STAT 6021:
PROJECT 2 REPORT

Aug 6, 2020

Hemani Choksi, Oretha Domfeh,
Paul Hicks, Sudhanshu Luthra

UNIVERSITY of VIRGINIA

×

UCI
Machine Learning Repository

SUMMARY

Project Goals

The primary goal of this project was to illuminate potential ways the vineyard could be more profitable by changing some of its wine making processes to lead to higher quality ratings. Our objectives included gaining an understanding of the various factors that lead to quality ratings. We expected that not all of the available data points held the same importance across the board and we wanted to discover these relationships and the relatively more important focus areas. We aimed to discover a small, simple but impactful set of options that warrant further analysis for potential changes, and aimed to result in only high confidence recommendations.

The Data

To inform the team's analysis, we developed a model using a data set provided from the University of Minho, Portugal regarding the "Vino Verde" wine. The data set contains technical information detailed in about ~1600 observations of wine quality that each include 11 numerical variables. We will go into the technical details later in this report regarding each of these factors and how they were considered in our analysis but in general each attribute relates to some form of the chemical processes in "Vino Verde's" make-up. The quality ratings fall in categories ranging from lowest ("3") to highest ("8") with the majority of observations in the mid-levels of 5 and 6. Our goal, as mentioned, would be to enable the vineyard owner to make changes to their chemical make-up to increase the wine ratings and potentially realize higher profits as a result.

Results/Conclusions

From the approximately 1600 discrete observations on quality associated each with eleven attributes, we were able to use quantitative and statistical methods to result in 2 major recommendations for the vineyard. We were able to develop an accurate model in the end that provides about an 82% accuracy level, and think there is real potential to gain higher quality ratings more consistently if the owner focuses on 2 of the 11 attributes, namely the amount of sulphates and alcohol. Sulphates show great potential in increasing the probability of gaining a higher quality rating by making incremental changes in the relevant levels, showing an increase in the probability of a higher quality wine rating by as much as 28 times by making a one unit change in the sulphates level during production and controlling other variables. The vineyard also could double the probability of gaining a high quality rating by a one unit change in the alcohol level. These are our two main outcomes from our exploratory data analysis and predictive model building. That said, each of these decisions needs a bit more context, including the recommendation that these options go through more cost/benefit analysis before wide adoption to ensure they are viable and have a strong business case.

DETAILED DESCRIPTION

The Dataset

Our data set utilizes publicly available data from the UCI Machine Learning Repository. This wine quality dataset is sourced from the University of Minho, Portugal, and contains information about physicochemical variables (inputs) and sensory variables (outputs) related to red and white variants of the Portuguese "Vinho Verde" wine.[2] A summary of the initial variables of interests we explored are listed in Table 1 below.

Variable Name:	Description
fixed acidity	(*g*(tartaric acid)/dm^3) [1]
volatile acidity	(*g*(acetic acid)/dm^3) [1]
citric acid	(g/dm^3) [1]
residual sugar	(g/dm^3) [1]
chlorides	(*g*(sodium chloride)/dm^3) [1]
free sulfur dioxide	(mg/dm^3) [1]
Total sulfur dioxide	(mg/dm^3) [1]
density	(g/cm^3) [1]
pH	pH of the wine [1]
sulphates	(*g*(potassium sulphate)/dm^3) [1]
alcohol	(vol.%) [1]
Quality	Sensory, based on scale of 0-10 [1]

All of the above mentioned predictors are quantitative continuous predictors, measuring physical/chemical properties of the wine. Of note, our response variable, “quality”, is categorical predictor with discrete values. The total amount of data points in this dataset were 1599. After pulling in the dataset, we ran a scatterplot of all the quantitative predictors and found the following results:

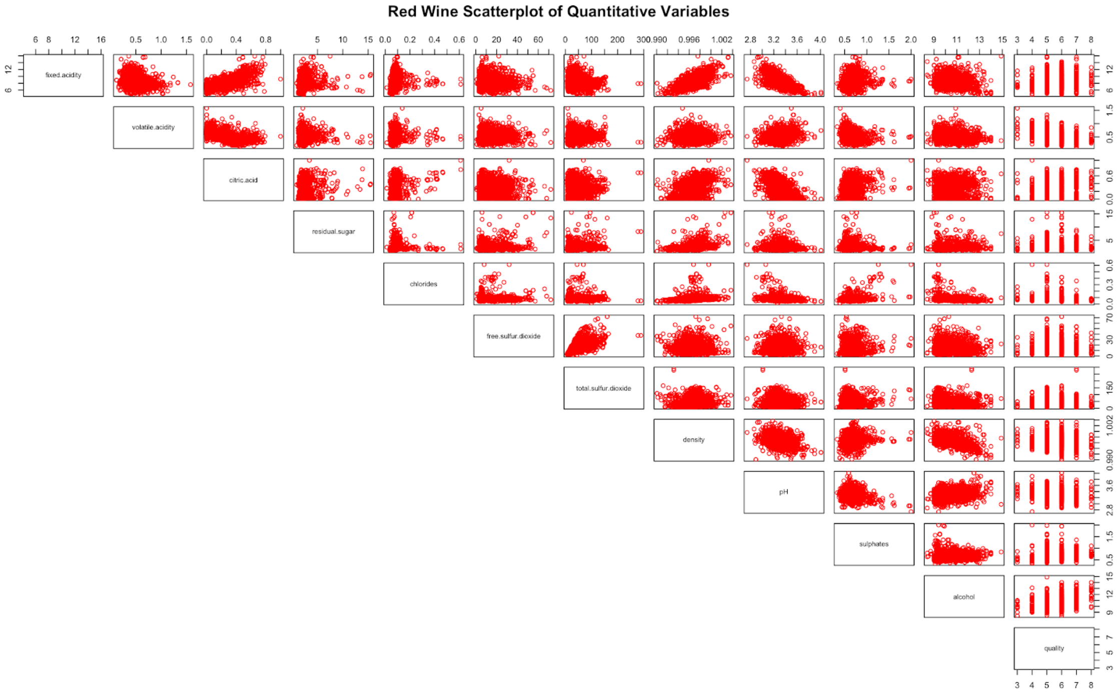
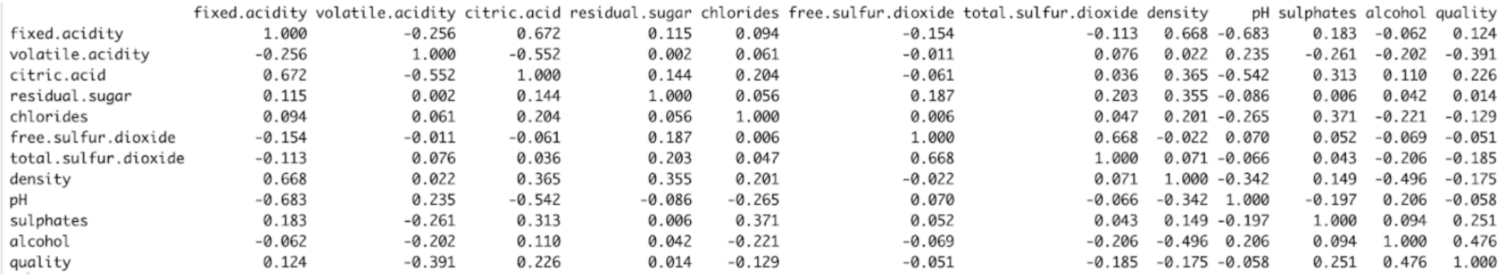


Chart 1.

Of note, one can observe from the scatterplot matrix that all predictors held at least some linear correlation to the response variable. We verified this by SLR for each predictor against the quality level. Further, we proceeded to conduct more in depth analysis regarding combinations of different predictors and to see if any predictors are correlated with each other.

Correlation Matrix (with All Predictors)

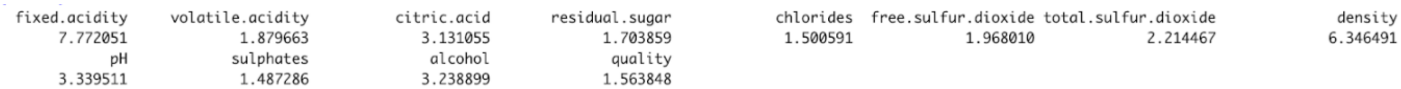


```
                     fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide total.sulfur.dioxide density     pH sulphates alcohol quality
fixed.acidity                1.000           -0.256       0.672          0.115     0.094              -0.154               -0.113   0.668 -0.683     0.183  -0.062   0.124
volatile.acidity            -0.256            1.000      -0.552          0.002     0.061              -0.011                0.076   0.022  0.235    -0.261  -0.202  -0.391
citric.acid                  0.672           -0.552       1.000          0.144     0.204              -0.061                0.036   0.365 -0.542     0.313   0.110   0.226
residual.sugar               0.115            0.002       0.144          1.000     0.056               0.187                0.203   0.355 -0.086     0.006   0.042   0.014
chlorides                    0.094            0.061       0.204          0.056     1.000               0.006                0.047   0.201 -0.265     0.371  -0.221  -0.129
free.sulfur.dioxide         -0.154           -0.011      -0.061          0.187     0.006               1.000                0.668  -0.022  0.070     0.052  -0.069  -0.051
total.sulfur.dioxide        -0.113            0.076       0.036          0.203     0.047               0.668                1.000   0.071 -0.066     0.043  -0.206  -0.185
density                      0.668            0.022       0.365          0.355     0.201              -0.022                0.071   1.000 -0.342     0.149  -0.496  -0.175
pH                          -0.683            0.235      -0.542         -0.086    -0.265               0.070               -0.066  -0.342  1.000    -0.197   0.206  -0.058
sulphates                    0.183           -0.261       0.313          0.006     0.371               0.052                0.043   0.149 -0.197     1.000   0.094   0.251
alcohol                     -0.062           -0.202       0.110          0.042    -0.221              -0.069               -0.206  -0.496  0.206     0.094   1.000   0.476
quality                      0.124           -0.391       0.226          0.014    -0.129              -0.051               -0.185  -0.175 -0.058     0.251   0.476   1.000
```

Table 1.

Above (**Table 1.**) is the correlation matrix with all of our predictors. We noted that none of our predictors had greater than 0.50 correlation with quality, but that alcohol was the strongest association with a moderately positive correlation with quality. We also noted that fixed acidity, volatile acidity, citric acid, density, free sulfur dioxide, total sulfur dioxide have moderately high correlations with each other, particularly between free and total sulfur dioxide. To reaffirm what we observed in the correlation matrix, we ran a calculation of VIF for the predictors.

VIF (Variance Inflation Factors)



```
 fixed.acidity     volatile.acidity          citric.acid       residual.sugar            chlorides  free.sulfur.dioxide total.sulfur.dioxide              density
      7.772051             1.879663             3.131055             1.703859             1.500591             1.968010             2.214467             6.346491
            pH            sulphates              alcohol              quality
      3.339511             1.487286             3.238899             1.563848
```

Table 2.

As you can see from this VIF table (**Table 2.**), it shows evidence for higher variance rates for fixed acidity, citric acid, density, pH, and alcohol due to collinearity. For example, the VIF for fixed acidity is 7.77, which tells us that the variance for fixed acidity is 7.77 times larger than it would have been without collinearity. Later down below, we conduct a partial F-test to see if we can drop chlorides, pH, free sulfur dioxide, and total sulfur dioxide from the model given issues with multicollinearity mentioned here and using the results from the box-plots to see which predictors are the most influential factors on the response (quality level) given their distributions.

Regsubsets

To do a more comprehensive analysis of different predictor combinations, the group utilized the regsubsets() function. This allowed us to see which predictors would create the best model for different criteria (adjusted R2, MSE, Cp, and BIC). The following predictors lead to the best first-order model using the different criteria:

Adjusted R2: volatile.acidity, citric.acid, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, pH, sulphates, alcohol

MSE: volatile.acidity, citric.acid, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, pH, sulphates, alcohol

Cp: volatile.acidity, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, pH, sulphates, alcohol

BIC: volatile.acidity, chlorides, total.sulfur.dioxide, pH, sulphates, alcohol

Moreover, we also used automated search procedures such as forward, both, and backward passes to aid in the variable selection process. Each of these methods rely on the AIC metric to iterate over potential solutions to either increase or drop the # of available predictors as the progress. Interestingly, all of the automated model selection algorithms resulted in the same predictors. For the forward pass the variables selected were alcohol, volatile.acidity, sulphates, total.sulfur.dioxide, chlorides, pH, free.sulfur.dioxide. Based on backward pass results the model chosen was volatile.acidity, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, pH, sulphates and alcohol. The “both” algorithm results are similar to the previous methods and we were able to converge on an initial candidate model relatively quickly. In this case, we chose the forward selection model in part because it added alcohol as the first predictor, which we know from the earlier correlation matrix was the strongest correlation with quality. In this manner, we thought we had the best chance of achieving an accurate model. Using predictors for the forward pass results: alcohol, volatile.acidity, sulphates, total sulfur dioxide, chlorides, pH, free.sulfur.dioxide. All of the predictors are significant, however, the model only explains about 36% of the variation in the wine quality.

Box Plots

From the forward, both, and backward auto selection models the key takeaway was the potential contributions of 7 of the 11 available predictors might make towards our quality level rating. To further gain insight into these most impactful predictors available in our data set to inform model building, we conducted a series of box plots relating the response to each of the 7 predictors. The strongest linear relationship between quality (shown on the x-axis) and predictor values is with the first

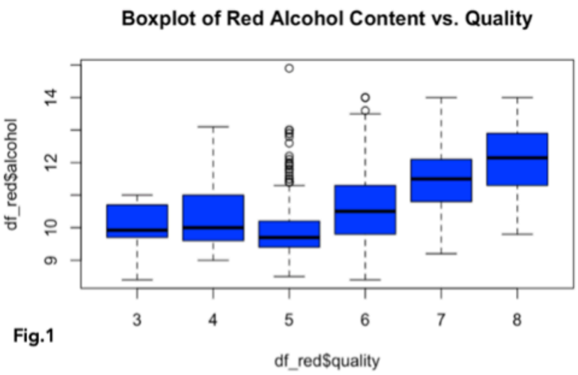


Fig.1

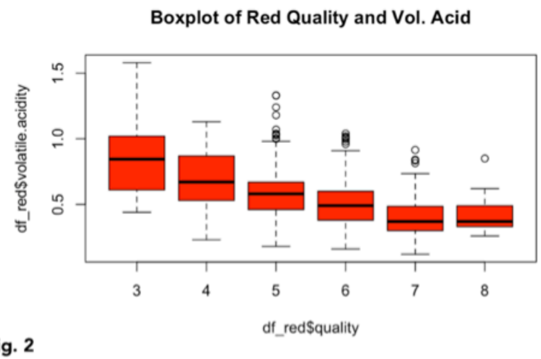


Fig. 2

two figures regarding **alcohol content** (**Fig.1**) and **volatile acidity (Fig. 2)**. Though less dramatic, sulphate levels (Fig. 3) appear correlated with quality ratings as well.

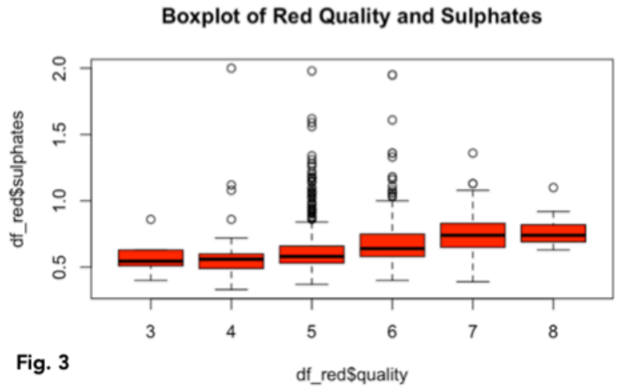
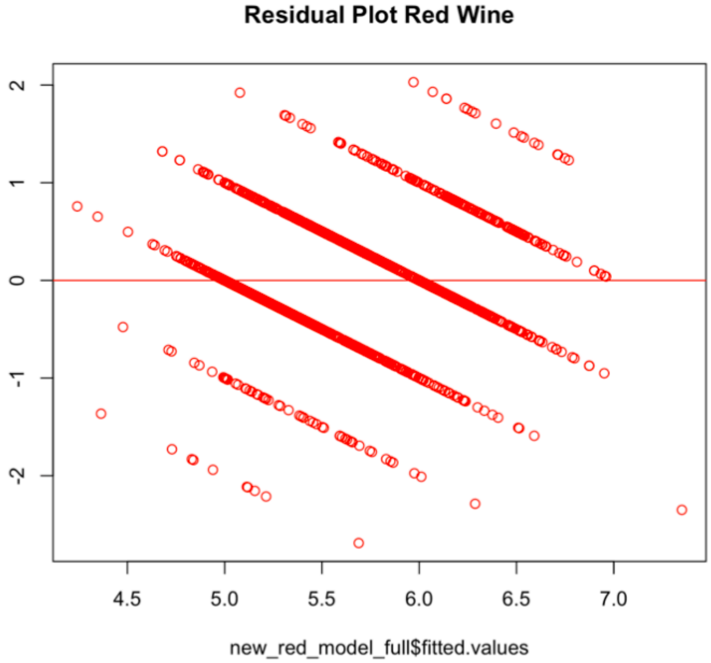


Fig. 3

Of these seven predictors, quality ratings appear sensitive to alcohol, volatile acidity, and sulphate. Conversely, quality rating was found to be less sensitive to chlorides, pH, free sulfur dioxide, and total sulfur dioxide. This realization is somewhat consistent with our results from the automatic model searching algorithm results and correlation matrix. We can also see that at quality levels 5 and 6 in several of the box plots, there appears to be a number of potential outliers though this is somewhat expected given that most of our observations fall in these two levels.

Given the potential to simplify our model as indicated by the box plots and in an attempt to improve our model accuracy, we conducted a partial F-test to consider dropping 4 predictors (chlorides, pH, free sulfur dioxide, and total sulfur dioxide). The ANOVA results indicated we could not reject the null



hypothesis because the p value was less than 0.5.

To gain more information on the full 7 predictor model, we conducted a residual plot (shown above), which suggested that we needed an alternative method to model our data, given the results shown above.

Final Model

Overall, based on a poor accuracy conclusion in our MLR model (~36% of the response is explained) and recognition of discrete response values shown in the histogram below, we saw an opportunity for an alternative approach.

The histogram plot (Chart 2.) of the various quality levels shed light on the potential for a binomial logistic model setting instead of a linear regression model. We noticed that in addition to the data having discrete levels associated with the quality ratings (3, 4, .., 8), the ratings were generally equally divided between levels 5 and 6. Additionally, the distribution of ratings was highest in 5 and 6. In an effort to balance between the different levels, we saw a natural delineation of high versus low responses between these levels 5 and 6. We carried that forward in the modeling process and conducted binomial logistic regression with these two response levels to more accurately fit our data.

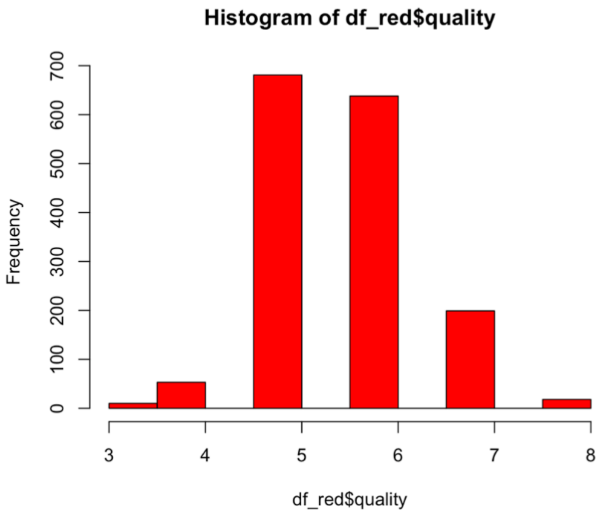


Chart 2.

For the binomial logistic regression model, we used the original 7 significant predictors from early EDA (alcohol, sulphates, volatile acidity, pH, free sulfur dioxides, total sulfur dioxides, and chlorides and got output from R (Table 3.) shown above. The results table highlighted in yellow above, each of the predictors had significance in their p-values except for pH. Using the Wald Test, we were able to drop pH from our model.

```
Call:
glm(formula = train_red$qual_level ~ train_red$alcohol + train_red$sulphates +  train_red$volatile.acidity +
train_red$chlorides + train_red$total.sulfur.dioxide +  train_red$free.sulfur.dioxide + train_red$pH, family =
"binomial",   data = train_red)

Deviance Residuals:
   Min      1Q   Median      3Q      Max
-2.8271  -0.8635   0.3454   0.8610   2.2652

Coefficients:
                                 Estimate      Std. Error     z value       Pr(>|z|)
(Intercept)                      -8.05417      2.06430        -3.902        9.55e-05 ***
train_red$alcohol                 0.66776      0.09524         7.011        2.36e-12 ***
train_red$sulphates              3.37808       0.63953         5.282        1.28e-07 ***
train_red$volatile.acidity       -3.05228      0.54795        -5.570        2.54e-08 ***
train_red$chlorides              -5.55027      1.92666        -2.881        0.00397 **
train_red$total.sulfur.dioxide   -0.01428      0.00367        -3.890        0.00010 ***
train_red$free.sulfur.dioxide    0.02645       0.01088        2.431         0.01504 *
train_red$pH                     0.43886       0.60232        0.729         0.46624
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

   Null deviance: 1104.88  on 798  degrees of freedom
Residual deviance:  857.65  on 791  degrees of freedom
AIC: 873.65

Number of Fisher Scoring iterations: 4
```

Table 3.

Our final logistic regression model is therefore the following (pi is the odds of having a high quality rating for the red wine):

$$\log(\pi/(1-\pi))=-6.78 + 0.68*\mathit{alcohol}+3.366*\mathit{sulphates} -2.92*\mathit{volatile.acidity}-5.85*\mathit{chlorides}$$

$$-0.015*\mathit{total.sulfur.dioxide}+0.028*\mathit{free.sulfur.dioxide}$$ [continued from previous line].

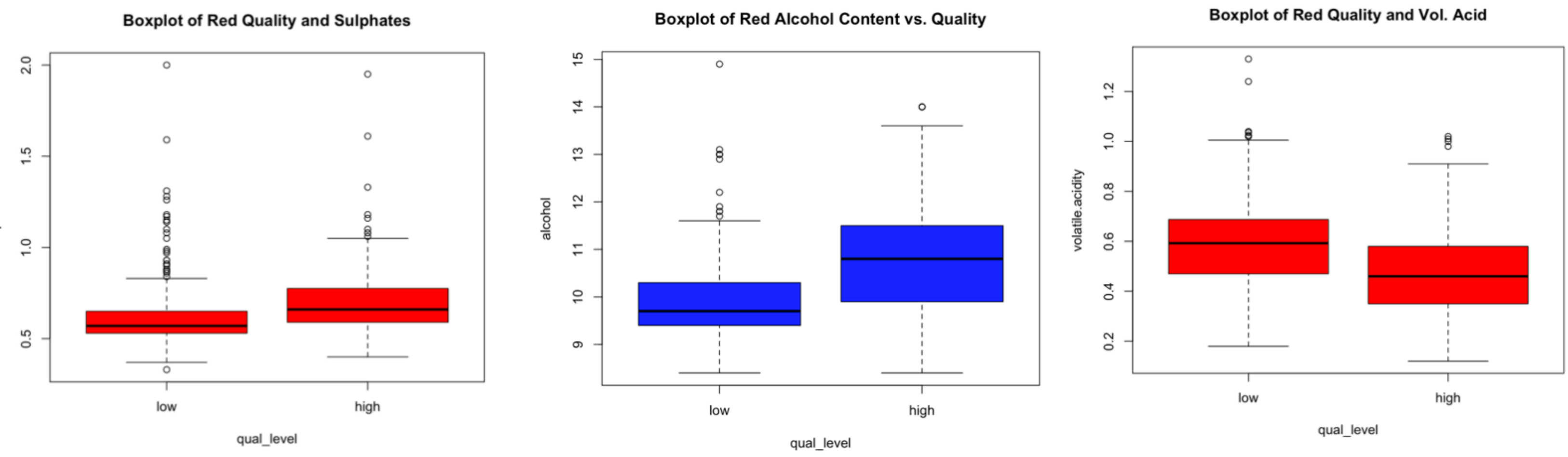


Fig. 5, 6, 7

We ended up running the box-plots again given change to a binomial response variable, and similar patterns remain: Alcohol, Sulphates, and Volatile Acid (Fig. 5, 6, 7) had the most influence on the response, quality level (qual_level) as shown in the shifts above.

To test the quality of our binomial logistic model we used the split half of the original data in our test set to validate our training model. Our resulting ROC Curve (Fig. 8) is shown below. We calculated an AOC value of 0.825, which again is an improvement in our model over our initial attempts at an MLR. The result of our confusion matrix (Table 3.) set at a 0.5 threshold showed us hat we have a false positive rating of (106/(106+262)), or 0.288 and a false negative rate of 0.222. Given the

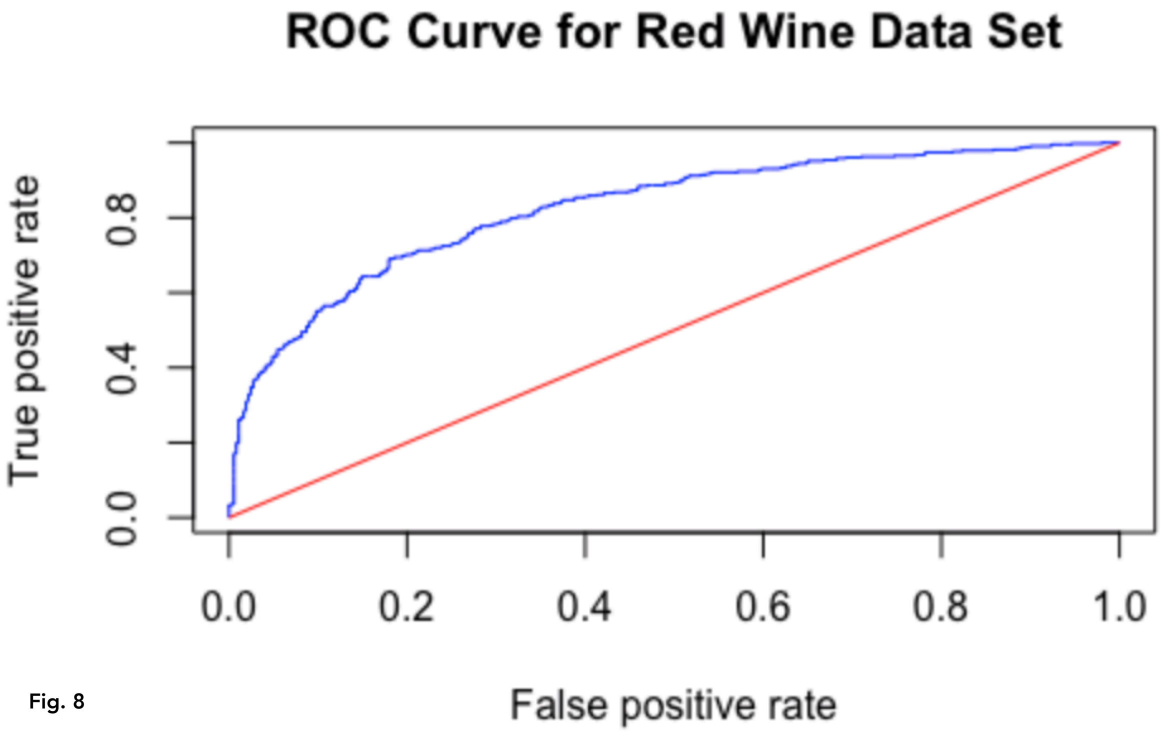


Fig. 8

financial implications of a high false positive rate, where conceivably the vineyard owner could make an investment decision that does not lead to an increase in the quality rating, we raised the threshold to 0.7 (results shown below).

Threshold 0.5

Quality Rating	FALSE	TRUE
"Low" Rating	262	106
"High" Rating	96	336

Threshold 0.7

Quality Rating	FALSE	TRUE
"Low" Rating	332	36
"High" Rating	197	235

At the cost of lowering the false positive rate down to (36/(36+332), or 0.098, we now have a false negative rating increase up to (197/(197+235)) or 0.456. This is an acceptable outcome given that we have increased confidence in any recommendations we will make using this model, which reduces potential overconfidence in any outcomes on increased quality after associated financial investments are made.

For a final check on our model's appropriateness for our data, we conducted a Deviance Goodness of Fit test on the overall model relative to our chosen predictors. ΔG2 = 219.66. Shown in our attached R code we obtained a p-value of 0 during our deviance goodness of fit test indicating that our model is useful in estimating the log odds of producing a high quality wine or conversely, a low quality wine. Combined with our ROC Curve, AOC value, and confusion matrix results we are in position to weigh various options for the vineyard owner.

Outlook

We now have potential options for the vineyard owner to consider future investments aimed at achieving higher quality ratings. The results fell into two distinct categories where some predictors had a higher potential impact on the quality rating, per change in the unit level for that predictor. Taking a look at these more impactful predictors first, we consider the impact of sulphates on the quality rating.

Sulphates

A one unit change in the sulphates level is expected to increase the odds of a higher quality rating by 28.95 times, when controlling for the other predictors. As a result, this is our strongest recommendation considering potential changes in the production process.

Alcohol content

In EDA, there was a large apparent association between the alcohol level and the quality rating. Using our model, we calculate the potential increase in the odds of gaining a high quality rating per unit increase in alcohol level as 1.97 times (or nearly double), when controlling the other predictors.

Free sulfur content

Our model indicates, we calculate the potential increase in the odds of gaining a high quality rating as a result of a one unit increase in free sulfur dioxide levels as 1.03 times, when controlling the other predictors.

Our remaining predictors did not offer these potential increases showing multiplying factors less than 1.0 due to their relatively smaller coefficients in our model

Recommendations

We recommend considering changes to both the sulphates and alcohol levels in the wine production process as they have the greatest probabilities of improving the overall quality ratings of your wine. That said, we do not know the viability of these options and the financial implications with such a choice and recommend weighing those two factors closely before proceeding. Moreover, we do not have access to the profits associated with the various quality levels, and are unable to provide quantitative analysis on the potential benefits of gaining higher ratings. Therefore, we recommend that a thorough cost benefit analysis is implemented based on the above recommendations.

Citations

1. Modeling wine preferences by data mining from physicochemical properties. (2009). *ELSEVIER*, 549. https://doi.org/10.1016/j.dss.2009.05.016.

2. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.