Wine Quality Prediction using Logistic Regression

Dhaval Thakur, Rushi Bhuva and Tejas Pandit

Introduction

We have been given red wine and white wine dataset from UCI's repository. The goal is to model wine quality based on physicochemical tests. These two datasets were created, using red and white wine samples. Each expert graded the wine quality between 0 (very bad) and 10 (very excellent). In this report, we would trace our steps, from getting familiar with our wine data, cleaning and transforming the data, choosing variables for regression model creation, looking for potential outliers, influential points, checking for assumptions, plotting relevant graphs and giving a conclusion through various statistical methods and measures of accuracy in quantitative form.

Summary of Original Dataset

The two datasets winequality_red (dataset A) and winequality_white (dataset B) has 1599 and 4898 records with 12 variables in each respectively. The 12 variables (along with their types) are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol and quality (score between 0 and 10). Quality is an **oridinal variable** and all other variables are **continuous**.

Data cleaning and Transformation

For this assignment first we check for missing values and we found none for both the dataset (A and B). After that our concern was to check for duplicate obseravtion present in any dataset and we found that in dataset A there were **1359** unique observations out of **1599** and in dataset B there were **3961** unique observations out of **4898**.

Summary of new Dataset

With the absence of missing values in either of the datasets, we build a new dataset called the final_dataset by using outer join and get a new dataset of **5320** rows with same variables as mentioned in the summary section, so that we can start our process of making a regression model for the prediction of quality of wine.

Although tastes vary from person to person and are probably unique, some wines are better than others, and most people would probably recognize a good wine from a bad one thus it gives motive to make quality a **binary variable**. Furthermore, as we have to create a regression model, on the final transformed dataset we created a new variable called **rating** with two levels - **below average**(rating<=5) wine and **above average** (rating>5) wine [2]. Total observation for "below average" wine is **1988** and total observation for "above average wine" is **3332** Thus, we build a Logistic model, since we do not use Linear Regression for binary classification.

Planning phase

Selection of Predictor variables

In the absence of subject-matter expertise of wine quality, stepwise regression (backward step method) can assist with the search for the most important predictors of the outcome of interest. We would be using Backward step method rather then forward step method because of **suppressor effects**, which occur when a predictor has an effect but only when another variable is held constant.

Variables Selection and Initial Model Creation

Initially, we took all the variables mentioned in the summary of the dataset following the methodology of step-down variable selection method [1], adding to that, our outcome variable "rating" has binary outcome: **below average**(rating<=5) or **above average**(rating>5). Thus, we would be using binomial family for the model.

Initial analysis of model:

Analysing the model from its summary, we see that the predictor variables alcohol, volatile acidity, sulphates, pH, Sulfur family (total, free) and residual sugar predictors with p-value **(Pr|Z| <0.05)** significance level. On the other hand, predictor variables: citric acid, fixed acidity, chlorides and density have p-values less then 0.05 making them contribute less towards this corresponding model.

We would use the Akaike information criterion (AIC) and the Bayes information criterion (BIC) to judge model fit.[5] We would be using this criteria instead of $\mathbf{R^2}$ because of the fact that every time we add a variable to the model, $\mathbf{R^2}$ increases. We want a measure of fit that we can use to compare two models which penalizes a model that contains more predictor variables.

Interpreting AIC (Akaike Information Criterion):

The AIC is an estimator of out-of-sample prediction error and thereby relative quality of statistical models for a given set of data. AIC estimates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model. [1]
Further, When the summary of Initial Logistic model is generated, we see that the **AIC** comes to be **5492.4**. Our next course of action is to improve this Regression model and try to obtain a low AIC or significant p-value for the predictors. Seeing the summary and the significance figures present in the output we start to optimize our predicting model by eleminating the predictors one by one in order of their non-significance p-value to see if the AIC decreases or not. Thus, we remove the variable having the least significant p-value in each iteration and check for new p-values and residuals. Fastforwarding, After 5 iterations **(AIC - 5492.4 -> 5490.5 -> 5490.3 -> 5488.9)** , we get a model with **AIC 5488.9**.

Deciding the final Regression model:

At the 5th iteration of the model(where we removed predictor variable: Chloride), upon measuring the AIC we get the value of 5490.7, which is an *increase* from the previous iterated model (AIC: 5488.9), thus we stick to our previous model and discard the new 5th iterative model. Thus, from this whole process pipeline, we get the minimum AIC of **5488.9** which is a significant improvement from wineModel 1 which had AIC of **5492.4**.

Interpretation of Logistic model

From the summary of the model we get the estimates of coefficients of predictor variables, and thus we can obtain logistic Equation for our model as [1]:

$$\mathbf{Y = -11.24 - 4.175x_1 + 0.04801x_2 - 2.057x_3 + 0.01885x_4 - 0.007201x_5 + 0.7994x_6 + 2.072x_7 + 0.09325x_8}$$

Where, Y(outcome variable): Rating, x_1: volatile acidity, x_2: residual sugar, x_3: chlorides, x_4: free sulfur dioxides, x_5: total sulfur dioxide, x_6:pH, x_7: sulphates, x_8: alcohol.

We see from the above logistic Regression equation that the coefficient of x_5 (total sulfur dioxide) (~0.007) seems to be contributing less to the overall odds of outcome variable(rating) as compared to other coefficients when the coefficient of x_5 increases by one unit. Also, we see that the Null Deviance comes to be 7031.8 (means the fit of the model when no variables are taken into consideration in the model) , and Residual deviance falls to 5470.9 accounting the effect of the predictor variables and we can initially say that this model is better then a model with no predictor variables.
Thus, further, to support this model, its AIC measure value is 5488.9 which is lower than that 5492.4, thus giving us another support to go further with this model.

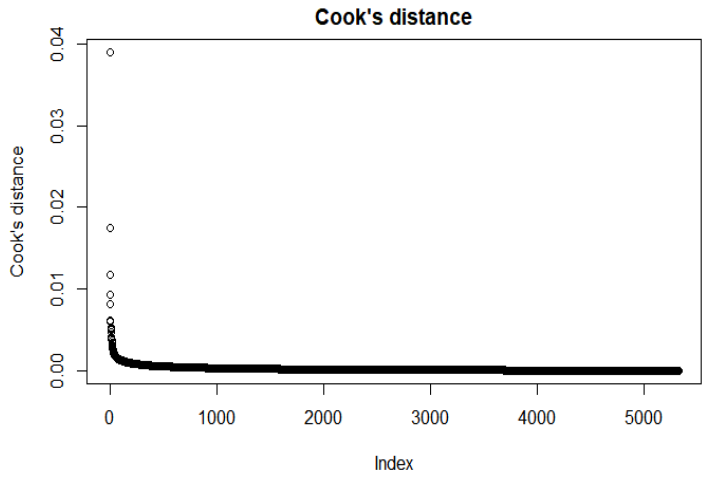

Figure 1

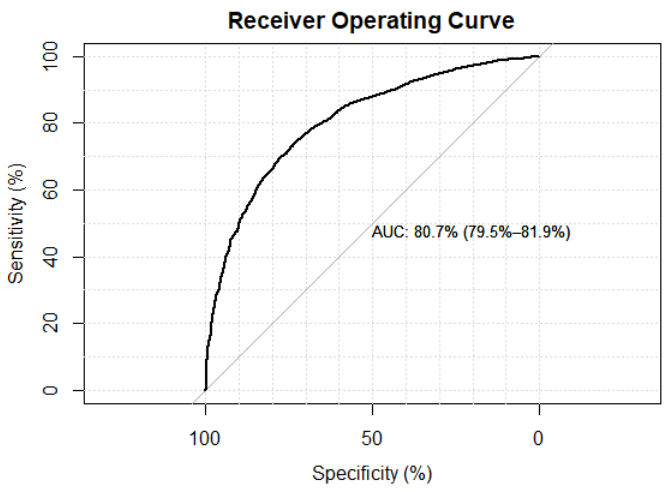

Figure 2

Establishing Hypothesis

Null hypothesis: The coefficients on the parameters (including interaction terms) of the logistic regression modeling of Quality of Wine as a function of fixed acidity, volatile acidity, free sulfur dioxide, total sulfur dioxide, density, sulphates and alcohol are zero.

Alternative hypothesis: The coefficients on the parameters (including interaction terms) of the logistic regression modeling of Quality of Wine as a function of fixed acidity, volatile acidity, free sulfur dioxide, total sulfur dioxide, density, sulphates and alcohol are non-zero and significant with all assumptions considered.

Assessing the regression model: Assumption and Generalization

In this section, we test the assumptions for the validity of our model creation. Generalization is a critical additional step, and if we find that our model is not generalizable, then we must restrict any conclusions based on the model to the sample used. First,we will look at how we establish whether a model is an accurate representation of the actual data.

- **Variable types:** To begin, one of the main assumptions of logistic regression is the appropriate structure of the outcome variable. Binary logistic regression requires the dependent variable to be binary and ordinal logistic regression requires the dependent variable to be ordinal. And we have created a new binary variable called rating. Thus, meeting the requirements for this assumption.
- **Multicollinearity:** There should be no perfect linear relationship between two or more of the predictors. So, the predictor variables should not correlate too highly. To check this assumption We would check VIF and tolerance statistics. Also, tolerance can be calculated through reciprocal of the VIF. (code present in R file)
  - **Interpreting the VIF and Tolerance:** For our current model we found out that the VIF values are all well below 10 and the tolerance statistics all well above 0.2. Also, the average VIF is very close to 1. Based on these measures we can safely conclude that there is no collinearity within our data.
- **Independent errors:** For any two observations the residual terms should be uncorrelated (or independent). This assumption can be tested with the Durbin–Watson test. (code present in R file)
  - **Outcome:** As a conservative rule, its suggested that value should be less than 1 or greater than 3 should definitely raise alarm bells. The closer to 2 that the value is, the better, and for the data above the value is 1.751, which is so close to 2 that the assumption has almost certainly been met.
- **Linearity:** The mean values of the outcome variable for each increment of the predictor(s) lie along a straight line.
  - **Outcome:** We did the Linearity logit variable and found that from our model only three variables are present where the interactions do not have significance values Pr(>|Z|) greater then 0.05.
  - **Further Insight:** Since the three variables which are failing the Linearity logit assumption have significant impact on our created model, we can't remove them and we can keep them in the model and not remove them. **Though** this failing of assumption would definitely have an impact on the accuracy of our model and generalizability would be limited.

Outliers

Outliers is one of the main concerns that can make the model inaccurate. We see the residuals from the dataset. We extract our fitted, residuals from our model and then we found out possible outliers, and that about 95% of standardized residuals should be between -1.96 and 1.96. (Methodology followed for outliers, [4])

We found out that outlier consists of approx. 2.46 percent of total data. Furthermore, 131 residuals are above or below 1.96 standard deviations. As this represents **2.462406%** of the observations, expected if the residuals are normal (5% of data is expected to be outside of 2 standard deviations). Thus, we need not to remove them and we shift our concern towards Influential cases.

Influential Cases

In our combined dataset there might be some data points that may exert undue influence on the model, rather than simply being very far away from the mean. We would be using Cook's distance which would be giving us a measure of the overall influence of a case on our model, we will not be using Studentized residuals as they do not provide any information about how a case influences the model as a whole. [1]

Thus, we calculated Cook's Distance on the developed model. Cook's distance was a maximum of `0.03906621`, far below the chosen cut-off value of 1 [Fig. 1]. We thus conclude that there are no influential cases.

Odd Ratios and Confidence Intervals for Odd Ratio

We can interpret the odds ratio in terms of the change in odds. From the odd Ratios calculated, We interpret in terms of change in odds. We observe that **sulphates** has maximum value **(7.93)** for odds ratio compared to all other predictor

variables thus it indicates that as this predictor increases, the odds of our outcome variable "rating" also increases, matching our regression equation.

Interpreting Confidence Intervals

From our calculated confidence intervals, we observe that it does not cross 1(the values at each end of the interval are greater than 1).This is important because values greater than 1 means that as the predictor variable increases,so do the odds of having good quality of wine.

Area Under Curve (AUC)

Predictive models like logistic regression gives a predicted probability of a positive for each individual based on the values of that individual's predictor values and not one decision rule. Area under Curve (AUC) tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, Higher the AUC, better the model is at distinguishing between two binary outcomes. [3] Also, the further the curve is from the diagonal line, the better the model is at discriminating between positives and negatives in general. We plotted the AUC curve using and found AUC of 80.9% for our model. [Fig. 2].

Conclusion

- A Binary Logistic Regression model has been created. The target/outcome variable is rating, whereas the attributes of physicochemical tests (volatile acidity, residual sugar, chlorides, free sulfur dioxides, total sulfur dioxide, pH, sulphates and alcohol) are the predictor variables.
- We built a Logistic Regression Model finally with an AIC of **5488.9** which is a significant improvement from our initial first iterative Logistic Model with an AIC of 5492.4.
- To furthermore support our analysis, Observation is made that from the model we get the p-values of each predictor variables is significant at 95 % level of confidence. For e.g. If we see the top two significant predictors (volatile acidity and alcohol) in our model, they have a p-value of **2x10^-16** for each which are significant at 0.05 level of significance.
- Moreover, according to AIC, all models are approximations to reality, and reality should never have a low dimensionality. Thus, we considered calculating **BIC** value also for the initial model, it came to be **5548.066** and for our final regression model we obtained a better BIC = **5543.296**, which clearly states that our final model is better than the initial one. [5]
- We calculated confidence intervals for each of the predicting variables present in our final model such as Volatile Acidity ($\mathbf{9.381102 * 10^{-03} - 9.500438 * 10^{-01}}$) ,alcohol (**2.358 - 2.741**) etc, all are greater then 1 thus giving us the insight that as our predictor variables increases so does the odds of a good quality wine. These intervals give us estimate that in our Logistic Regression model, we have 95% of confidence that the intercept of the dependent variables would lie between these stated bounded intervals.
- It can be depicted from the AUC curve (Figure 1) we got AUC as **80.7%** for our model when we ran Goodness of fit test for our binomial regression model (more in references) to furthermore support for the accuracy of the model.

Thus, we reject our Null Hypothesis and accept the alternative hypothesis stated in above section.

References

[1] A. M. FIELD, *DISCOVERING STATISTICS USING R.* SAGE PUBLICATIONS, 2013.

[2] "Wine-Tasting by Numbers: Using Binary Logistic Regression to Reveal the Preferences of Experts | Minitab." [Online]. Available: https://www.minitab.com/en-us/Published-Articles/Wine-Tasting-by-Numbers--Using-Binary-Logistic-Regression-to-Reveal-the-Preferences-of-Experts/. [Accessed: 30-Mar-2020].

[3] "What Is an ROC Curve? - The Analysis Factor." [Online]. Available: https://www.theanalysisfactor.com/what-is-an-roc-curve/. [Accessed: 30-Mar-2020].

[4] K. Sarkar, H. Midi, and S. Rana, "Detection of outliers and influential observations in binary Logistic regression: An empirical study," *J. Appl. Sci.*, vol. 11, no. 1, pp. 26–35, 2011, doi: 10.3923/jas.2011.26.35.

[5] "modeling - Is there any reason to prefer the AIC or BIC over the other? - Cross Validated." [Online]. Available: https://stats.stackexchange.com/questions/577/is-there-any-reason-to-prefer-the-aic-or-bic-over-the-other. [Accessed: 29-Mar-2020].