Gil Graybill – BAN 525
Assignment 2 – Penalized Logistic Regression and Wine Quality
Professor Cetin Ciner
May 31, 2020

Introduction

Wine has been with mankind for millennia, and all cultures enjoy it. The serving of wine indicates a special occasion or a time for slowing down and enjoying the moment. Winemaking can be a source of national pride. When I toured a German winery many years ago, the winemaker was asked what she thought of French wines. She replied, “The French are very good. They don’t even need grapes to make wine.” Many prospective winemakers have been disappointed when the same grapes that make great wine in one region make what amounts to vinegar in theirs. The question then comes up, “What is it that makes wine good?”

Some say that because wine tasting is subjective and consistent standards are not applied (https://io9.gizmodo.com/wine-tasting-is-bullshit-heres-why-496098276) that it must be possible to objectively measure how good or bad a wine is based on its physical and chemical properties. While it would be neat to find out what makes a superior wine (if that criteria could even be determined), this assignment will be looking into what make a bad wine.

The dataset we are looking at is 6497 rows consisting of human expert rankings of a large number of *vinho verde* wine samples (while the translation is “green wine” it actually means “young wine”, or wine that is being released 3-6 months after grapes are harvested) . The input variables, based on physicochemical tests, are described in the Appendix. They are: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density 9, pH, sulphates, alcohol level, and quality.

The predictor variable of “quality” has been converted from a 0-10 ordinal score to a “Good/Bad” categorical variable. 246 of the rows are classified as “Bad”, or 3.7%. Because it is categorical, we will be using logistic regression for the analysis. Additionally, we’ll be using penalized logistic regression with the hope of shrinking the coefficients that don’t contribute to the model towards zero.

We will be using 5 types of analysis on this dataset:

(1) OLS (Ordinary Least Squares): This type of logistic regression model calculates the best fit through the data points that minimizes the squared differences between the observed values and the fitted values. All variables end up in the final equation.
(2) Lasso (Least Absolute Shrinkage and Selection Operator) is a logical regression that aims to shrink data values by penalizing them, with the hope of eliminating variables that don't contribute to the model, thereby removing unneeded detail from the model.
(3) Adaptive Lasso runs OLS behind the scenes first to get an idea of what variables appear to be the most important. Those variables are penalized less, giving them a better chance of making it into the model when Lasso is run.
(4) Elastic Net combines penalties obtained from Lasso and Ridge to generate more zero valued coefficients. The square of the coefficient is used. The result can be a model where highly correlated predictors are all included in the model.
(5) Adaptive Elastic Net runs OLS first to favor more likely predictors, hence penalizing the weak performers more.

Analysis and Model Comparison

OLS regression will find any relationships that exist between response variable and the potential dependent variables. One disadvantage for OLS is that if there are many variables that are highly correlated, OLS may model random errors. This is our "simple" benchmark analysis.

Penalized regressions add a penalty function to the coefficient estimates, and how that penalty function is defined determine the type of regression analysis. With Lasso regression the coefficients (weights) are penalized by using the *absolute values of the weights*. Lasso performs variable shrinking and selection at the same time. A problem with Lasso is that when there are highly correlated variables, Lasso will just pick one of them.

With Ridge regression (which we are not directly using for this assignment) the coefficients are penalized using the *square of the weights*. While it may seem that the square of the weights is bigger than the absolute value, it is actually smaller because normalization of the data makes the coefficients less than 1.

Elastic Net uses both Lasso and Ridge to determine the penalty function to capture anything that lands "in the net". With the analogy, the small weights (big penalties) go through the net, leaving only the larger ones behind, even if they are highly correlated. The advantage/disadvantage is that more variables may be included in the model, which could lead to either a more precise model or overfitting.

Use of the “Adaptive” option jumpstarts the variable shrinking process by running OLS regression first and penalizing those that appear to have no predictive value. The hope is that the OLS regression was not incorrectly minimalizing those variables.

For cross-validation we employed a 60-20-20 split for training-validation-test split. This gave us 3898 rows of training data, 1299 rows of validation data, and 1300 rows of test data, split up the data randomly.

“Quality” was chosen as the response variable and the rest of the variables were tested as dependent variables using JMP Pro 15. The calculated value for the response variable, the probability that a wine was classified as “Bad”, was determined and recorded and the resulting formulas determined using all 5 methods.

The best way to compare results for a logistical regression model is the check the “Area Under Curve”, or AUC. The value consists of the ratio of “True Positives” vs. the ratio of “True Negatives”, plotted against all possible cutoff values for success. The higher the AUC, the better the model is. Because we are using cross-validation, we’ll be looking at the results for the Test data. The results are as follows:

Cross-Validation	Predictor	AUC
Training	OLS	0.7819
Training	Lasso	0.7819
Training	Adaptive Lasso	0.7819
Training	Elastic Net	0.7819
Training	Adaptive Elastic Net	0.7819
Validation	OLS	0.7731
Validation	Lasso	0.7731
Validation	Adaptive Lasso	0.7731
Validation	Elastic Net	0.7730
Validation	Adaptive Elastic Net	0.7731
Test	OLS	0.8131
Test	Lasso	0.8131
Test	Adaptive Lasso	0.8130
Test	Elastic Net	***0.8132***
Test	Adaptive Elastic Net	0.8130

Within each level of cross-validation, the regression models all came up with similar values for AUC. This tells us that we have a pretty stable dataset and that almost any reasonable analysis will most likely come up with similar results. For the Test data, Elastic Net is slightly better with an AUC of 0.8132, which is considered excellent (An AUC of 0.5 would be no better than a coin flip).

Interpretation

The formula that was determined using Elastic Net was:

Probability of BAD rating = (-320.302446540103) + -0.0379967107608063 * fixed acidity + 5.74214287213481 * volatile acidity + -0.382999850901012 * citric acid + -0.143580997104869 * residual sugar + 3.19222136765852 * chlorides + -0.0273847998846288 * free sulfur dioxide + -0.00327871785106214 * total sulfur dioxide +319.620148759056 * density + 0.0664069807355946 * pH + -0.426624825837152 *sulphates + -0.0119201573118092 * alcohol + Match(:color, "red", -3.91758469629474, "white", 0)

When running variable importance against these variables, we get the following:

Variable	Main Effect	Total Effect
Density	0.41	0.638
Residual sugar	0.113	0.259
Volatile Acidity	0.098	0.227
Free Sulfur Dioxide	0.078	0.193
Color	0.057	0.149
Chlorides	0.007	0.016

Density is a positive influencer of bad wines, and it accounts for 64% of the total effect. Alcohol has a density of 0.789 g/cc, whereas water is 1.0 g/cc, meaning that you'd expect the density to be less than water. However, dissolved sugar will add to the density. Analyzing "density" shows that the values of the middle 80% are [0.99067, 0.9984]. If a wine seems too dense it means that it has a lot of sugar and less alcohol. Who'd want that?

Residual sugar, which is the amount of sugar leftover after the fermentation process, is a positive influencer of bad wine, with a total effect value of 26%. During the fermentation process, sugars from the grapes combine with yeast to make alcohol. If there is no sugar leftover (a "dry" wine) it would mean that there might still be some potential fermentation that has been missed. Too much sugar indicates excess sweetness, which correlates with what we saw with density.

Volatile acidity is the amount of acetic acid in wine, which can account for a vinegar taste. It is a positive predictor in this model, with a total effect of 23%. And most people don't like to drink vinegar.

Free sulfur dioxide is a negative predictor, with a total effect of 19% . Because the proper amount prevents microbial growth and the oxidation of wine, having too little of it is not desirable, unless you want microbes and rust with your vinegar.

Color is a categorical variable (red, white), and it has a total effect of 14% . The color "Red" has a negative effect on the model, indicating that red wines have a lower tendency to be classified as "Bad" compared to white ones.

I ran "Analyze" against the predicted "Quality" values, and only 14 observations were found to be "Bad" for each model. Since 246 observations were classified as "Bad", it seems like trying to create a model using this data may not be a good idea. Apparently there are factors besides physiochemical ones that play into expert evaluations of wine, which does appear to make winetasting subjective in nature. Is that a bad thing (https://www.winespectator.com/articles/matt-kramer-wine-subjectivity) ?

We were asked to make a prediction for a new wine sample with the following data:

> Fixed acidity=8.9, Volatile acidity= .90, Citric Acid=.35, pH=3.20, Residual sugar=4.8, Cholorides=.09, Free sulphur dioxide=4, total sulphur dioxide=40, density=.99, sulphates=.65, alcohol=9.0, color=red

Using our Elastic Net model, it would be classified as "Good", with a probability of being "Bad" of 0.01954, or less than 2%.

Appendix: Descriptions of Columns in dataset

fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily)
volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
citric acid: found in small quantities, citric acid can add 'freshness' and flavor to wines
residual sugar: the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet
chlorides: the amount of salt in the wine
free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine
total sulfur dioxide:amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
density: the density of water is close to that of water depending on the percent alcohol and sugar content
pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant
alcohol: the percent alcohol content of the wine
quality: output variable (based on sensory data, score between 0 and 10)