Predicting Wine Quality

A Logistic Regression Approach using Physiochemical Tests

Intro to Data Science: Capstone Project, November 2018
Krystle Lee

Project Overview

- Objective
- Background
- Data Set Details
- Data Wrangling
- Exploratory Data Analysis
- Analysis Approach
- Model Evaluation and Metrics
- Final Takeaways

Objective: Use physiochemical tests to predict the quality of wine

- Empower winemakers to make cost effective decisions
- Give winemakers insight into their wine to make improvements
- Quality of wine can determine the price point. Better quality can lead to more revenue.
- Eliminate and reduce resource expenditures on wines that are predicted as bad quality
- Repurpose bad quality wines by creating wine blends
- Provide customers with a better wine experience
- Give winemakers sense of pride in the product they are producing

Background

- Wine is produced through the process of turning grapes into wine by fermentation. Grapes are picked off the vines, fermented into alcohol and bottled. The longer the wine ages in the bottle tends to lead to a better wine.
- There are many varieties of grapes and even more wines
- Most wines that are scored by critics are given a score based on the 50-100 scale introduced by Robert M. Parker Jr.

5 Basic characteristics of wine:

1. Sweetness
2. Acidity
3. Tannin
4. Alcohol
5. Body

Wine Quality Data Set

Data set from UCI Machine Learning Repository (Link)

Data Set Characteristics	Overall	
Type	Portuguese "Vinho Verde"	
Observations	6,497 (Red: 1,599; White: 4,898)	
Missing values	None	
Attributes	12 1 - fixed acidity 2 - volatile acidity 3 - citric acid 4 - residual sugar 5 - chlorides 6 - free sulfur dioxide 7 - total sulfur dioxide 8 - density	9 - pH 10 - sulphates 11 - alcohol Output variable (based on sensory data): 12 - quality (score between 0 and 10)
Year Provided	2009-10-07	

Data Wrangling: Prepping the Data Set

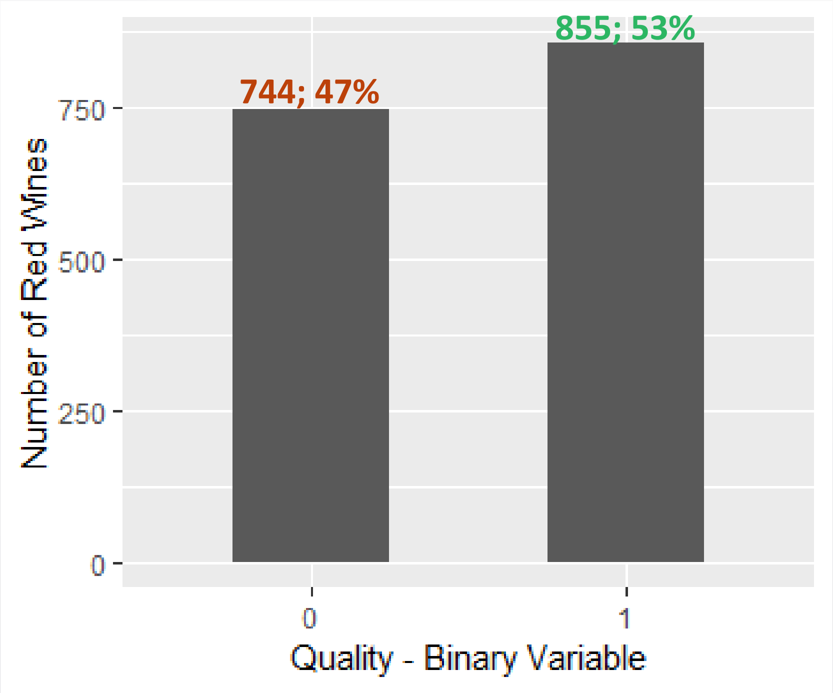
- Since no missing values in the data set, not much data wrangling was needed to be performed to get the data set ready for analysis
- Red and white wines are appreciated for their different qualities and are therefore going to be kept as separate data sets to focus on physiochemical properties specific to the wine color
- Wine quality rating was transformed into binary values to be used in a logistic regression analysis

Quality	Quality Rating	Binary Variable
Bad	0 – 5	0
Good	6 – 10	1

Exploratory Data Analysis: Taking a closer look

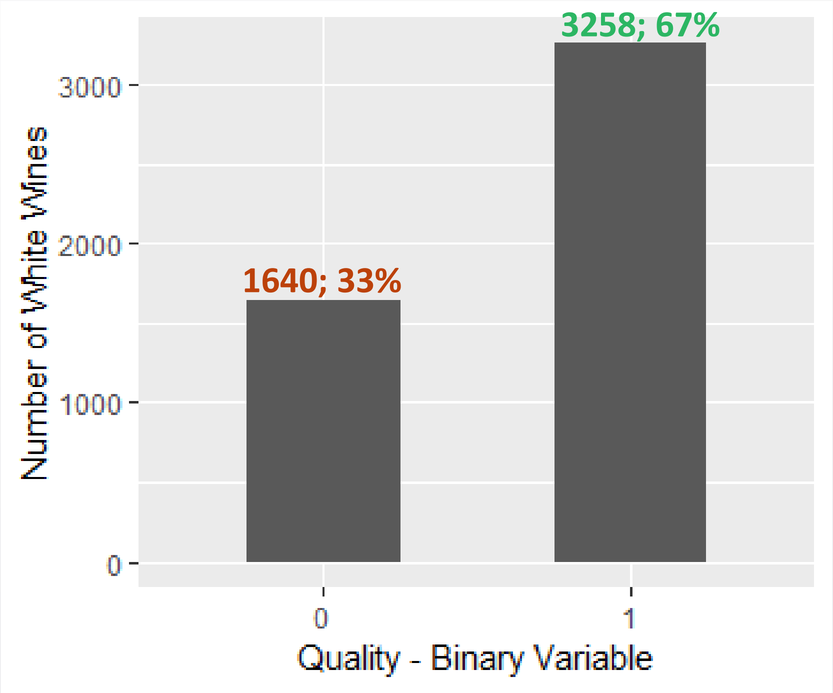
- Bar charts

Red Wines



Red wine data set is close to balanced

White Wines



White wine data set has an imbalanced of good and bad quality wines

Exploratory Data Analysis: Taking a closer look

- Structure of Data Sets

> str(red)
'data.frame': 1599 obs. of 12 variables:
$ fixed.acidity : num 7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
$ volatile.acidity : num 0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
$ citric.acid : num 0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
$ residual.sugar : num 1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
$ chlorides : num 0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
$ free.sulfur.dioxide : num 11 25 15 17 11 13 15 15 9 17 ...
$ total.sulfur.dioxide: num 34 67 54 60 34 40 59 21 18 102 ...
$ density : num 0.998 0.997 0.997 0.998 0.998 ...
$ pH : num 3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
$ sulphates : num 0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
$ alcohol : num 9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
$ binary : num 0 0 0 1 0 0 0 1 1 0 ...

> str(white)
'data.frame': 4898 obs. of 12 variables:
$ fixed.acidity : num 7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
$ volatile.acidity : num 0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
$ citric.acid : num 0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
$ residual.sugar : num 20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
$ chlorides : num 0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
$ free.sulfur.dioxide : num 45 14 30 47 47 30 30 45 14 28 ...
$ total.sulfur.dioxide: num 170 132 97 186 186 97 136 170 132 129 ...
$ density : num 1.001 0.994 0.995 0.996 0.996 ...
$ pH : num 3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
$ sulphates : num 0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
$ alcohol : num 8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
$ binary : num 1 1 1 1 1 1 1 1 1 1 ...

Exploratory Data Analysis: Taking a closer look

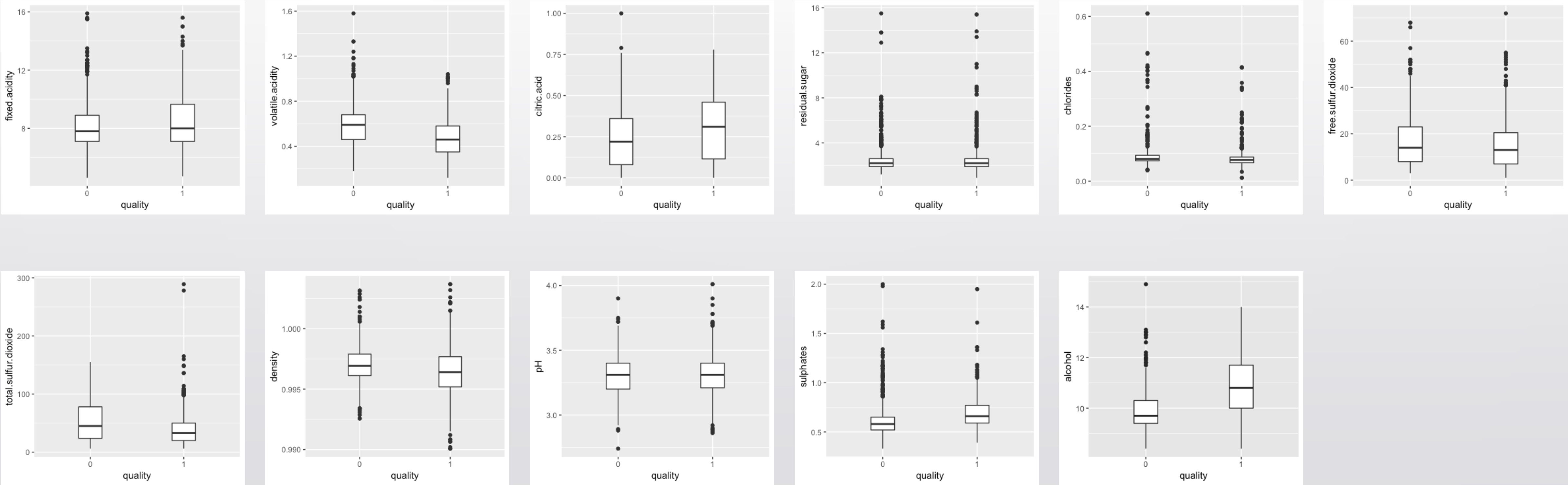
- Summary of Data Sets

Red Wine Summary	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
fixed.acidity	4.6	7.1	7.9	8.32	9.2	15.9
volatile.acidity	0.12	0.39	0.52	0.5278	0.64	1.58
citric.acid	0	0.09	0.26	0.271	0.42	1
residual.sugar	0.9	1.9	2.2	2.539	2.6	15.5
chlorides	0.012	0.07	0.079	0.08747	0.09	0.611
free.sulfur.dioxide	1	7	14	15.87	21	72
total.sulfur.dioxide	6	22	38	46.47	62	289
density	0.9901	0.9956	0.9968	0.9967	0.9978	1.0037
pH	2.74	3.21	3.31	3.311	3.4	4.01
sulphates	0.33	0.55	0.62	0.6581	0.73	2
alcohol	8.4	9.5	10.2	10.42	11.1	14.9
binary	0	0	1	0.5347	1	1

White Wine Summary	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
fixed.acidity	3.8	6.3	6.8	6.855	7.3	14.2
volatile.acidity	0.08	0.21	0.26	0.2782	0.32	1.1
citric.acid	0	0.27	0.32	0.3342	0.39	1.66
residual.sugar	0.6	1.7	5.2	6.391	9.9	65.8
chlorides	0.009	0.036	0.043	0.04577	0.05	0.346
free.sulfur.dioxide	2	23	34	35.31	46	289
total.sulfur.dioxide	9	108	134	138.4	167	440
density	0.9871	0.9917	0.9937	0.994	0.9961	1.039
pH	2.72	3.09	3.18	3.188	3.28	3.82
sulphates	0.22	0.41	0.47	0.4898	0.55	1.08
alcohol	8	9.5	10.4	10.51	11.4	14.2
binary	0	0	1	0.6652	1	1

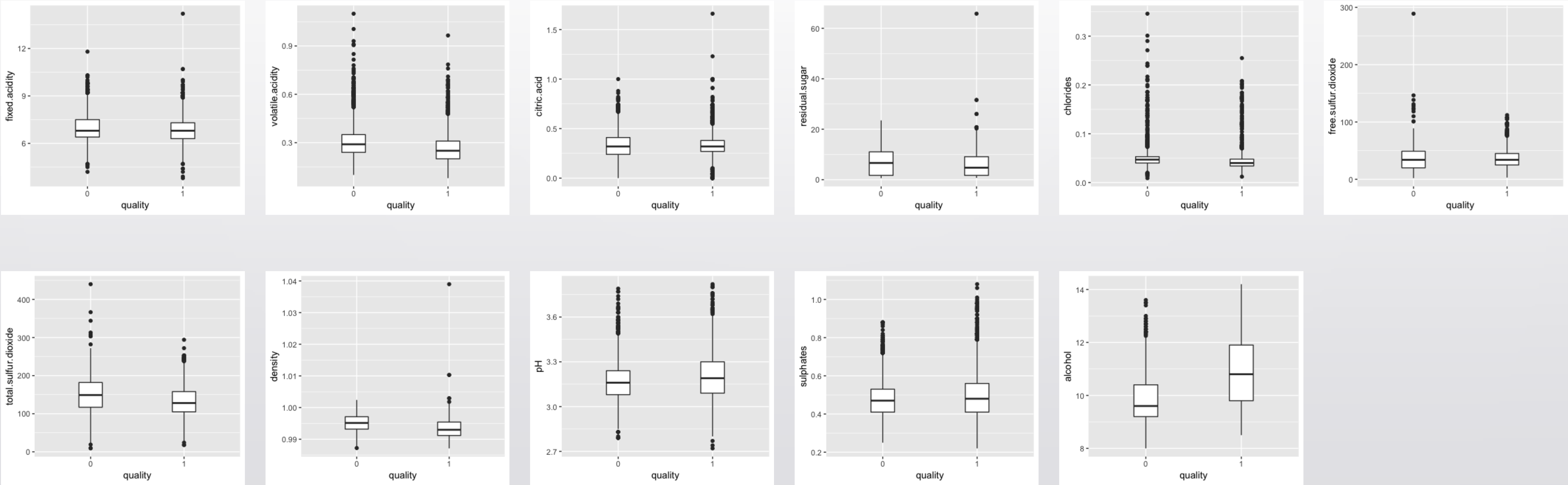
Exploratory Data Analysis: Taking a closer look

- Red Wine Box Plots



Exploratory Data Analysis: Taking a closer look

- White Wine Box Plots



Exploratory Data Analysis: Taking a closer look

- Red Wine Correlation Table

	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	binary
fixed.acidity	1.00	-0.26	0.67	0.11	0.09	-0.15	-0.11	0.67	-0.68	0.18	-0.06	0.10
volatile.acidity	-0.26	1.00	-0.55	0.00	0.06	-0.01	0.08	0.02	0.23	-0.26	-0.20	-0.32
citric.acid	0.67	-0.55	1.00	0.14	0.20	-0.06	0.04	0.36	-0.54	0.31	0.11	0.16
residual.sugar	0.11	0.00	0.14	1.00	0.06	0.19	0.20	0.36	-0.09	0.01	0.04	0.00
chlorides	0.09	0.06	0.20	0.06	1.00	0.01	0.05	0.20	-0.27	0.37	-0.22	-0.11
free.sulfur.dioxide	-0.15	-0.01	-0.06	0.19	0.01	1.00	0.67	-0.02	0.07	0.05	-0.07	-0.06
total.sulfur.dioxide	-0.11	0.08	0.04	0.20	0.05	0.67	1.00	0.07	-0.07	0.04	-0.21	-0.23
density	0.67	0.02	0.36	0.36	0.20	-0.02	0.07	1.00	-0.34	0.15	-0.50	-0.16
pH	-0.68	0.23	-0.54	-0.09	-0.27	0.07	-0.07	-0.34	1.00	-0.20	0.21	0.00
sulphates	0.18	-0.26	0.31	0.01	0.37	0.05	0.04	0.15	-0.20	1.00	0.09	0.22
alcohol	-0.06	-0.20	0.11	0.04	-0.22	-0.07	-0.21	-0.50	0.21	0.09	1.00	0.43
binary	0.10	-0.32	0.16	0.00	-0.11	-0.06	-0.23	-0.16	0.00	0.22	0.43	1.00

Exploratory Data Analysis: Taking a closer look

- White Wine Correlation Table

	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	binary
fixed.acidity	1.00	-0.02	0.29	0.09	0.02	-0.05	0.09	0.27	-0.43	-0.02	-0.12	-0.09
volatile.acidity	-0.02	1.00	-0.15	0.06	0.07	-0.10	0.09	0.03	-0.03	-0.04	0.07	-0.23
citric.acid	0.29	-0.15	1.00	0.09	0.11	0.09	0.12	0.15	-0.16	0.06	-0.08	0.00
residual.sugar	0.09	0.06	0.09	1.00	0.09	0.30	0.40	0.84	-0.19	-0.03	-0.45	-0.09
chlorides	0.02	0.07	0.11	0.09	1.00	0.10	0.20	0.26	-0.09	0.02	-0.36	-0.18
free.sulfur.dioxide	-0.05	-0.10	0.09	0.30	0.10	1.00	0.62	0.29	0.00	0.06	-0.25	0.00
total.sulfur.dioxide	0.09	0.09	0.12	0.40	0.20	0.62	1.00	0.53	0.00	0.13	-0.45	-0.17
density	0.27	0.03	0.15	0.84	0.26	0.29	0.53	1.00	-0.09	0.07	-0.78	-0.27
pH	-0.43	-0.03	-0.16	-0.19	-0.09	0.00	0.00	-0.09	1.00	0.16	0.12	0.08
sulphates	-0.02	-0.04	0.06	-0.03	0.02	0.06	0.13	0.07	0.16	1.00	-0.02	0.05
alcohol	-0.12	0.07	-0.08	-0.45	-0.36	-0.25	-0.45	-0.78	0.12	-0.02	1.00	0.38
binary	-0.09	-0.23	0.00	-0.09	-0.18	0.00	-0.17	-0.27	0.08	0.05	0.38	1.00

Analysis approach: Training and Testing Sets

- Separating the data set into a training and testing set provides a way to evaluate the performance of the model when new observations are introduced
- If the model is over fitted, it will perform well with the training set and poorly with the testing set.
- For each wine color, the data was randomly split by the number of observations into a training (67%) and testing (33%)
- Split method:
  - sample.split from caTools package

	Training Set (67%)	Testing Set (33%)	Total Wines
Red	1,071	528	1,599
White	3,282	1,616	4,898

Analysis approach: glm model

Logistic Regression – All variables

- With each training set, created a glm (generalized linear model) using all variables in the data set
- Red:
  - glm(formula = binary ~ ., family = binomial, data = redrain)
- White:
  - glm(formula = binary ~ ., family = binomial, data = whiteTrain)

Logistic Regression with Stepwise Method

- Incorporated stepwise method to help identify the independent variables that attributed to a high performing model
- Red:
  - glm(formula = binary ~ fixed.acidity + volatile.acidity + chlorides + free.sulfur.dioxide + total.sulfur.dioxide + sulphates + alcohol, family = binomial, data = redTrain)
- White:
  - glm(formula = binary ~ volatile.acidity + residual.sugar + free.sulfur.dioxide + density + pH + sulphates + alcohol, family = binomial, data = whiteTrain)

Analysis approach (Red): Investigate glm summary

- Analyzed the model using the summary function
- Took note of each models AIC score and moved forward with the lowest

RED GLM with all variables

```
Call:
glm(formula = binary ~ ., family = binomial, data = redTrain)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4731  -0.8624   0.3159   0.8308   2.1643
Coefficients:
                      Estimate   Std. Error   z value    Pr(>|z|)
(Intercept)           1.29E+02     9.56E+01     1.351      0.1768
fixed.acidity         2.04E-01     1.21E-01     1.691      0.09085 .
volatile.acidity     -2.46E+00     5.58E-01    -4.407      1.05E-05 ***
citric.acid          -6.19E-01     6.61E-01    -0.936      0.34909
residual.sugar        7.36E-02     6.73E-02     1.095      0.2736
chlorides            -3.54E+00     1.91E+00    -1.847      0.06468 .
free.sulfur.dioxide   1.31E-02     9.93E-03     1.315      0.18846
total.sulfur.dioxide -1.21E-02     3.35E-03    -3.599      0.00032 ***
density              -1.40E+02     9.76E+01    -1.435      0.15136
pH                    2.08E-01     8.70E-01     0.239      0.81148
sulphates             2.76E+00     5.52E-01     5.001      5.71E-07 ***
alcohol               8.11E-01     1.25E-01     6.508      7.64E-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 1479.5  on 1070  degrees of freedom
Residual deviance: 1127.7  on 1059  degrees of freedom
AIC: 1151.7
Number of Fisher Scoring iterations: 4
```

RED GLM with stepwise method

```
Call:
glm(formula = binary ~ fixed.acidity + volatile.acidity + chlorides +
    free.sulfur.dioxide + total.sulfur.dioxide + sulphates +
    alcohol, family = binomial, data = redTrain)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2958  -0.8542   0.3242   0.8383   2.2209
Coefficients:
                      Estimate    Std. Error   z value   Pr(>|z|)
(Intercept)          -9.608401    1.098417     -8.747    < 2e-16 ***
fixed.acidity         0.063392    0.044532      1.424      0.155
volatile.acidity     -2.330716    0.453931     -5.135    2.83E-07 ***
chlorides            -3.925939    1.72687      -2.273      0.023 *
free.sulfur.dioxide   0.014464    0.009622      1.503      0.133
total.sulfur.dioxide -0.012346    0.003021     -4.086    4.39E-05 ***
sulphates             2.536938    0.526708      4.817    1.46E-06 ***
alcohol               0.917475    0.085443     10.738    < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 1479.5  on 1070  degrees of freedom
Residual deviance: 1131.1  on 1063  degrees of freedom
AIC: 1147.1
Number of Fisher Scoring iterations: 4
```

Move forward with GLM using stepwise method due to lower AIC value

Analysis approach (White): Investigate glm summary

- Analyzed the model using the summary function
- Took note of each models AIC score and moved forward with the lowest

WHITE GLM with all variables

Call:
glm(formula = binary ~ ., family = binomial, data = whiteTrain)
Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.6902	-0.8957	0.4440	0.7977	2.5535

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	3.91E+02	9.08E+01	4.3	1.71E-05	***
fixed.acidity	7.80E-02	8.89E-02	0.878	3.80E-01	
volatile.acidity	-6.39E+00	5.13E-01	-12.439	<2.00E-16	***
citric.acid	3.77E-01	3.65E-01	1.034	3.01E-01	
residual.sugar	2.06E-01	3.41E-02	6.036	1.58E-09	***
chlorides	9.27E-01	2.06E+00	0.451	6.52E-01	
free.sulfur.dioxide	8.97E-03	3.43E-03	2.616	0.00889	**
total.sulfur.dioxide	-1.33E-03	1.48E-03	-0.897	3.70E-01	
density	-4.04E+02	9.20E+01	-4.385	1.16E-05	***
pH	1.34E+00	4.48E-01	2.993	0.00277	**
sulphates	2.32E+00	4.47E-01	5.192	2.08E-07	***
alcohol	5.43E-01	1.18E-01	4.598	4.26E-06	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 4185.0 on 3281 degrees of freedom
Residual deviance: 3309.2 on 3270 degrees of freedom
AIC: 3333.2
Number of Fisher Scoring iterations: 5

WHITE GLM with stepwise method

Call:
glm(formula = binary ~ volatile.acidity + residual.sugar + free.sulfur.dioxide + density + pH + sulphates + alcohol, family = binomial, data = whiteTrain)
Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.6893	-0.9029	0.4432	0.8029	2.5434

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	3.29E+02	5.59E+01	5.873	4.28E-09	***
volatile.acidity	-6.60E+00	4.95E-01	-13.337	<2.00E-16	***
residual.sugar	1.82E-01	2.24E-02	8.117	4.78E-16	***
free.sulfur.dioxide	7.35E-03	2.74E-03	2.679	0.00739	**
density	-3.40E+02	5.60E+01	-6.067	1.30E-09	***
pH	9.67E-01	3.13E-01	3.093	0.00198	**
sulphates	2.20E+00	4.34E-01	5.074	3.90E-07	***
alcohol	6.25E-01	8.23E-02	7.593	3.13E-14	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 4185.0 on 3281 degrees of freedom
Residual deviance: 3312.4 on 3274 degrees of freedom
AIC: 3328.4
Number of Fisher Scoring iterations: 5

Move forward with GLM using stepwise method due to lower AIC value

Model Evaluation and Metrics (Red Training Set): Predicted values, average probabilities, confusion matrix, ROCR, AUC

- Created prediction model and identified average probability of bad and good quality wines

Range of predicted probabilities

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.02852	0.29656	0.51166	0.53501	0.79278	0.99562

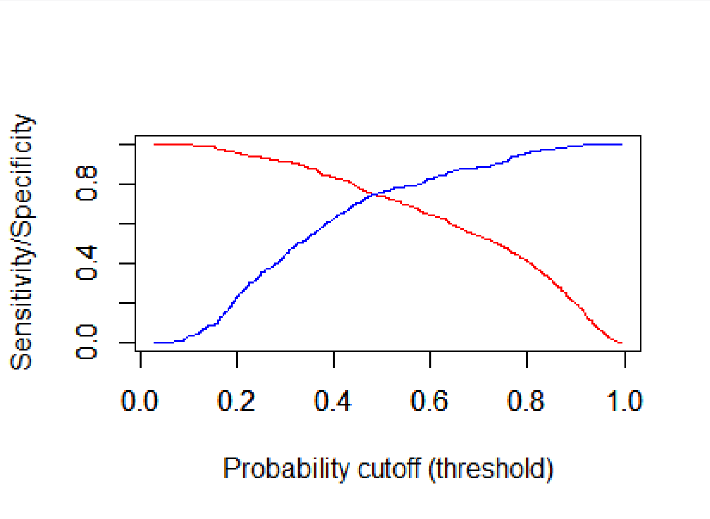
Average predicted probabilities of each quality

0	1
0.378402	0.671127

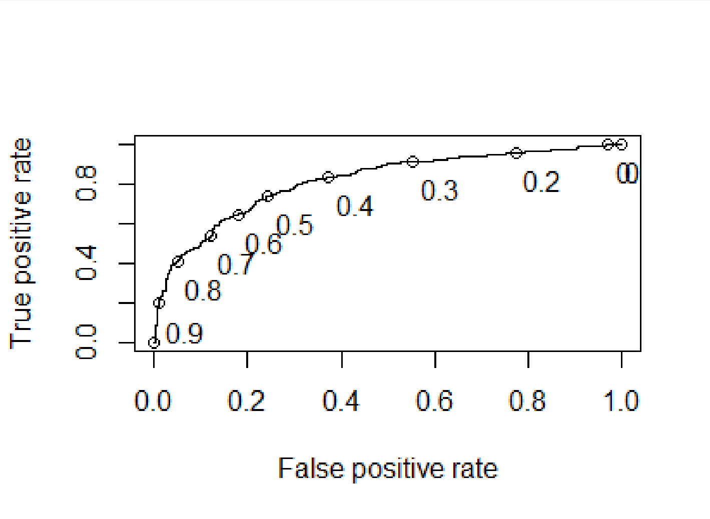
Confusion matrix using >50% as threshold value

	FALSE	TRUE
0	377	121
1	150	423
Sensitivity	74%	
Specificity	76%	
Accuracy	75%	

Sensitivity (red line) vs Specificity (blue line)



Area Under the Curve (AUC) = 81.4%



Model Evaluation and Metrics (Red Testing Set): How does training model do against testing data set?

- Apply prediction model from training set to testing to see how well it does on new observations

Range of predicted probabilities

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.02688	0.28701	0.51637	0.52566	0.77709	0.98596

Average predicted probabilities of each quality

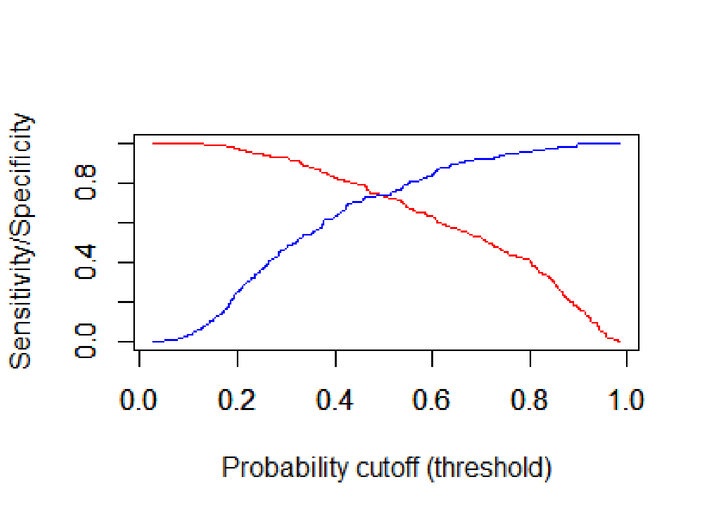
0	1
0.3650934	0.6657357

Confusion matrix using >50% as threshold value

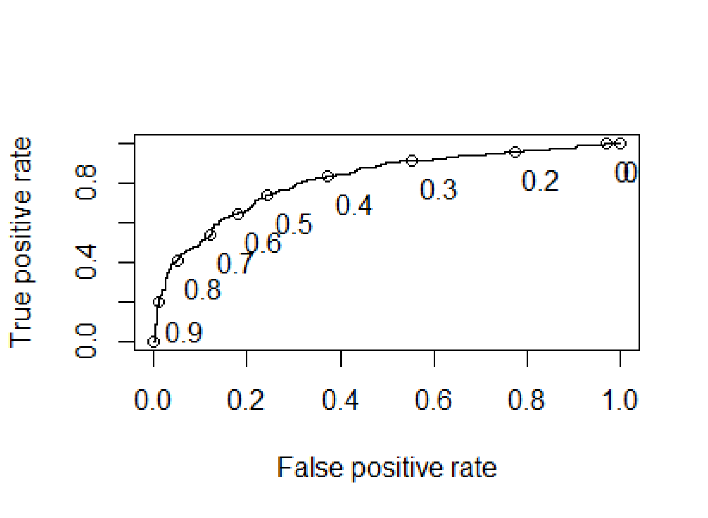
	FALSE	TRUE
0	181	65
1	71	211
Sensitivity	73%	
Specificity	74%	
Accuracy	73%	

All rates went down when applying test set but by less than 2%

Sensitivity (red line) vs Specificity (blue line)



Area Under the Curve (AUC) = 82.8%



Model Evaluation and Metrics (White Training Set): Predicted values, average probabilities, confusion matrix, ROCR, AUC

- Created prediction model and identified average probability of bad and good quality wines

Range of predicted probabilities

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.002416	0.502754	0.705134	0.665143	0.86871	0.992237

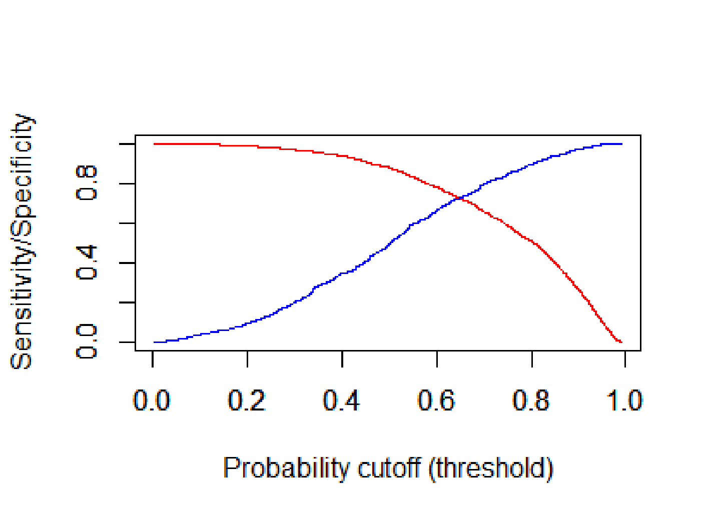
Average predicted probabilities of each quality

0	1
0.5011334	0.7477116

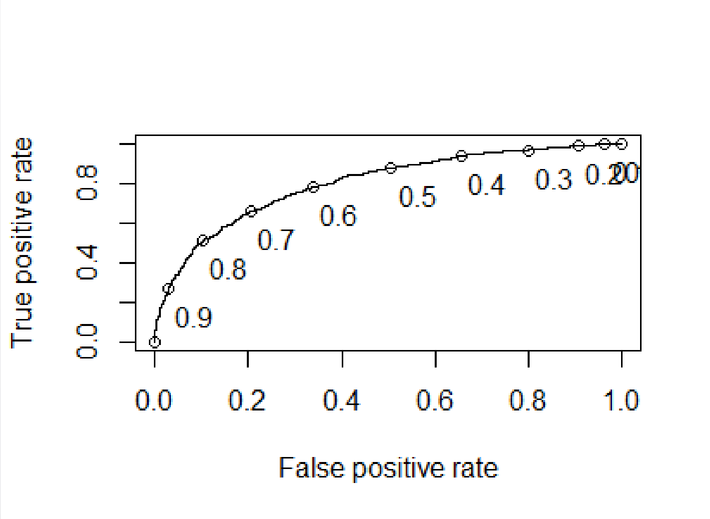
Confusion matrix using >50% as threshold value

	FALSE	TRUE
0	543	556
1	267	1916
Sensitivity	88%	
Specificity	49%	
Accuracy	75%	

Sensitivity (red line) vs Specificity (blue line)



Area Under the Curve (AUC) = 80%



Model Evaluation and Metrics (White Testing Set): How does training model do against testing data set?

- Apply prediction model from training set to testing to see how well it does on new observations

Range of predicted probabilities

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0009725	0.5063071	0.71176	0.6657905	0.8660665	0.9926941

Average predicted probabilities of each quality

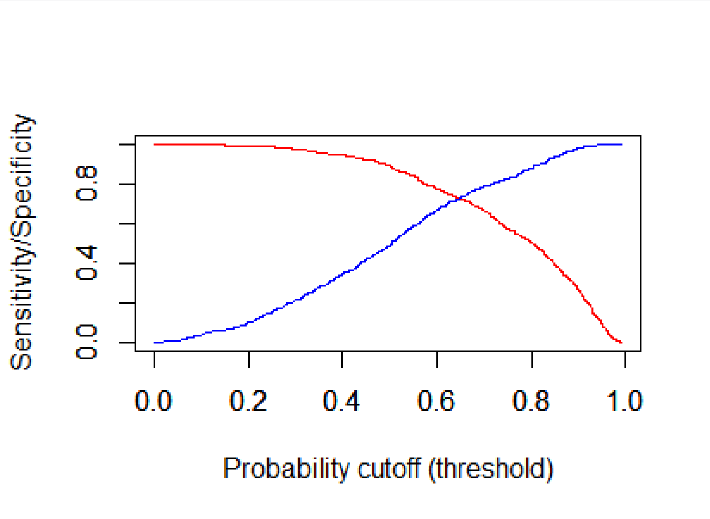
0	1
0.4997906	0.7493309

Confusion matrix using >50% as threshold value

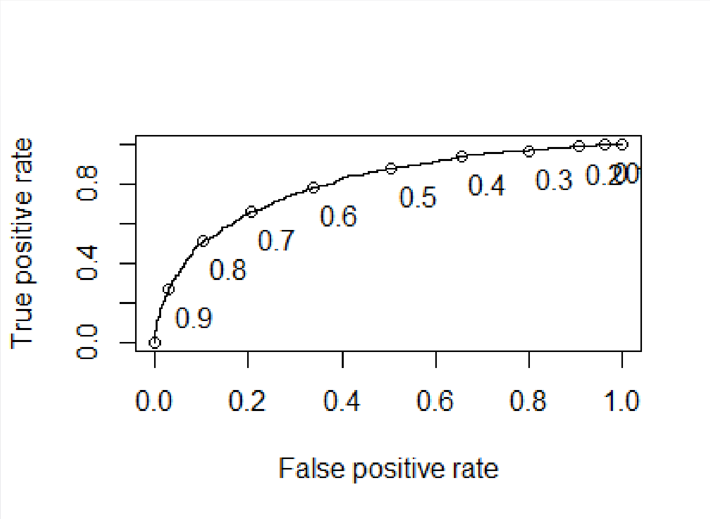
	FALSE	TRUE
0	264	277
1	118	957
Sensitivity	89%	
Specificity	49%	
Accuracy	76%	

Sensitivity and accuracy rates improved slightly when applying test set

Sensitivity (red line) vs Specificity (blue line)

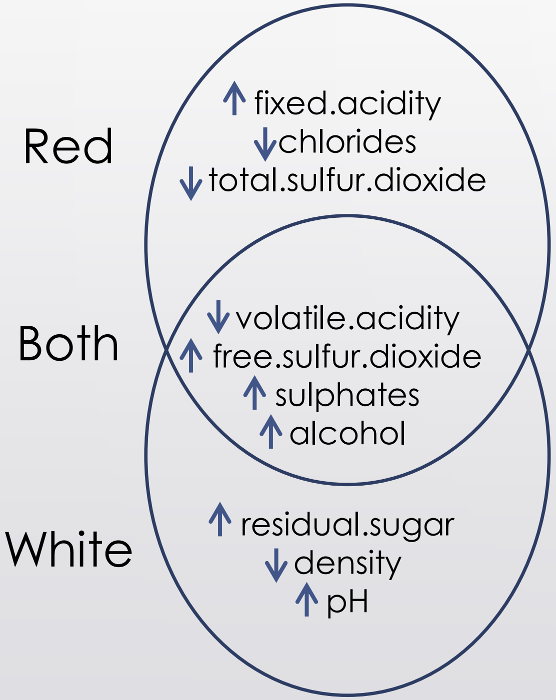


Area Under the Curve (AUC) = 80.2%



Final Takeaways

Good quality wines tend to show below physiochemical values



	Red Wine	White Wine
(Intercept)	-9.608401	3.29E+02
fixed.acidity	0.063392	
volatile.acidity	-2.330716	-6.60E+00
citric.acid		
residual.sugar		1.82E-01
chlorides	-3.925939	
free.sulfur.dioxide	0.014464	7.35E-03
total.sulfur.dioxide	-0.012346	
density		-3.40E+02
pH		9.67E-01
sulphates	2.536938	2.20E+00
alcohol	0.917475	6.25E-01

- Moderation and balance is key in creating good quality wines
- The physiochemical properties that have the most significant impact on wine quality tie into the basic characteristics of wine:
    1. Sweetness (White)
    2. Acidity(Both)
    3. Tannin (Red)
    4. Alcohol (Both)
    5. Body (Both)
- With this prediction model, winemakers will have a better understanding of their wines and produce better quality wines