

Wine Quality

Problem Statement: The goal behind this project is to be able to predict the quality of wine from its chemical properties. The 11 features that I had are as follows: 'alcohol', 'density', 'total sulfur dioxide', 'free sulfur dioxide', 'volatile acidity', 'fixed acidity', 'chlorides', 'residual sugar'. This kind of prediction would be helpful in business in choosing wine when buying in bulk. It could help predict what wines are underpriced or what wines are overpriced.

Data Wrangling: Wine-quality dataset didn't need any data wrangling but I did check for any null or na values and whether the data types were in the expected format. All data exist in the best formats. I am only considering the white wine data since it contains the most observations and, as advised by my mentor, contains enough observations to carry out a fulfilling investigation. In the dataset of only white wine we see that only quality ratings of 3 - 9 exist.

Data Story:

Histogram Analysis:

- A greater 'fixed acidity' correlates to a lower quality wine.
- A 'volatile acidity' greater than 0.5 correlates with lower quality wine.
- Citric acid' levels above 0.6 or below 0.2 corresponds to a lower quality wine.
- There seems to be no correlation of 'residual sugar' to the quality of wine.

- A low 'chloride' content correlates to a higher quality wine.
- Lower 'free sulfur dioxide' content correlates with lower quality wine.
- Where 'total sulfur dioxide' less than 75, we see a high chance of low quality wine.
- We see that a lower 'density' correlates to a better quality wine.
- The higher the pH the better the chance it will be a high quality wine.
- As for the 'sulphates' there was no noticeable correlation.
- A higher 'alcohol' content correlates to a higher quality rating and vice versa.

Mean, Median, Range Analysis:

- If the 'alcohol' content is less than 8.6 it is a 100% that the wine is of mid or low quality.
- A 'density' greater than 1.00040 corresponds most of the time to a mid or low quality wine.
- A 'total sulfur dioxide' content greater than 229 or less than 34 corresponds 100% of the time to a mid or low quality wine.
- A 'free sulfur dioxide' content greater than 108 corresponds 100% to a mid or low quality wine. A 'free sulfur dioxide' content greater than 131 corresponds 100% to a low quality wine.
- 100% chance if 'chloride' levels are above 0.135 it corresponds to a mid or low quality wine.
- 'Residual sugar' above 19 correlates to mid or low quality wine.

- 'Citric acid' above 0.75 or below 0.29 is a 100% chance of being mid to low quality wine.

Inferential Statistics: Below I calculated the 90% mean or median confidence intervals values for the 11 features of wine quality. (High Quality - ratings 9, 8, 7 | Mid quality - 5, 6 | Low Quality: 3, 4)

Alcohol mean 90% confidence intervals and absolute analysis

- High Quality: 10.77 - 12.07, mean = 11.42
- Mid Quality: 9.73 - 10.88, mean = 10.27
- Low Quality: 9.67 - 10.74, mean = 10.17
- Alcohol < 8.6 corresponds to low or mid quality wine

Density median 90% confidence intervals and absolute analysis

- High Quality: 0.99059 - 0.9935, mean = 0.99185
- Mid Quality: 0.9925 - 0.99634, mean = 0.99438
- Low Quality: 0.9927 - 0.99583, mean = 0.99421
- Density > 1.0006 corresponds to mid quality wine
- Density < 0.9872 corresponds to high quality wine

Total Sulfur Dioxide median 90% confidence intervals and absolute analysis

- High Quality: 104.5 - 144.5, mean = 122.59

- Mid Quality: 115.5 - 172.0, mean = 140.83
- Low Quality: 93.5 - 175.0, mean = 124.62
- Total sulfur dioxide < 34 and > 229 corresponds to a mid or low quality wine
- Total sulfur dioxide > 344 corresponds to a low quality wine

Free Sulfur Dioxide median 90% confidence intervals and absolute analysis

- High Quality: 27.0 - 42.0, mean = 33.42
- Mid Quality: 25.5 - 47.5, mean = 34.63
- Low Quality: 11.5 - 34.0, mean = 19.56
- Free sulfur dioxide content > 108 corresponds to a mid or low level wine
- Free sulfur dioxide > 131 corresponds to a low quality wine

Chloride mean 90% confidence intervals and absolute analysis

- High Quality: 0.0332 - 0.045, mean = 0.03817
- Mid Quality: 0.0397 - 0.0646, mean = 0.04771
- Low Quality: 0.0411 - 0.0733, mean = 0.05059
- Chloride content > 0.135 corresponds to mid or low quality wine
- Chloride content < 0.012 corresponds to mid quality wine

Residual Sugar median 90% confidence intervals and absolute analysis

- High Quality: 2.15 - 8.18, mean = 4.04
- Mid Quality: 3.0 - 10.9, mean = 5.95
- Low Quality: 1.55 - 8.4, mean = 3.45

- Residual sugar < 0.8 corresponds to mid or low quality wine
- Residual sugar > 17.55 corresponds to a mid or high quality wine

Citric Acid mean 90% confidence intervals and absolute analysis

- High Quality: 0.288 - 0.371, mean = 0.326
- Mid Quality: 0.281 - 0.414, mean = 0.338
- Low Quality: 0.237 - 0.399, mean = 0.308
- Citric acid level > 0.74 corresponds to a mid or low quality wine
- Citric acid < 0.29 corresponds to all quality ratings 7 and below

Volatile Acidity mean 90% confidence intervals and absolute analysis

- High Quality: 0.222 - 0.32, mean = 0.265
- Mid Quality: 0.234 - 0.335, mean = 0.277
- Low Quality: 0.301 - 0.48, mean = 0.376
- Volatile acidity > 0.76 corresponds to mid or low quality wine
- Volatile acidity < 0.11 corresponds to mid or high quality wine

Fixed Acidity mean 90% confidence intervals and absolute analysis

- High Quality: 6.33 - 7.13, mean = 6.724
- Mid Quality: 6.46 - 7.34, mean = 6.876
- Low Quality: 6.62 - 7.84, mean = 7.181
- Fixed acidity > 9.2 corresponds to low quality wine

Sulphates mean 90% confidence intervals and absolute analysis

- High Quality: 0.44 - 0.58, mean = 0.5
- Mid Quality: 0.44 - 0.55, mean = 0.49
- Low Quality: 0.42 - 0.54, mean = 0.48
- Sulphates < 0.25 or > 0.87 corresponds to mid or high quality wine
- Sulphates < 0.23 or > 1.06 corresponds to a high quality wine

pH mean 90% confidence intervals and absolute analysis

- High Quality: 3.14 - 3.3, mean = 3.22
- Mid Quality: 3.11 - 3.26, mean = 3.18
- Low Quality: 3.1 - 3.27, mean = 3.18

In summary the findings from the inferential statistics suggest that a number of features were strong indicators for different quality tiers of wine. The following is a list of the observed tiers that seemed logical to split the data into.

- 'Alcohol' and 'density' are features that do well in a binary split of data and can strongly indicate whether the wine is high quality or low/mid quality.
- 'Total sulfur dioxide' does well in a binary split of data and can indicate whether wine is of mid quality or low/high quality.

- 'Free sulfur dioxide', 'volatile acidity', and 'fixed acidity' are three features that does well in a binary split of data and can indicate what wine is low quality or mid/high quality.
- 'Chlorides' and 'residual sugar' are the only two features that does well in splitting the data three ways and can indicate relatively well (compared to all other features) what wine is either high quality, mid quality, or low quality.

Machine Learning: I took the combination of features that were strong indicators of particular tiers of wine and created machine learning models using the KNN classifier, Naive Bayes classifier, Multinomial Logistic Regression classifier, GradientBoosting classifier and the Random Forest classifier. It turns out that the best classifier when taking into consideration of the different tiers of wine quality was the Random Forest classifier when it considered 9 out of 11 of the features rather than 2 out of 11 features. Recognizing this pattern I decided to include all 11 features and got the highest accuracy score of about 65%. I then proceeded to tune all parameters of the Random Forest classifier to further optimize the accuracy of the predictions.

In optimizing the accuracy score I went and tuned the `n_estimators`, `max_features`, `criterion`, and `bootstrap` parameters by running a for loop over all possibilities and checking the corresponding accuracy scores. The following list shows my findings for the best parameters for the Random Forest classifier:

- `n_estimators`: 1150
- `max_features`: 'auto'
- `criterion`: 'entropy'

- bootstrap: True

All of the tuning yielded an accuracy score of 68.3%.

I talked to my mentor and he suggested a 'target encoding' approach to increase the accuracy score significantly. For target encoding I took a train-test data split and built high, mid, and low quality distributions over all features with the training data. I now went into every row of data and for each feature I encoded every feature's probability of belonging to a high, mid, or low quality wine. I then went ahead and did the same tests above and found that the best classifier was once again the random forest classifier with the following parameters:

- N_estimators: 1450
- Max_features: auto
- Criterion: entropy
- Bootstrap: false

With target encoding I came to an accuracy score of 84.23%!