**Predictive Analytics**

Table of Contents

[1. Purpose 3](#_Toc52621586)

[2. Methodology 3](#_Toc52621587)

[2.1a Characteristics of variables 3](#_Toc52621588)

[2.1b Correlation of variables against quality 3](#_Toc52621589)

[2.2 Multilinear regression with all variables 3](#_Toc52621590)

[2.3 Stepwise Multilinear Regression with all variables 4](#_Toc52621591)

[2.4 Nonlinear Regression 4](#_Toc52621592)

[2.5 Nonlinear Stepwise Regression 4](#_Toc52621593)

[2.6 Scoring of all models & predictive accuracy 4](#_Toc52621594)

[3. Analysis 4](#_Toc52621595)

[4. Limitations 5](#_Toc52621596)

[5. Appendix 6](#_Toc52621597)

# 1. Purpose

The purpose of this report is to predict the wine quality score of red wine based on specific physicochemical properties. The ability to predict wine quality score has several applications. Firstly, it allows winemakers to judge how their wine may turn out in term of quality score, early in the process. That forecasted score then enables the intervention in quality improvements and proper marketing mix (advertising budget, packaging & pricing) of the product based. Furthermore, the prediction of the wine quality score also helps to better plan the winemaking process in the upcoming season to optimize the balance between costs and quality score.

# 2. Methodology

1. Exploratory Data Analysis of wine dataset
   1. Characteristics of each variable
   2. Correlation between variables
2. Multilinear regression with all variables
3. Multilinear regression with all variables (stepwise)
4. Nonlinear regression
5. Nonlinear regression (stepwise)
6. Scoring of all models & predictive accuracy

## 2.1a Characteristics of variables

Each wine has multiple physiochemical variables which affect overall wine quality, and its score. Below are the average value and standard deviation of these variable for different attributes of red wine.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **FA** | **VA** | **CA** | **RS** | **Ch** | **FSD** | **TSD** | **Density** | **pH** | **Sulphates** | **Alc** |
| **Avg** | 8.32 | 0.53 | 0.27 | 2.54 | 0.09 | 15.88 | 46.47 | 1.00 | 3.31 | 0.66 | 10.42 |
| **St dev σ** | 1.74 | 0.18 | 0.19 | 1.41 | 0.05 | 10.46 | 32.90 | 0.00 | 0.15 | 0.17 | 1.07 |

## 2.1b Correlation of variables against quality

The below provides a view of which variables are negatively or positively correlated with quality score. It appears ‘Volatile Acidity’ is most negatively correlated whilst ‘Alcohol’ is most positively correlated with the score.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **FA** | **VA** | **CA** | **RS** | **Ch** | **FSD** | **TSD** | **Density** | **pH** | **Sulphates** | **Alc** |
| 0.12 | -0.39 | 0.23 | 0.01 | -0.13 | -0.05 | -0.19 | -0.18 | -0.06 | 0.25 | 0.48 |

## 2.2 Multilinear regression with all variables

The multilinear regression with all variables produced an adjusted r-squared of 0.3495, as detailed in Appendix 1. There are several variables which the model deems to be statistically insignificant in predicting the quality score, such as ‘Fixed Acidity’, ‘Citric Acid’, ‘Residual Sugar’ and ‘Wine Density’.

Exploring Appendix 2, specifically the fitted values vs actual values, it’s evident that the model tends to overestimate at the low range of the quality score and underestimate at the high range of the quality score. Apart from that, other residual plots do not seem to indicate any nonlinear relationships between quality score and the explanatory variables.

## 2.3 Stepwise Multilinear Regression with all variables

Outlined in Appendix 3, the stepwise regression slightly improves on the prior model with an adjusted r-squared of 0.3511, highlighting the advantage of utilizing less irrelevant parameters. Despite the improved adjusted r-squared, the stepwise multilinear regression still suffers from the same overestimate and underestimate issues of the quality score.

Similarly, the new set residual plots, as depicted in appendix 4, also do not seem to indicate any nonlinear relationships. However, further analysis is still recommended.

## 2.4 Nonlinear Regression

Outlined in Appendix 5, the nonlinear regression with quadratic and logarithm terms on the seven significant variables (VA, Ch, FSD, TSD, pH, Sulphates, Alc) highlighted by a multilinear stepwise regression improves the adjusted r-squared to 0.3814. But the improvement of adjusted r-squared is not significant, and the issue of overestimating and underestimating remain.

## 2.5 Nonlinear Stepwise Regression

Outlined in Appendix 7, the stepwise regression on the nonlinear model improves the adjusted r-squared to 0.3819. The primary concern of systematic over and under estimation still exits, and improvement of r-squared is meagre.

## 2.6 Scoring of all models & predictive accuracy

All four models have been scored and they perform closely in term of predicting the quality score of wine using existing physiochemical properties. The loss functions metrics in the table below demonstrate their predictive strength against each other.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **RMSE** | **MAE** | **MAPE** | **MASE** | **# of Parameters** |
| **Linear** | 0.58593 | 0.46498 | 8.66332 | 0.72938 | 12 |
| **Stepwise** | 0.58721 | 0.46462 | 8.65633 | 0.72881 | 8 |
| **Nonlinear** | 0.58349 | 0.46077 | 8.61997 | 0.72278 | 18 |
| **Nonlinear plus stepwise** | 0.58434 | 0.46105 | 8.61332 | 0.72321 | 13 |

# 3. Analysis

The four models used in predicting wine qualities all scored extremely closely, despite the nonlinear models being more complicated through the use of quadratic and logarithmic terms. In this instance, it is safe to use any of the four reported metrics as there are no extreme values which can affect the RMSE and no values of 0, which would ordinarily affect the MAPE.

The two models which had the best predictive power were the nonlinear ones (without and with stepwise), and in the search for the model which best predicts quality of the wine, either of the two nonlinear models should be selected.

In the nonlinear (without stepwise) model, a prediction deviates on average 0.583 points away from true quality according to RMSE. It varies 0.461 from true quality using the MAE measure. A forecast on average deviates by 8.62% to actual quality scores according to MAPE and the MASE of 0.72 highlights it is worth running a prediction model as it is approximately 28% better than running a naïve prediction.

However, if there were a need to derive causal effects, it would be recommended to use the linear stepwise model as it can explain the majority of what influences the quality score and use fewer parameters.

To explain how wine quality is affected by the physicochemical properties of wine, the linear stepwise model will be referenced due to the simplicity of explanation. ‘Free Sulphur Dioxide’, ‘Sulphates’ and ‘Alcohol’ all appear to have a positive effect on the score of red wine. ‘Volatile Acidity’, ‘Chlorides’, ‘Total Sulfur Dioxide’ and ‘pH’ all have a negative effect on the quality of the wine.

These results appear to follow the correlation matrix to a certain extent, except for ‘Free Sulphur Dioxide’ which has a negative association with quality score in the correlation matrix.

The other physiochemical properties have been excluded from the model, likely due to either being correlated with other variables or being statistically insignificant in predicting the quality of the wine.

# 4. Limitations

Based on adjusted R-squared, all models could barely explain less than 40% of the variation of wine quality score. Furthermore, they tend to overpredict at the low range of observed scores and underestimate at the high range of observed scores. These issues could be because of several reasons.

Firstly, all the data available that goes into scoring wine may not be explained purely by physicochemical properties. For example, there are elements of human judgement that goes into choosing a quality score for wine which is not captured in the dataset. There may also be other data points which have not been recorded or unable to be measured but are crucial in determining wine quality.

Secondly, it is possible the regression model is not well suited to predict wine quality from physicochemical properties. The regression models tend to be parametric, which means they follow a smooth pattern. Other non-parametric models, which tends to be more flexible, may do a much better job at predicting quality score from the available variables.

Another limitation is that the analysis and scoring do not place any weighting on over or under predictions. Consequently, winemakers may use different metrics to measure the predictive accuracy of modelling so that a user-defined loss function may be more suited in that instance.

Despite all that, each industry may have its benchmark in term of how precisely a statistic model can fit into the available and make the prediction. It may very well be that the standard for the wine industry is usually having adjusted R-squared below 40%.

# 5. Appendix

**Appendix 1** – Multilinear Regression Summary

|  |
| --- |
| Call:  lm(formula = QS ~ ., data = trainData)  Residuals:  Min 1Q Median 3Q Max  -2.65050 -0.38391 -0.04515 0.45702 2.01409  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 1.490e+01 2.405e+01 0.620 0.53569  FA 1.242e-02 2.998e-02 0.414 0.67877  VA -1.021e+00 1.398e-01 -7.304 4.91e-13 \*\*\*  CA -1.296e-01 1.687e-01 -0.768 0.44262  RS 6.203e-03 1.749e-02 0.355 0.72284  Ch -2.059e+00 4.711e-01 -4.371 1.34e-05 \*\*\*  FSD 4.145e-03 2.502e-03 1.656 0.09791 .  TSD -3.561e-03 8.351e-04 -4.264 2.16e-05 \*\*\*  Density -1.024e+01 2.455e+01 -0.417 0.67667  pH -5.783e-01 2.193e-01 -2.637 0.00846 \*\*  Sulphates 8.651e-01 1.329e-01 6.508 1.09e-10 \*\*\*  Alc 2.908e-01 3.040e-02 9.564 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.6634 on 1267 degrees of freedom  Multiple R-squared: 0.3551, Adjusted R-squared: 0.3495  F-statistic: 63.41 on 11 and 1267 DF, p-value: < 2.2e-16 |

**Appendix 2** – Multilinear Residual Plot

Chart, scatter chart

Description automatically generated

**Appendix 3** – Stepwise Multilinear Regression Summary

|  |
| --- |
| Call:  lm(formula = QS ~ VA + Ch + FSD + TSD + pH + Sulphates + Alc,  data = trainData)  Residuals:  Min 1Q Median 3Q Max  -2.65513 -0.37980 -0.04544 0.46399 2.02174  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 4.7312900 0.4552581 10.393 < 2e-16 \*\*\*  VA -0.9654987 0.1151049 -8.388 < 2e-16 \*\*\*  Ch -2.1511320 0.4473199 -4.809 1.70e-06 \*\*\*  FSD 0.0045550 0.0024584 1.853 0.0641 .  TSD -0.0037041 0.0007855 -4.715 2.68e-06 \*\*\*  pH -0.5871845 0.1336715 -4.393 1.21e-05 \*\*\*  Sulphates 0.8437438 0.1278251 6.601 5.99e-11 \*\*\*  Alc 0.2972283 0.0189738 15.665 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.6626 on 1271 degrees of freedom  Multiple R-squared: 0.3546, Adjusted R-squared: 0.3511  F-statistic: 99.77 on 7 and 1271 DF, p-value: < 2.2e-16 |

**Appendix 4** - Stepwise Multilinear Regression Residual Plot

A picture containing cat, looking, sitting

Description automatically generated

**Appendix 5** – Nonlinear Regression Summary

*Nonlinear regression with quadratic and logarithm terms added for seven statistically significant predictors identified by the previous stepwise: VA, Ch, FSD, TSD, pH, Sulphates, Alc.*

|  |
| --- |
| Call:  lm(formula = QS ~ VA + I(VA^2) + I(log(VA)) + Ch + I(Ch^2) +  I(log(Ch)) + TSD + I(TSD^2) + I(log(TSD)) + FSD + I(FSD^2) +  I(log(FSD)) + pH + Sulphates + I(Sulphates^2) + I(log(Sulphates)) +  Alc, data = trainData)  Residuals:  Min 1Q Median 3Q Max  -2.69813 -0.38174 -0.03291 0.43054 2.02515  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 7.372e+00 2.572e+00 2.866 0.004224 \*\*  VA 2.278e+00 1.967e+00 1.158 0.247177  I(VA^2) -1.418e+00 8.460e-01 -1.676 0.094036 .  I(log(VA)) -7.178e-01 5.291e-01 -1.357 0.175103  Ch 7.050e+00 3.620e+00 1.948 0.051662 .  I(Ch^2) -9.158e+00 4.714e+00 -1.943 0.052239 .  I(log(Ch)) -7.194e-01 2.624e-01 -2.741 0.006210 \*\*  TSD -1.260e-02 4.270e-03 -2.952 0.003218 \*\*  I(TSD^2) 3.716e-05 1.420e-05 2.616 0.008995 \*\*  I(log(TSD)) 1.776e-01 1.315e-01 1.351 0.177011  FSD -4.246e-03 1.737e-02 -0.245 0.806872  I(FSD^2) 3.004e-05 2.133e-04 0.141 0.888014  I(log(FSD)) 1.334e-01 1.455e-01 0.917 0.359365  pH -7.846e-01 1.348e-01 -5.820 7.45e-09 \*\*\*  Sulphates -6.912e+00 3.181e+00 -2.173 0.029958 \*  I(Sulphates^2) 1.071e+00 8.354e-01 1.282 0.199905  I(log(Sulphates)) 4.542e+00 1.373e+00 3.308 0.000966 \*\*\*  Alc 2.778e-01 1.958e-02 14.184 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.6469 on 1261 degrees of freedom  Multiple R-squared: 0.3896, Adjusted R-squared: 0.3814  F-statistic: 47.35 on 17 and 1261 DF, p-value: < 2.2e-16 |

**Appendix 6** – Nonlinear Regression Residual Plot

A picture containing looking, cat, sitting

Description automatically generated

**Appendix 7** – Stepwise Nonlinear Regression Summary

|  |
| --- |
| Call:  lm(formula = QS ~ I(VA^2) + Ch + I(Ch^2) + I(log(Ch)) + TSD + I(TSD^2) + I(log(TSD)) + I(log(FSD)) + pH + Sulphates + I(log(Sulphates)) + Alc, data = trainData)  Residuals:  Min 1Q Median 3Q Max  -2.74200 -0.38463 -0.03049 0.43293 1.94216  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 6.146e+00 1.163e+00 5.283 1.50e-07 \*\*\*  I(VA^2) -6.305e-01 9.359e-02 -6.736 2.46e-11 \*\*\*  Ch 5.977e+00 3.449e+00 1.733 0.08332 .  I(Ch^2) -7.837e+00 4.530e+00 -1.730 0.08389 .  I(log(Ch)) -6.528e-01 2.498e-01 -2.614 0.00906 \*\*  TSD -1.316e-02 3.992e-03 -3.298 0.00100 \*\*  I(TSD^2) 3.884e-05 1.382e-05 2.811 0.00502 \*\*  I(log(TSD)) 1.925e-01 1.204e-01 1.600 0.10994  I(log(FSD)) 8.972e-02 4.451e-02 2.016 0.04403 \*  pH -7.719e-01 1.333e-01 -5.791 8.81e-09 \*\*\*  Sulphates -2.895e+00 5.623e-01 -5.149 3.04e-07 \*\*\*  I(log(Sulphates)) 2.869e+00 4.184e-01 6.857 1.10e-11 \*\*\*  Alc 2.796e-01 1.939e-02 14.420 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.6467 on 1266 degrees of freedom  Multiple R-squared: 0.3877, Adjusted R-squared: 0.3819  F-statistic: 66.79 on 12 and 1266 DF, p-value: < 2.2e-16 |