Predicting Wine Quality & Color from Physiochemical Properties using Statistical Modeling

Vivek Sahukar

Course IDS 702 | Modeling & Representation of Data

ABSTRACT

The wine datasets are used to predict the quality and color from 11 physiochemical properties of the wine. Quality is an ordinal variable which ranges from 1 (lowest) to 10 (highest). Color is a categorical variable which has two levels: Red and White. All physiochemical properties are treated as continuous predictor variables. Initial exploratory data analysis has been done to get a better understanding of predictors and efficiently deal with multicollinearity and outliers. Regression techniques have been employed to predict the quality of the wine. Diagnostics have been done to check the consistency of the model. Interactions among predictors and transformations of the explanatory variables had been done to improve the plain vanilla model. An attempt has been made to use advanced machine learning models for increasing prediction accuracy in this classification problem. At last, the model choice has been suggested by comparing explanatory power vs. prediction accuracy.

Contact: Vivek Sahukar | Duke University | Masters in Interdisciplinary Data Science
Mobile: 1-984-209-6183 | Email: vivek.sahukar@duke.edu

INTRODUCTION

- Why there is a need for a model for wine tasting despite the presence of experienced wine tasters? Wine tasting is very tough and subjective. The data from already tried and tasted wine can be used to build a model which can predict the quality. Though these models are not a substitute for an experienced and nuanced human wine connoisseur, they would assist him/her in the decision-making process
- Scientific models are aloof to any human judgment bias. They justify the oenologist's wine evaluations. These models help in improving the quality (by reducing human error) and speed up the decision-making process.
- Measuring the impact of physiochemical properties on final wine quality can help in improving the production process. The wine producer does not have to wait to test the particular wine quality in the market. Thus, wine producer can freely experiment with winemaking process and shorten the response feedback time in the supply chain operation cycle.
- Further application of such models is target marketing, where consumer preferences can be modeled in the niche or profitable markets.

OBJECTIVES

1. Do the physiochemical properties of the wine differ by color? (Use logistic regression to predict the color of the wine)
2. How accurately can the category of the wine be predicted from its physiochemical properties? (Use multinomial logistic regression model)
3. How much be the accuracy of prediction improved by using black box models such as Support Vector Machines and Random Forest, while sacrificing the interpretability of the model?

PREDICTORS

Wine Physiochemical Property	Unit	Description
fixed acidity	g(tartaric acid) / dm^3	Measurement of total concentration of titratable acids & free hydrogen ions in wine (non-volatile acids)
volatile acidity	g(acetic acid) / dm^3	Refers to amount of acetic acid in wine; is an indicator of spoilage or errors in manufacturing process; too much causes vinegar taste
citric acid	g/dm^3	Acts as a preservative and added to increase acidity or complement flavors or add 'freshness'. Too much can ruin the taste
residual sugar	g/dm^3	Refers to the natural grape sugars that are leftover of fermentation. Gives rise to the sweet taste
chlorides	g(sodium chloride) / dm^3	Indicates "saltiness". Influenced by the environment, cultivation practices and grape variety. Right amount makes wine savory
free sulfur dioxide	mg/dm^3	Free form of SO2, prevents microbial growth and oxidation.
total sulfur dioxide	mg/dm^3	Most common preservative to prevent negative effects of exposure to air. Acts as sanitizing agent killing unwanted microorganisms. Too much becomes evident
density	g/cm^3	Also known as specific gravity, used to measure alcohol concentration. Sweeter wines have higher densities generally.
ph	no unit (continuous) range: 0-14	Measure of hydrogen ion concentration. Acidic (pH < 7) | Neutral (pH = 7) | Basic (pH > 7)
sulphates	g(potassium sulphate)/dm^3	Used to correct mineral deficiencies in water during brewing. Also adds bit of 'sharp' taste
alcohol	vol.%	Refers to the percent alcohol content in wine

DATASETS

The datasets pertain to the red and white variants of the vinho verde wine, which belongs to the Minho (northwest) region of Portugal. The period of the data corresponds to May 2004 to February 2007. Automated samples were taken from the laboratory for wine testing. Each sample was blind tasted by a minimum of 3 expert wine tasters. The grading of the wine is discrete from 0 (very bad) to 10 (excellent). The final quality score of the wine sample is median of all the scores assigned by different experts. Each observation in the dataset corresponds to quality score and measure of physiochemical properties. Since there are few data points in lower and higher quality, the dataset has been divided into three types of quality: Low = 1 (< 6), Medium =2 (= 6), High = 3 (> 6). Moreover, multinomial logistic regression does not work very well if there are more than 4 or 5 levels.

Variables	**Type**	**Count**	**Type**	**Range**
Quality	Outcome	1	Ordinal	1 / 2/ 3
Color	Outcome	1	Categorical	Red – 0 | White - 1
Physiochemical properties	Predictor	11	Continuous	Discussed ahead in summary statistics

Wine Quality Wine Dataset	1	2	3	Total Observations	Missing Values
Red	744	638	217	1599	0
White	1640	2198	1060	4898	0

LOGISTIC REGRESSION MODEL – SUMMARY

glm(formula = color ~ . - quality + as.factor(quality), family = binomial, data = wr.m)

Predictors	Estimate	Std. Error	z value	Pr(>|z|)	2.5%	97.5%	exp(estimate)	exp(2.5%)	exp(97.5%)
(Intercept)	4.23	0.33	12.72	0.00	3.58	4.88	68.65	35.77	131.65
fixed.acidity	0.40	0.23	1.70	0.09	-0.06	0.86	1.49	0.94	2.36
volatile.acidity	-6.36	1.05	-6.07	0.00	-8.42	-4.31	0.00	0.00	0.01
citric.acid	2.61	1.17	2.22	0.03	0.31	4.91	13.59	1.36	135.73
residual.sugar	0.95	0.10	9.41	0.00	0.75	1.15	2.59	2.12	3.16
chlorides	-22.26	3.98	-5.60	0.00	-30.05	-14.47	0.00	0.00	0.00
free.sulfur.dioxide	-0.07	0.01	-4.70	0.00	-0.09	-0.04	0.94	0.91	0.96
total.sulfur.dioxide	0.05	0.00	10.68	0.00	0.04	0.06	1.05	1.04	1.06
density	-1859.00	191.40	-9.71	0.00	-2233.99	-1483.67	0.00	0.00	0.00
pH	1.83	1.42	1.28	0.20	-0.96	4.62	6.22	0.38	101.39
sulphates	-2.94	1.26	-2.33	0.02	-5.41	-0.47	0.05	0.00	0.63
alcohol	-1.87	0.28	-6.64	0.00	-2.42	-1.31	0.15	0.09	0.27
as.factor(quality)2	-0.20	0.37	-0.54	0.59	-0.92	0.52	0.82	0.40	1.69
as.factor(quality)3	-0.38	0.56	-0.68	0.50	-1.48	0.72	0.68	0.23	2.05

Number of Fisher Scoring iterations: 9
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 7250.98 on 6496 degrees of freedom
Residual deviance: 428.35 on 6483 degrees of freedom
AIC: 456.35

Deviance Residuals				
Min	1Q	Median	3Q	Max
-5.66290	0.00110	0.01860	0.05720	6.76490

Interpretation

- As variable (such as residual sugar) is increased by 1 unit, the log-odds for color 'white' increase by 0.95 or odds for color 'white' increase by a factor of 2.59 (= $e^{0.95}$), confidence interval (2.12, 3.16)
- β_0 = 4.23 is the log-odds for color 'white' at the mean of one physiochemical property.
- In other words, the odds for color 'white' at the mean of one predictor is 68.65

MULTINOMIAL LOGISTIC REGRESSION MODEL – RED WINE

multinom (formula = quality ~ ., data = red.m)

	Quality = 2								
Predictors	Coefficients	Std. Errors	z value	Pr(>|z|)	2.5	97.5	exp(coeff)	exp(2.5)	exp(97.5)
(Intercept)	0.06	0.06	0.87	0.38	-0.07	0.18	1.06	0.93	1.20
fixed.acidity	0.09	0.06	1.47	0.14	-0.03	0.22	1.10	0.97	1.24
volatile.acidity	-3.06	0.49	-6.23	0.00	-4.02	-2.10	0.05	0.02	0.12
citric.acid	-1.42	0.57	-2.48	0.01	-2.55	-0.30	0.24	0.08	0.74
residual.sugar	0.01	0.05	0.32	0.75	-0.08	0.10	1.01	0.93	1.11
chlorides	-2.99	1.58	-1.90	0.06	-6.09	0.10	0.05	0.00	1.10
free.so_2	0.02	0.01	2.76	0.01	0.01	0.04	1.02	1.01	1.04
total.so_2	-0.02	0.00	-5.25	0.00	-0.02	-0.01	0.98	0.98	0.99
density	-2.66	0.01	-310.63	0.00	-2.68	-2.65	0.07	0.07	0.07
pH	-0.43	0.60	-0.72	0.47	-1.61	0.74	0.65	0.20	2.10
sulphates	2.37	0.45	5.23	0.00	1.48	3.26	10.68	4.39	25.97
alcohol	0.79	0.08	10.24	0.00	0.64	0.94	2.20	1.89	2.56
	Quality = 3								
Predictors	Coefficients	Std. Errors	z value	Pr(>|z|)	2.5	97.5	exp (coeff)	exp(2.5)	exp(97.5)
(Intercept)	-2.06	0.15	-13.81	0.00	-2.35	-1.77	0.13	0.10	0.17
fixed.acidity	0.34	0.09	3.74	0.00	0.16	0.52	1.41	1.18	1.68
volatile.acidity	-4.72	0.85	-5.58	0.00	-6.38	-3.07	0.01	0.00	0.05
citric.acid	-0.45	0.93	-0.49	0.63	-2.27	1.37	0.64	0.10	3.93
residual.sugar	0.25	0.07	3.82	0.00	0.12	0.38	1.29	1.13	1.47
chlorides	-10.76	3.54	-3.04	0.00	-17.69	-3.82	0.00	0.00	0.02
free.so_2	0.03	0.01	1.85	0.06	0.00	0.05	1.03	1.00	1.05
total.so_2	-0.03	0.01	-4.85	0.00	-0.04	-0.02	0.97	0.96	0.98
density	-267.31	0.02	-15420.29	0.00	-267.34	-267.27	0.00	0.00	0.00
pH	-0.07	0.95	-0.08	0.94	-1.93	1.79	0.93	0.15	5.97
sulphates	5.38	0.63	8.50	0.00	4.14	6.63	217.88	62.92	754.47
alcohol	1.31	0.11	11.66	0.00	1.09	1.54	3.72	2.98	4.64

Summary

Residual Deviance	2426.159
AIC:	2474.159

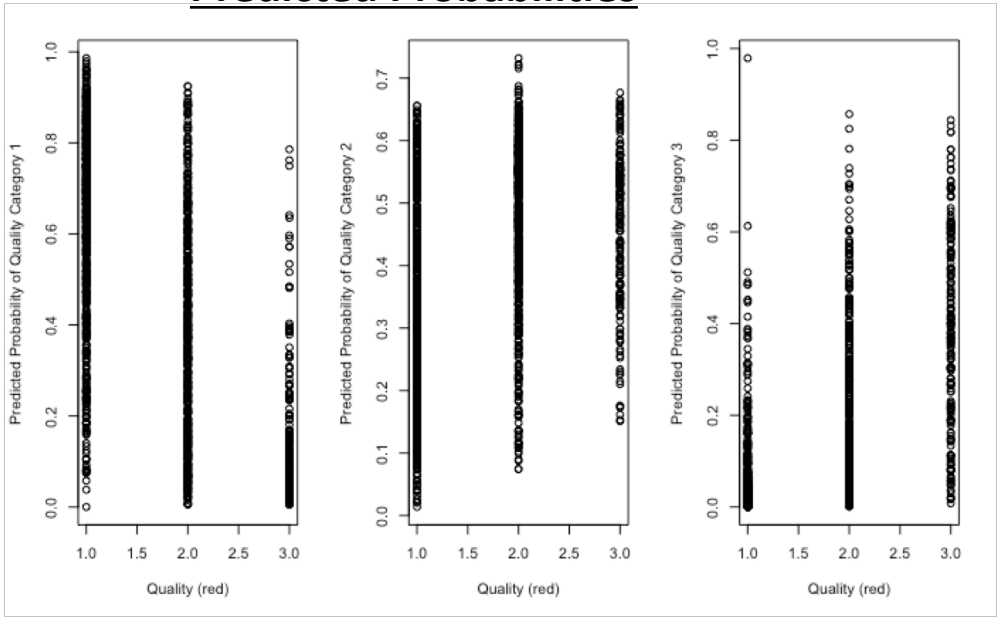
MULTINOMIAL LOGISTIC REGRESSION MODEL – RED WINE

Confusion Matrix

Quality	Original Categories		
Predicted categories	1	2	3
1	580	226	14
2	156	363	124
3	8	49	79

Accuracy = 63.91%

Predicted Probabilities



Diagnostics

- All the predictors have been mean centered before fitting the model.
- Binned residual plots are fine. There is no definite pattern in these plots across all levels of quality.
- By inspecting the model summary, it can be seen that variables: fixed acidity, residual sugar, citric acid, & pH are not significant (having p-value > 0.05). Change in deviance test is used to check whether these variables are useful predictors or not.
- All except pH are relevant. However, prediction accuracy decreased from 63.91% to 63.79%. Also, binned plots grew worse for alcohol and free sulfur dioxide.
- All except pH are relevant. Moreover, prediction accuracy decreased slightly from 63.91% to 63.79% when pH is dropped. Also, binned plots grew worse for alcohol and free SO2. Therefore all the predictors are kept in the model for the better explanatory power of the model.

MULTINOMIAL LOGISTIC REGRESSION MODEL – RED WINE

Interpretations

- Level 1 (Bad Quality) is the baseline in this model.
- For each 1 unit increase in any physiochemical property, say total SO2, the log odds of red wine being medium quality (level 2) instead of being a bad quality (level 1) increase by a factor of -0.02.
- For each 1 unit increase in any physiochemical property, say total SO2, the odds of red wine being medium quality (level 2) instead of being a bad quality (level 1) increase by a factor of 0.98 (= e-0.02).
- The 95% confidence interval limits for this multiplicative increase are given by (coeff. +/- 1.96 * S.E.). The coefficients and 95% confidence intervals are already exponentiated here.
- Similarly, for each 1 unit increase in any physiochemical property, say total SO2, the odds of red wine being high quality (level 3) instead of being level 1 increase by a multiplicative factor of 0.97 [Confidence interval : (0.96, 0.98)]

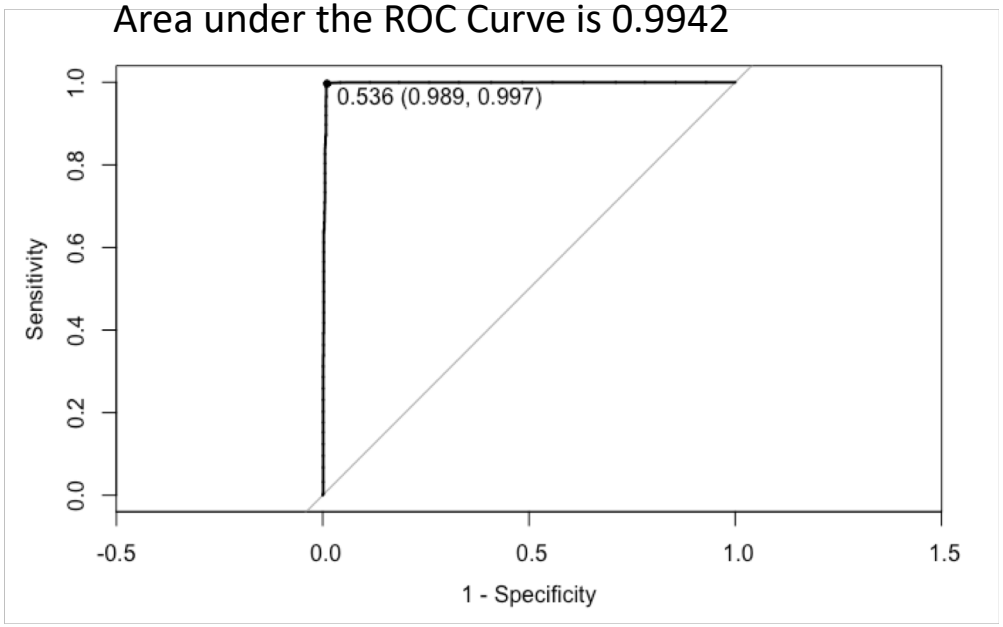
Discussion

- By looking at some residual plots, transformations (both log and squaring) and interactions (such as between free SO2 & total SO2, fixed acidity & volatile acidity since these pairs of variables are scientifically related). However, strong interaction effects were not found (p-values were insignificant at 95% confidence level). However, the prediction accuracy improved from 63.91% to 64.29% only. Still, the model with no interactions and transformations is chosen for easier interpretability.
- Outliers: There was no useful effect of transformations & interactions on outliers. The effect of outliers could be seen in binned residual plots. Therefore, complete regression is rerun again after removing outliers. Now, the number of observations dropped from 1599 to 1212. The prediction accuracy increased from 63.91% to 65.18%, while binned residuals plots remained the same. Almost 25% data has been lost after removing outliers, causing severe loss of information. Hence, outliers are included for analysis, and the original multinomial logit model is accepted without any transformations or interactions.

LOGISTIC REGRESSION MODEL

ROC Curve

Area under the ROC Curve is 0.9942



Confusion Matrix

Predicted Color \ Actual Color	1	2
Red (0)	1582	17
White (1)	16	4882

Best Threshold	0.536
Accuracy	99.49%

Diagnostics

- All predictors have been mean centered.
- Binned plots work well here since both the datasets are large.
- There are no systematic patterns in the binned plots of residuals vs. the predictors. Hence the model describes the data very well.
- Moreover, the prediction is very accurate without any transformation or interaction.
- Therefore, the logistic regression model mentioned previously is used for predicting the color of the wine using the physiochemical properties of the wine.

Discussion

- Prediction accuracy for the color of the wine from its physiochemical properties is very high.
- No transformations or interactions have been used in this model, as the accuracy of prediction is very high.
- Only ‘Quality’ is not significant for prediction of color.
- Hence, the color of the wine is inherently different due to physiochemical properties. Therefore, in further analysis, separate modeling has been done for red and white wine datasets.

MULTINOMIAL LOGISTIC REGRESSION MODEL – WHITE WINE

multinom (formula = quality ~ ., data = white.m)

Summary

	Quality = 2								
Predictors	Coefficients	Std. Errors	z value	Pr(>|z|)	2.5	97.5	exp(coeff)	exp(2.5)	exp(97.5)
(Intercept)	0.59	0.04	14.87	0.00	0.51	0.67	1.81	1.67	1.95
fixed.acidity	-0.05	0.05	-1.05	0.29	-0.15	0.04	0.95	0.86	1.05
volatile.acidity	-6.17	0.42	-14.55	0.00	-7.01	-5.34	0.00	0.00	0.00
citric.acid	0.19	0.31	0.63	0.53	-0.41	0.80	1.21	0.66	2.22
residual.sugar	0.12	0.01	14.62	0.00	0.11	0.14	1.13	1.11	1.15
chlorides	1.59	1.64	0.96	0.33	-1.64	4.81	4.88	0.19	122.70
free.so_2	0.01	0.00	2.75	0.01	0.00	0.01	1.01	1.00	1.01
total.so_2	0.00	0.00	-0.91	0.36	0.00	0.00	1.00	1.00	1.00
density	-164.93	0.01	-16829.07	0.00	-164.95	-164.91	0.00	0.00	0.00
pH	0.55	0.27	2.02	0.04	0.02	1.09	1.74	1.02	2.98
sulphates	1.45	0.36	4.08	0.00	0.75	2.15	4.27	2.12	8.58
alcohol	0.69	0.05	15.03	0.00	0.60	0.79	2.00	1.83	2.19
	Quality = 3								
Predictors	Coefficients	Std. Errors	z value	Pr(>|z|)	2.50	97.50	exp (coeff)	exp(2.5)	exp(97.5)
(Intercept)	-0.60	0.06	-10.54	0.00	-0.71	-0.48	0.55	0.49	0.62
fixed.acidity	0.48	0.07	7.17	0.00	0.35	0.61	1.61	1.41	1.83
volatile.acidity	-8.04	0.59	-13.66	0.00	-9.20	-6.89	0.00	0.00	0.00
citric.acid	-0.57	0.46	-1.24	0.22	-1.47	0.33	0.57	0.23	1.39
residual.sugar	0.37	0.01	31.12	0.00	0.35	0.39	1.45	1.42	1.48
chlorides	-11.05	3.95	-2.80	0.01	-18.78	-3.31	0.00	0.00	0.04
free.so_2	0.01	0.00	3.87	0.00	0.01	0.02	1.01	1.01	1.02
total.so_2	0.00	0.00	-0.64	0.52	0.00	0.00	1.00	1.00	1.00
density	-745.63	0.03	-26056.11	0.00	-745.69	-745.58	0.00	0.00	0.00
pH	3.57	0.36	9.87	0.00	2.86	4.28	35.60	17.52	72.34
sulphates	3.21	0.43	7.54	0.00	2.38	4.05	24.86	10.78	57.29
alcohol	0.68	0.06	11.38	0.00	0.56	0.80	1.98	1.76	2.22

Residual Deviance	8571.07
AIC:	8619.07

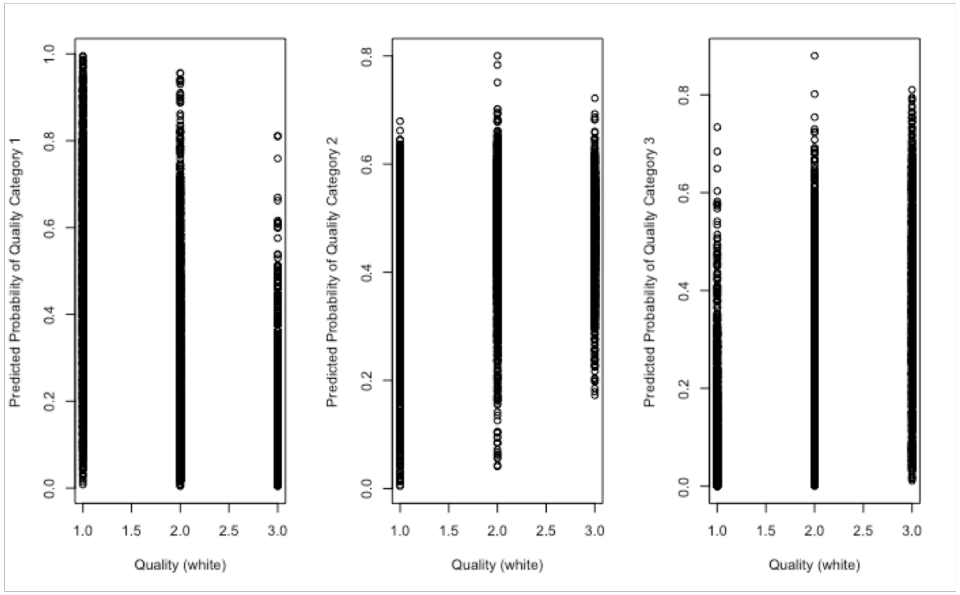
MULTINOMIAL REGRESSION MODEL – WHITE WINE

Confusion Matrix

Quality	Original Categories		
Predicted categories	1	2	3
1	944	458	59
2	671	1506	634
3	25	234	367

Accuracy = 57.51%

Predicted Probabilities



Diagnostics

- All the predictors have been mean centered before fitting the model.
- Binned residual plots are fine. There is no definite pattern in these plots across all levels of quality.
- By inspecting the model summary, it can be seen that variables: fixed acidity, citric acid, chlorides, & total sulfur dioxide are not significant (having p-value > 0.05). Change in deviance test is used to check whether these variables are useful predictors or not.
- All except citric acid, & total sulfur dioxide are relevant. Moreover, prediction accuracy increased slightly from 57.51% to 57.84% when these two variables are dropped. However, there is no improvement in binned residual plots. Therefore all the predictors are kept in the model for the better explanatory power of the model.

MULTINOMIAL LOGIT MODEL – WHITE WINE

Interpretations

- Level 1 (Bad Quality) is the baseline in this model.
- For each 1 unit increase in any physiochemical property, say residual sugar, the log odds of white wine being medium quality (level 2) instead of being a bad quality (level 1) increase by a factor of 0.12.
- For each 1 unit increase in any physiochemical property, say residual sugar, the odds of white wine being medium quality (level 2) instead of being a bad quality (level 1) increase by a factor of 1.13 (= e0.12).
- The 95% confidence interval limits for this multiplicative increase are given by (coeff. +/- 1.96 * S.E.). The coefficients and 95% confidence intervals are already exponentiated here.
- Similarly, for each 1 unit increase in any physiochemical property, say residual sugar, the odds of white wine being high quality (level 3) instead of being level 1 increase by a factor of 1.45 [Confidence interval : (1.42, 1.48)]

Discussion

- By looking at some residual plots, transformations (both log and squaring) and interactions (such as between free SO2 & total SO2, fixed acidity & volatile acidity since these pairs of variables are scientifically related). Even though strong interaction effects were found (p-values were significant at 95% confidence level), the prediction accuracy improved from 57.51% to 58.78% only. Therefore, the model with no interactions and transformations is chosen for easier interpretability.
- Outliers: There was no useful effect of transformations & interactions on outliers. The effect of outliers could be seen in binned residual plots. Therefore, complete regression is rerun again after removing outliers. Now, the number of observations dropped from 4898 to 4074. The prediction accuracy dropped to 56.67%, with no additional improvement in plots of binned residuals vs. predictors. More than 16% data has been lost after removing outliers leading to a severe loss of information. Hence, outliers are included for analysis, and the original multinomial logit model is accepted without any transformations or interactions

EXPLORATORY DATA ANALYSIS

Summary Statistics of Explanatory Variables

RED WINE	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
fixed.acidity	4.60	7.10	7.90	8.32	9.20	15.90
volatile.acidity	0.12	0.39	0.52	0.53	0.64	1.58
citric.acid	0.00	0.09	0.26	0.27	0.42	1.00
residual.sugar	0.90	1.90	2.20	2.54	2.60	15.50
chlorides	0.01	0.07	0.08	0.09	0.09	0.61
free.sulfur.dioxide	1.00	7.00	14.00	15.87	21.00	72.00
total.sulfur.dioxide	6.00	22.00	38.00	46.47	62.00	289.00
density	0.99	1.00	1.00	1.00	1.00	1.00
pH	2.74	3.21	3.31	3.31	3.40	4.01
sulphates	0.33	0.55	0.62	0.66	0.73	2.00
alcohol	8.40	9.50	10.20	10.42	11.10	14.90

WHITE WINE	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
fixed.acidity	3.80	6.30	6.80	6.86	7.30	14.20
volatile.acidity	0.08	0.21	0.26	0.28	0.32	1.10
citric.acid	0.00	0.27	0.32	0.33	0.39	1.66
residual.sugar	0.60	1.70	5.20	6.39	9.90	65.80
chlorides	0.01	0.04	0.04	0.05	0.05	0.35
free.sulfur.dioxide	2.00	23.00	34.00	35.31	46.00	289.00
total.sulfur.dioxide	9.00	108.00	134.00	138.40	167.00	440.00
density	0.99	0.99	0.99	0.99	1.00	1.04
pH	2.72	3.09	3.18	3.19	3.28	3.82
sulphates	0.22	0.41	0.47	0.49	0.55	1.08
alcohol	8.00	9.50	10.40	10.51	11.40	14.20

Observations

- Predictor values greater than Q3 + 1.5 IQR are considered outliers. The levels of physiochemical properties in wine deviate from average values, that gives rise to a peculiar taste of wine. Here the outliers are not due to an anomaly in data entry. Instead, the outliers represent essential values of the predictor variables.
- The correlation coefficients are not very high in red wine dataset, so there is no problem of multicollinearity. There seems to be some correlation between citric acid & volatile acidity, pH & citric acid. Further effects would be investigated by using interactions among them in the regression models.
- The correlation coefficient is unusually high between density & alcohol in the white wine dataset. Both variables were dropped from the model one at a time. However, neither the binned plots improved nor the prediction accuracy. Rest there is no multicollinearity issue in white wine dataset.

RED WINE - CORRELATION MATRIX

Predictor variables	fixed. acidity	volatile. acidity	citric. acid	residual. sugar	chlorides	free. SO2	total. SO2	density	pH	sulphates	alcohol
fixed. acidity	1.00	-0.26	0.67	0.11	0.09	-0.15	-0.11	0.67	-0.68	0.18	-0.06
volatile.acidity	-0.26	1.00	-0.55	0.00	0.06	-0.01	0.08	0.02	0.23	-0.26	-0.20
citric.acid	0.67	-0.55	1.00	0.14	0.20	-0.06	0.04	0.36	-0.54	0.31	0.11
residual.sugar	0.11	0.00	0.14	1.00	0.06	0.19	0.20	0.36	-0.09	0.01	0.04
chlorides	0.09	0.06	0.20	0.06	1.00	0.01	0.05	0.20	-0.27	0.37	-0.22
free. SO2	-0.15	-0.01	-0.06	0.19	0.01	1.00	0.67	-0.02	0.07	0.05	-0.07
total. SO2	-0.11	0.08	0.04	0.20	0.05	0.67	1.00	0.07	-0.07	0.04	-0.21
density	0.67	0.02	0.36	0.36	0.20	-0.02	0.07	1.00	-0.34	0.15	-0.50
pH	-0.68	0.23	-0.54	-0.09	-0.27	0.07	-0.07	-0.34	1.00	-0.20	0.21
sulphates	0.18	-0.26	0.31	0.01	0.37	0.05	0.04	0.15	-0.20	1.00	0.09
alcohol	-0.06	-0.20	0.11	0.04	-0.22	-0.07	-0.21	-0.50	0.21	0.09	1.00

WHITE WINE - CORRELATION MATRIX

Predictor variables	fixed. acidity	volatile. acidity	citric. acid	residual. sugar	chlorides	free. SO2	total. SO2	density	pH	sulphates	alcohol
fixed. acidity	1.00	-0.02	0.29	0.09	0.02	-0.05	0.09	0.27	-0.43	-0.02	-0.12
volatile.acidity	-0.02	1.00	-0.15	0.06	0.07	-0.10	0.09	0.03	-0.03	-0.04	0.07
citric.acid	0.29	-0.15	1.00	0.09	0.11	0.09	0.12	0.15	-0.16	0.06	-0.08
residual.sugar	0.09	0.06	0.09	1.00	0.09	0.30	0.40	0.84	-0.19	-0.03	-0.45
chlorides	0.02	0.07	0.11	0.09	1.00	0.10	0.20	0.26	-0.09	0.02	-0.36
free. SO2	-0.05	-0.10	0.09	0.30	0.10	1.00	0.62	0.29	0.00	0.06	-0.25
total. SO2	0.09	0.09	0.12	0.40	0.20	0.62	1.00	0.53	0.00	0.13	-0.45
density	0.27	0.03	0.15	0.84	0.26	0.29	0.53	1.00	-0.09	0.07	-0.78
pH	-0.43	-0.03	-0.16	-0.19	-0.09	0.00	0.00	-0.09	1.00	0.16	0.12
sulphates	-0.02	-0.04	0.06	-0.03	0.02	0.06	0.13	0.07	0.16	1.00	-0.02
alcohol	-0.12	0.07	-0.08	-0.45	-0.36	-0.25	-0.45	-0.78	0.12	-0.02	1.00

NON PARAMETRIC MODELS

1. Two machine learning models viz. Support Vector Machines (SVM) & Random Forests are used without fine tuning any parameters.
2. The datasets were divided into training data (70%) and test data (30%).

Comparison of Prediction Accuracy of Quality Category			
	Multinomial	SVM	Random Forest
Red Wine	63.91%	65.25%	73.52%
White Wine	57.51%	53.12%	69.50%

- Random Forest Model performs better than SVM on test data.

Why not linear regression?: Though many studies have considered outcome variable (quality) as a continuous variable (ranging from 1 to 10) and applied multiple linear regression model to analyze the effect of properties on quality. There are two shortcomings with this approach:

1. Since a human tester assigns the wine quality score in discrete numbers, so it would not be appropriate to convert categorical variable to continuous variable, it would violate the testing condition itself.
2. It is challenging to rate the human perception of complex wine tastes, and it would be indifferentiable between say 5.4 and 5.5. Discretization of the score will automatically come in if the scores have to be interpretable and accepted on a universal scale by the wine industry.

CONCLUSION

1. Physiochemical properties differ by the color of the wine.
2. Quality cannot be predicted with high accuracy given the physiochemical properties.
3. Outliers cannot be removed from the dataset for a better fit of the model.
4. However, such a classification model should be used to aid the quality checking process.
5. If the interpretation is the primary goal, then parametric models viz. the multinomial regression model should be used.
6. However, if the focus is to get the highest possible prediction accuracy, then advanced deep learning models such as random forest and convolutional neural networks should be used.

FUTURE STUDY

1. Fine tune the random forest model to improve the prediction accuracy.
2. Use more datasets on different types of wine from different regions to train and test the models.
3. Explore more about outliers by understanding the relationship between different wine properties and taste. Search novel ways to deal with outliers.

REFERENCES

1. Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, Elsevier, 47(4): 547-553.
2. Ivan, Morgan (2017, January 24). Classification using Random Forest in R. Retrieved from https://en.proft.me
3. Rashid, Miadad (2015, May 17). Wine Quality Exploration. Retrieved from http://rstudio-pubs-static.s3.amazonaws.com/
4. Published on STAT 897D. Analysis of Wine Quality Data. Retrieved from https://onlinecourses.science.psu.edu/ stat857/node/223
5. Hariharan, Ashwin (2017, February 7). Game of Wines. Retrieved from https://medium.freecodecamp.org/using-data-science-to-understand-what-makes-wine-taste-good-669b496c67ee

ACKNOWLEDGEMENTS

Course Instructor: Dr. Jerry Reiter, Professor, Statistical Science, Duke University

Course Teaching Assistants: Andrew Cooper & Jingyi Zhang

Contact: Vivek Sahukar | Duke University | Masters in Interdisciplinary Data Science

Mobile: 1-984-209-6183 | Email: vivek.sahukar@duke.edu