Aqeel Padaria, Qiru Pan, Ibrahim Oner, Mazin Rafi

Professor Yaqing Chen

Regression and Time Series Analysis

Due Date: December 12th, 2022

**A Regression Analysis of White Wines; using R to predict Higher Quality Beverages and Identify key Impact Factors**

ABSTRACT

From box wines at grocery stores to highly exclusive Rieslings, wines can take on a range of different qualities. Often, these qualities are a result of subjective sensory evaluation by experts with decades of experience, which somewhat limits the abilities of vintners to easily classify new vintages. On the other hand, physicochemical data can be easily gathered and analyzed - the only challenge then is translating this into the more nuanced world of wine quality.

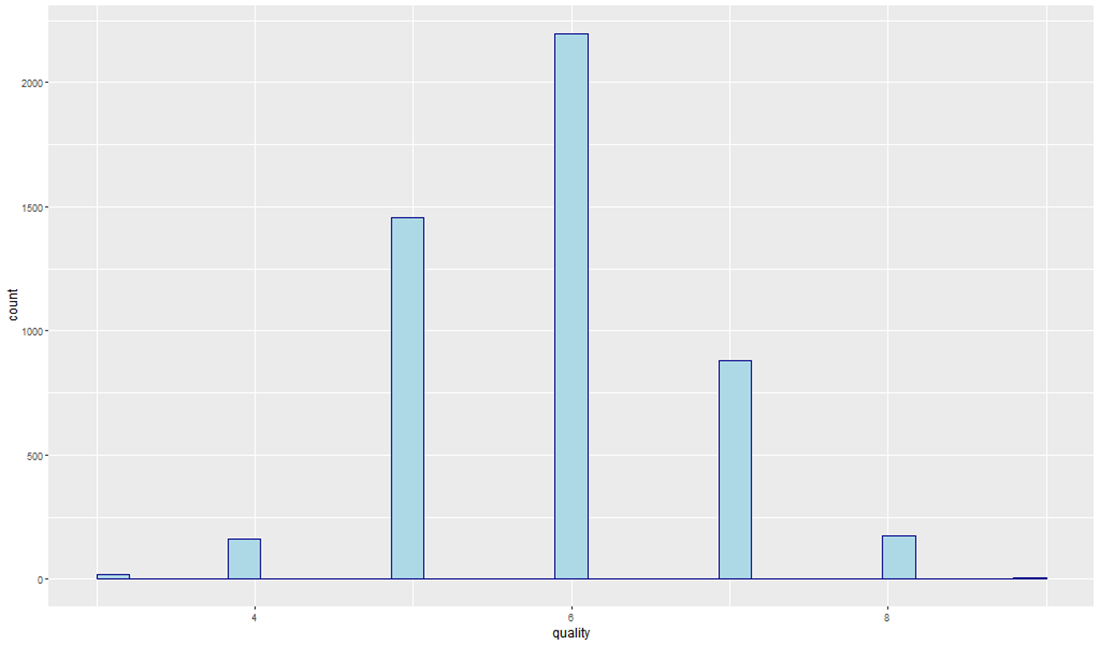
This paper aims to present an approach to predict the quality of white wines by building a simple yet effective model. The intent is to provide a streamlined, automatable way for vintners to assess the quality of their wines, which would enable them to much more efficiently plan distribution, evaluation, and pricing criteria.

The dataset used here is related to the Portuguese Vinho Verde wine - specifically, white variants. For more information, please consult the dataset description and/or the citation.

We begin with initial explorations of the dataset. After implementing techniques to select appropriate variables, we then explore regression and categorization of quality using key predictors, and evaluate our models on the basis of our objective (our primary criterion is accuracy on withheld data).

DESCRIPTION OF DATASET:

The dataset contains 4898 observations and 12 variables. 11 variables are based on physicochemical tests, while the response is based on sensory data. See Table 1 in the appendix for the data dictionary. The chart below shows the distribution of quality (the response).



VARIABLE SELECTION/MODEL FITTING:

Performing a stepwise linear regression on this dataset creates a model with the following predictors: **alcohol, free acidity, volatile.acidity, residual.sugar, free.sulfur.dioxide, density, pH and sulfates.**

Additionally, we ran the Boruta algorithm, which is a Random-Forest based algorithm to identify unimportant variables, and it did not recommend any removals.

We explored the data for outliers and influential points, but treating these did not help.

LINEAR REGRESSION

First, we performed a linear regression with all the predictors included and with treating the quality value as numeric.

| F-Statistic | Adjusted R2 | AIC | BIC |
| --- | --- | --- | --- |
| 174.34 | 0.28 | 11113.48 | 11197.94 |

The full model showed a few notable issues. The adjusted-R2 value is low. There were 553 leverage points, 5 outliers, and 1 influential point. The VIF factors and correlation plot show issues of multicollinearity. Alongside this, the QQplot and Shapiro-Wilk’s test showed that the residuals do not follow a normal distribution. The Durbin-Watson Test (D=1.6206) also showed a positive autocorrelation between error terms.

We use the model recommended by stepwise regression and Boruta to fit a linear regression on the data, treating quality as numeric.

| F-Statistic | Adjusted R2 | AIC | BIC |
| --- | --- | --- | --- |
| 239.73 | 0.28 | 11108.29 | 11173.25 |

Overall, this appears to be a better fit under multiple criteria (namely AIC, as the dataset observed is quite large), albeit not by much. When compared against withheld data, the predictive accuracy of this model (in terms of MAPE) is 0.1029947. This is extremely low accuracy. Based on the hatvalues, there are 236 leverage points, 5 outliers, and one influential observation. Unfortunately, many of the present issues found in the previous model are largely unchanged, such as multicollinearity through high average VIF values, non-normal residuals, and a positive autocorrelation between error terms.

Overall, the model fit does not show good promise in predicting power. This is possibly due to the ordinal nature of the response variable, so our next step is to approach the problem as a categorization.

CATEGORIZATION OF THE DATA:

The response variable is in ordinal form. We have ratings on a scale of 0 to 10 (although the actual ratings range from 3 to 9). While there is a clear ranking in place, it is less clear how this rank scales from a numeric perspective, so an ordinal scale is the better interpretation.

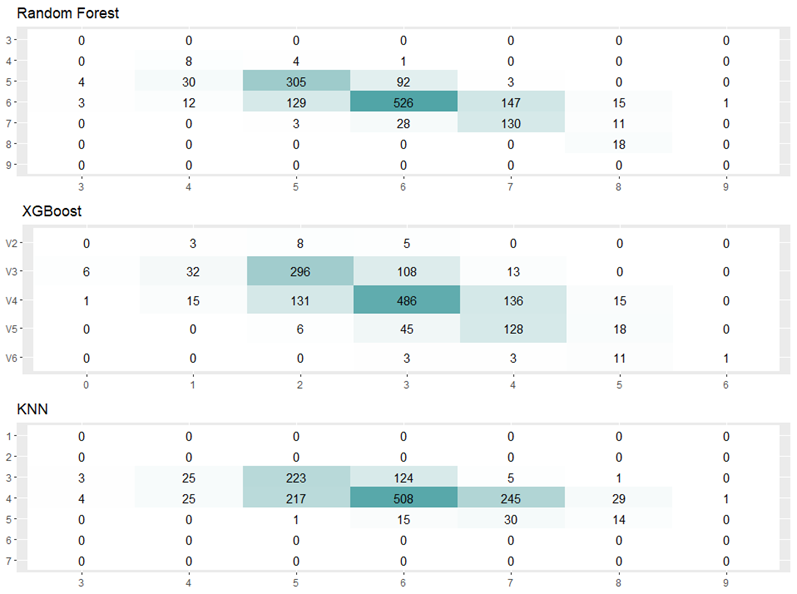
There are three main approaches to categorization of the quality:

1. Each value is considered its own category. This leads to problems since the extreme values are lower represented than the more central ones.
2. The quality is divided into low (<6), medium(=6), high(>6). Here, the data is somewhat evenly divided, with medium quality taking up 44% of the dataset and high being 22%. Here, however, we also explore discarding the medium quality wines in an attempt to more accurately differentiate between low and high quality wines.
3. The quality is divided into low(<=6) and high(>6). Here, the low quality makes up 78%.

CLASSIFICATION

Approach 1

Taking the direct approach of using each ordinal level as a category, we attempted to fit several models to the data. These models included Random Forest and KNN.



Here, while the overall model accuracy seemed relatively high, this was actually driven by high performance in the more central classes (primarily 5, 6 and 7). Meanwhile, more extreme classes saw up to 80% misclassification rates.

Approach 2

Now, we take all quality values below 6 as “low”, and all above 6 as “high”. Initial models show a lot of overlap among medium and the others, so the medium wines are set aside for now.



Using Random Forest and a train-test split of 70-30, we are able to achieve a classification accuracy of 91%. In addition, the sensitivity (considering high quality as the positive, as it is the lower frequency class) is 89%. Thus, this model is able to strongly differentiate between low and high quality wines.

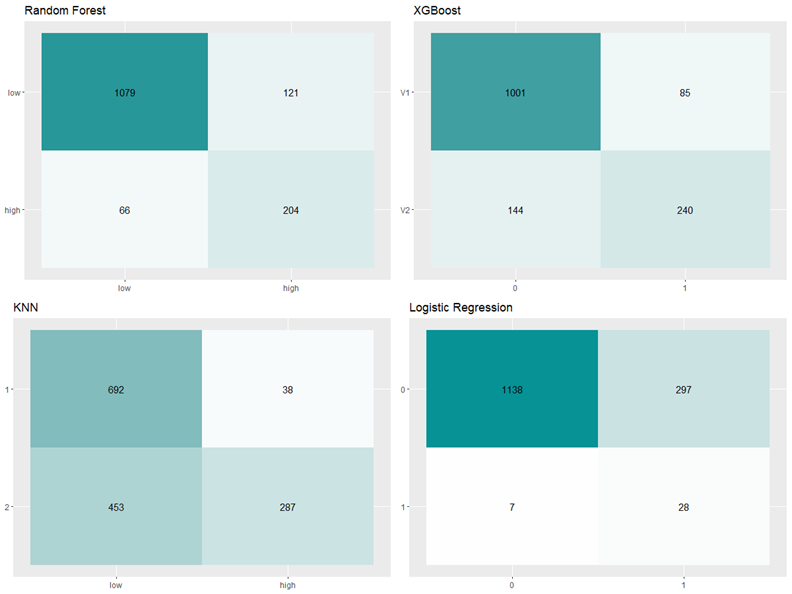
If we now introduce the medium quality wines to this Random Forest model, it splits them relatively evenly between both classes (with about 56% being classified as low quality).

Approach 3

Now, we are no longer setting aside the medium quality wines - instead, we include them in the low quality bracket.

Now, we have 78% of the wines classified as “low” quality, while the remaining are “high”. This represents a class imbalance. While it may not seem that extreme, it does throw off most popular classification algorithms.

As such, we employ oversampling on the train set to bring the class distribution to around even.



With this approach, we now implement Random Forest as a classifier. This now gives us 63% sensitivity, 75% precision, and 94% specificity.

The tradeoff, of course, is that we have transitioned to binary classification, which loses some of the granular information - however, this classifier is still useful, as we will discuss in the next section.

DISCUSSION

Our final model can be Random Forest based on Approach 2 or 3. To recap,

We take our ordinal response variable, and divide it into two classes. All wines with quality below 6 are classified as “low”, while those higher are “high”. For Approach 2, wines with quality 6 are dropped - for Approach 3, they are included as “low”.

The predictors are chosen through stepwise regression, and validated using Boruta. Our final set of predictors is alcohol, free acidity, volatile.acidity, residual.sugar, free.sulfur.dioxide, density, pH and sulfates. These predictors are all numeric.

Due to the relatively lower frequency of high quality wines in Approach 3, we implement oversampling to ensure additional representation. We add additional sampled rows of high quality wines, bringing the ratio from 78-22 to 52-48.

The Random Forest classifier for both approaches takes the following parameters:

* The forest grows 2000 trees, to ensure each row is sampled a few times.
* Each tree randomly selects about 3 variables as candidates, as per the default.
* Each tree draws a sample the same size as the dataset with replacement.
* We have not set a maximum number of nodes, so trees grow as far as possible.

CONCLUSION

Our best final model achieves 91% classification accuracy when classifying wines as high or low quality (excluding medium), while an alternate model predicts high quality wines with 75% precision. This allows vintners to easily identify high quality wines, allowing them to manage pricing, set aside candidates for competitions, better market their products, and other factors that relate to the quality of their offerings, without needing to rely on manual evaluations.

The tradeoff for both approaches 2 and 3 is that the 0-10 scale of quality has been reduced into binary classification - more, one in which 45% (of the wines in the dataset, at least) are either dropped or clubbed with another 33%. As such, when attempting to differentiate between low and medium quality wines, this model will not be useful.

A potential solution is to use a form of nested classification - reclassify all those in the “low” bracket of Approach 3 using another model. Expansion of data using other variables such as lactic acid and yeast content may also be considered. Such a task is outside the scope of this analysis, however, preliminary investigations have shown that there may be some promise to this.

REFERENCES

Dataset - Wine Quality

Created by: Paulo Cortez (Univ. Minho), Antonio Cerdeira, Fernando Almeida, Telmo Matos and Jose Reis (CVRVV) @ 2009

Citation:

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. ISSN: 0167-9236.

Available at: [@Elsevier] http://dx.doi.org/10.1016/j.dss.2009.05.016

[Pre-press (pdf)] http://www3.dsi.uminho.pt/pcortez/winequality09.pdf

[bib] http://www3.dsi.uminho.pt/pcortez/dss09.bib

<https://reneues.files.wordpress.com/2010/01/an-introduction-to-generalized-linear-models-second-edition-dobson.pdf> (See 8.4.2 for information on logistic odd regression models.)

<https://data.library.virginia.edu/fitting-and-interpreting-a-proportional-odds-model/> (See

**Proportional odds:** an ordering constraint and compare against baseline odds.

Slopes are the same for others.

<https://users.stat.ufl.edu/~aa/ordinal/R_examples.pdf>

An introduction to generalized linear models. Dobson Annette

8.4 Proportional Odds Cumulative Logit Model

[Machine-Learning-with-R-datasets/whitewines.csv at master](https://github.com/stedy/Machine-Learning-with-R-datasets/blob/master/whitewines.csv)

<https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityWhites.csv> (Automatic Download link for the dataset)

Appendix - Tables

Table 1 - Data Dictionary

| Column | Format | Description | State | Notes |
| --- | --- | --- | --- | --- |
| Fixed acidity | Numeric | A nonvolatile acid that makes up a majority of wine content (does not evaporate quickly). | Predictor |  |
| Volatile acidity | Numeric | Acids that evaporate quickly (most common in white wines is acetic acid). | Predictor |  |
| Citric acid | Numeric | Citric acid component | Predictor |  |
| Residual sugar | Numeric | Natural grape sugars in wine. Often called the “sweetness” of wine. | Predictor |  |
| Chlorides | Numeric | Salt within the wine. | Predictor |  |
| Free sulfur dioxide | Numeric | Unreacted sulfur dioxide in wine. | Predictor |  |
| Total sulfur dioxide | Numeric | Sum of free and combined sulfur dioxide in the sample. | Predictor |  |
| Density | Numeric | Overall density. | Predictor | (g / cm^3) |
| pH | Numeric | Acidity of a solution. A solution with a pH lower than 7 is considered acidic. | Predictor |  |
| Sulphates | Numeric | Ions present from the fermentation process. Useful in preserving wine. | Predictor |  |
| Alcohol | Numeric | Total alcohol content. | Predictor | (% by volume) |
| Quality | Ordinal | Average of a score from 0-10 from 3 wine experts. | Response | Range 3-9 |