Introduction

The consumption of wine around the world has been steadily on the rise through out the last decade. The growths come from not just major wine producing countries but also minor countries. This means that international trading of wine has been on the uptick to keep up with the overall demands.

Traditionally, wines are usually classified instead of rated for their quality. Wines produced in different continents have different systems of classifying their products. Europe countries kept the traditional of appellation classification. Newer producing countries such as the US and Australia are more favorable to class by grape variety. However, classification is not always a strong indicator of quality. The emergence of wine critic and review publications revolutionized how wines are marketed, by given sensory assessment scores.

Although sensory assessment is the most direct way to rate the quality of wine, it has limitation on the amount of wines each individual assessor can assess due to sensory fatigue and difficulty to calibrate sensory assessment among the group. Therefore, building a prediction model based on sensory assessment database is a great way to broadly implement sensory preference ratings beyond human limitation.

Wholesalers who are purchasing wines can use prediction modeling to assist their professional assessors on wine quality rating. This method also addresses the challenges of assessing international producers that requires oversea shipment of wine samples. Furthermore, winemakers can use prediction modeling to determine the quality of their wine for marketing purposes.

The data

The white wine quality dataset is acquired from the UCI Machine Learning Repository which was originally published by the authors of *Modeling wine preferences by data mining from physicochemical properties*. The dataset was already compiled and cleaned, so very little pre-processing is needed. It consists of nearly 4900 samples of white Portuguese Vinho Verde, which are wines from a particular wine region in Portugal. The data contains 11 physicochemical measurements as attributes variables and 1 sensory preference score as target variable.

The attributes are common analytical tests taken by winemakers. For the samples in the dataset, the attribute variables are measured as finished products. The target variable is a sensory preference score ranged from 0 to 10. Each sample is evaluated by a group of three tasters and the median score is taken.

The description of attributes are:

1. Fixed acidity: the measurement of tartaric acid in the wine, which contributes to the main acidity in taste.

2. Volatile acidity: the measurement of acetic acid, which contributes undesirable scent of vinegar.

3. Citric acid: the measurement of citric acid, which is an additive in minute quantity to boost acidity in wine. It is also added in standardized quantity to remove excess iron and copper from wine.

4. Residual sugar: the amount of dissolved sugar present.

5. Chlorides: the amount of salt (sodium chloride) in the wine.

6. Free sulfur dioxide: the measurement of unbounded form of sulfur dioxide. Free sulfur dioxide have the property of inhibiting microbial growth (turning alcohol into vinegar) and the oxidation of wine, both of which have negative impact to quality.

7. Total sulfur dioxide: the measurement of both free and bounded sulfur dioxide.

8. Density: density measurement of the wine, which can affect the mouthfeel of tasting.

9. pH: measures the chemical acidity of the wine.

10. Sulphates: the measurement of potassium sulphates, which may be related to the presence of fertilizer and an indicator of fermentable nutrition in grapes

11. Alcohol: the percent of alcohol measured in wine.

The target variable of the entire dataset is distributed from value of 3 to 9. The majority of samples have values falling between 5 and 7, only a fraction of samples on either end of the spectrum. Since the project is to create a prediction model for wine quality, the model would be more efficient by reducing the number of classes from 7 to 3. Therefore, the dataset is given a new target variable for multi-class classification. The new variable have three labels, which are poor, mediocre, and desirable. Rows with the original target variable value between 3 to 5 are labeled as poor, the ones with value of 6 are labeled as mediocre, and the rest with value of 7 or above are labeled as desirable. The new target variable have more balanced labels, each are between 21 to 44 percent of the entire dataset. Therefore, resampling is not required to offset extremely unbalanced labels.

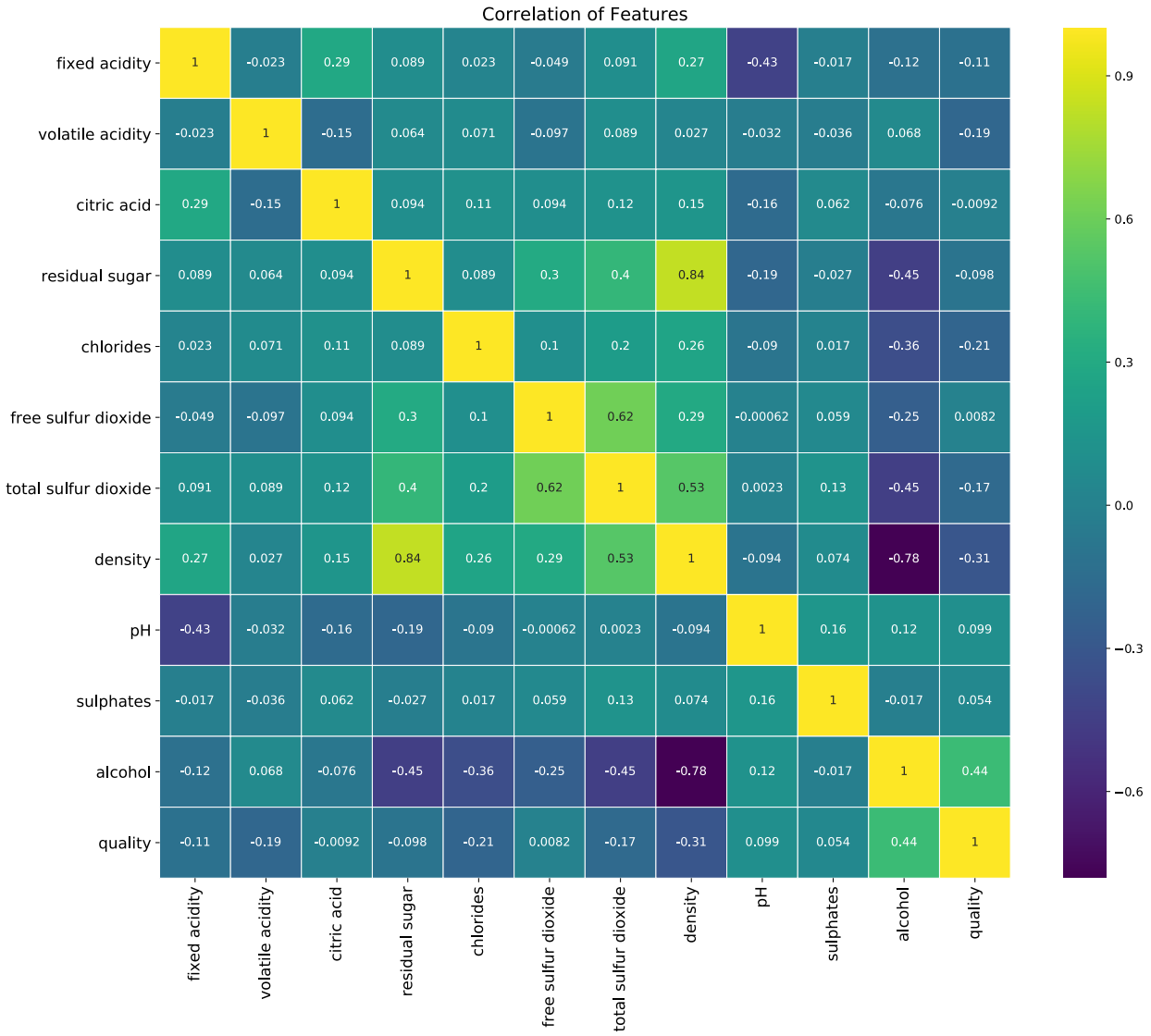
The variance of features like residual sugar, free and total sulfur dioxide are much higher than the rest of the features. These features are normalized by taking log of the features. In addition, the distribution of residual sugar appears to be highly skewed. The skewness is also resolve by taking log of the feature. Besides normalizing features with high variance, the entire dataset is also rescaled to optimize machine learning algorithms that use distance to make prediction, which are K-nearest neighbor and support vector machine.

Data Exploration

Most of the features are distributed somewhat symmetrically and unimodal. The distributions are not completely symmetrical due to existence of many outliers.

Overall, there is not a single variable that is determinative of the target variable. The highest correlation between all attributes and wine quality is alcohol content, which the correlation coefficient is 0.44. Empirical cumulative distribution functions of each new target variable label are constructed to prove that the correlation is statistically significant.

Furthermore, there are many attributes that are highly correlated due their physical and chemical properties. Examples are residual sugar increases the density of water while alcohol has less density than water, therefore residual sugar, alcohol, and density are highly correlated. Since total sulfur dioxide is the measurement of both free and bounded forms of sulfur dioxide, the attributes are also correlated.



Modeling

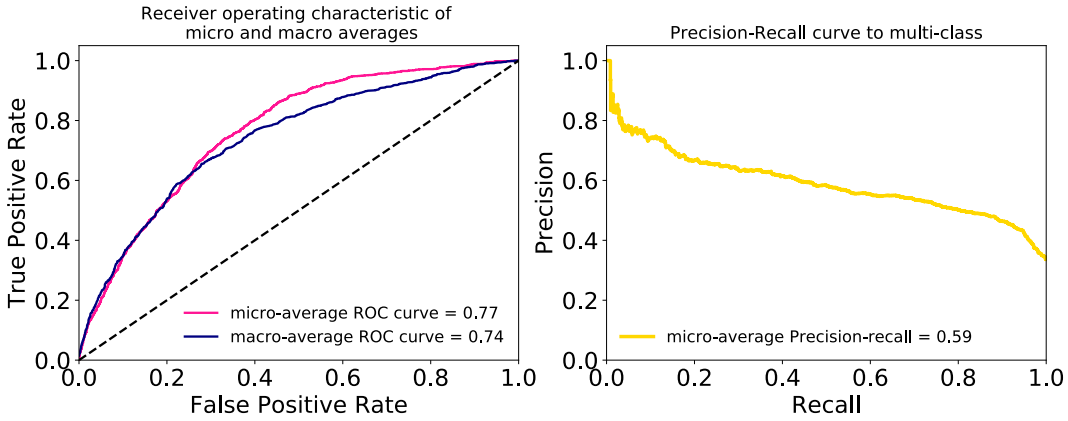
The model prediction is based on the dataset with existing target variable of 3 labels, therefore supervised learning algorithms are best for the job of multi-class classification. Python is the language used in the machine learning portion, and scikit learn is the main library used to perform machine learning. As supplement, numpy and pandas are used to manipulate dataset into series and data frames. Matplotlib and seaborn libraries are also used for generating visualizations.

The dataset is divided into 67% of training set and 33% of test set. Multiple machine learning algorithms are used for building the best model for this project, these algorithms are logistic regression, random forest, K-nearest neighbor, support vector machines, and Gaussian naive Bayes. To optimize each model, both grid search and 5-fold cross validation are used to optimize the model hyperparameters.

Logistic Regression:

This popular algorithm is used for this multi-class classification by using one-vs-rest method to fit each label to the rest as binary classification. The hyperparameters tuned in this algorithm are penalty and C, which is the regularization strength. The result can be seen below as both micro and macro average ROC curves. There is a slight difference between micro and macro averages for ROC curve. Since the proportion of each class is not very balanced, we will take micro-average ROC curve as the preferred metric for evaluation. The ROC AUC seems pretty promising, but the PR AUC is significantly lower due to low recall of desirable class.

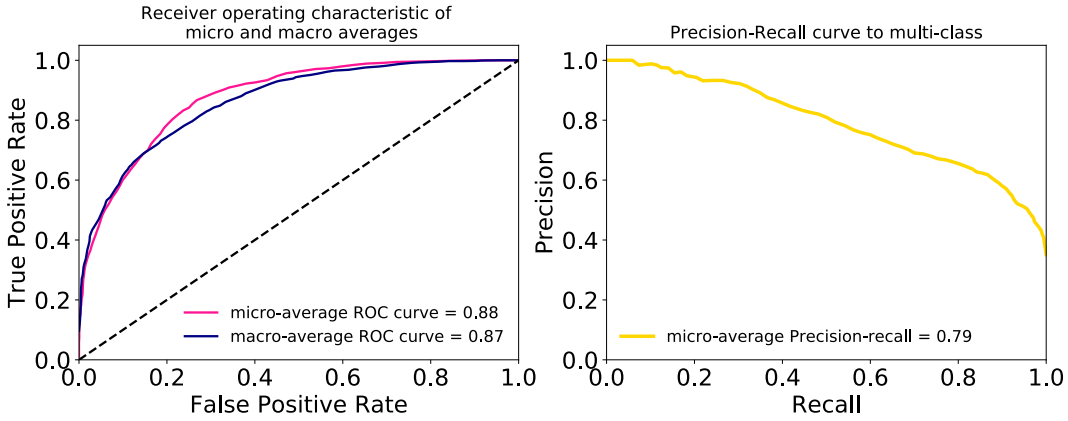
Class	Values			ROC AUC	PR AUC
	Precision	Recall	F1-score	0.77	0.59
0	0.60	0.59	0.59	Best score	Log loss
1	0.53	0.68	0.60	0.58	0.89
2	0.63	0.29	0.39		



Random Forest:

Unlike logistic regression that predicts each sample by regression, random forest uses decision tree to make decision splits to separate samples into classes that looks like a tree. The algorithm also produces feature importance that influences its decision making. In the training grid search, the hyperparameters selected for optimization are max depth of tree, criterion, max features for decision split, minimum number of samples required to split a node, and number of trees in used per model. The prediction made by random forest model performs well, with recall values well above 0.6 for each class.

Class	Values			ROC AUC	PR AUC
	Precision	Recall	F1-score	0.88	0.79
0	0.72	0.73	0.72	Best score	Log loss
1	0.67	0.73	0.70	0.69	0.68
2	0.79	0.62	0.69		



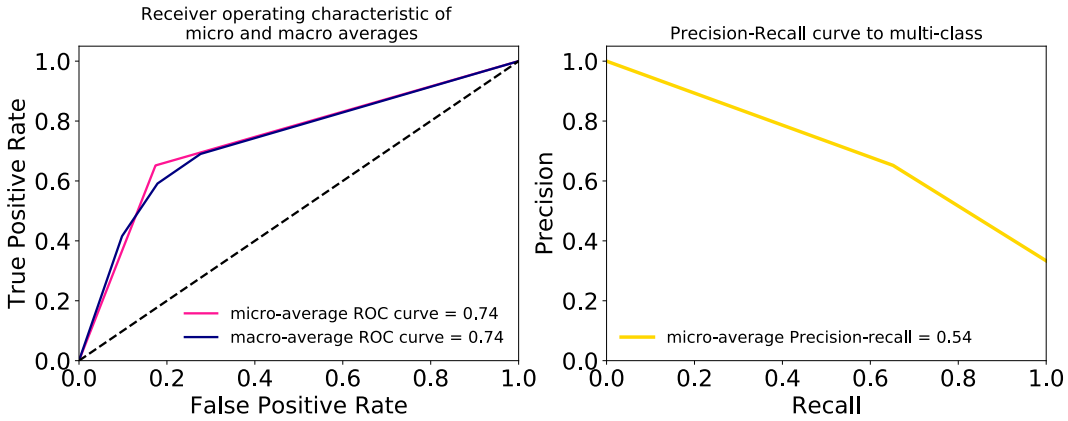
Importance	
alcohol	0.124239
density	0.106224
volatile acidity	0.100776
total sulfur dioxide	0.090634
free sulfur dioxide	0.088809
residual sugar	0.086608
chlorides	0.086587
pH	0.082699
citric acid	0.082001
sulphates	0.076000
fixed acidity	0.075422

Surprisingly, the importances of each feature are relatively close despite density and alcohol are the only two features with either positive or negative correlation to quality above 0.3. Therefore, no features are dropped as redundant for the algorithm.

K-Nearest Neighbor

The K-nearest neighbor algorithm uses a data point's nearest neighbors to make prediction. The algorithm relies heavily on correctly scaled dataset. The only hyperparameter that KNN needs to be optimized is the number of neighbors. The optimized KNN model also have recall values above 0.6 across all classes. However, the micro-average AUC PR curve for the model is fairly low. The log loss is also abnormally high at a value of 12.02.

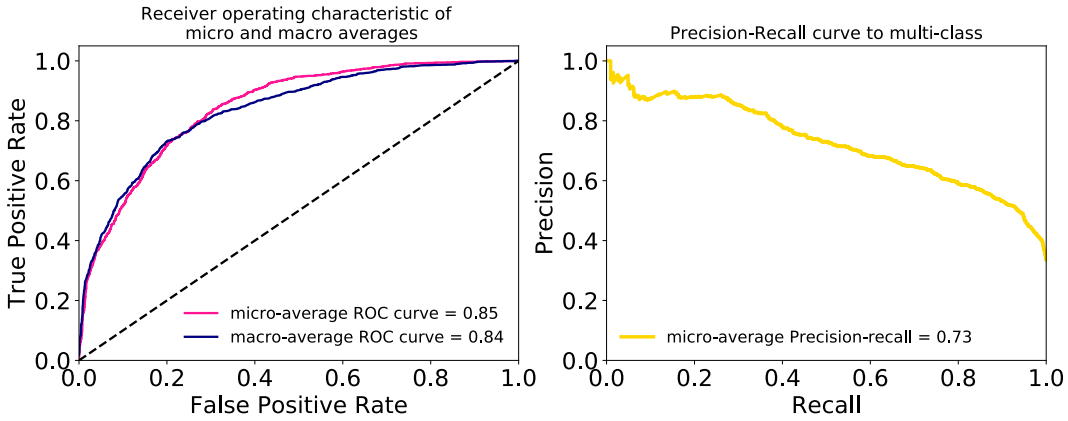
Class	Values			ROC AUC	PR AUC
	Precision	Recall	F1-score	0.74	0.54
0	0.65	0.69	0.67	Best score	Log loss
1	0.65	0.63	0.64	0.65	12.02
2	0.65	0.65	0.65		



Support Vector Machine

SVM is also another algorithm that uses distance between data points to make prediction. For this multi-class classification, one-vs-one scheme is used for modeling. Support vector machine has unique property of making predictions based on support vectors and do no assign probability estimation. In this case, the probability estimate is enabled in the implementation to calculate log loss.

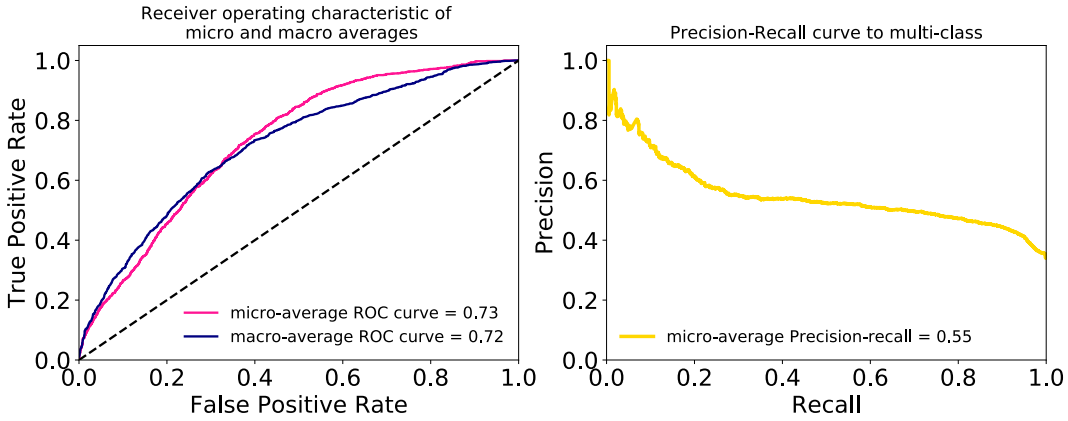
Class	Values			ROC AUC	PR AUC
	Precision	Recall	F1-score	0.85	0.73
0	0.68	0.64	0.66	Best score	Log loss
1	0.60	0.73	0.66	0.65	0.74
2	0.76	0.47	0.58		



Gaussian Naive Bayes

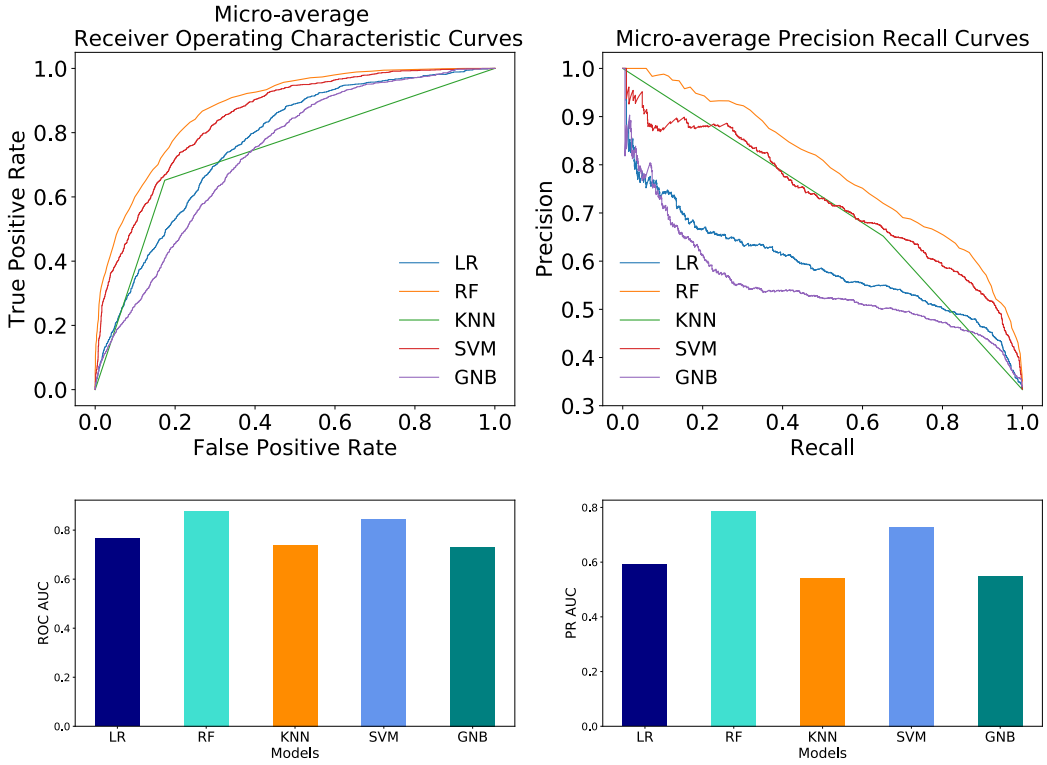
Gaussian Naive Bayes applies Bayes' theorem with the assumption of that all features are independent to each other. The algorithm uses prior probabilities and present data to make predictions, so there are no hyperparameter that needs to be optimized. This model produces poor prediction for mediocre wines.

Class	Values			ROC AUC	PR AUC
	Precision	Recall	F1-score	0.73	0.55
0	0.57	0.67	0.62	Best score	Log loss
1	0.54	0.38	0.45	0.52	1.07
2	0.45	0.60	0.51		



Model Comparison/Finding:

The dataset is relative small enough that training speed differences are not considerable for model optimization. Overall, random forest outperforms all other models in both ROC AUC and PR AUC. The figure below depicts the micro-average curves for all models. The worst performing models are Gaussian Naive Bayes and K-nearest neighbor with the lowest ROC AUCs and PR AUCs.



Conclusion:

The model prediction has been tested to have potential as an alternative tool for wine industry to sort wines into tiers by replicating human sensory assessment. To increase the model complexity, the dataset can include other designated origin or single varieties. The addition of designated origins would also introduce different durations of barrel aging and production methods as new features.

In addition, the target variable of the dataset used in this project is the median scores of multiple human assessors. If the raw sensory assessment data of each assessor can be included in the dataset, it would be interesting to explore the relationship between outliers and high degree of discrepancy among assessors.