a model with fewer predictors: alcohol, volatile acidity, sulphates, fixed acidity, and chlorides. Table 3 shows an output of all coefficient estimates, robust standard errors, 95% confidence intervals, and p-values. We find that all of the predictors in the model are significant at the 0.05-level. We return a BIC value of 623.2, which is fairly close to but slightly higher than the AIC value found in our primary model. While we cannot directly compare these criteria, we do prefer models with lower AIC/BIC values. To see how the AIC and BIC values for each iteration of the process for both the primary and second model, see figure 2.

Error Rate

We use the randomly chosen test set in order to calculate the misclassification error rate. For our primary model, we find an error rate of 0.13. When using the model fit in our sensitivity analysis, we note an error rate of 0.14. These are very close, but the primary model does a slightly better job at classification than the second model.

Figure 2 is a graphical representation of the ROC curves for each model: we see that the curves are fairly similar and even cross multiple times. The area under the curve for the primary model is 85%, while for the second model, it's 84%. These numbers are very close but once again indicate a better fit using our primary model.

Discussion

We find that the primary model using AIC performs slightly better than the model using BIC. This is not surprising, given that the primary model includes more predictors, and as such we expect a lower AIC value. However, the model using BIC may be more practical for making inference on, given that it only includes six predictors at only a small cost to the misclassification error rate, which is roughly 1% higher than that of the primary model. This difference in the number of predictors selected using AIC versus BIC is unsurprising, as we expect BIC to put a higher penalty for adding additional predictors to the model than AIC.

Limitations

This project aims to model wine quality given physicochemical attributes of the wine; however, the quality scores used to create the outcome in this analysis are based on human preference, which is very subjective, and as such there may be some differences or human error when assessing quality. Additionally, while the dataset consists of a large number of wines, we find that a large proportion of wines (86.4%) are considered poor quality, meaning our data is not balanced. We also note that some information was removed from this dataset, such as grape type and wine brand, which may have been useful predictors of wine quality if they were known.[3]

Implications

The development of a reliable predictive model for wine quality is valuable for both producers and consumers of wine as well as regulatory and certification agencies. The predictors that we include in our model are all metrics that are readily available as part of the wine certification process, and the implementation of such a model could be a useful supplement to the wine evaluation process as an objective analytical test of quality. This model can also help us understand the important physicochemical qualities of a good wine, which could prove to be a useful piece of knowledge for wine producers.

Tables and Figures

Table 1: Total counts of wines and descriptive statistics (mean and standard deviation) of variables stratified by wine quality and overall

	Overall (n=1599)	Poor (n=1382)	Good (n=217)
Fixed acidity (g(tartaric acid)/dmˆ3)	8.32 (1.74)	8.24 (1.68)	8.85 (2.00)
Volatile acidity (g(acetic acid)/dmˆ3)	0.53 (0.18)	0.55 (0.18)	0.41 (0.14)
Citric acid (g/dmˆ3)	0.27 (0.19)	0.25 (0.19)	0.38 (0.19)
Residual sugar (g/dmˆ3)	2.54 (1.41)	2.51 (1.42)	2.71 (1.36)
Chlorides (g(sodium chloride)/dmˆ3)	0.09 (0.05)	0.09 (0.05)	0.08 (0.03)
Free sulfur dioxide (mg/dmˆ3)	15.87 (10.46)	16.17 (10.47)	13.98 (10.23)
Total sulfur dioxide (mg/dmˆ3)	46.47 (32.90)	48.29 (32.59)	34.89 (32.57)
Density (g/cmˆ3)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
pH	3.31 (0.15)	3.31 (0.15)	3.29 (0.15)
Sulphates (g(potassium sulphate)/dmˆ3)	0.66 (0.17)	0.64 (0.17)	0.74 (0.13)
Alcohol (vol.%)	10.42 (1.07)	10.25 (0.97)	11.52 (1.00)

Table 2: Foward stepwise selection model estimates with AIC

	Estimate	Exponentiated Estimate	Robust SE	95% Upper	95% Lower	P Value
Intercept	299.321	9.84e+129	129.161	1.12e+20	8.62e+239	0.013
Alcohol	0.820	2.271	0.151	1.689	3.054	<0.0001
Volatile Acidity	-2.755	0.064	0.863	0.012	0.345	0.001
Sulphates	3.567	35.400	0.601	10.895	115.017	<0.0001
Fixed Acidity	0.394	1.483	0.118	1.176	1.870	<0.0001
Chlorides	-13.131	0.000	4.526	0.000	0.014	0.012
Total Sulfur Dioxide	-0.018	0.982	0.008	0.966	0.999	0.004
Density	-315.133	0.000	129.915	0.000	0.000	0.01
Residual Sugar	0.253	1.288	0.090	1.080	1.537	0.007
Free Sulfur Dioxide	0.023	1.023	0.017	0.990	1.057	0.131

Table 3: Foward stepwise selection model estimates with BIC

	Estimate	Exponentiated Estimate	Robust SE	95% Upper	95% Lower	P Value
Intercept	-14.981	0.000	1.913	0.000	0.000	<0.0001
Alcohol	1.118	3.057	0.121	2.410	3.879	<0.0001
Volatile Acidity	-3.628	0.027	0.811	0.005	0.130	<0.0001
Sulphates	2.556	12.881	0.508	4.757	34.881	<0.0001
Fixed Acidity	0.231	1.260	0.064	1.112	1.428	<0.0001
Chlorides	-14.763	0.000	5.220	0.000	0.011	0.007

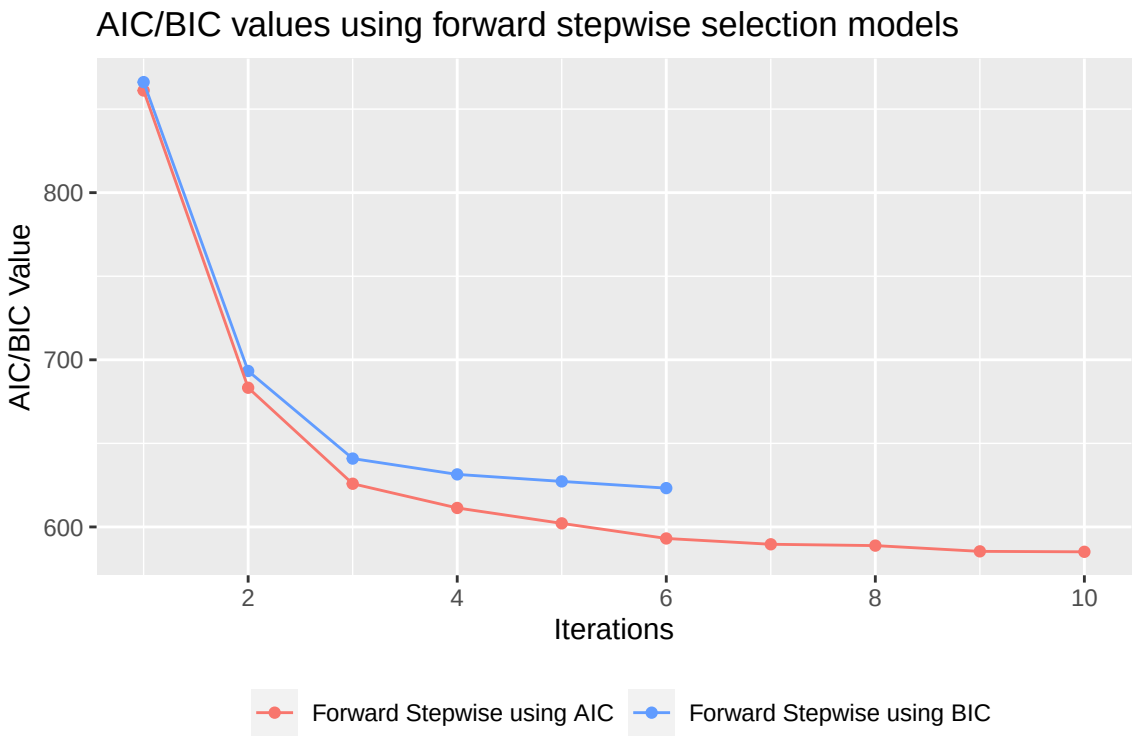

Figure 1: This figure compares the AIC or BIC values at each step of the foward stepwise selection process of adding predictors from the wine dataset.

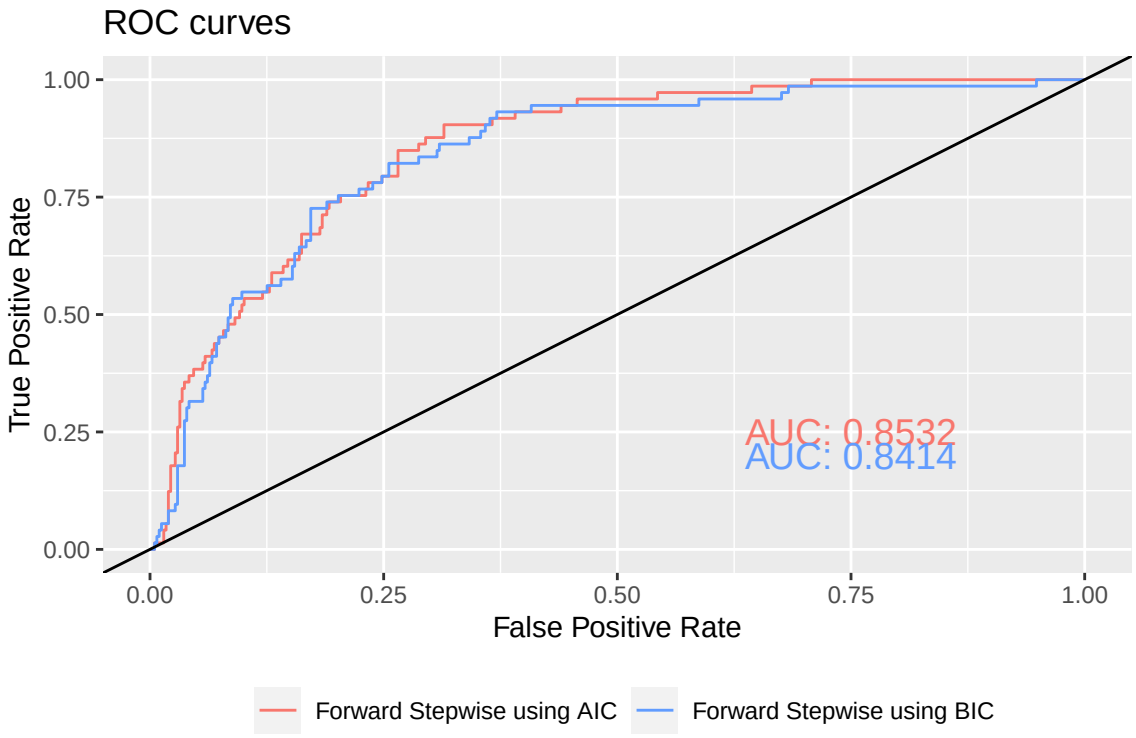

Figure 2: This graph compares ROC curves for forward stepwise using AIC and BIC on the wine dataset.

References

1. Cortez P, Cerdeira A, Almeida F, Matos T, Reis J. Modeling wine preferences by data mining from physicochemical properties. ScienceDirectt. https://doi.org/10.1016/j.dss.2009.05.016. Published 2009.
2. "What are Wine Scores?" Wine-Searcher. https://www.wine-searcher.com/wine-scores. Published 2022.
3. "Wine Quality Data Set." UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/wine+quality. Published 2009.