Quality Assessment - Wines

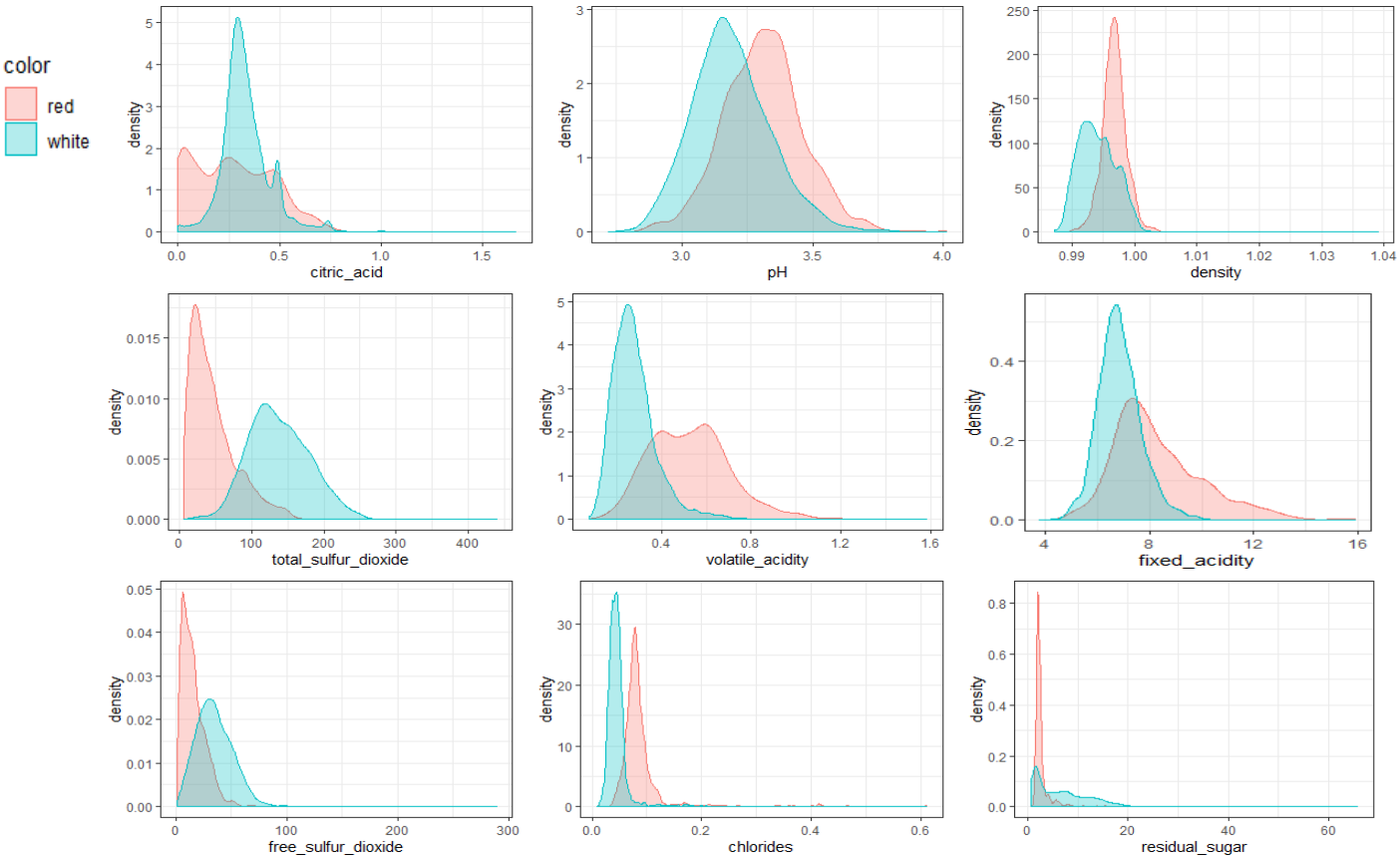
Shabnam Hajian

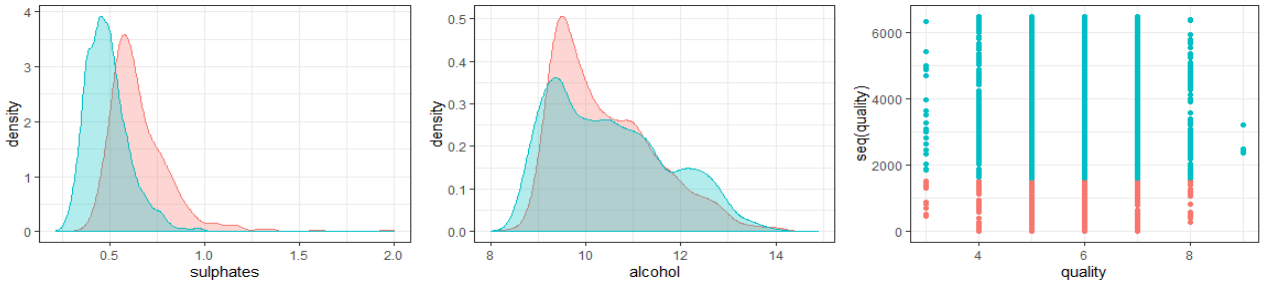
2020-04-22

Summary of the dataset

In this mini-project, we will work on two separate datasets that have been collected from red and white wines samples of "Vinho Verde" in the north of Portugal for quality assessment[1]. The data comes from two separate datasets that included related data of two different wine colours, red and white. The *red* dataset contains *1599* and the *white* one includes *4898* observations. Moreover, both data sets have *12* variables before any changes. To have a general idea about these variables, **fixed_acidity** is the combination of 4 to 5 acids, **volatile_acidity** is the amount of acetic acid in wine, **citric_acid** add freshness and flavour to wines, **residual_sugar** is the amount of sugar remaining after fermentation stops, **chlorides** is the amount of salt in the wine, **free_sulfur_dioxide** is a form of SO2 that exists in wine to prevents microbial growth and the oxidation of the wine, **total_sulfur_dioxide** is the amount of free and bound forms of S02, **density** is the relationship between the mass and size, **pH** a number on a scale of 0 to 14 which shows how acid or alkaline a substance is, **sulphates** which acts as an antimicrobial and antioxidant, **alcohol** the amount of alcohol in the wine, and **quality** which is a score between 0(very bad) and 10(very excellent).

Generally, except quality, all other variables are continues. The *fixed_acidity*, *volatile_acidity*, *sulphates* and *alcohol* data show positive skewness in this dataset. In addition, *citric_acid*, *free_sulfur_dioxide*, *density*, and *pH* looks approximately normal. In *residual_sugar* the majority of data are below 3 which is the median of this dataset and, in *chlorides* variable, most of our data is below 0.1. plus, the *total_sulfur_dioxide* data look bimodal. *quality* as the most important variables, is an integer variable with the minimum=3, maximum=9, mode=6 in this dataset. The below charts give a general idea about these variables in two different colors for two wine color datasets. We also used frequency table for the "quality" to achieve more information about this variable.





In this project, first we will apply **binary logistic regression** to find the *wine color* and then we will apply **ordinal logistic regression** to indicate the *wine quality*. Since we knew that the red and white wines might have different behavior, we try to combine these two logistic regression models to create a general model for wine quality assessment.

Analysis

This part will be split into three subparts. The first part*(A)* contains changing in raw data to turn into a good format, then part *B* is about determining the wine color and the last part*(C)* is about finding a model to determine the wine quality.

A) Format data

First of all, since we need to combine these two files into one, in order to make sure that we will not miss information about the color of our final product, we will add a new variable and call it **color** and fill it "red" or "white" based on the information we have from the dataset; now, the data is "Tidy". After these transformations, we looked for the missing data and find no missing information in the dataset. As a final step, we use visual scatter plots and check the minimum and maximum of the variables. The main purpose of this plotting is to check data for any data entry errors. By looking at graphs, it seems that we only need to consider *free_sulfur_dioxide* that shows a maximum of 289 which is approximately twice of the second-largest value(146.5) in this dataset. Plus, we know that the range of *free_sulfur_dioxide* for wine is usually below 150. Thus, we consider this point as a strange observation. As we do not have any access to the original data to check this observation, we would eliminate this case from our data set by assuming it as a data entry error. Now we can conclude that our data is tidy and clean with no data entry error in 6496 and ready to analyze. We will control outlier and influencer points after we build our models.

B) Predicting the wine color

In this section, we are going to predict the color of the wine by applying *binary logistic regression*. First, we need to make sure that we can apply binary logistic regression:

1- The outcome variable is *color*; this data is categorical and it is in only two categories, "white" or "red".
2- There is not incomplete information problem as we have a full combination of variables.
3- Data have overlap(we can conclude this from the summary of data set graphs).

As a result, we are able to use binary logistic regression in this case. We would check 3 more assumptions of binary logistic regression after we build the model. There is another important note about correlation between predictor variables. "total_sulfur_dioxide" and "free_sulfur_dioxide" are highly correlated to each other(cor=0.72); "density" also shows correlation of 0.55 and -0.69 with "residual_sugar" and "alcohol" respectively. We knew that "total_sulfur_dioxide" is a better indicator in the study of the wine[2]. Thus, from now, we will only consider the "total_sulfur_dioxide" variable. Furthermore, we knew that the wine density depends on the amount of sugar and alcohol in the wine. As a result, we will not consider density in the rest of our analysis as we know that studying "residual_sugar" and "alcohol" is enough. Moreover, we will not use the quality in this prediction model, since we want to find wines' color in all quality levels and then use it in the quality assessment. Now, we will try to find a model to predict wine color based on the remaining variables.

In this dataset, most of our data is white, so we will use "white" as the baseline category. We run the model based on the *forced entry* method and use all 9 predictor variables to predict the color. After running the model, we need to check the significance of the predictor variables. In the model summary, only "citric_acid" p-value is not significant and all other variables are significate at 5% level of significate. Furthermore, we calculate odds ratio, which can be used as the effect size too, as a more reliable method to check the significance of the model input variables. Results show that only "citric_acid" has overlap with one and all the other eight predictor variables are significant, though their effect size might be different. As the next step, we will use the backward stepwise method, and check whether eliminating "citric_acid" will improve the model or not. We run the new model with 8 predictor variables and this time we found that all coefficients show significant p-value at 5% level of significate, with no overlap with one in coefficients odds ratio.

On the other hand, we can use *AIC* of the model results to compare two models. AIC value is a measure of the quality of the model and it is desirable in lower amounts. The AIC of the second model increased in a small amount(from 625.61 in model 1 to 625.99 in model 2). Thus, we need to use the chi-square test to check the null hypothesis of *these two models are not significantly different in predicting wines' color* and the alternating hypothesis as *these two models are different in predicting wines' color*. The p-value of this test is 0.12 which is greater than 0.05 and thus we could not reject the null hypothesis. This means we can conclude that the increase in the model's AIC is not significant. Since the second model has fewer predictor variables and all of these variables are significate, we will continue with the second model to predict the wine color. The other important point is to check how much the model including the predictor variables is better than the model without those predictors. Thus, we want to use chi-square to test following hypotheses:

H_0: Model is not significantly better than chance, in predicting the wines' color
H_1: Model is significantly better than the chance, in predicting the wines' color

The Chi-square p-value of this test is 0 which means that we could reject the null hypothesis with the 5% significate level and use this model as a better model than the chance to predict wine color.

The next step is to check for outliers and influencers; so we need to examine residuals. We will use "standardize residuals" to check for outliers and the "cook's distance" to check for influential points. The results show that 4% of our mode standardize residuals are out of the 2 standard deviation which is less than 5% (comes from predicted amount of data, out of two standard deviations by assuming a normal distribution for residuals). Then we will use the cook's distance and find a maximum of 0.08 for this measure which is considerably below one. Based on these two numbers, we will continue with the assumptions of no outliers at 95% confidence level and no influential points in this model.

Now as the last step in this section, we need to check the binary logistic regression model assumptions:

1- *Linear relationship between continuous predictors and the logit of the outcome variable:* To check this assumption, first we need to make some new variables as the interaction terms of each of the continuous predictor variables with its log. In this model, we used eight predictor variables which all were continuous. Then we will make a new model with all 16 variables(the 8 main variables and the interaction terms), to check whether this assumption is held or not we only need to look at the interaction terms(new variables which we had made) and check their p-value. Any interaction that is significant(p-value(Z)<alpha), indicates that the main variable has violated the assumption of linearity with the outcome logit. Results show four of the log interaction terms are significate at 5% level of significance. In other words "volatile_acidity", "chlorides", "total_sulfur_dioxide", and "pH" variables are violated the linearity assumption in this model. Thus, there is a limitation on the generalization of this model.

2- *Independence of errors:* The test results indicate d=1.5 which not close to 2(as the threshold) and the test p-value=0 which is smaller than 0.05. As a result, we reject the null hypothesis of this test(H_0: the errors are significantly independent) at the 5% level of significance and conclude that the

model violated the assumption of independence of residuals and thus we have limitations in the model generalization.

3- - *multicollinearity(predictor variables should not correlate highly):* we use tolerance(=1/VIF) to check multicollinearity between predictor variables. Since the results show all numbers are above 0.2(minimum tolerance in the model is 0.6) and the VIF mean is considerable below 10(model VIF mean=1.4), we conclude there is no multicollinearity problem and continue with the assumption of no multicollinearity in this model.

To conclude this section, we found a model to predict the wine color with eight significate predictor variables by using binary logistic regression and the below table shows a summary of the results of the wines' color prediction model. We did not find any concern about the outliers and influencer points at 5% level of significate in the model, however, by checking assumptions, we realized that two of the three model assumptions are violated and we have a limitation in the generalization of this model.

## Variables	Estimate	std_error	Odds_lower_bound	Odds_ratio	Odds_Upper_bound
## (Intercept)	-45.09597	3.651291	1.576e-23	2.601e-20	2.647e-17
## fixed_acidity	1.39677	0.124679	3.198e+00	4.042e+00	5.219e+00
## volatile_acidity	11.23457	0.743564	1.851e+04	7.570e+04	3.432e+05
## residual_sugar	-0.10988	0.038188	8.170e-01	8.959e-01	9.522e-01
## chlorides	34.55286	3.261825	1.895e+12	1.014e+15	7.064e+17
## total_sulfur_dioxide	-0.05567	0.003397	9.393e-01	9.458e-01	9.519e-01
## pH	9.59728	0.831876	3.003e+03	1.472e+04	7.862e+04
## sulphates	8.88702	0.913644	1.306e+03	7.237e+03	4.660e+04
## alcohol	-0.34842	0.107500	5.704e-01	7.058e-01	8.699e-01

C)Predicting the wines' quality

In this section, we will use the results of the color model in the last part to create a prediction model for wine quality by applying *ordinal logistic regression*(also known as "proportional odds logistic regression"). We choose ordinal logistic regression over the multinominal logistic regression based on the ordinal characteristics of our outcome variable, quality. The multinomial logistic regression does not preserve the ranking information in the dependent variable when returning the information on the contribution of each independent variable and thus it is not a good choice when the outcome has the rank characteristic as the "quality" in this dataset.

Right now the outcome variable is stored as a score between 0 to 10. To make modelling easier, we will divide this rank score into three larger categories as **poor**(ranks 0-4), **fair**(ranks 5-7), and **excellent**(ranks 8-9) and store it in the new variable as "**label**". Now by choosing *label* as the outcome variable, we have data for all possible outcome variables. One important note is, this new variable(label) should store as an ordinal variable to be able to run ordinal logistic regression. Now we will check two assumptions of using ordinal logistic regression:

1- The outcome variable is *label*, which is the ordinal categorical variable with three categories(more than two) in the order that excellent is the best category and poor is the worst one.
2- All predictor variables are either continuous, categorical or ordinal.

As these two basic assumptions hold, we are able to use ordinal logistic regression and we will check the other two assumptions after building the model. Plus, we know we don't have any data entry errors or missing value as we had checked them in the last section. As the last step before modelling, we would partition data into two groups of *train*(75% of data) and *test*(25% of data) to being able to test the consistency of the model. This split let us apply the results we obtained from the trained group into the test group and evaluate out of sample errors.

To modelling the quality of the wine, we will use the study which has been done by Cortez et al., about the predictor variable in the quality assessment. Based on that report we knew that *sulphates*, *alcohol*, and *volatile_acidity* play an important role in predicting the quality of wines. We also know that *citric_acid*, and *residual_sugar* are important variables, but they are more important when the wine color is "white".

Now as regards we know which of the input variables are important, we enter them into the model based on the *hierarchical* method. The important note about the *citric_acid* and *residual_sugar* is that they should represent in the model as an interaction variable with the wine color. On the other hand, we have mentioned before that the purpose of this report is to combine two models of predicting wine color and wine quality together. Thus, instead of using the color column from the first dataset, we will use the fitted value of the model we have made in the prior part. The important note is the color models' probabilities could not enter directly into this new model and we need to change and store them into a new categorical(factor) variable as "**colormodel"**.

After running the model, the summary of results shows only the coefficients "value", "standard error" and their "t value"; however we can easily calculate the "p-value" based on the amount of prepared "t-value". After calculating the p-values, we can see that although the "citric_acid" shows a smaller p-value when the wine color is white, this p-value is still not significant at 5% level of significate. As a result, for the next step, we will apply the backward stepwise method and eliminate the interaction variable of the "citric_acid" and "colormodel". Then we will check p-values, and the amount of model improvement according to the changes of AIC.

After running a new model, we can see that the model coefficients are all significate at 5% level of significate. The model AIC also decrease in a small amount(from 2704.269 in the first model to the 2701.108 in the second model) which indicates the second model as a better model to predict wine quality. Thus, we will accept the second model for wine quality with the following summary.

Variables	Estimate	std_error	Odds_lower_bound	Odds_ratio	Odds_Upper_bound
volatile_acidity	-3.16277081	0.38626505	0.019902	0.04230835	0.09056606
sulphates	1.00413867	0.42873279	1.182358	2.72955522	6.31943270
alcohol	0.65913785	0.05496913	1.736947	1.93312497	2.15488124
colormodelred:residual_sugar	0.20578622	0.06267885	1.085496	1.22849054	1.38621444
colormodelwhite:residual_sugar	0.07051915	0.01312430	1.045465	1.07306511	1.10097871
poor|fair	3.28661392	0.64093111	NA	NA	NA
fair|excellent	10.63056501	0.70848718	NA	NA	NA

The above table presents a summary of the quality model. In this table, data are shown in two parts. The first part is the *coefficients* of the model, we show the "estimated" value and the "standard error" of each coefficient. We also calculated the odds ratio for these variables to have a better idea about them. The interpretation of this part is:

> *We expect for one unit increase in "Variables" parameter, the expected value of "label"(outcome variable) change by the amount of "Estimate" in the log-odds scale, given that all other variables in the model are held constant.*

The second part is the *intercepts* of this model and can interpret as follow:
poor|fair *the log of odds of having wine quality of "poor" versus having a wine quality of "fair" or "excellent".*
fair|excellent *the log of odds of having wine quality of "poor" or "fair" versus having wine quality of "excellent".*

As we mentioned before, one of the advantages of splitting the sample into two parts and modelling for the training part is that we could check the strength of the model by the test part. In this report, we will use the *confusion matrix* and *misclassification error*. The confusion matrix gives a visual inspection of the performance of the ordinal logistic regression model. We can also calculate the misclassification error of the model to have a mathematical control. First, we calculate the misclassification error for the train data(75% of data that we had made the model based on them) and then compare it with the misclassification error of the test data(the other 25% of data). The misclassification error for train data is 6.97% and for the test data, this error is 6,43%. The errors of the trained data and the test data do not show major difference which indicates the consistency of the model we built for the wine quality.

As the last step, we need to check the other two assumptions of the ordinal logistic regression model.

1-No multicollinearity between dependent variables: To test this assumption, we need to perform "Variance Inflation Factor(VIF)" test. The ordinal logistic regression model uses a categorical variable as an outcome. Therefore, VIF will not be sensible. To solve this problem, we usually need to transfer categorical variables to a numeric dummy variable. However, because in this report we already have the numeric variable for the label(quality), we can use the "quality" variable in this part. We will fit a multiple linear regression by using "quality" as a numeric variable and all other variables in the main model. The results of the test show all VIF small(between 1 and 2) and the mean of VIF as 1.3 which is considerably below 10(as a threshold). Thus, we conclude that there is no multicollinearity problem in this model and the assumption is met.

2-Proportional odds(parallel regression): This assumption means that each pair of outcome groups should have the same relationship. If each pair of outcome groups have the same relationship, there is only one set of the coefficient that describes this relationship. As a result, the model will be unique. This assumption can be checked by the "**Brant's test"** with the null hypothesis as: "parallel regression assumption holds". An important note in using this test is, we should eliminate the interaction variables since this test could not accept such variables. After making changes and running the test, the result shows significate p-values for all variables and "Omnibus"(as a measure for the whole model) at 5% level of significate(all p-values are less than 0.05). Therefore, we conclude that the proportional odds assumption is violated and this model has a limitation in generalization.

Conclusion

In this report, we used ordinal logistic regression to predict wine quality based on the most critical variables that had been founded by the study of Cortez et al., and employed the binary logistic regression model for the wines' color. First, we tried to predict wine color as we knew that it is an essential variable in quality evaluation. Then, we applied the color prediction model results in our second model to predict wine quality. After building models, we looked for outliers, influencers, and the strength of our models. We also checked model assumptions to found out about the models' capability in generalization.

Potential follow-up: First, about the color model, we used forced entry and backward stepwise methods. However, to do such modelling, it is better to study reasoning and create a model based on the hierarchical method. Moreover, in this model we found excessive amounts of standardize-residuals out of the three standard deviations; eliminating these far observations can increase model accuracy. Then, in the quality model, we manually split our data into a single test set(train sample and test sample) and evaluated out of sample error once. However, in this process, the presence or absence of the outliers can considerably change the out of sample ambiguity. Using cross-validation approaches will give us a more precise estimate of the out of sample errors. In addition, cross-validation can help us to study the influencers and outliers in the model more accurately. Furthermore, we found out that both of our models are violated some of their assumptions, this means that we have a generalization problem for both models. Some changes, for instance, removing outliers or doing causation analysis before modelling will improve the model's ability in prediction. Re-modelling or applying better approaches also will help to have stronger models with generalization merit. As a final note, there is a strong role for the color model in this report as it has been used in the second model too. Thus, any improvement in the color model will affect the quality model as well and a better color model can give us a better model to predict the quality of wines.

References

[1] 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.

[2] Field, A. P., Miles, J., & Field, Z. (2012). Discovering statistics using R/Andy Field, Jeremy Miles, Zoë Field.

Some of information or ideas retrieved from:

https://www.extension.iastate.edu/wine/total-sulfur-dioxide-why-it-matters-too

http://rstudio-pubs-static.s3.amazonaws.com/330833_fcecaf54338b4aecb79a3f46aff2f054.html

https://www.researchgate.net/publication/311919082_The_Classification_of_White_Wine_and_Red_Wine_According_to_Their_Physicochemical_Qualities

https://www.edureka.co/community/46062/different-types-of-logistic-regression

https://medium.com/evangelinelee/ordinal-logistic-regression-on-world-happiness-report-221372709095

https://www.youtube.com/watch?v=qkivJzjyHoA

https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/

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

https://towardsdatascience.com/implementing-and-interpreting-ordinal-logistic-regression-1ee699274cf5

https://stackoverflow.com/questions/6988184/combining-two-data-frames-of-different-lengths

https://medium.com/analytics-vidhya/a-guide-to-machine-learning-in-r-for-beginners-part-5-4c00f2366b90

https://www.r-bloggers.com/how-to-perform-ordinal-logistic-regression-in-r/

https://stats.stackexchange.com/questions/43551/how-to-test-for-outliers-in-an-mlogit-model-in-r

https://stats.stackexchange.com/questions/58772/brant-test-in-r

https://www.youtube.com/watch?v=pXJb3u6Dlw4