Linear Regression Models Predicting Wine Quality Based on Physio-chemical Predictors

Connor Kordes - JCK160130
Nathan Tompkins - NAT180002
University of Texas at Dallas
CS 4372

1 Project and Dataset Selection

Dataset link: https://archive.ics.uci.edu/ml/datasets/Wine+Quality

As this project was a regression project, the group chose the Wine Quality dataset in the UCI ML repository. Beyond this, there were two files – one for red wine and one for white wine – and for simplicity the group isolated the data to solely red wines. In the red wine data set there were 1599 observations. Each observation has the following predictors – descriptions are included as they are not in the UCI description:

- Fixed Acids - Acids in the wine that do not rapidly evaporate, such as: tartaric, citric, malic, and, succinic
- Volatile Acids - Acids in the wine that rapidly evaporate. The main volatile acid is acetic acid, which has the smell and taste of vinegar. Ethyl acetate is also a volatile acid in wine that smells like nail polish.
- Citric Acid - A fixed acid in wine that can give freshness to the wine.
- Residual Sugar - The amount of sugar that remains after the fermentation process. Wines with more than 45 grams/liter are sweet wines.
- Chlorides - The amount of salt in the wine
- Free Sulfur Dioxide - The amount of Free Sulfur in the wine. This prevents microbial growth and oxidation of the wine.
- Total Sulfur Dioxide - The amount of free and bound sulfur in the wine.
- Density - The density of the wine. This value should be close to the density of water, but can be changed depending on the alcohol/sugar content.
- PH - The value corresponding to the acidity or basicity of the wine. Most wines are between 3-4.
- Sulphates- An additive to the wine which can increase Sulfur Dioxide levels.
- Alcohol- The percent alcohol content in the wine.

The data sets' target variable is Quality. A score that is an integer number between one and ten. Moreover, this data set could be used in a multiclass classification scenario with 10 levels, but can also be used for Regression as the target variable, Quality, has a quantitative meaning. In other words each level is discernible to a regressor compared to other scenarios where the target is something like color and encoding the different levels has no meaningful quantitative interpretation.

2 Regression Model Building

2.1 Pre-Processing

The data was uploaded to a public github repository here under the Linear Regression directory https://github.com/M0nster5/CS4372-Projects. From this, all that was necessary was the *read_csv* function and the RAW url for the dataset specifying the separator as a ";".

Once the dataset was properly converted to a data frame the group looked for null values within each column of the dataframe. The results are shown below.

```
fixed acidity           0
volatile acidity        0
citric acid             0
residual sugar          0
chlorides               0
free sulfur dioxide     0
total sulfur dioxide    0
density                 0
pH                      0
sulphates               0
alcohol                 0
quality                 0
```

Furthermore, the data was free of null values.

To further analyze, the group was interested in looking at the type of each attribute and wanted to ensure that all the data would be interpretable by a linear regression model. Additionally, it is important to note that with this dataset every column (predictors and target) should have a quantitative type. The results of the types are below.

```
fixed acidity           float64
volatile acidity        float64
citric acid             float64
residual sugar          float64
chlorides               float64
free sulfur dioxide     float64
total sulfur dioxide    float64
density                 float64
pH                      float64
sulphates               float64
alcohol                 float64
quality                   int64
```

From this the group saw that every attribute had a numeric type and was properly interpreted by the *read_csv* function for a regression task.

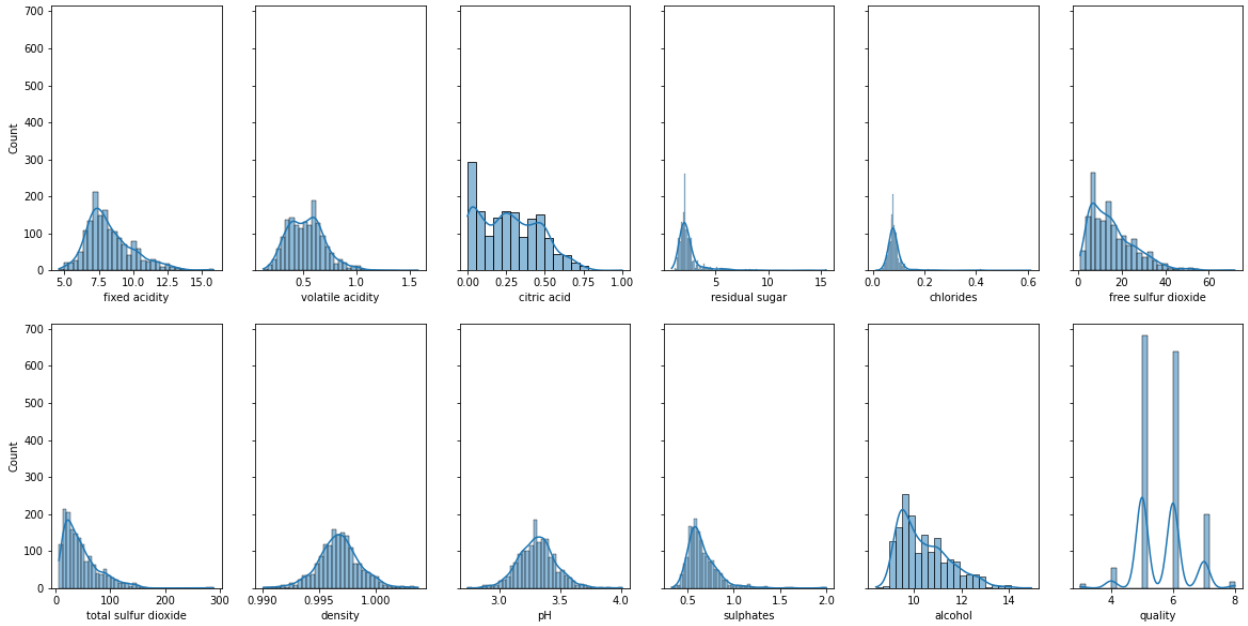
Next, the group wanted a statistical summary of the attributes. This is possible using the *df.describe* method. The output is shown below.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000
mean	8.319637	0.527821	0.270976	2.538806	0.087467	15.874922	46.467792	0.996747	3.311113	0.658149	10.422983	5.636023
std	1.741096	0.179060	0.194801	1.409928	0.047065	10.460157	32.895324	0.001887	0.154386	0.169507	1.065668	0.807569
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6.000000	0.990070	2.740000	0.330000	8.400000	3.000000
25%	7.100000	0.390000	0.090000	1.900000	0.070000	7.000000	22.000000	0.995600	3.210000	0.550000	9.500000	5.000000
50%	7.900000	0.520000	0.260000	2.200000	0.079000	14.000000	38.000000	0.996750	3.310000	0.620000	10.200000	6.000000
75%	9.200000	0.640000	0.420000	2.600000	0.090000	21.000000	62.000000	0.997835	3.400000	0.730000	11.100000	6.000000
max	15.900000	1.580000	1.000000	15.500000	0.611000	72.000000	289.000000	1.003690	4.010000	2.000000	14.900000	8.000000

Some interesting takeaways from this statistical output is that the quality, the target, which should range from 1-19 has a minimum of 3 and a maximum of 8. Additionally, the alcohol

percentage has a tighter range from eight to fourteen (typical of red wine). After looking at the other variables, their statistics seemed consistent with red wines.

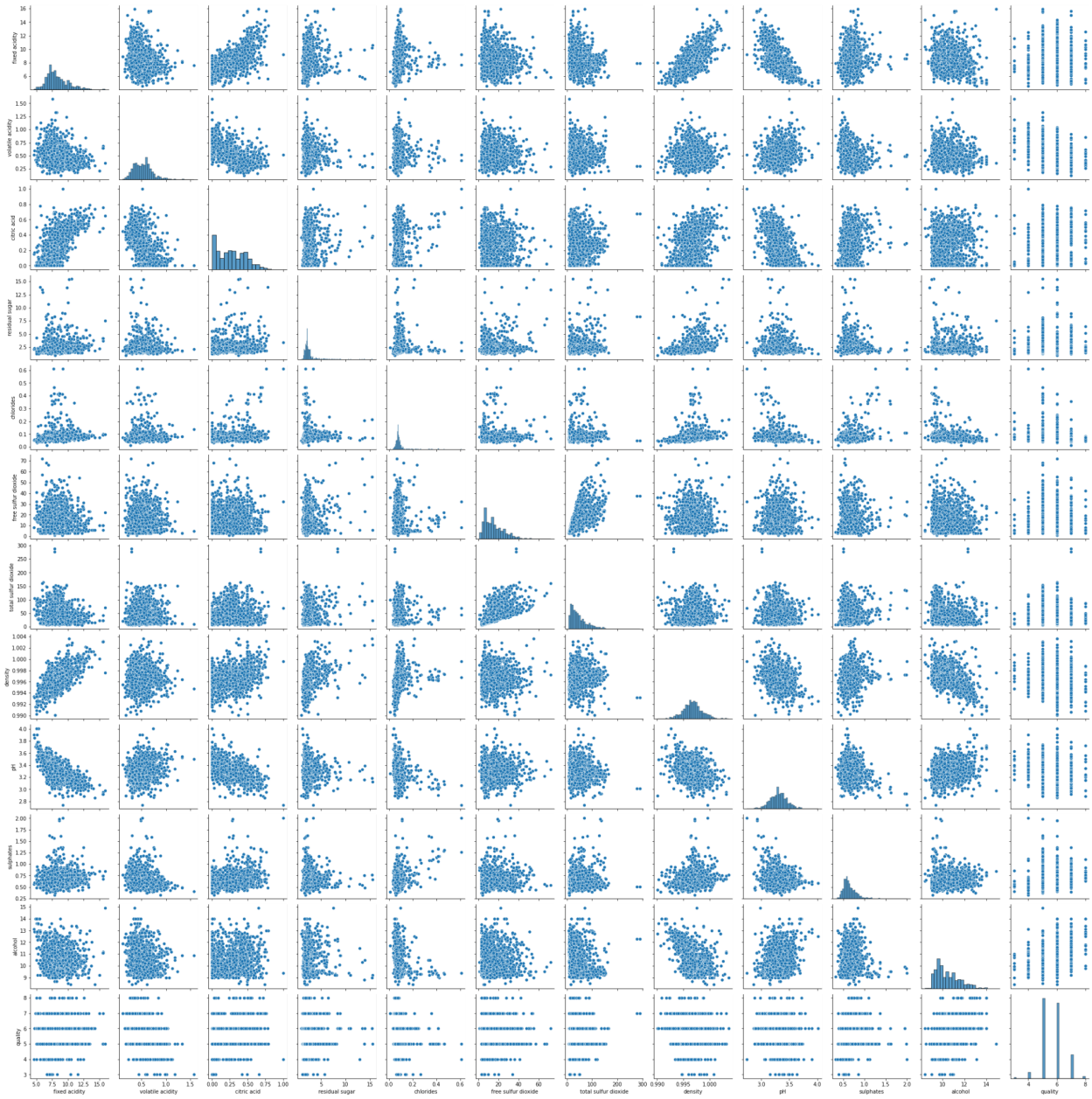
Beyond the mere statistics, the group wanted to get a visual sense of the distributions of the different predictors and the target. They created a histogram for each predictor and graphed them together below.



Starting at the top left we can see fixed acidity is slightly right skewed, likely due to the fact that red wines are typically less acidic. Similarly for citric acid is largely right skewed. Since this attribute has to do with the freshness of the wine the first bin is likely composed of higher quality wines which we can see are abundant from the quality distribution. Finally, free sulfur dioxide is right skewed, likely due to the fact that there are more “quality” wines in this dataset. Most other distributions are normal with the exception being alcohol. Alcohol is typically in this distribution for red wine.

Once this initial analysis was done. The attributes were standardized to preserve their original distributions. This was accomplished through the use of a MinMax scaler. With this, exploratory graphs were still done on the initial data set to improve interpretability of quality scores.

After standardizing the attributes (including the target), the group performed an exploratory analysis on the data to see how different attributes related to the target. To start this exploration, the group created a pairplot of all of the attributes and targets shown below.



With this, the group discovered possible relationships within the predictors listed below

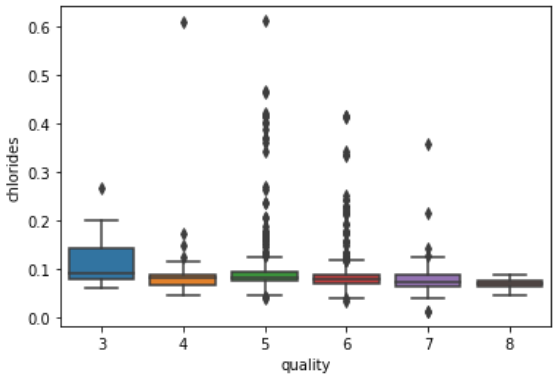
- Fixed Acids vs Volatile Acids
- Citric Acid vs Fixed Acids
- Free Sulfur Dioxide vs Total Sulfur Dioxide
- Sulphates vs Free Sulfur Dioxide
- Sulphates vs Total Sulfur Dioxide
- Density vs Sugar
- Density vs Alcohol
- PH vs Fixed Acids
- PH vs Volatile Acids

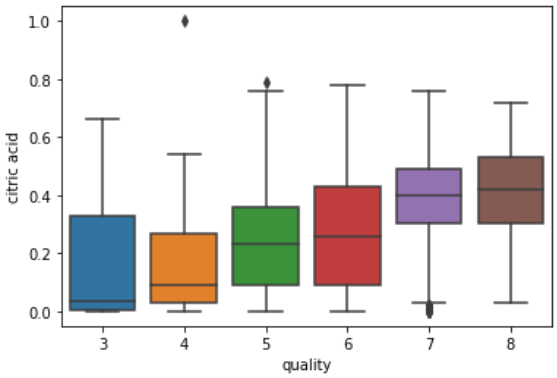
- PH vs Citric Acids

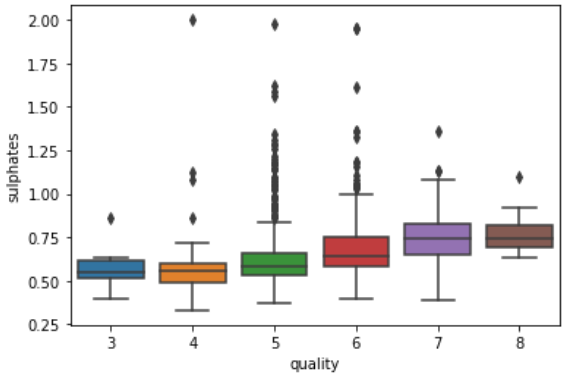
Looking at predictors and the target the group saw the following relationships:

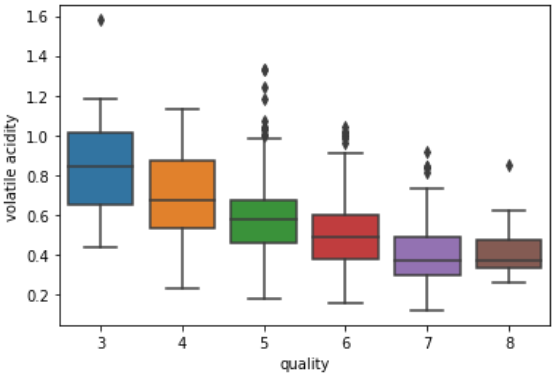
- Alcohol vs Quality
- Volatile Acidity vs Quality
- Sulfates vs Quality
- Citric Acid vs Quality
- Chlorides vs Quality

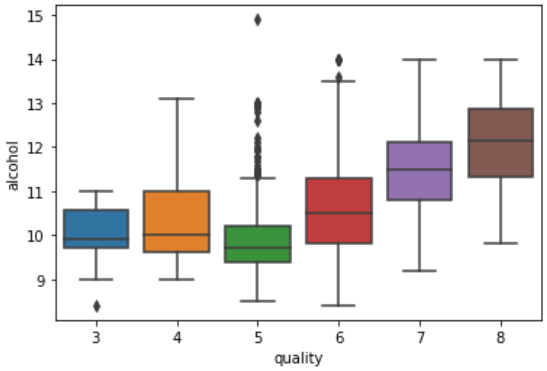
The group was unsatisfied with solely doing a pairplot to explore the data. Additionally they didn't want to include interaction terms in the model so they further analyzed the above predictor target relationships. To do this, the group created boxplots against each predictor and the target quality. This was possible as quality takes on integer values. These plots are included below.











Furthermore, after this exploratory analysis, several attributes were selected for linear regression and the rest were omitted from the model. These attributes are:

- alcohol
- volatile acidity
- sulphates
- citric acid
- chlorides

From here all that was necessary before model construction was separating the data in to train and test subsets. The test subset was randomly generated using the scikit-learn *train_test_split* method and was chosen to be 20% of the entire data set. The train subset was composed of all values except those in the test subset consisting of 80% of the data. This concludes the preprocessing section.

2.2 Model Construction

This section will cover multiple linear regression models using both stochastic gradient descent (SGD) and ordinary least squares (OLS).

2.2.1 SGD

Performing stochastic gradient descent with the scikit-learn library allows for a multitude of parameters. Each of these parameters must be tuned to obtain the best model for the test data. The group decided to look at each of the following parameters on the specified ranges

- `'alpha': [0.00001, 0.0001, .001, .01]`
- `'max_iter': [10000, 5000, 1000, 500]`
- `'learning_rate': ['constant', 'optimal', 'invscaling']`

The *alpha* parameter specifies the step size in stochastic gradient descent. In layman's terms it is how drastically the optimizer will step in the direction of the gradient. On the other hand, *max_iter* is the max iterations that will be run on the training data – with each iteration contributing to one step in the direction of the gradient. Finally, *learning_rate*, specifies how the learning rate will change within iterations. The three options have the following definitions

- *constant*: `eta = eta0`
- *optimal*: `eta = 1.0 / (alpha * (t + t0))` where t0 is chosen by a heuristic proposed by Leon Bottou.
- *invscaling*: `eta = eta0 / pow(t, power_t)`

Now that the hyperparameters are understood, the group had to create a way to look at every possible combination of these parameters or their cartesian product. To do this, the group utilized the *ParameterGrid* function from sklearn. From here the group looped through the parameter

grid and passed in the arguments to a *SGDRegressor* model. The best model's output is displayed below along with its hyperparameters.

```
{'alpha': 0.0001, 'learning_rate': 'optimal', 'max_iter': 5000}
Train Statistics:
Mean Absolute Error: 0.10636864305950068
Mean Squared Error: 0.018119332392685646
R2 Score: 0.30572050516479676

Test Statistics
Mean Absolute Error: 0.10148032860332515
Mean Squared Error: 0.01667540511980989
R2 Score: 0.35762490246462597
```

From this we can see that the MSE is 0.01668 meaning on average the model is off by about .016 units, however it is important to remember that the data is scaled so this is actually a pretty large MSE. Additionally, The R2 score is 0.3576 meaning there is a large amount of variance in the data that is not explained by the model. Furthermore, this dataset does not appear to follow the linear assumption and might be better with a lower bias model. Finally the equation of the model is shown below.

$$\hat{Y} = 0.52 + 0.46(alc) - 0.32(va) + 0.21(s) - 0.001(ca) - 0.26(chl)$$

2.2.2 OLS

Now rather than SGD regression we will be performing Ordinary Least Squares regression with the *stats_models* package. We do not have to tune any hyperparameters for OLS, so the results are below.

OLS Regression Results

Dep. Variable:	quality	R-squared:	0.336
Model:	OLS	Adj. R-squared:	0.334
Method:	Least Squares	F-statistic:	129.0
Date:	Mon, 19 Sep 2022	Prob (F-statistic):	1.23e-110
Time:	13:02:41	Log-Likelihood:	778.81
No. Observations:	1279	AIC:	-1546.
Df Residuals:	1273	BIC:	-1515.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.4763	0.018	25.786	0.000	0.440	0.513
alcohol	0.3523	0.024	14.882	0.000	0.306	0.399
volatile acidity	-0.3409	0.037	-9.136	0.000	-0.414	-0.268
sulphates	0.2834	0.042	6.817	0.000	0.202	0.365
citric acid	0.0264	0.024	1.114	0.265	-0.020	0.073
chlorides	-0.2164	0.052	-4.133	0.000	-0.319	-0.114

Omnibus:	15.519	Durbin-Watson:	2.039
Prob(Omnibus):	0.000	Jarque-Bera (JB):	19.336
Skew:	-0.170	Prob(JB):	6.33e-05
Kurtosis:	3.498	Cond. No.	18.4

The group started by analyzing the output. First off, the coefficient values show the numerical value of the relationship between the regressor and the target value (a unit increase in the predictor corresponds to the coefficient increase in the response. When the coefficient has a negative value, we see that it has an inverse relationship with the target value. The standard error shows the average error in estimating our coefficients and allows for the creation of confidence intervals. Our standard errors are all relatively low which shows that our regressors are fairly accurate.

The t-value shows the value on a t-distribution in which the probability of a value falling past this region is small enough for the p-value to reject the null hypothesis as it is unlikely those values occurred by chance. All of our p-values are 0 (or extremely close to 0) except citric acid. Since the p-value is greater than 0.05 it can likely be removed from the model to better generalize against new observations. The R squared value shows the percentage accuracy that our whole model correlates to the data. The adjusted R squared value is similar to the R squared value, however it takes into account the number of regressors and the value of each regressor to the model. Both R squared values are relatively close to each other at .336 and .334 respectively. This R squared value is low, meaning a large amount of the variance is not explained by the model. In other words the model is not that good, however when given a real data set, this can be expected. The F statistic is similar to the T statistic however it is testing the whole model. The p value of our F statistic is very low, so we can say that our model passes the null hypothesis.

Finally we construct the equation of the linear model.

$$\hat{Y} = 0.47 + 0.35(alc) - 0.34(va) + 0.28(s) + 0.026(ca) - 0.22(chl)$$

There is a slight difference between this and the SGD regressor most noticeable is the positive coefficient of citric acid. Beyond this, the result is very similar to the SGD regressor.

3 Summary and Future Work

While the results of these models were not spectacular, this can be attributed both to the loose correlation amongst predictors and the response and the rigidity or high bias for a linear model. To combat this high bias, higher order terms could be investigated as well as interaction terms. Additionally, a residual analysis would be useful in validating normality and variance assumptions. Finally, treating this as a multiclass classification problem could yield better results.

Sources

https://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity
https://extension.psu.edu/volatile-acidity-in-wine
https://rstudio-pubs-static.s3.amazonaws.com/57835_c4ace81da9dc45438ad0c286bcbb4224.html