Analysis of Wine Quality

Bethany Austhof and Adam Pratt

Erdös Institute Data Science Boot Camp

Table of Contents

Table of Contents

Data Set

- Data set of 12 variables, 11 objective variables and a quality score
- Containing 6497 observations of wine, 4898 white and 1599 red

Variables

- **Fixed Acidity**: Acidity naturally occurring in grapes
- **Volatile Acidity**: Acidity created in the fermentation process
- **Citric Acid**: Found in grapes, but can also be added to stave bacteria
- **Residual Sugar**: Sugar naturally occurring grapes that remains after the fermentation process
- **Chlorides**: Gives wine it's salinity which enhances sweet flavors and dulls acidity
- **Free Sulfur Dioxide**: Prevent bacterial growth

Variables

- **Total Sulfur Dioxide**: Free sulfur dioxide plus sulfur dioxide bound to other chemicals. Added sulfur dioxide reduces aromas.
- **Density**: Concentration of alcohol, sugar and glycerol
- **pH**: Acidity of wine
- **Sulphates**: Can be added to preserve freshness
- **Alcohol**: Percentage of alcohol by volume
- **Quality**: Ranked 1-10 with 10 being the best

Table of Contents

Summary of Data: Red Wine

Variable	Min	Mean	Max
Fixed Acidity	4.600	8.3200	15.900
Volatile Acidity	0.1200	0.5278	1.5800
Citric Acid	0.0000	0.2710	1.0000
Residual Sugar	0.9000	2.539	15.5000
Chlorides	0.0120	0.08747	0.6110
Free Sulfur Dioxide	1.0000	15.8700	72.0000
Total Sulfur Dioxide	6.0000	46.4700	289.0000
Density	0.9901	0.9967	1.0037
pH	2.7400	3.311	4.0100
Sulphates	0.3300	0.6581	2.0000
Alcohol	8.4000	10.4200	14.9000
Quality	3.0000	5.636	8.0000

Table: Summary of Red Wine Features

Summary of Data: White Wine

Variable	Min	Mean	Max
Fixed Acidity	3.800	6.855	14.200
Volatile Acidity	0.0800	0.2782	1.1000
Citric Acid	0.0000	0.3200	1.1000
Residual Sugar	0.6000	6.391	65.8000
Chlorides	0.0090	0.04577	0.3460
Free Sulfur Dioxide	2.0000	35.3100	289.0000
Total Sulfur Dioxide	9.0000	138.400	440.0000
Density	0.9871	0.9940	1.0390
pH	2.720	3.188	3.820
Sulphates	0.2200	0.4898	1.0800
Alcohol	8.00	10.51	14.20
Quality	3.000	5.878	9.000

Table: Summary of White Wine Features

Histogram

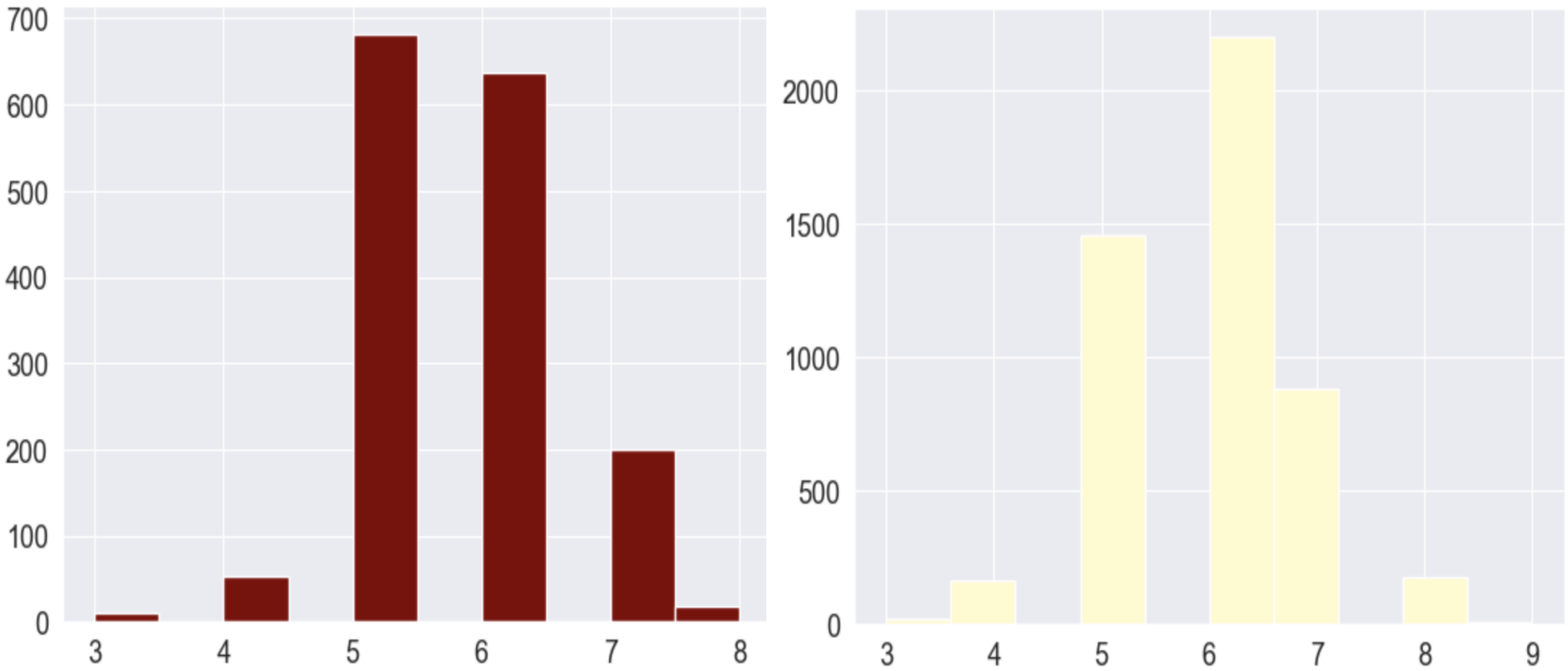

Figure: Histogram of Quality Scores for Red (left) and White (right) Wines

Table of Contents

The Model

- For both red and white wine data sets we randomly sampled 80% of the data to be training data and 20% to be testing data
- We had the 11 predictor variables and dependent variable quality score
- We used a Gaussian regression model to fit the data, and then using predict we analyzed the accuracy of our predictions

Significant Factors

Variable	Red Wine Coefficient	White Wine Coefficient
Intercept	3.5983	2.9219
Fixed Acidity	0.0174	-0.0219
Volatile Acidity	-0.6964	-1.4588
Citric Acid	-0.0667	-0.1226
Residual Sugar	0.0202	0.0160
Chlorides	-1.5925	-0.8385
Free Sulfur Dioxide	0.0041	0.0031
Total Sulfur Dioxide	-0.0040	-0.0010
Density	0	0
pH	-0.2113	0.0087
Sulphates	0.7691	0.4373
Alcohol	0.2607	0.2874

Table: Coefficients for Regression Models Predicting Quality Score

Performance Review: Red Wine

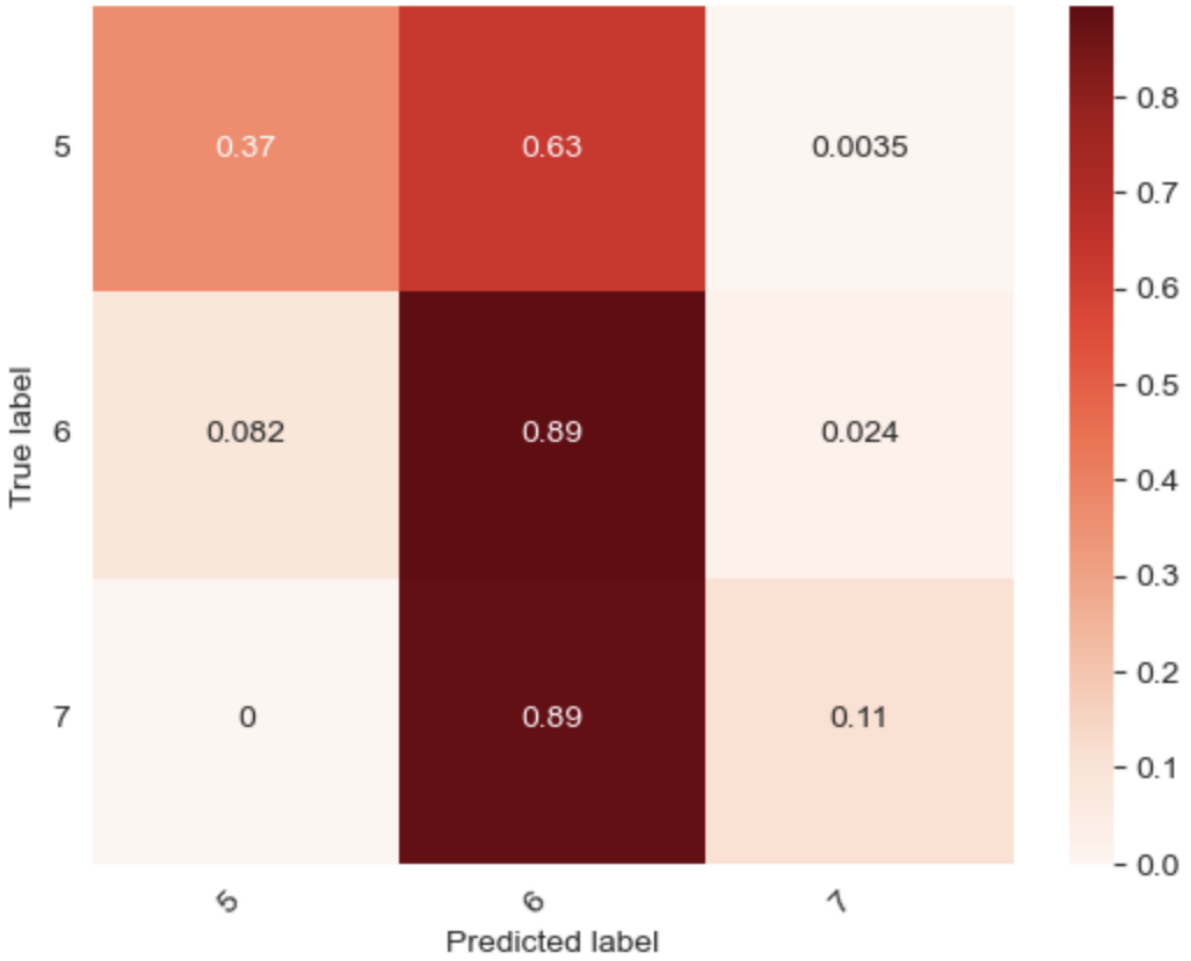

Figure: Red Wine Quality Score Regression Model Results (Percentage Correct)

Performance Review: White Wine

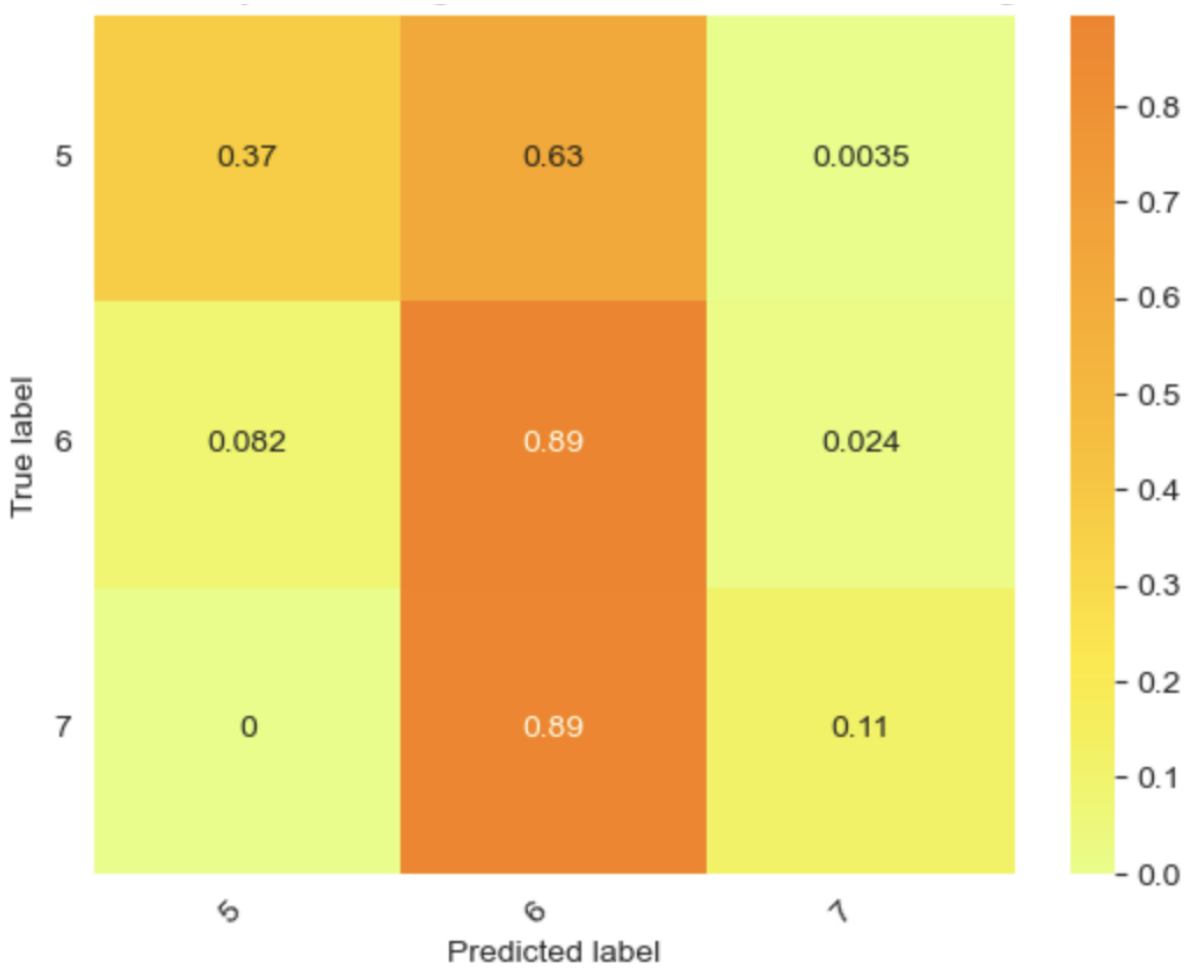

Figure: White Wine Quality Score Regression Model Results (Percentage Correct)

Problems of Analysis

- When initially running the model the most appropriate distribution to use seemed to be a Multinomial
- However, with very few data observations for high and low quality scores, we didn't get a reliable model
- Red Wine Model Prediction Accuracy: 63.1%
- White Wine Model Prediction Accuracy: 58.4%
- We intend to look further into adjusting our model to see if we can increase accuracy