Data Science 2: Statistics for Data Science

Predicting Red Wine Quality

Velagala, Priyanka

Table of Contents

1.0 Overview

The chosen dataset from UCI's Machine Learning Repository has a database of various red wines from the north of Portugal where each wine is broken down in terms of their physicochemical properties and assigns them a quality based on sensory data. In total, there are 11 physicochemical properties that are assessed for each wine, the properties are- fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol content.

2.0 Objectives

The main objective is to build a model that leverages the above listed physicochemical properties and predicts the quality of a wine. A secondary objective in analysis is to understand from the given factors, which factor most and least influences the quality of wine. The model could potentially be used by vendors to help set a fair market price for a wine based on their characteristics. In addition, in identifying the most and least important factors, this could serve to alter growing and processing conditions to help produce higher quality wine which may increase profitability for wine makers.

3.0 Data Preparation/Cleaning

Databricks was the choice of development platform as it is an industry standard tool for manipulating and analyzing large datasets. All analysis was conducted using python, with addition of pandas and statsmodels libraries.

The procured dataset was a standard machine learning dataset from UCI's machine learning repository. The dataset was available as a comma separated values file which was loaded into the development environment with ease. The data quality was quite good, it had a total of 1599 rows of data with no null values for any of the data points. In addition, all values were of numerical data types so there was no need to employ techniques to scrub the input dataset.

4.0 Understanding the data

The list below provides an overview of the input variables along with a brief description:

1. Fixed acidity - amount of non-volatile acidity
2. Volatile acidity - amount of acetic acid
3. Citric acid - amount of citric acid
4. Residual sugar - amount of sugar after fermentation
5. Chlorides - amount of salt

6. Free sulfur dioxide - amount of free sulfur dioxide that exists in equilibrium between molecular sulfur dioxide and bisulfite ion
7. Total sulfur dioxide - amount of free and bound forms of sulfur dioxide
8. Density - density relative to water
9. pH - level of acidity/alkalinity (0-very acidic, 14-basic)
10. Sulphates - additive that contributes to the level of sulphur dioxide
11. Alcohol - percent alcohol content

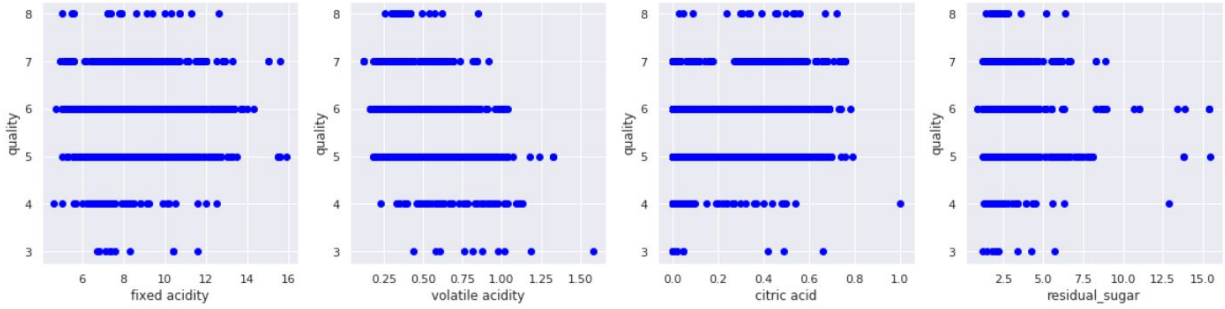
The output variable of interest is the quality of wine which has a score of 0-10 based on sensory data.

5.0 Analysis

Multiple linear regression was used to build the predictive model. As all available input variables were numerical, the raw data could be used as is without additional transformations. A stepwise approach with forward selection was used to build the model where the correlation coefficients of each input variable were used to determine the order in which variables were added at each iteration.

5.1 Scatter plot of independent variables

The graphs below show the quality of wine plotted as a response to each of the 11 independent variables. For each input variable, the quality is a discrete value for a specific range of the input variable, where the ranges overlap for different levels of wine quality. While the scatter plots below show the relationship between the input and output variable is non-linear, this doesn't negate the possibility that a combination of two or more of these variables may be used in a multiple linear regression to produce a reasonably accurate prediction model.



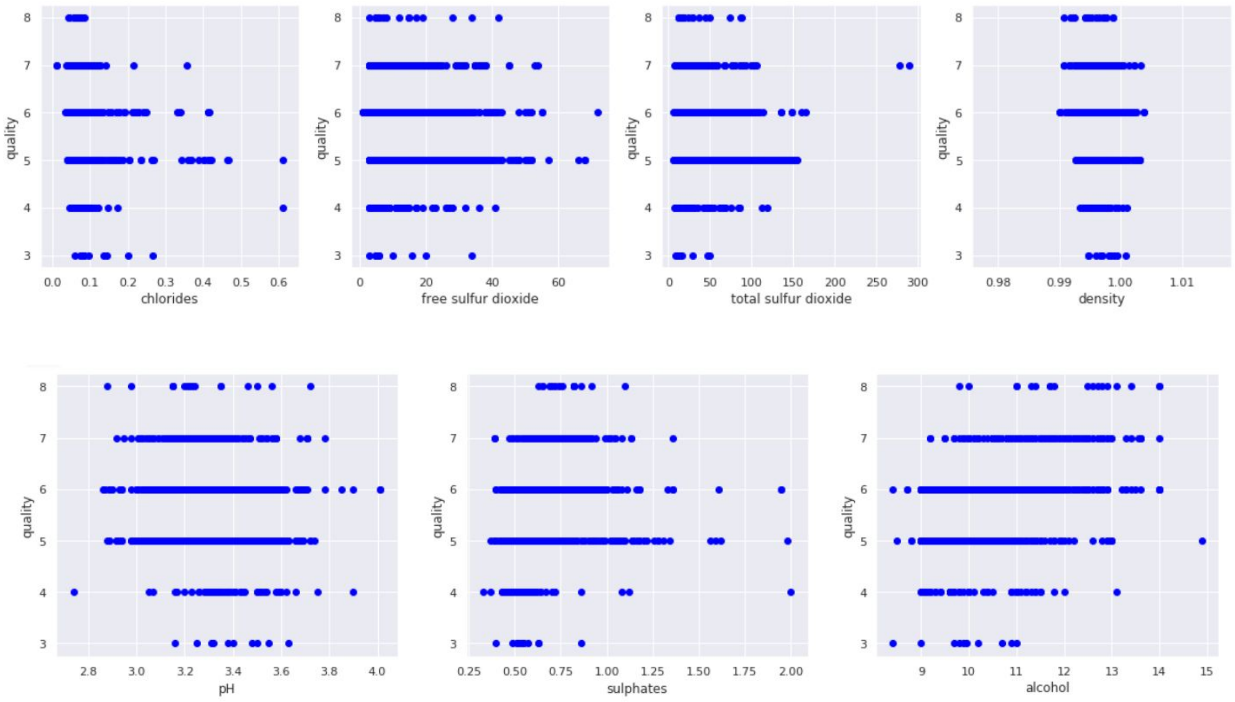


Figure 5.1 - Quality of wine as a function of each input variable

5.2 Finding the correlation coefficients

Prior to proceeding with setting up the model, the correlation coefficient of each independent variable was calculated to assess the extent of linear dependence of wine quality on the variable. This served to provide the order in which the variables would be added to the regression model and evaluate whether each variable better/worse explained the variability in output as a result of the input variable (i.e. R-squared). The table below provides a list of the input variables along with their correlation coefficients to the output variable.

Input Variable	Correlation Coefficient
Fixed Acidity	0.124052
Volatile Acidity	0.390558
Citric Acid	0.226373
Residual Sugar	0.013732
Chlorides	0.128907
Free Sulfur Dioxide	0.050656

Total Sulfur Dioxide	0.185100
Density	0.174919
pH	0.057731
Sulphates	0.251379
Alcohol	0.476166

Table 5.2 - Correlation Coefficients between input and output variables

5.3 Performing stepwise multiple linear regression using forward selection

For the multiple linear regression model, the ordinary least squares model from the statsmodels library was used. The first variable with which the model was run is alcohol. As can be seen in Table 5.2, as this variable had the highest correlation to quality, it would be highly unlikely that this factor would be excluded from the final model. After this, each of the above variables were added iteratively and the adjusted R-squared was evaluated to decide whether the variable should be part of the final model. If the adjusted R-squared yielded a lower value after adding the variable than without it, the variable was discarded. If it yielded the same or a better adjusted R-squared value, the variable was retained for the next iteration. Following this methodology, variables were added in order of descending correlation coefficients. The table below shows the R-squared value, the adjusted R-squared value and whether each variable was retained or discarded (denoted by ✔ and ✘, respectively). The final model was that which was obtained in iteration 11. The summary of model parameters can be found in Figure 7.1 in the Appendix.

Iteration	**Variable**	**R-squared**	**Adjusted R-squared**	**Keep Variable?**
1	Alcohol	0.227	0.226	✔
2	Volatile Acidity	0.317	0.316	✔
3	Sulphates	0.336	0.335	✔
4	Citric Acid	0.336	0.334	✘
5	Total Sulfur Dioxide	0.334	0.342	✔
6	Density	0.334	0.342	✔

7	Chlorides	0.352	0349	✔
8	Fixed Acidity	0.347	0.345	✘
9	Free Sulfur Dioxide	0.353	0.350	✔
10	pH	0.360	0.356	✔
11	Residual Sugar	0.360	0.356	✔

Table 5.3 - R-squared value of multiple linear regression model at each iteration

5.4 Testing the model

To test the accuracy of the model, the parameter values from the final model (iteration 11) were used to calculate the predicted values (see Appendix A - Table 7.1 for list of model parameter values).

As the quality of wine in the dataset is a discrete value between 0 and 10 and the multiple linear regression model expectedly didn't conform to this, the results were further transformed by rounding the predicted value to the nearest integer prior to assessing accuracy. By employing this method, the model was able to accurately predict the correct value of wine quality 59.6% of the time.

5.5 Evaluating the model

After transforming the predicted output to produce a discrete value between 0 and 10, the residuals were calculated to examine the distribution. The graph below shows the residual plot where the distribution appears to be normal and the mean centered at 0, meaning that the model is not very biased.

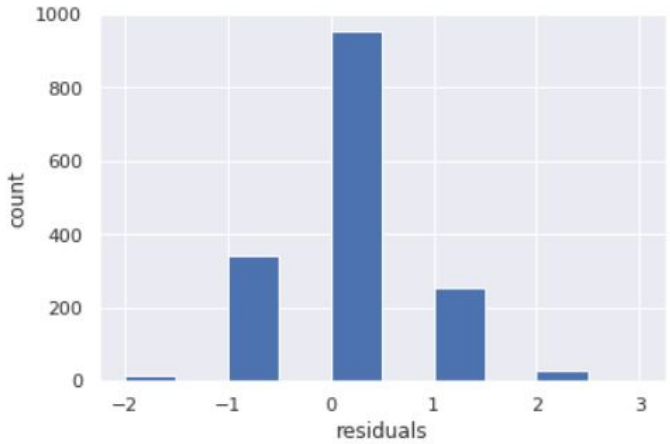


Figure 5.5 - Residual plot

A residual and influence plot (see Figure 7.2, 7.3 in Appendix) was used to determine whether dropping outliers would help improve the model, however after removing roughly 12 outliers, the model showed a marginal improvement in R-squared and only about 1% improvement in number of accurate predictions. Therefore, ultimately it was decided to not drop these data points as the improvement was marginal at the cost of excluding potentially meaningful data points.

6.0 Conclusion

With respect to the primary objective, despite the scatter plots of each independent variable vs. the quality of wine which showed a non-linear relationship, the final multiple linear regression model yielded an impressive accuracy rate of 59.6% after transforming the predicted result to a whole number to allow comparison with the expected value. The model had an adjusted R-squared value of 0.356, where the following attributes from the input dataset were incorporated - alcohol, volatile acidity, sulphates, total sulfur dioxide, density, chlorides, free sulfur dioxide, pH, and residual sugar.

In addition, to identify the factor that most influenced the quality of wine, the t-test values of each independent variable can be examined from the model summary (from Figure 7.1 in Appendix). Of the factors that were significant and incorporated into the model, the following list orders the attributes from most to least influential - alcohol, volatile acidity, sulphates, total sulfur dioxide, chlorides, pH, residual sugar and density. Both citric acid and fixed acidity were rejected from the model as they yielded a worse R-squared value when included in the model than not. It is also important to note that density and residual sugar while included in the model didn't necessarily improve the R-squared value.

Overall, while a multiple linear regression model provided reasonably accurate predictions, it's worth mentioning that there are factors other than physicochemical properties that affect the quality of red wine. This may include - grape variety, environmental factors in which the grapes were cultivated, as well as manufacturing processes among other factors.

In addition, as the output variable of interest, that is, the quality of wine, is a variable that is assessed based on sensory information they may be an inherent bias and level of subjectivity that can't be dissociated from the final result as respondents in the study may have preferences that affect their judgement of quality. For example, respondents with a preference for smoother wines may rank wines with higher residual sugar higher, or respondents with a preference for oaky flavours might rank wines with a higher citric acid content lower (as this characteristic adds freshness).

Other issues with the dataset mostly deal with multicollinearity which can be problematic when used in linear regression models. For example, in the final model, density is an input variable that influences the quality. However residual sugar and alcohol content are also factors that influence

both quality of wine and density (higher the alcohol content, lower the density). Other examples of confounding variables can be found by examining the scatter plot in the supporting notebook. For example, volatile acidity and pH (higher the acidity, lower the pH). Another example is including both free sulphur dioxide, sulphates and total sulfur dioxide where an increase in the former two variables would naturally result in an increase in the total value.

7.0 References

[1] Ray Isle Updated June 07 and R. Isle, “What Are Red Blends, Really?,” *Food & Wine*, 07-Jun-2017. [Online]. Available: https://www.foodandwine.com/wine/red-wine/best-red-blends-drink-now. [Accessed: 15-Dec-2020].

[2] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.
Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

[3] U. C. I. M. Learning, “Red Wine Quality,” *Kaggle*, 27-Nov-2017. [Online]. Available: https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009. [Accessed: 15-Dec-2020].

[4]https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/1813937161281675/2602243913710292/3514329913855704/latest.html

8.0 Appendix

OLS Regression Results

Dep. Variable:	quality	R-squared:	0.360
Model:	OLS	Adj. R-squared:	0.356
Method:	Least Squares	F-statistic:	99.22
Date:	Mon, 14 Dec 2020	Prob (F-statistic):	4.56e-147
Time:	01:20:44	Log-Likelihood:	-1570.1
No. Observations:	1599	AIC:	3160.
Df Residuals:	1589	BIC:	3214.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	12.0575	12.009	1.004	0.316	-11.498	35.613
alcohol	0.2810	0.020	13.912	0.000	0.241	0.321
volatile_acidity	-1.0128	0.101	-10.035	0.000	-1.211	-0.815
sulphates	0.9024	0.113	7.989	0.000	0.681	1.124
total_sulfur_dioxide	-0.0036	0.001	-5.143	0.000	-0.005	-0.002
chlorides	-2.0491	0.399	-5.133	0.000	-2.832	-1.266
free_sulfur_dioxide	0.0049	0.002	2.279	0.023	0.001	0.009
density	-7.5668	11.865	-0.638	0.524	-30.839	15.706
pH	-0.4920	0.121	-4.063	0.000	-0.730	-0.254
residual_sugar	0.0115	0.013	0.855	0.393	-0.015	0.038

Figure 7.1 - Model Summary (from Iteration 11)

Parameter Type	Model Parameter	Value
Intercept	Intercept	12.05750
Slope	Alcohol	0.281002
Slope	Volatile Acidity	-1.012792
Slope	Sulphates	0.902443
Slope	Total Sulfur Dioxides	-0.003569
Slope	Chlorides	-2.049094
Slope	Free Sulfur Dioxide	0.004874
Slope	Density	-7.566836
Slope	pH	-0.491991
Slope	Residual Sugar	0.011505

Table 7.1 - Model Parameter Values (from Iteration 11)

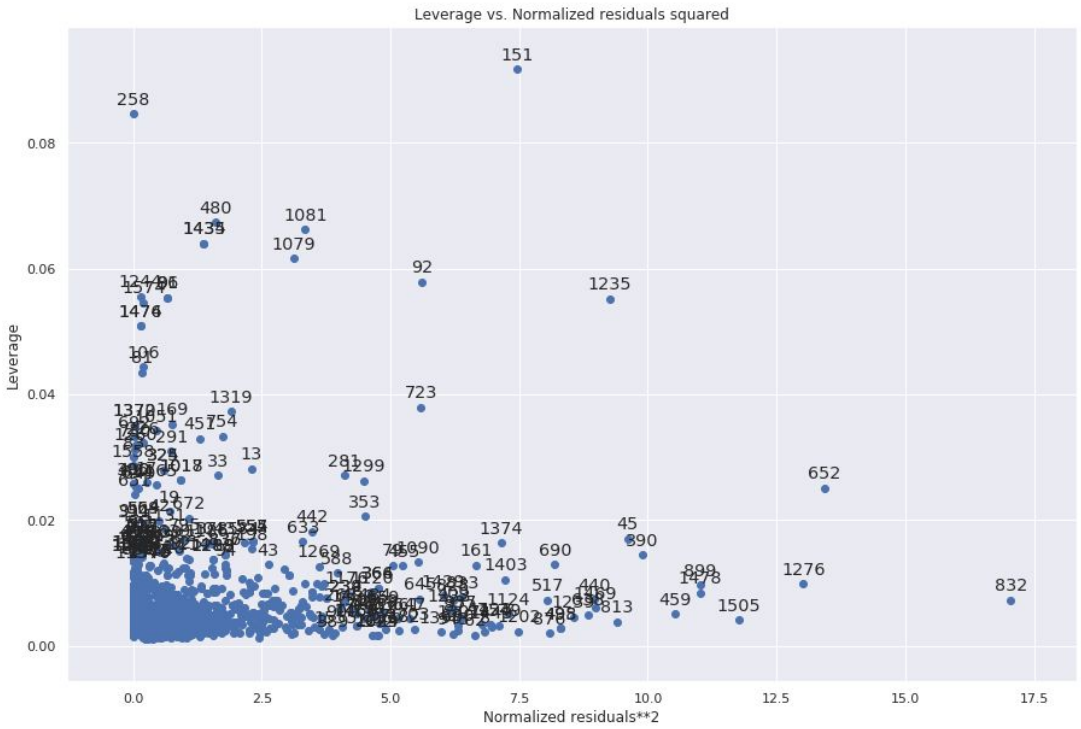


Figure 7.2 - Leverage Normalized residual plot

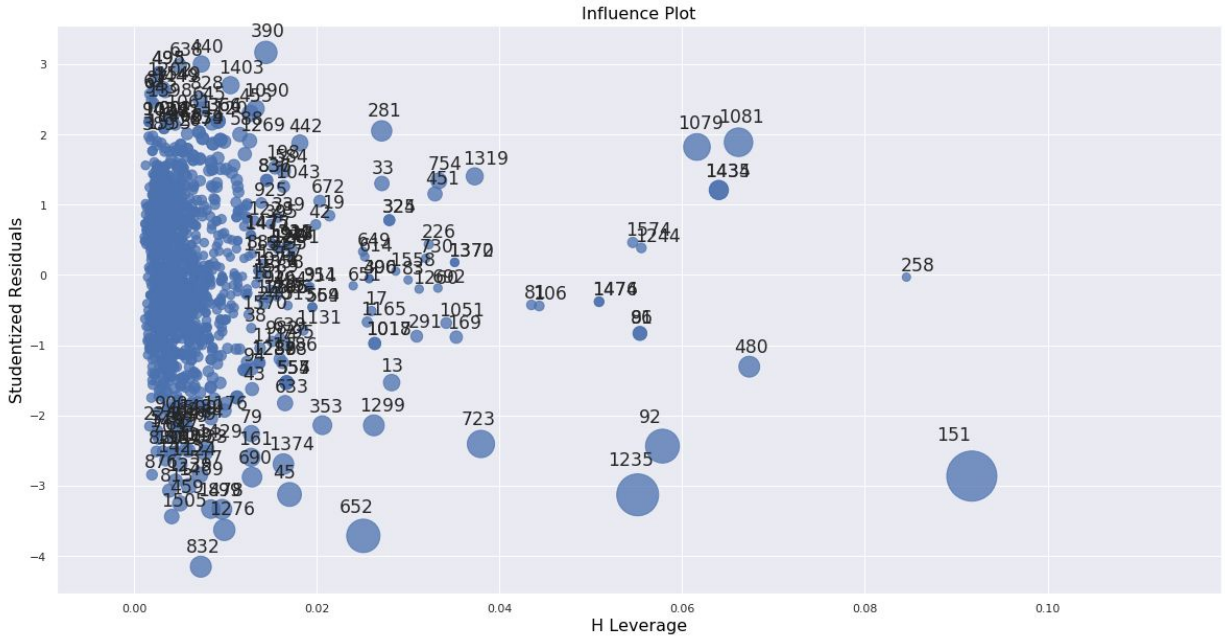


Figure 7.3 - Influence Plot