Abstract

It is the job of an oenologist to find the ideal mix of ingredients for wine, one that will please consumers and business owners alike. Previous research has shown that alcohol, acidity, and residual sugar are the four main factors that influence wine quality in general. In this paper, we analyze quality data of the red and white variants of the Portuguese Vinho Verde wine known for their lower alcohol level and higher acidity relative to other wines. Using several machine learning methods, we select, from a list of eleven variables, the few that most acutely affect wine quality. With our ideal method of Generalized Additive Models, we conclude that volatile acidity and alcohol are most significant in predicting the quality of the Vinho Verde wines. Our research could be used by Portuguese winemakers in determining which aspects of the wine production process they should focus on most to optimize the quality of their wines. It may also inform wine store owners about the factors most likely to affect wine sales.

Introduction

We set out to find the best set of input variables that help predict the wine quality for both red and white Vinho Verde wines. We use both parametric and non-parametric model selection methods. The parametric methods are multiple linear, Ridge, and Lasso regression, and the non-parametric methods are principal component regression (PCR), local regression, and generalized additive models (GAMs). Within each method, we found the model with the lowest 10-fold cross validation error and then compared these “candidate models” across methods, discovering that GAMs yielded the optimal results.

Description of Subjects

The data is separated into two smaller datasets, each based on one variant of the Portuguese Vinho Verde wine being assessed, whether red or white. Both datasets have twelve columns. Eleven are the physicochemical input variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, and alcohol. The twelfth is the response variable, wine quality, which was determined by taking the median value of at least three sensory evaluations from wine experts that were measured on a scale from 0 (very bad) to 10 (excellent). The two datasets differ in the number of observations: there are 1599 in the red wine dataset and 4898 in the white wine dataset. Both datasets are pre-processed, so no further data cleaning is required.

We employed the likelihood ratio test to compare the intercept-only model with all possible single-predictor models, removing the predictors with p-values greater than 0.05. For this test, the null hypothesis states that the model has a good fit, and the alternative asserts that it does not. In the red wine dataset, we remove residual sugar since it has a p-value of 0.5829. In the white wine dataset, we removed both citric acid and free sulfur dioxide, which have p-values of 0.5192 and 0.5680, respectively.

As a last step before running multiple linear regression, we constructed the correlation matrix of the predictors in each dataset to determine whether collinearity was present. We planned to remove the variable with a correlation coefficient greater than 0.8 from each dataset. For the red wine, none of the remaining variables has a correlation coefficient that exceeds 0.8, so we keep all ten predictor variables. In white wine, residual sugar has a high correlation with density exceeding 0.8. Therefore, we removed residual sugar from the white wine dataset for having high collinearity.

Results

Multiple Linear Regression

The first machine learning method we used was multiple linear regression (MLR). MLR serves as a simple model to predict the response variable using multiple predictor variables. This method is often easily interpretable and can fit any dataset reasonably well.

Before selecting the predictors to be fitted in an MLR model, we remove the variables deemed insignificant by the chi-square test we conducted earlier and focus on the predictors that have a significant association with the response variable *quality*. Specifically, we remove *residual.sugar* from the red wine dataset, and *citric.acid* and *free.sulfur.dioxide* from the white wine dataset. For simplicity, we ignore models with interaction terms.

Using a best subset algorithm[1], we determined the lowest 10-fold cross-validation error model for red wine to be *quality* ~ *volatile.acidity* + *chlorides* + *free.sulfur.dioxide* + *total.sulfur.dioxide* + *pH* + *sulphates* + *alcohol* (CV error = 0.4210929), and that for white wine to be *quality* ~ *fixed.acidity* + *volatile.acidity* + *total.sulfur.dioxide* + *density* + *sulphates* + *alcohol* (CV error = 0.5837895). These models are easily interpretable, but in fact quite surprising: when rating the quality of red and white wines, the customers are influenced by very different factors that are often imperceptible.

Refitting the models selected above on the original datasets, the best MLR models with coefficients can be found in Appendix II. Interestingly, we observed that for both red wines and white wines, volatile acidity has a negative effect on the quality rating, while sulphates and alcohol have positive effects on the quality rating.

Ridge Regression

For each type of wine, we created a model containing all the predictors (after removing the insignificant predictors in each dataset) by using a parametric machine learning model called ridge regression. Ridge regression shrinks the coefficient estimates, reducing their variance significantly. Due to the bias-variance tradeoff, the reduction in variance comes with an increase in bias. The goal of ridge regression is to find the best tuning parameter that makes the RSS small while shrinking the coefficient estimates toward zero. We first use cross-validation to find the best tuning parameter by calling the built-in function cv.glmnet() in the glmnet package. The

[1]For each dataset, the five models with the lowest CV error can be found in Appendix I.

function performs a ten-fold cross-validation by default and we set a random seed to make the results reproducible. The minimum tuning parameter for red wine is 0.03844171, while that of white wine is 0.03857224. After we obtain the minimum tuning parameter values, we call the glmnet() function to fit the ridge regression using these values.

The resulting coefficients for both wines are small numbers, but none of the coefficients is zero. The best ridge regression models with coefficients are can be found in Appendix III. Both models indicate that fixed acidity, total sulfur dioxide, and density have different effects on the wine quality for the two types of wine. For example, total sulfur dioxide and density are negatively correlated with red wine quality, while they are positively correlated with white wine quality.

Ridge regression performs reasonably well on both datasets, yielding CV errors of 0.4137 (red) and 0.5839 (white). However, it includes all the predictors in the final model, making model interpretation more difficult.

Lasso Regression

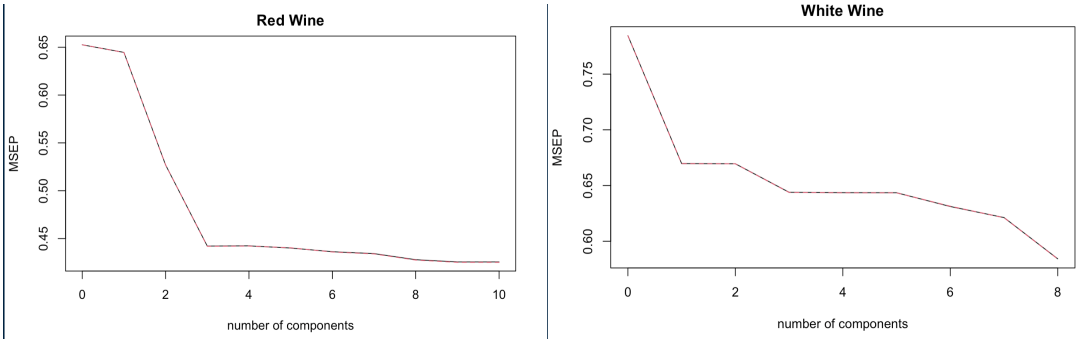
Another parametric machine-learning method we use is Lasso regression. Both Lasso and ridge regression shrink the coefficient estimates to reduce variance, but the Lasso uses a different penalty term that can force some of the coefficient estimates to be exactly zero, effectively performing variable selection. As with ridge regression, we first used the cv.glmnet() function to find the best tuning parameter through cross-validation. For the Lasso models, we set the α value to 1 instead of 0. We also set a random seed before we ran the ten-fold cross-validation. The minimum tuning parameter for red wine is 0.006412464, while that of white wine is 0.0006899222. We fit the Lasso regression with these tuning parameter values, and the results yield two sparse models. For the red wine, three variables are dropped (set to zero) from the model: fixed acidity, citric acidity, and density. On the other hand, the sparse model for white wine only dropped one variable, citric acidity. The best Lasso models with coefficients can be found in Appendix IV.

By looking at both models, we observed that total sulfur dioxide has different effects on the wine quality. Total sulfur dioxide is negatively correlated with the quality of red wine, while it is positively correlated with the quality of white wine.

Both Lasso regression models keep the same predictors. The model for red wine dropped two predictors (citric acidity and free sulfur dioxide), while the model for white wine did not drop any predictors. These results show that the same factors influence wine quality for both red and white wine, but the impacts of certain predictors are different. For instance, increasing the amount of total sulfur dioxide improves the quality of white wine while reducing the quality of red wine. Thus, we can see that the factors that determine the quality of red wine differ from those for red wine.

Principal Component Regression

Now, we create a new model using a parametric machine learning method: principal component regression (PCR). PCR is an effective method used to solve high-dimensional problems. Although it does not perform variable selection, it's usually very good at reducing the number of dimensions in the model and can model the response reasonably well.



We begin by fitting the models using the pcr() function in R and analyzing the optimal number of principal components in the model in order to achieve the lowest cross-validation error. As shown by the graphs above, the CV error of the model for red wine levels off after using eight principal components, and for white wine does not level off after using the maximum of eight principal components. Therefore, we obtain an 8-component model for predicting the quality of red wines and an 8-component model for predicting the quality of white wines.

However, principal component regression is not very meaningful in this situation. It does not perform model selection as expected nor does it sufficiently reduce the number of dimensions. In fact, the optimal number of principal components is almost as large as the total number of predictors. Because the number of observations is significantly greater than the number of predictors in both of our datasets, PCR is likely not a meaningful way to build our

desired models. Moreover, PCR typically works best when there exists high multicollinearity between a large number of variables; however, according to our correlation matrix, the majority of the correlation coefficients are smaller than 0.3, and we have already removed the variables with extremely high collinearity, so PCR is not able to fully utilize its strength for our datasets. For this reason, we decide to try out two other non-parametric methods.

Local Regression

Local regression is a powerful approach that combines the simplicity of linear regression with the flexibility of nonlinear regression. The tuning parameter for local regression is the span, which controls the proportion (from 0 to 1) of points near the input point the fitting algorithm will use for the regression, allowing for a kind of moving average. However, it is computationally expensive and only allows 1-4 predictors to be fit at once. It also has difficulty predicting the response for input values far away from those it was trained on. We begin by finding all combinations of three predictors from the list of predictors of the quality of red wine, of which there are 11, and from the list of predictors of the quality of white wine, of which there are 9. Our function of choice is loess(), as it allows us to force the outlying observations containing the minimum or maximum for each predictor into the training set. Varying the number of predictors from 1 to 3, the predictors themselves, and the span from 0.1 to 1 in intervals of 0.1, we fit 1,750 models on the red wine quality training set and 920 models on the white wine quality training set.

We conclude that the best model[2] for the quality of red wine contains two predictors, volatile acidity and sulphates and uses a span value of 1.0, yielding a CV error of about 0.4109. We find this result surprising because of the model's apparent sparsity, which typically results in large bias and low variance, a suboptimal tradeoff. The best model for the quality of white wine also contained two predictors, volatile acidity and total sulfur dioxide, but used a lower span value of 0.6 and yielded a larger CV error of around 0.5629.

Generalized Additive Models

The generalized additive model (GAM) is another flexible nonparametric machine learning method that allows us to model non-linear relationships, i.e., splines, for each predictor

[2] For each dataset, the five models with the lowest CV error can be found in Appendix V.

and add them together. Its tuning parameters are the degrees of freedom for each predictor, which determine the curviness of the spline for each predictor. We run 3,500 models on the red wine dataset and 1,840 models on the white wine dataset, varying the number and names of the predictors included and the degrees of freedom. For simplicity, we assign the spline for each predictor in the model the same degrees of freedom (k, for k a varying integer). We also forced the observations with the two most outlying observations (minimum and maximum) for every predictor into the training set so that the GAMs would be able to predict the response at similarly extreme input values.

The optimal model[3] for the red wine dataset contains the predictors volatile acidity, sulphates, and alcohol, all with 19 degrees of freedom, and yields a CV error of about 0.4110. Volatile acidity and alcohol are also included in the optimal model for the white wine dataset, but the predictor sulphates is replaced by total sulfur dioxide, the degrees of freedom are lower at 16, and the resulting CV error is higher at 0.5604.

Discussion

Model Cross-Comparison

For each method, we consider the model with the lowest CV error and employ a two-step approach to cross-assess their fits. First, we place all the models in the same loop and train them on the same data. Next, we use ten-fold cross-validation to estimate the mean-squared test error for each of the models[4], averaging the individual errors obtained using seed(k), with k an integer ranging from 1 to 100. The optimal method for the red wine dataset, yielding a CV error of 0.4082, is a GAM with the predictors volatile acidity, sulphates, and alcohol, all possessing 19 degrees of freedom. The optimal method for the white wine dataset, yielding a CV error of 0.5604, is also a GAM with the predictors volatile acidity, total sulfur dioxide, and alcohol, holding 16 degrees of freedom.

GAMs are the preferred method for both datasets, followed closely by MLR in the red wine dataset (with a CV error 0.0042 greater) and by Lasso regression in the white wine dataset (with a CV error 0.0220 greater). MLR also shows promise on the white wine dataset, merely

[3] For each dataset, the five models with the lowest CV error can be found in Appendix VI.
[4] For each dataset, the complete ranking of models with ascending CV error can be found in Appendix VII.

adding 0.0004 units of CV error to the Lasso method. Similarly, Lasso regression performs reasonably well on the red wine dataset, with a CV error only 0.0011 units greater than that of MLR. In both datasets, Lasso regression performs better than Ridge regression, and the difference is 8 times as large in the white as in the red wine dataset at 0.0016 versus 0.0002. PCR and local regression perform the worst on both datasets, fitting the new data very poorly compared to the other four methods. The inferiority of these methods is more apparent in the white wine dataset, in which their average CV is about 0.1129 units higher than that of the GAM.

This result does not come as a surprise: PCR and local regression, like most non-parametric methods, suffer from the curse of dimensionality, so they are only able to withstand a maximum of four predictors and lose information when making predictions. GAMs, on the other hand, although also non-parametric, are able to overcome this problem and therefore dominate all the other models.

Conclusion

We conclude that our GAMs are best for predicting the quality of both red and white wines. They provide evidence that both volatile acidity and alcohol are useful in predicting the quality of Vinho Verde wine, with sulphates especially useful for red wine and total sulfur dioxide useful for white wine. This research can be extended by analyzing the data using other machine learning tools and examining a larger set of predictors collected over a longer time period. Other predictors that may impact customers' ratings of wine quality include chlorides, free sulfur dioxide, and pH for red wine, and fixed acidity and density for white wine. Given the industrialization of wine-making, it may be worthwhile to determine whether, and if so, how, the factors affecting the wine ratings evolve over time.

References

Annie. (2022, November 18). *Vinho Verde Wine Guide: Portugal's prominent wine*. Wineries Guide & Wine Tips. Retrieved December 3, 2022, from https://sonomawinegarden.com/vinho-verde-wine/

Twohig, A. (2009, May 28). *Wine-tasting 101: The Four Factors*. Press Banner. Retrieved December 3, 2022, from https://pressbanner.com/wine-tasting-101-the-four-factors/

Wine Quality Data Set. UCI Machine Learning Repository: Wine quality data set. (n.d.). Retrieved December 3, 2022, from https://archive.ics.uci.edu/ml/datasets/wine+quality

Appendix

I. Multiple Linear Regression Model Selection

Top 5 Red Wine Models				
Iteration	Model	AIC	BIC	CV Error
957	quality ~ volatile.acidity + chlorides + free.sulfur.dioxide + total.sulfur.dioxide + pH + sulphates + alcohol	3158.977	3207.371	0.4210929
1008	quality ~ volatile.acidity + citric.acid + chlorides + free.sulfur.dioxide + total.sulfur.dioxide + pH + sulphates + alcohol	3159.839	3213.610	0.4219134
813	quality ~ volatile.acidity + chlorides + total.sulfur.dioxide + pH + sulphates + alcohol	3162.701	3205.719	0.4228839
993	quality ~ fixed.acidity + volatile.acidity + chlorides + free.sulfur.dioxide + total.sulfur.dioxide + pH + sulphates + alcohol	3160.961	3214.732	0.4229068
1024	quality ~ fixed.acidity + volatile.acidity + citric.acid + chlorides + free.sulfur.dioxide + total.sulfur.dioxide + density + pH + sulphates + alcohol	3163.470	3227.996	0.4233165

Top 5 White Wine Models				
Iteration	Model	AIC	BIC	CV Error
232	quality ~ fixed.acidity + volatile.acidity + total.sulfur.dioxide + density + sulphates + alcohol	11267.99	11319.97	0.5837895
256	quality ~ fixed.acidity + volatile.acidity + chlorides + total.sulfur.dioxide + density + pH + sulphates + alcohol	11264.21	11329.18	0.5838441
182	quality ~ fixed.acidity + volatile.acidity + density + sulphates + alcohol	11268.93	11314.40	0.5839304
253	quality ~ fixed.acidity + volatile.acidity + total.sulfur.dioxide + density + pH + sulphates + alcohol	11269.31	11327.78	0.5840748
252	quality ~ fixed.acidity + volatile.acidity + chlorides + density + pH + sulphates + alcohol	11265.74	11324.21	0.5841103

II. Multiple Linear Regression Model with Coefficients

Red Wine:

$$\hat{quality} = 4.430099 - 1.012753(volatile.acidity) - 2.017814(chlorides) + 0.005077(free.sulfur.dioxide) - 0.003482(total.sulfur.dioxide) - 0.482661(pH) + 0.882665(sulphates) + 0.289303(alcohol)$$

White Wine:

$$\hat{quality} = -43.23 - 0.09724(fixed.acidity) - 2.116(volatile.acidity) + 0.0005265(total.sulfur.dioxide) + 46.07(density) - 0.2988(sulphates) + 0.4131(alcohol)$$

III. Ridge Regression Model with Coefficients

Red Wine:

$$\hat{quality} = 20.107919837 + 0.022325775(fixed.acidity) - 1.021693384(volatile.acidity) - 0.083060965(citric.acid) - 1.802714149(chlorides) + 0.004040670(free.sulfur.dioxide) - 0.003018585(total.sulfur.dioxide) - 16.081276651(density) - 0.358054856(pH) + 0.866200900(sulphates) + 0.266945413(alcohol)$$

White Wine:

$$\hat{quality} = -23.64733 - 0.08664864(fixed.acidity) - 1.934014(volatile.acidity) - 1.812097(chlorides) + 0.0005144724(total.sulfur.dioxide) + 27.05839(density) - 0.02246347(pH) + 0.3279441(sulphates) + 0.3503117(alcohol)$$

IV. Lasso Linear Regression Model with Coefficients

Red Wine:

$$\hat{quality} = 4.257251206 - 1.028144571(volatile.acidity) - 0.014172392(citric.acid) - 1.77444116(chlorides) + 0.003053570(free.sulfur.dioxide) - 0.002849922(total.sulfur.dioxide) - 0.414998628(pH) + 0.835388998(sulphates) + 0.286745476(alcohol)$$

White Wine:

$$\hat{quality} = -41.01034 - 0.1011538(fixed.acidity) - 2.080723(volatile.acidity) - 1.418902(chlorides) + 0.0005640614(total.sulfur.dioxide) + 44.25769(density) - 0.06784844(pH) + 0.3113071(sulphates) + 0.4011643(alcohol)$$

V. Local Regression Model Selection

Top 5 Red Wine Models			
Iteration	Model	Span	CV Error
1190	quality~volatile.acidity+sulphates	1.0	0.4109127
1189	quality~volatile.acidity+sulphates	0.9	0.4183140
1188	quality~volatile.acidity+sulphates	0.8	0.4202619
1187	quality~volatile.acidity+sulphates	0.7	0.4208661
1186	quality~volatile.acidity+sulphates	0.6	0.4220263

Top 5 White Wine Models			
Iteration	Model	Span	CV Error
656	quality~volatile.acidity+total.sulfur.dioxide	0.6	0.5628795
203	quality~volatile.acidity	0.3	0.5631776
705	quality~volatile.acidity+pH	0.5	0.5638768
706	quality~volatile.acidity+pH	0.6	0.5639380
659	quality~volatile.acidity+total.sulfur.dioxide	0.9	0.5644978

VI. General Additive Model Selection

Top 5 Red Wine Models			
Iteration	Model	DF	CV Error
2379	quality ~ s(volatile.acidity) + s(sulphates) + s(alcohol)	19	0.4109705
2378	quality ~ s(volatile.acidity) + s(sulphates) + s(alcohol)	18	0.4110045
2380	quality ~ s(volatile.acidity) + s(sulphates) + s(alcohol)	20	0.4110260
2377	quality ~ s(volatile.acidity) + s(sulphates) + s(alcohol)	17	0.4111290
2376	quality ~ s(volatile.acidity) + s(sulphates) + s(alcohol)	16	0.4113425

Top 5 White Wine Models			
Iteration	Model	DF	CV Error
1316	quality ~ s(volatile.acidity) + s(total.sulfur.dioxide) + s(alcohol)	16	0.5603862
1315	quality ~ s(volatile.acidity) + s(total.sulfur.dioxide) + s(alcohol)	15	0.5603983
1317	quality ~ s(volatile.acidity) + s(total.sulfur.dioxide) + s(alcohol)	17	0.5603984
1318	quality ~ s(volatile.acidity) + s(total.sulfur.dioxide) + s(alcohol)	18	0.5604280
1314	quality ~ s(volatile.acidity) + s(total.sulfur.dioxide) + s(alcohol)	14	0.5604417

VII. Best Models for Cross Comparison

(a) Red Wine

Method	**Model**	**CV Error**[5]
Generalized Additive Model	gam(quality ~ s(volatile.acidity, df = 19) + s(sulphates, df = 19) + s(alcohol, df = 19), data)	0.4082425
Multiple Linear Regression	lm(quality ~ volatile.acidity + chlorides + free.sulfur.dioxide + total.sulfur.dioxide + pH + sulphates + alcohol, data)	0.4124466
Lasso Regression	glmnet(X, y, alpha = 1, lambda = 0.006412464)	0.4135951
Ridge Regression	glmnet(X, y, alpha = 0, lambda = 0.03844171)	0.4137496
Principal Component Regression	pcr(quality~., ncomp = 8, data)	0.4981279
Local Regression	loess(quality~volatile.acidity+sulphates, span = 1.0, data)	0.5027779

(b) White Wine

Method	**Model**	**CV Error**
Generalized Additive Model	gam(quality ~ s(volatile.acidity, df = 16) + s(total.sulfur.dioxide, df = 16) + s(alcohol, df = 16), data)	0.5603338
Lasso Regression	glmnet(X, y, alpha = 1, lambda = 0.0006899222)	0.5823121
Multiple Linear Regression	lm(quality ~ fixed.acidity + volatile.acidity + total.sulfur.dioxide + density + sulphates + alcohol, data)	0.5827139
Ridge Regression	glmnet(X, y, alpha = 0, lambda = 0.03857224)	0.5839236
Principal Component Regression	pcr(quality~., ncomp = 8, data)	0.6287342
Local Regression	loess(quality ~ volatile.acidity + total.sulfur.dioxide, span = 0.6, data)	0.7176594

[5] CV Error is computed using 10 fold cross-validation; we average the results using seed(k), with k from 1 to 100.