Discovering the Quality of Wine

Kasey Moran

Table of Contents

Executive Summary

Objective

The wine industry is a growing worldwide industry with a lot of opportunity. Wine has a culture of its own and the demand is ever growing. This report will describe what factors contribute to the quality of wine. These factors can be of high significance to a wine producer looking to improve the quality of their wine or a seller looking to sell only the best wines to their consumers. In 2021 the total worldwide sales of wine were $354.7 billion[1]. Capitalizing on the wine market and determining what wine features to select for production or sales could have a lucrative return.

Key Findings

- Alcohol percentage was found to be the most impactful variable on wine quality
- Volatile acidity, chlorides, total sulfur dioxide and pH had a negative effect on wine quality
- Free sulfur dioxide, sulphates and alcohol had a positive effect on wine quality

Conclusion and Recommendation

When creating a batch of wine, the most important factors to account for are alcohol, volatile acidity, sulphates, chlorides, and total sulfur dioxide. pH and free sulfur dioxide also have an impact but are much smaller. These insights can help a wine producer create a higher quality bottle of wine to increase their revenues. It can also lend insight to a wine seller or distributer looking to select only the finest quality of wines to sell. I recommend focusing on all these factors when trying to create a high-quality wine, if possible or in the least focusing on the three most impactful factors of alcohol, volatile acidity, and sulphates.

Introduction

There are many different types of wine on the market, but not all are of good quality. For a wine producer or a seller to consumers, it is important to know what factors will indicate a quality bottle of wine. This report will aim to identify those factors based on the data of red wines. I will use three models to determine which factors are the most impactful to a wine's quality.

The data being used for this analysis is from the UCI machine learning repository. The data points are on the red wines of the Portuguese Vino Verde varietal from 2009. There are 4898 total rows of data and 12 attributes described as follows.

Table 1. The variables names and their descriptions

Quality	The output variable. Quality is on a scale of 0-10, 0 being the worst.
Fixed acidity	All acids found in wine
Volatile acidity	Amount of acetic acid in the wine
Citric Acid	Amount of citric acid, which is an additive for acidity in wine
Residual Sugar	The amount of sugar remaining in the wine after fermentation
Chlorides	The amount of chlorides(salt) in the wine
Free sulfur dioxide	Amount of free sulfur dioxide in the wine
Total sulfur dioxide	Amount of free and bound sulfur dioxide in the wine
Density	The density depending on water in the wine
PH	The pH of the wine on a scale of 0 to 14
Sulphates	Amount of the wine additive sulphates in the wine
Alcohol	Percentage of alcohol in the wine

Methodology and Findings

Before any modeling was performed the data was evaluated for preprocessing. First, I checked for any missing values so that they could be removed before analysis. Then, since I planned to use regression and a neural network model, I decided to scale the data. Originally each feature was on a different scale so to improve performance and make understanding easier the dataset was scaled and normalized for both the multiple linear regression and neural network models. After this was complete, I ran three different models. All predictors were numerical and did not need to be converted into dummy variables.

The first model I decided to use was a multiple linear regression model. Since there are multiple variables, I thought being able to see individual variable importance would be insightful. I started with a multiple linear regression model including all variables with quality as the output.

Image 1. Multiple linear regression model with all variables

```
Call:
lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +
    residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
    density + pH + sulphates + alcohol, data = wine.df)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5378 -0.0733 -0.0094  0.0904  0.4050

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.54251    0.03019  17.971  < 2e-16 ***
fixed.acidity          0.05648    0.05864   0.963   0.3357
volatile.acidity      -0.31641    0.03536  -8.948  < 2e-16 ***
citric.acid           -0.03651    0.02944  -1.240   0.2150
residual.sugar         0.04769    0.04381   1.089   0.2765
chlorides             -0.22453    0.05023  -4.470 8.37e-06 ***
free.sulfur.dioxide    0.06193    0.03083   2.009   0.0447 *
total.sulfur.dioxide -0.18478    0.04125  -4.480 8.00e-06 ***
density               -0.04871    0.05893  -0.827   0.4086
pH                    -0.10507    0.04867  -2.159   0.0310 *
sulphates              0.30606    0.03819   8.014 2.13e-15 ***
alcohol                0.35906    0.03443  10.429  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1296 on 1587 degrees of freedom
Multiple R-squared:  0.3606,    Adjusted R-squared:  0.3561
F-statistic: 81.35 on 11 and 1587 DF,  p-value: < 2.2e-16
```

As seen in the results above only volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphates, and alcohol were statistically significant at a 5% level. Out of those variables, sulphates, alcohol, and free sulfur dioxide showed a positive effect on wine quality. Volatile acidity, chlorides, total sulfur dioxide and pH showed a negative effect on wine quality. The top three variables that showed the largest impact on quality were alcohol, sulphates, and volatile acidity.

To determine the best multiple linear regression model, I used the backward method next. The backward method allowed me to find the optimal model with the lowest AIC shown below. This method dropped the variables of fixed acidity, citric acid, residual sugar, and density.

Image 2. MLR using the backward method

```
Step:  AIC=-6527.77
quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
    total.sulfur.dioxide + pH + sulphates + alcohol

                      Df Sum of Sq    RSS     AIC
<none>                              26.701 -6527.8
- free.sulfur.dioxide   1    0.0958 26.797 -6524.0
- pH                    1    0.2829 26.984 -6512.9
- total.sulfur.dioxide  1    0.4315 27.133 -6504.1
- chlorides             1    0.4324 27.134 -6504.1
- sulphates             1    1.0824 27.784 -6466.2
- volatile.acidity      1    1.6927 28.394 -6431.5
- alcohol               1    4.9793 31.681 -6256.4
```

Image 3. Results of optimal MLR model

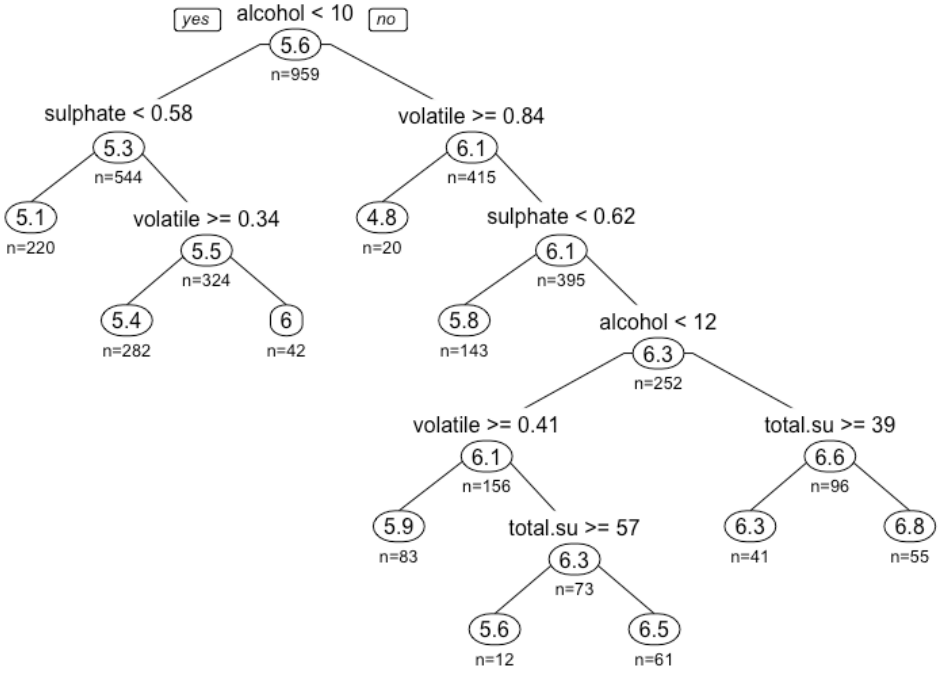
```
Call:
lm(formula = quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
    total.sulfur.dioxide + pH + sulphates + alcohol, data = wine.df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.53784 -0.07351 -0.00931  0.09216  0.40591

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.53349    0.01881  28.369  < 2e-16 ***
volatile.acidity      -0.29572    0.02945 -10.043  < 2e-16 ***
chlorides             -0.24173    0.04763  -5.076 4.31e-07 ***
free.sulfur.dioxide    0.07210    0.03018   2.389    0.017 *
total.sulfur.dioxide -0.19710    0.03887  -5.070 4.43e-07 ***
pH                    -0.12260    0.02986  -4.106 4.23e-05 ***
sulphates              0.29481    0.03671   8.031 1.86e-15 ***
alcohol                0.37609    0.02183  17.225  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1295 on 1591 degrees of freedom
Multiple R-squared:  0.3595,    Adjusted R-squared:  0.3567
F-statistic: 127.6 on 7 and 1591 DF,  p-value: < 2.2e-16
```

The results of the optimal model show that all seven remaining variables are statistically significant at the 5% level. Volatile acidity, chlorides, total sulfur dioxide and pH have a negative effect on quality while free sulfur dioxide, sulphates and alcohol have a positive effect. The optimal model is:

Quality = .5335 – 0.2957*VolatileAcidity* – 0.2417*chlorides* + 0.0721*freeSulfurDioxide* -0.1971*totalSulfurDioxide* - 0.1226*pH* + 0.2948*Sulphates* + 0.3761*Alcohol*

The second model I decided to use was a decision tree. The benefit of a decision tree is that it will provide a simple to understand visual overview for predicting wine quality. I used the non-scaled data for the regression tree so the numbers at each node were easier to understand in the context to a non-expert. I started with a regression tree using all variables and then pruned the tree according to the cp level with the lowest xerror. This cp level was .01 and the pruned tree is shown below.

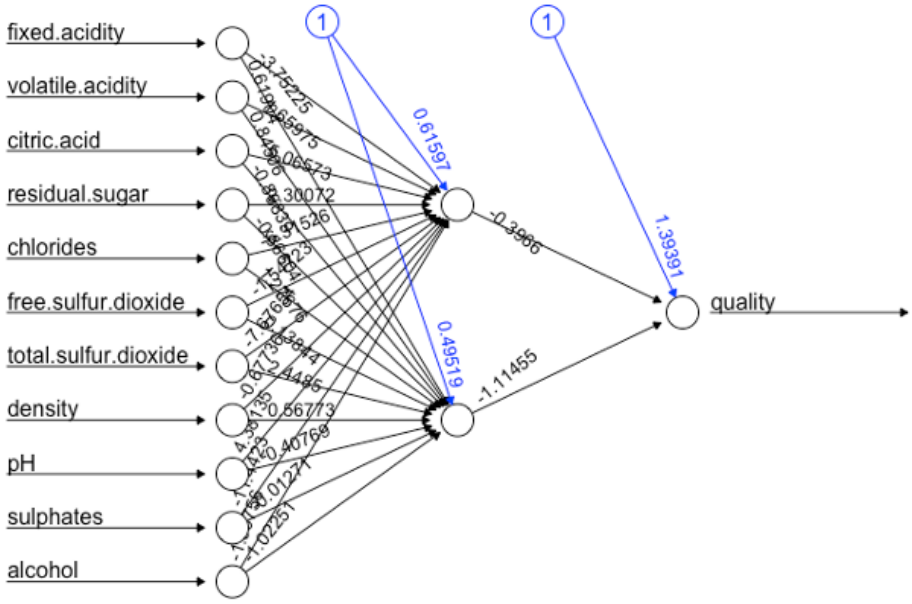
Image 4. Pruned regression tree



The regression tree first splits based on if the alcohol is above or below 10%. If alcohol is less than 10% the tree will next split on if sulphate is less than .58. If alcohol is greater than or equal to 10% the tree will next split on if volatile acidity is greater than or equal to .84.

The last model used was a neural network model. I decided to use a neural network of one hidden layer and two nodes. All eleven independent variables were included in the model shown next.

Image 5. Neural Network Model



In evaluating each of the three models, I believe that the best model would be the multiple linear regression created from the backward method. This model included only those variables that were statistically significant to the quality of wine. I also believe that the regression tree is a highly valuable model since it is easy to ready and simple to someone looking to predict the quality of wine. Alcohol showed to be the most impactful on quality in the multiple linear regression and the regression tree. Sulphates and volatile acidity were the next two variables that had the biggest influence on quality. In the multiple linear regression model alcohol had an estimate of .3761, Volatile acidity of -0.2957 and sulphates of 0.2948.

Conclusion

Out of the eleven original independent variables, I found seven to be statistically significant to impact the quality of wine. These are volatile acidity, chlorides, total sulfur dioxide, pH, free sulfur dioxide, sulphates, and alcohol. Free sulfur dioxide had the smallest impact and p-value but was still significant. The most impactful in order of effect was alcohol, volatile acidity, sulphates, chlorides, and then total sulfur dioxide. Alcohol has the largest positive effect on quality which makes sense because wines that are higher in alcohol tend to be more supple,

full bodied and even dense or chewy. The volatile acidity showed the largest negative effect on quality which is to be expected because this can give a sharp vinegary taste and smell to the wine which is not desired. Chlorides, which indicate how salty a wine is also produced an expected negative affect on quality.

There are some limitations of the data and models. The adjusted r-squared for the final multiple linear regression model was .3567, which indicates that 35.67% of the variability in the outcome is explained by the model. Because of this lower percentage I would recommend additional analysis adding more variables to see if we could explain more of the variability of quality. Also, the data was only on a specific red varietal of wine. I would recommend analyzing different varietals of red wine to see if more variability could be explained.

For a wine producer these findings may help them focus on factors so that they can produce a higher quality wine and increase their revenue. For a wine seller or importer, this can allow them to select wine with the factors that improve quality and avoid wines that may hinder the quality to ensure high numbers of sales. Worldwide wine consumption in 2021 was around 23.6 billion liters[2], so capitalizing on figuring out which factors produce the best quality wine is worthwhile.

References

[1] *Revenue of the wine market worldwide by country 2018*. (n.d.). Statista. https://www.statista.com/forecasts/758149/revenue-of-the-wine-market-worldwide-by-country

[2] *25+ Exquisite Wine Industry Statistics [2023]: Market Trends + Consumption Statistics*. (2023, March 9). Zippia. https://www.zippia.com/advice/wine-industry-statistics/#Global_Wine_Industry_Statistics

Rising Alcohol Levels: How Winemakers are Adjusting. (2017, October 24). SevenFifty Daily. https://daily.sevenfifty.com/taking-control-of-alcohol-levels-in-wine/#:~:text=Wines%20that%20are%20higher%20in

Volatile Acidity in Wine. (n.d.). Extension.psu.edu. Retrieved April 22, 2023, from https://extension.psu.edu/volatile-acidity-in-wine#:~:text=Volatile%20acidity%20(VA)%20is%20a%20measure%20of%20the%20wine