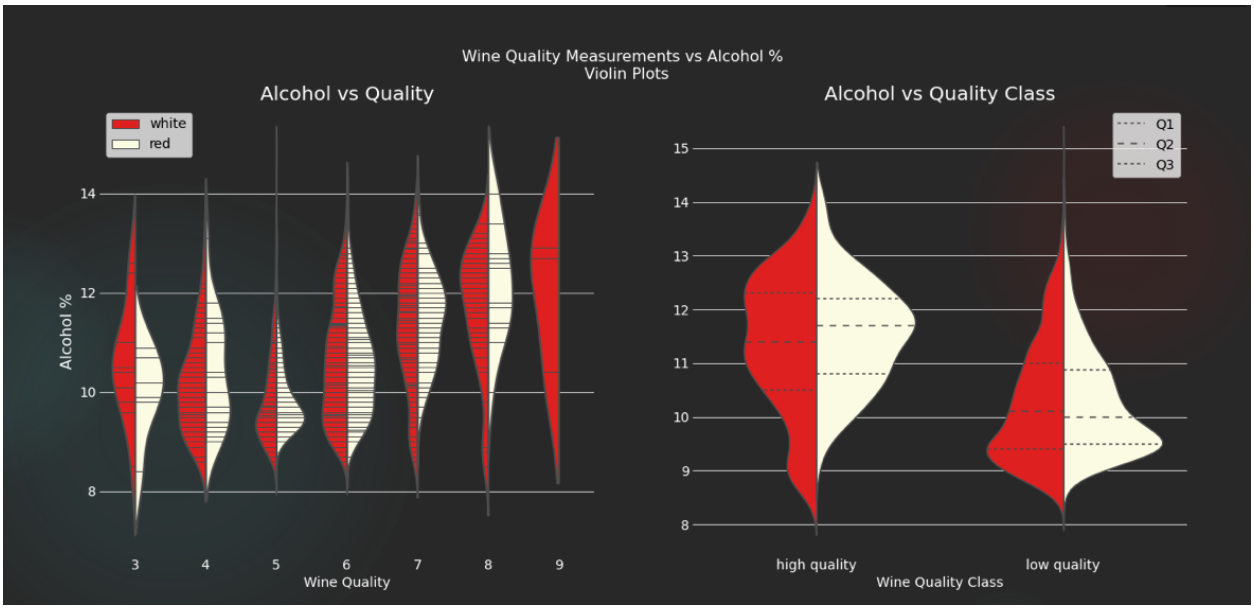
Wine Quality Classification

Tyler Poore



Introduction

Wine quality is a complicated metric that takes into account the wine's color, terroir (i.e. it's growing conditions), the balance, complexity, length of time on the palette, and the distinctiveness of the flavor notes of the wine (.e.g terms like oak, cherry, vanilla, and tobacco are common red wine notes, whereas whites tend to have citrus notes).

In this project, I have downloaded a dataset containing red and white vinho verde wines from Portugal for purposes of predicting wine quality based on physicochemical tests using machine learning.

To give a reason as to why this is important, I want to start with a quote from Maynard Anderw Amerine, who was a pioneering researcher in the cultivation, fermentation, and sensory evaluation of wine that helped to establish the metric upon which wines are judged.

"Wines quality is easier to detect than define"

This makes sense, if you enjoy the flavor of a soda for example, it is very easy to say whether or not we enjoy it, but knowing why we enjoy it is another matter entirely.

Similarly, quality is primarily a subjective measurement strongly influenced by extrinsic factors. While there is somewhat of a general consensus among wine connoisseurs as to what constitutes wine quality, that is, it is a subjective measurement that comes from extensive experience in wine tasting.

There are some quantifiable aspects to this howerer, and for example in white wines, acidity and alcohol/sugar should match; if the acidity is not enough compared to the wine's sugar content level, you run the risk of having a wine that is too cloying.
For reds, tannins, acidity and alcohol should all be in balance. Negative quality factors, such as off-odors, are generally easier to identify and control. Positive quality factors tend to be more elusive.

The hypothesis is that the physicochemical properties of a wine (e.g. pH and alcohol content) which can be controlled at the winemaking stage, can be used to predict the overall quality of a wine. If this hypothesis is correct, then we can figure out which of the wine's properties are most important in optimizing the output of high quality wines by a winemaker.

The Features
Vinho verde is a medium-alcohol wine from the Minho (northwest) region of Portugal particularly appreciated due to its freshness.

In this project, I am tasked with predicting the quality of Vinho Verde wines (ranging from 0 for

the lowest quality wines to 10 for the highest quality wines), given the physicochemical components of wines that are introduced into the product during the wine-making process.

Tartaric acid is the primary acid in wine grapes. It's probably the most durable acid in a wine, and it resists much of the effects of other acids. That's why it's called a fixed acid. That makes it one of the most important parts in stabilizing a wine's ultimate color and flavor profile.

Features	Red Wines	White Wines
	Mean	Mean
Fixed acidity (tartaric acid g/L)	8.3	6.9
Volatile acidity (acetic acid g/L)	0.5	0.3
Citric acid (g/L)	0.3	0.3
Residual sugar (g/L)	2.5	6.4
Chlorides (sodium chloride g/L)	0.08	0.05
Free sulfur dioxide (mg/L)	14	35
Total sulfur dioxide (mg/L)	46	138
Density (g/mL)	0.996	0.994
pH	3.3	3.1
Sulphates (potassium sulphate g/L)	0.7	0.5
Alcohol (vol.%)	10.4	10.4

Table 1. Wine Physicochemical Features

Citric acid has a minor presence in wine, but a noticeable one nonetheless. The quantity of citric acid in wine is about 1/20th that of tartaric acid. It's mostly added to wines after fermentation due to yeast's tendency to convert citric acid to acetic acid. It has an aggressive acidic taste, is often added by winemakers to increase a wine's total acidity, and should be added very cautiously.

Sulfur dioxide is added to wine during winemaking to prevent the microbial spoilage, oxidation and color changes due to undesirable enzymatic and non-enzymatic reactions. High concentrations of sulfur dioxide affect the final quality of the wine, mainly the smell and the taste and can inhibit the malolactic fermentation.

Sulfur dioxide (SO_2) as potassium thiosulphate was used as a microbial growth inhibitor during the current fruit juice substrate preparation to inhibit the growth of some yeast species and the majority of bacteria related to wine spoilage. Due to antiseptic and antioxidant properties on the final wine, sulfur dioxide (SO_2) is the most versatile and efficient additive than other additives such as dimethyl dicarbonate (DMDC) used during winemaking.

Business Case

By being able to predict a wine's quality from a given set of physicochemical parameters, wine-makers can reasonably produce high quality wines with the given features as inputs to the model. This can help a wine-maker to plan ahead to produce a high quality wine as well, by using the predicted parameters to obtain the best physicochemical values as predicted by the model.

Data Wrangling

The wine data data set was available in two .csv files from the [UCI Machine Learning Repository](), split into red and white wines. The datasets were loaded into pandas DataFrames and examined.

Dealing with Data Size Issues

To best reduce bias, I want to set aside a portion of the test data as early as possible. However, I can see there are some issues that need to be addressed first due to the size of the data.

Figure 1. Red and White Wine Quality Distributions

The target for this classification problem is the quality feature of the dataframe with scores ranging from 0 - 10 (bad - good). Plotting the distribution of these data as shown in Figure 1 shows that I am limited by the red wine data size, and that the target data values need to be categorized. As I need to classify the wines into high and low qualities, I decide that any wine with a quality score of 7 or greater will define the decision boundary between high and low quality wines. Figure 2 shows the distributions of counts for each of these two classes, and that there is a class imbalance that also needs to be addressed.

I decided to set aside 20% of the data for the white wines, and 40% of the data for the red, stratifying the target class. My reasoning behind this is that the test data will have 87 samples of the minority class (high quality wines) at this high of a split, and due to the limited data, I need to have enough to train a model that can recognize the minority class. The test data are then saved and the current

Figure 2. Red and White Wine Class Distributions

data are used for data exploration, model training, and validation.

Dataset Features

Each dataset has 11 features and 1 target, where the features are physicochemical descriptors of the wines. These data were published by Cortez et al[1], of red and white *vinho verde* samples from Portugal for classification purposes.

Data Discrepancies

All of the features are of type float and don't contain any missing data. However, roughly 20%[2] of the data are duplicated. In general, it is good practice to remove duplicated data on the training set so that the model can better generalize to the full dataset. I can always test whether or not duplicates affect model performance down the pipeline, but for now I will remove them moving forward.

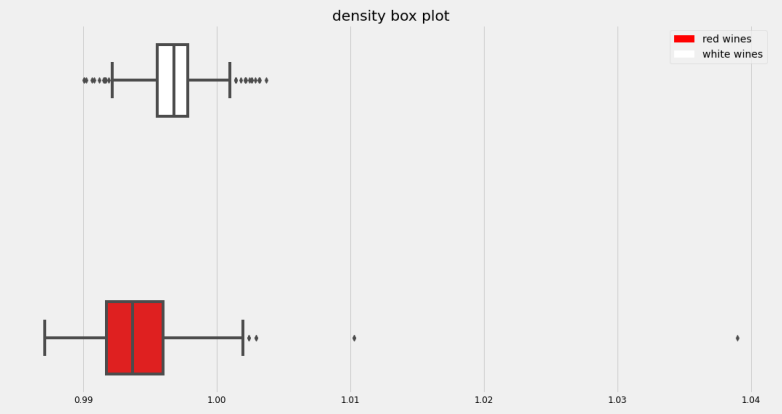


Figure 3. Box Plot Showing Outliers in Data

I next wanted to look at outliers in the data. Figure 3 shows a boxplot of the red and white wine densities with clear outliers present in the red wines dataset. For visualization purposes I will remove them to better understand the spread of the data[3]. However I plan to wait on removing them for modeling purposes until after I do feature selection.

Feature-Target Relationships

I wanted to see if the target feature had any strong dependencies from the physicochemical features in the dataset. This could allow me to better understand the differences between high and low quality wines as well as to aid in model training and prediction.

Figure 4. Violin Plots of Quality/Class vs. Alcohol

Figure 4 shows a violin plot of red and white wine

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, Elsevier, 47(4):547-553, 2009.

[2] 17.66% duplicated rows for red wines, 23.66% for whites

[3] Outliers were removed with a *z-score* < *3*

quality vs alcohol content. It is readily apparent that high quality wines have a higher alcohol content than lower quality wines. The differences in the means of the data for each class (as indicated by the dashed - - lines). Shows how readily apparent this difference is, with the average high quality of red wines being a bit higher than that of the white wines.

Feature-Feature Relationships

Some of the features are obviously linear combinations of other features in the dataset. The density feature for example, is given by Equation 1.

$$\textbf{Eq. 1}\ \varrho = m \div V$$

ϱ = Density of the solution in g/cm^3

m = mass of the solution in grams (g)

V = Volume of the solution in cubic centimeters (cm^3)

As 9 of the 11 features in the dataset are are the chemicals given in weight-per-volume dissolved in the wine [4], these features will naturally be collinear with density. Figure 5 is a heatmap of the pearson correlations between the features with a min heatmap visualization threshold of pearson = 0.2.

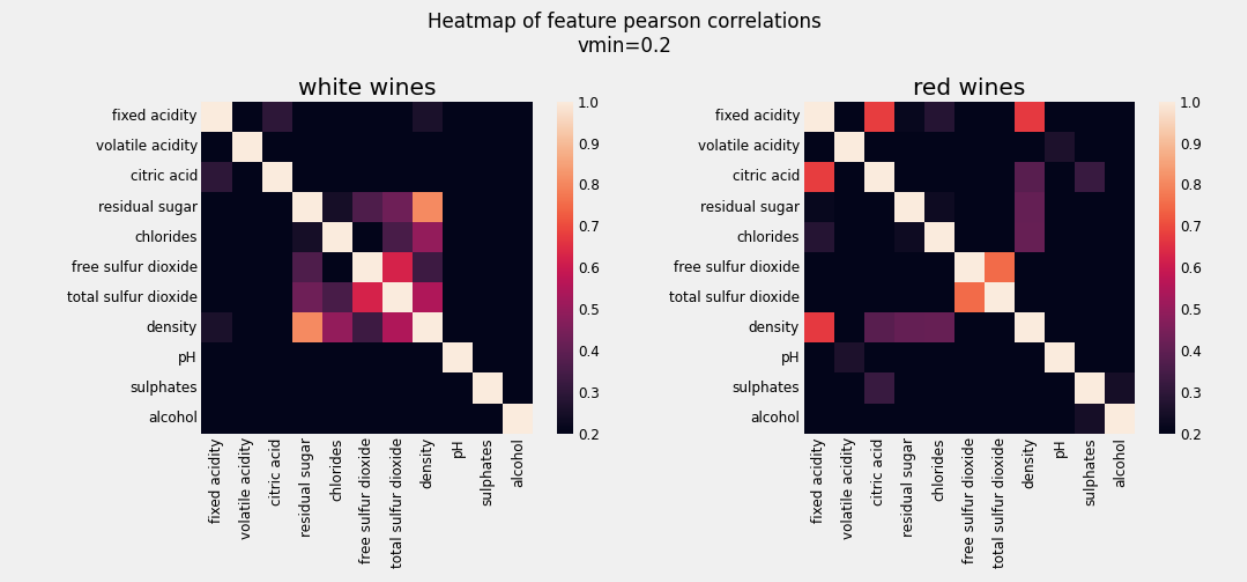


Figure 5. Heatmaps of Wine Features

For both red and white wines, it's easy to see that density is highly correlated with residual sugar in the white wines, and fixed acidity in the red wines. As the density of the wine represents the total weight contribution of all the components dissolved in the wine, it can be

[4] The weight contribution of the alcohol to the wine can be readily estimated using the percent alcohol by volume feature and the density of ethanol (0.789 g/mL).

inferred that the sugar makes up a greater overall weight percentage of white wines than in red wines, and that the same can be true for the fixed acidity for red wines over that of white. I can't say whether or not these features are more or less important as they are simply weight measurements, but it's possible that they may be of greater significance given their greater abundance in each of the wines. I will look at feature importances next to get a general idea of the wines.

Feature Importance

Given the problem, it would be advantageous for the employer to know which of the physicochemical features in the dataset contribute the most to model predictions. I will aim to use permutation feature importance using the best models selected for each dataset. Before that is done, I'd like to get a quick idea as to how a Random Forest Classifier will predict feature importances on the training dataset.

Red Wines

Figure 6 shows the importances for the red wines dataset after alcohol and density were removed. Alcohol was removed as it is by far the most important feature and thus dwarfs other features. Density was removed as multicollinear with many other features in the dataset.

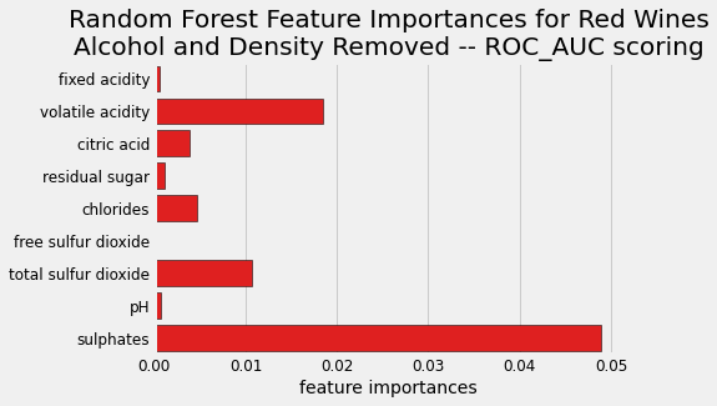


Figure 6. Predicted Red Wine Feature Importance

The most important features predicted on the training data by a random forest classifier for the red data are sulfates, volatile acidity, and total sulfur dioxide.

White Wines

The features that were removed from the red wines dataset were also removed from the white wines, and feature importance was similarly predicted on the training data using a random forest classifier, shown in Figure 7.
The model predicts chlorides to be most important, which comes from salt in the wine. I expected residual sugar and volatile acidity to be here, but I will see how the final model's predictive features' importances compare.

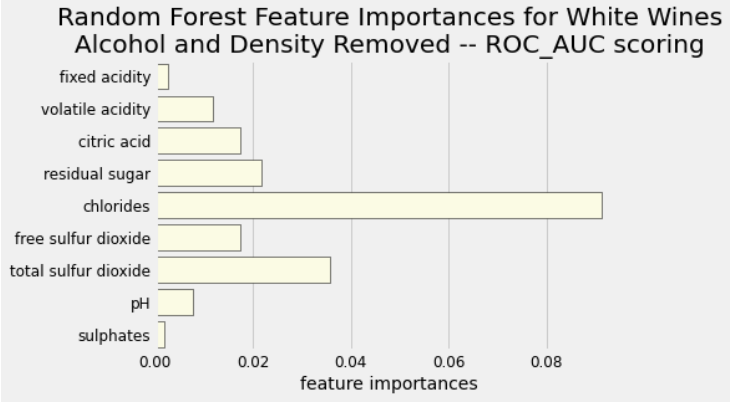


Figure 7. Predicted White Wine Feature Importance

Modeling

My strategy in addressing the problem of classifying each of the wine datasets by their quality classes was to develop a machine learning pipeline that would allow me to quickly prototype several classification models, as well as allowing me to experiment with various data preprocessing strategies and hyperparameter tuning.

At the end of this pipeline, I will blend each of the models that I've trained to obtain the best performance across the range of models tested. This *throw everything and see what sticks* approach is a nice way to optimize performance as well as to gain insights into the modeling by seeing how each of the models do, comparatively.

Frank Ceballos has a well-written and comprehensive article on stacking classifiers and their uses in boosting performance. They are best used when an increase in performance can't be achieved through any further data acquisition, feature engineering, or preprocessing and all typical classification model approaches have been exhausted. The downside to these types of models is that they can be computationally prohibitive. Luckily, I am dealing with fairly small datasets where this approach makes sense.

Choice of Metrics

As this is a binary classifier, I elected to use ROC AUC as an comparative measure of model performance. I will select the final model based off of the precision-recall scores and whatever is appropriate for the situation (e.g. excluding models with very long fit times, etc.).

Pipeline

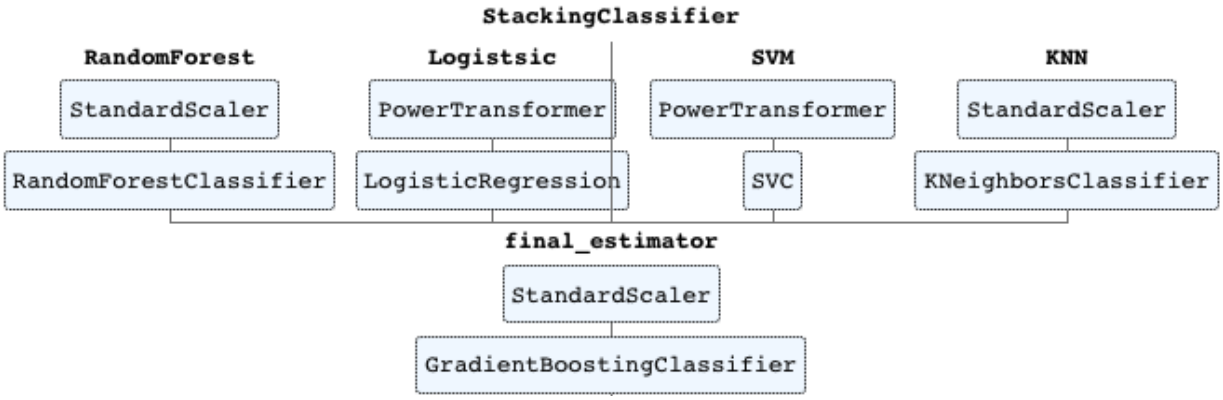


Figure 8. Stacking Classifier Pipeline

The general pipeline for the wine classifications shown above in Figure 8. This approach uses a stacking classifier that blends the predictions made from several other classifiers by having

those predictions passed as inputs for a final estimator (a gradient boosting classifier in this case).

The data were preprocessed by first undergoing feature selection and then applying Yeo-Johnson transformations for the linear estimators, otherwise the standard scaler was used to normalize the variance in the feature data.

Results and Discussion

The test roc auc test scores for each of the classifiers was evaluated by a 10-fold cross validation. The results are shown below in Figure 9.

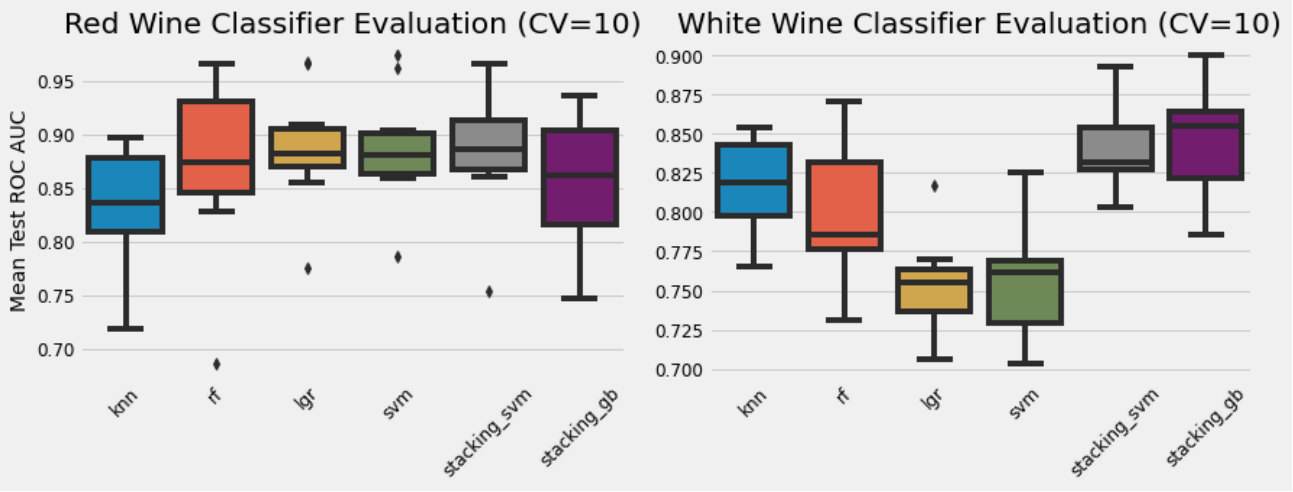


Figure 9. Mean ROC AUC Test Scores for Different Classifiers

The classifiers with the best mean test roc auc scores are the stacking classifiers in both datasets, which is to be expected. However, the best evaluator for the red wines data uses a support vector machine[5] classifier for its final estimator, whereas the white wines feature a gradient boosting classifier[6].

High and low quality wines in the white dataset are not so readily linearly separable and are therefore more difficult for a linear classifier to classify. Because the KNN classifier uses localized proximity to make predictions, localized groupings of high quality wines are able to be better predicted.

Plotting the decision boundaries and evaluating the models provides insight into how the model performs as well as how to tune it in a more intuitive way.

---

[5] Final estimator parameters SVC(C=100.0, class_weight='balanced', degree=4, gamma=3.16e-05, kernel='rbf')

[6] Final estimator parameters GradientBoostingClassifier(learning_rate=0.0017,, loss='exponential', max_features='sqrt', min_samples_leaf=0.17, min_samples_split=0.4, n_estimators=7222)

Red Wines

The mean results for the classification models are presented in Table 2 with roc auc train and test scores and modeling times. What is immediately apparent is that the higher performing models are those that tend to form a single, continuous decision boundary-one in which there is a clear distinction between the predicted positive and negative classes. Estimators that form many small, discrete decision boundaries tend to perform more poorly in relation.

Red Wine Classification Results

	test_score	train_score	fit_time	score_time
logisticregression	0.883624	0.891002	0.071392	0.008965
final_estimator__svc	0.883340	0.958254	2.243921	0.059419
svc	0.882143	0.890641	0.119681	0.009832
randomforestclassifier	0.869474	0.888827	0.481239	0.039913
final_estimator__gradientboostingclassifier	0.860632	0.953089	16.743882	0.058780
kneighborsclassifier	0.857320	1.000000	0.008534	0.014739

Table 2. Mean Train Test Scores for Red Wines

Stacking Classifier Comparisons

The stacking classifier takes the predictions of the estimators and sends them as inputs for a final estimator. This results in the final estimator's decision boundary that "blends" features present in the input estimators' decision boundaries. Appendix A1 # shows the decision boundaries for red wines following PCA decomposition into two principal components for visualization. The SVM final estimator predicts a separation between high quality wines and low quality wines in the 2D-plane, and the randomness associated in the boundary are from the input estimators.

Conversely, the gradient boosting classifier creates multiple islands that lend themselves to a lower ROC AUC score, but overall performs better in precision by predicting a higher ratio of high quality wines over low ones than the svc final estimator. It should be noted that stacking classifiers tend to overfit as can be seen in Table 1, and reducing the complexity of the model by dropping out some of the input estimators, or introducing some form of regularization into the final estimator will be the next step.

Feature Importance

I elected to use the stacking classifier with the gradient boost as the final estimator in examining the feature importance. Note that the negative values in Figure 10 are features the model deems unimportant in through permutation feature importance [that can appear in smaller datasets.](#) The most important features according to the model are alcohol, followed by sulfates and volatile acidity.

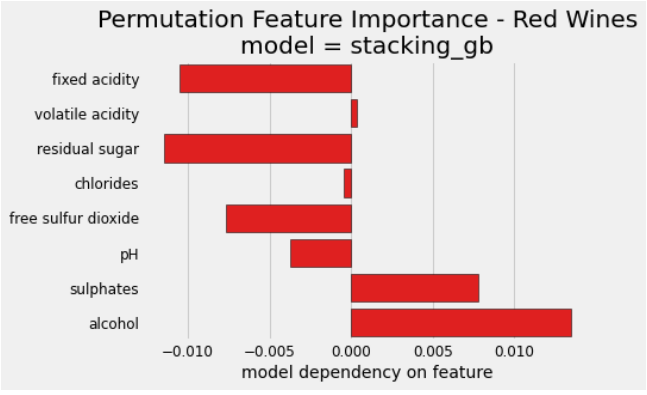


Figure 10. Red Wine Permutation Feature Importance

White Wines

Table 3 has the mean ROC AUC scores for the test data in the white wines. Compared to the red wines dataset, the white wines is 3x larger, and possibly as a result there is greater overlap between the high and low quality wines *see Appendix B.* Contrasting this as well, classifiers on the white wines dataset that tend to perform well are those that form multiple less-defined decision boundaries instead of the well-defined ones that did well in the red wines dataset.

White Wine Classification Results

	test_score	train_score	fit_time	score_time
final_estimator__svc	0.843383	0.999171	3.639952	0.080966
final_estimator__gradientboostingclassifier	0.840537	0.999781	32.763109	0.100103
kneighborsclassifier	0.817029	1.000000	0.012711	0.022624
randomforestclassifier	0.799165	0.820584	0.421152	0.033649
svc	0.755247	0.766366	0.433983	0.015245
logisticregression	0.754939	0.765170	0.078012	0.008902

Table 3. Mean Train Test Scores for White Wines

Stacking Classifier Comparisons

To get a better idea as to why this idea, I turn again to the Appendix B where the decision boundaries for the two stacking classifiers for the white wines are plotted following PCA decomposition for ease of visualization (Figures B1, B3). Using SVM as a final estimator creates a decision boundary that subdivides the plane into high quality and low quality wines. While this plot looks reasonable, there are still many misclassifications for both classes that results in a decent recall (as some high quality wines are found in the low quality region), and moderate precision (there are far too many low quality wines in the high quality class region).

Using the gradient boosting classifier as a final estimator boosts the precision to 68%. The downside of this is that the fit time for the model is 10x longer than the already long svc stacking classifier. This highlights the limitation of the stacking classifier in that it is computationally more expensive than the other classifiers used. What is seen is what was advertised, however, and using it results in a 3% in the roc auc test score after cross-validation[7].

Feature Importance

Permutation feature importance was done using the stacking classifier with the gradient boost final estimator. Figure 11 shows that alcohol is the most important feature followed by pH, sulphates, free sulfur

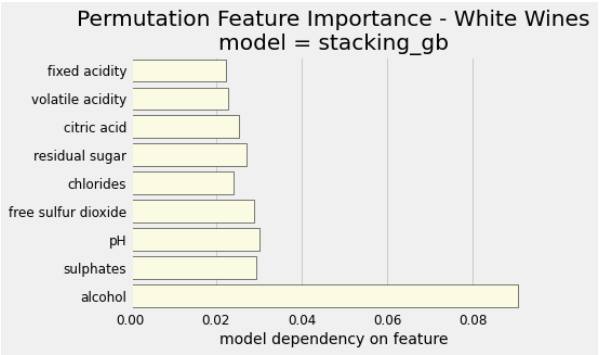


Figure 11. White Wine Permutation Feature Importance

[7] CV = 20

dioxide, and residual sugar. I was surprised to see that the chlorides were not considered as important to the model as the random forest indicated in Figure 7. Many of the features have similar scores and may indicate that the model needs to be further tuned or I need to revise my feature selection strategy.

Conclusions

The models I built for the client predict the wine qualities of red and white wines from Portugal given physicochemical properties of the wines, and these models are used to assess the importance of the physicochemical features using permutation feature importance.

For red wines, the most important qualities are the alcohol content and the sulfates in the wine. Sulfates are added to wines to prevent spoilage, and as the third most important feature, volatile acidity, is an indicator of spoilage, the most important features for high quality red wines are those that affect wine spoilage.

For white wines, the model predicts that alcohol is the most important, followed by pH and residual sugar. I am less confident about these predictions than I am of the red due to the complexity of the model and the more complicated dataset for the white wines. Despite this, the stacking classifier with a gradient boosting final estimator has a relatively high precision score of 0.69 on the test dataset, meaning wines predicted to be of high quality will be correct on average ~70% of the time. In comparison, the best red wine stacking classifier would be making those same predictions ~50% of the time.

The stacking models were used because they can boost the performance of a model. And while that is seen with the white wines dataset, they have 10-100x longer fit times and tend to overfit the data. If I were to present this to my client, I would need to make sure that these are issues that need to be addressed. It would be feasible if a 3% increase in the precision score of the model used would forecast profits that justify the operating costs.

Next Steps

- Apply a neural network and see how it compares
- Review feature importances for white wines model
- Improve model precision scores for red wines
- Try and get more data from client
- Do a cost-benefit-analysis on model implementation

Appendix A: Red Wines

Figure A1. Stacking Classifier Red Wines Decision Boundary (SVM)

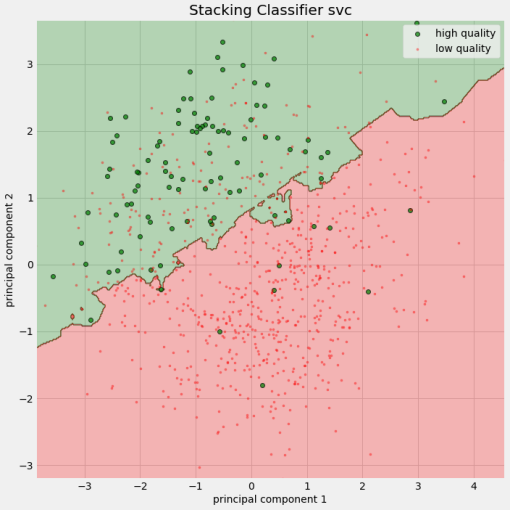


Figure A2. Stacking Classifier Red Wines Confusion Matrix (SVM)

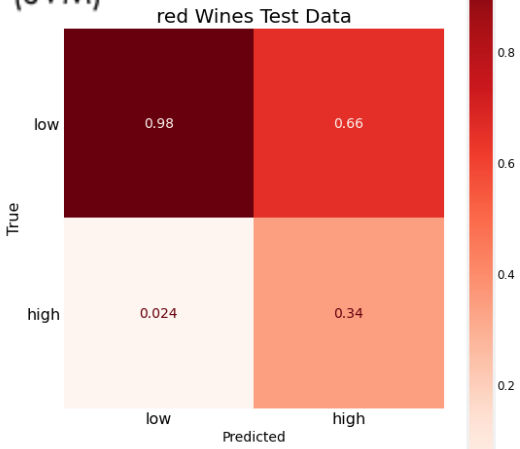


Table A1. Stacking Classifier Red Wines Classification Report (SVM)

	precision	recall	f1-score	support
0	0.98	0.73	0.84	553
1	0.34	0.89	0.49	87
accuracy			0.75	640
macro avg	0.66	0.81	0.67	640
weighted avg	0.89	0.75	0.79	640

Figure A3. Stacking Classifier Red Wines Decision Boundary (Gradient Boost)

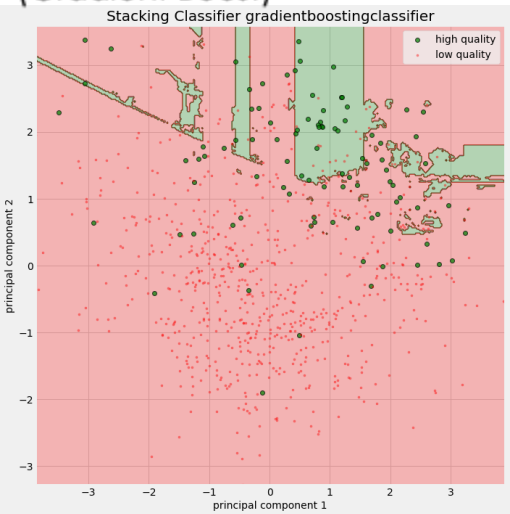


Figure A4. Stacking Classifier Red Wines Confusion Matrix (Gradient Boost)

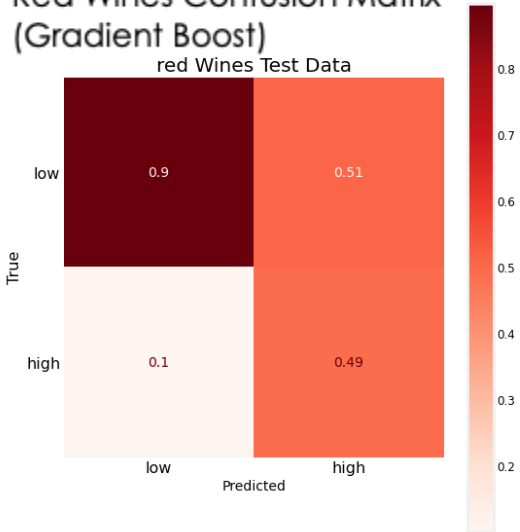


Table A2. Stacking Classifier Red Wines Classification Report (Gradient Boost)

	precision	recall	f1-score	support
0	0.90	0.95	0.92	553
1	0.49	0.32	0.39	87
accuracy			0.86	640
macro avg	0.70	0.63	0.66	640
weighted avg	0.84	0.86	0.85	640

Appendix B: White Wines

Figure B1. Stacking Classifier White Wines Decision Boundary (SVM)

Figure B2. Stacking Classifier White Wines Confusion Matrix (SVM)

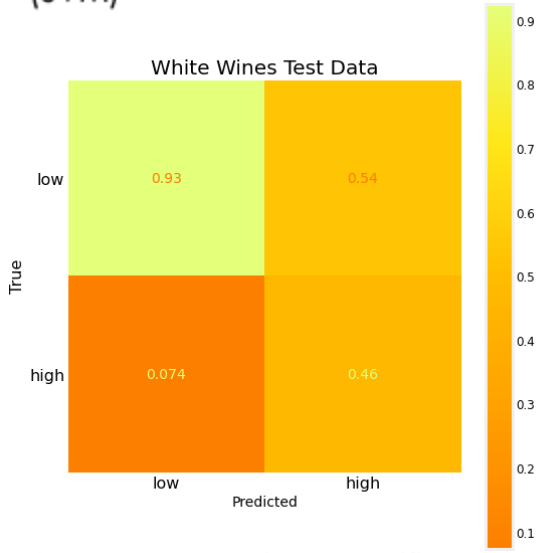


Table B1. Stacking Classifier White Wines Classification Report (SVM)

	precision	recall	f1-score	support
0	0.93	0.75	0.83	768
1	0.46	0.78	0.58	212
accuracy			0.76	980
macro avg	0.70	0.77	0.71	980
weighted avg	0.83	0.76	0.78	980

Figure B3. Stacking Classifier White Wines Decision Boundary (Gradient Boost)

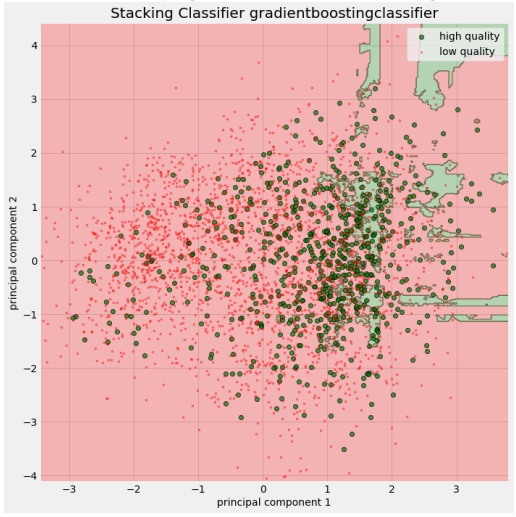


Figure B4. Stacking Classifier White Wines Confusion Matrix (Gradient Boost)

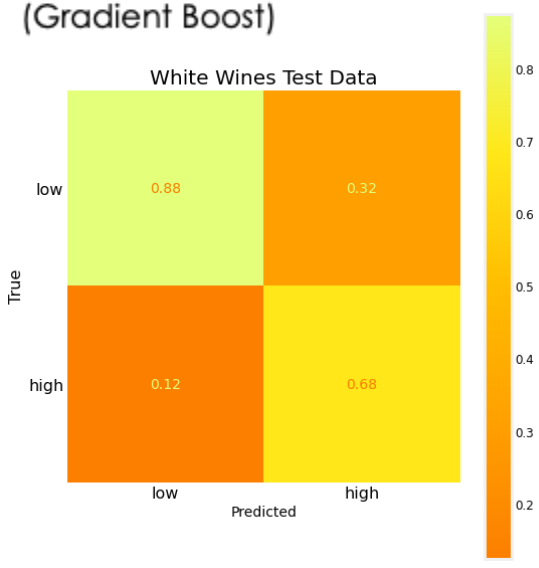


Table B2. Stacking Classifier White Wines Classification Report (Gradient Boost)

	precision	recall	f1-score	support
0	0.88	0.93	0.90	768
1	0.68	0.52	0.59	212
accuracy			0.84	980
macro avg	0.78	0.73	0.75	980
weighted avg	0.83	0.84	0.84	980