Wine quality prediction project

Yuchen Hua
Data Science Institution, Brown University

Github Link: https://github.com/yuchen996/data1030-final-project

- **Introduction**

Dataset: The datasets I am using for this project contain two main parts, one is red wine and the other is white wine. The two datasets contain both physicochemical properties and sensory graded wine quality. White wine dataset has 4898 data points, and red wine has 1599 points. Each has 11 input features and "quality" as target variable.

Regression problem: The problem I want to investigate is that which physicochemical feature has an important influence on the wine quality. The target variable is "quality", which is sensory data from wine experts' evaluation. 0 means the quality is bad and 10 means excellent. This is originally a regression problem as I am interested in predicting numerical measurement of wine quality and how features affect the quality. By studying this problem, it might help improve wine quality, as well as figure out market needs. It would also be interesting to investigate if red and white wine have similar most important features affecting wine quality.

- **Explorary Data Analysis (EDA)**

Explore target variable

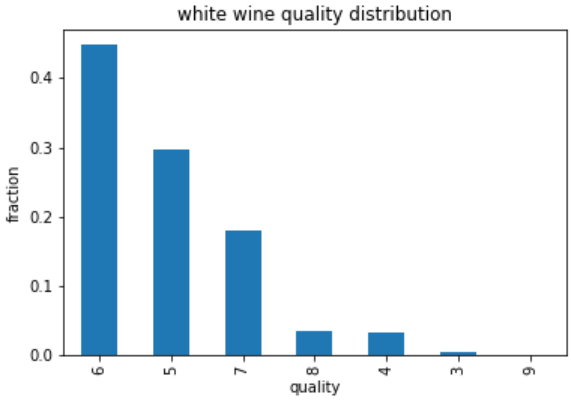


Figure 1 white wine quality distribution

According to the figures, the data is imbalanced as most of the wine quality are centered around normal scale (around 5 and 6), just a few are extremely bad (around 3) and extremely good (around9). Since red and white have the same pattern so that I just include one figure, except white wine has quality class 9.

Input variables distribution

As the input 11 features are all float types, it is appropriate to use MinMax or Standard Scaler for preprocessing. There are two kinds of distribution of features. One is tail-distributed and the other is normal distribution as showing in figure 2. I have used Standard Scaler for all input features.

Features 'alcohol', 'chlorides', 'citric acid', 'fixed acidity', 'free sulfur dioxide', 'residual sugar', 'sulphates', 'total sulfur dioxide', 'volatile acidity' are tailed distributed.

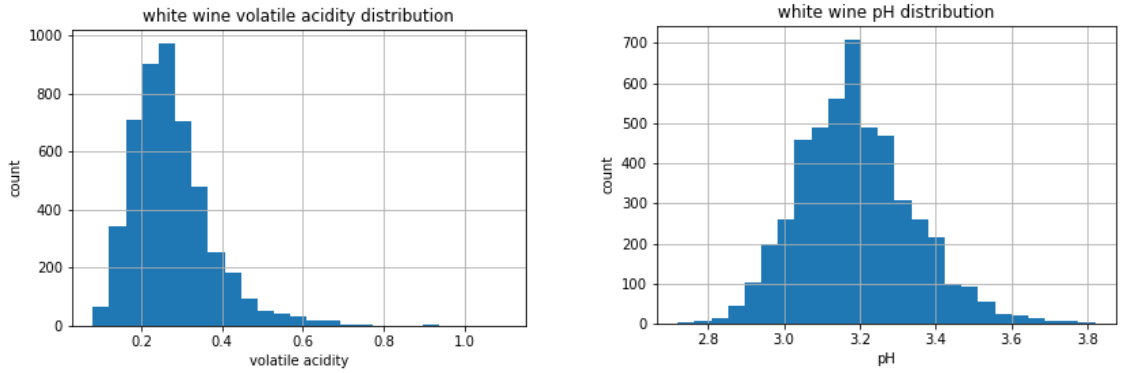
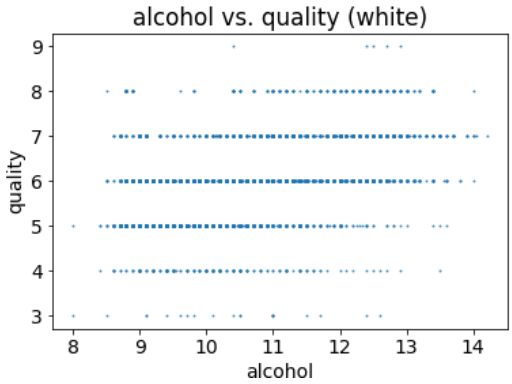


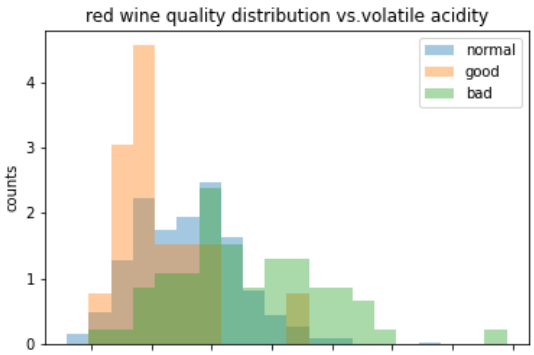
Figure 2 white wine features “volatile acidity” and “pH” distribution

Relations between each feature and target variable



The most linear related featured with target variable I found is alcohol. The data is ordered and rounded up to integers, but we could still see a linear relation between alcohol and quality.

Figure 3 white wine "alcohol" and "quality" relationship



For direct visualization, I classify the quality bad, normal and good. Except alcohol feature we have seen may have an influence on quality, another potential candidate is volatile acidity feature. From the figure we can see lower volatile acidity, better quality the wine is.

Figure 4 red wine quality distribution vs. "volatile acidity"

- **Methods**

Models: The models applied in the project are Lasso regression, Ridge Regression, ElasticNet, Random Forest, KNN (k nearest neighbors), SVR (support vector machine regressor) and XGBoost Regression. Each model’s performance will be presented in results session.

Splitting strategies: For investigating the influence due to randomly splitting, I compared several splitting strategies. The data is highly imbalanced, so that stratified is applied. I used cross-validation k-fold splitting method setting n_folds = 5. Also, I used 10 random states to minimize the effect of randomly splitting. I compared splitting with and without cross validation on XGBoost. I found cross-validation gives a better result.

Preprocessing: As the 11 features are all numerical features, I applied Standard Scaler to features for preprocessing. The data is iid, though it is imbalanced, each splitting used stratified method.

Parameter tuning: For each model, I tuned hyperparameters by gridsearchCV or ParameterGrid. For Lasso and Ridge regression, the tuned parameters is regularization parameter α. In ElasticNet, α and "l1_ratio" are tuned. In Random Forest," max_depth" and "max_features" are tuned as hyperparameters. In SVR, γ and C are tuned as hyperparameters, where γ is the kernel coefficient for default kernel 'rbf', and C controls the regularization. In KNN model, n_neighbors and weights are tuned as hyperparameters. Last but not least, in XGBoost regression, α, γ and max_depth are tuned as hyperparamertes.

Metrics: As predict a numerical number is a regression problem, the metrics are chosen accordingly as R2(R-square) and MSE(mean squared error), where R2 measures the data portion of which can be interpreted by the model, and MSE measures the ability of accurately predict the target variable.

In conclusion, the pipeline evolves data preprocessing (scale the data points as well as using cross-validation method split data into train, validation and test sets), model selection, parameter tuning, evaluation and interpret the models. I have chosen two models as Random Forest and KNN who perform the best with best selected hyperparameters, and also investigate the feature importance in the two models to interpret the models and analyze how to use the results help improve wine quality.

- **Results**
- **Model performance**:

I used DummyRegressor predicting mean value of quality as baseline. As I chose "mean" as baseline strategy, it is not surprising R-square for both white wine and red wine sets are 0. The baseline MSE for white wine dataset is 0.78, and for red wine is 0.65.

The results of each model are showing below in the table 1. Several models perform good. RandomForest has the highest R square score as 0.49 with a standard deviation as 0.04, and MSE is 0.33, which is the best among all models. After calculating many times as used different random states and cross validation folds, I found the best parameters as max_depth = 100, and max_features = 0.5. Compared to baseline MSE 0.65 for red wine, the RF model improved 0.32 MSE, which is 49.2%. For white wine, Random Forest R2 is 0.48 and MSE 0.38, which improved MSE 0.4 and 51.2%. KNN and XGB models perform not bad

too, but these two have a higher standard deviation fluctuation with higher MSE. Besides, XGB takes much longer for training. Linear regression models perform not too good on the dataset, as Ridge, Lasso and ElasticNet both have around 0.35 R2 and 0.41 MSE. Also, SVR takes too long to train too, and performs same as linear models.

model	R2	R2 std	MSE	MSE std	best parameter(s)
Ridge	0.34	0.033	0.41	0.021	alpha = 61.05
Lasso	0.35	0.041	0.42	0.014	alpha = 0.01
ElasticNet	0.34	0.022	0.42	0.021	l1 ratio = 0.99, alpha = 0.006
RandomForest	0.49	0.04	0.33	0.02	max depth = 100, max features = 0.5
SVR	0.37	0.049	0.41	0.032	C = 0.439, gamma = 0.072
KNN	0.45	0.04	0.35	0.026	weights = distance, n neighbors = 50
XGB	0.46	0.06	0.34	0.023	alpha = 0.1, lambda = 0.001, max depth = 30

Table 1 Red wine model performance

model	R2	R2 std	MSE	MSE std	best parameter(s)
Ridge	0.28	0.023	0.58	0.018	alpha = 61.05
Lasso	0.26	0.017	0.57	0.011	alpha = 0.001
RandomForest	0.48	0.049	0.38	0.012	max depth = 100, max features = 0.5
KNN	0.47	0.03	0.38	0.019	weights = distance, n neighbors = 30

Table 2 White Wine model performance

- **Splitting method comparison**
  I have tried simple splitting and cross-validation on the red wine dataset using the same model XGBoost. Without cross validation, the model has R2 highest as 0.381, although with cross validation to minimize the error due to randomly splitting can reach R2 highest as 0.525. It is obvious that cross-validation improved model performance.
- **Feature importance**
  For better understanding the model and prediction result, I investigate both global importance and local importance. Except global feature importance, the project also analyzed features interaction influence on model prediction.
  Permutation importance
  From figure 5, we can see the most important three features affecting model are “alcohol”, “sulphates” and “volatile acidity” on red wine dataset.

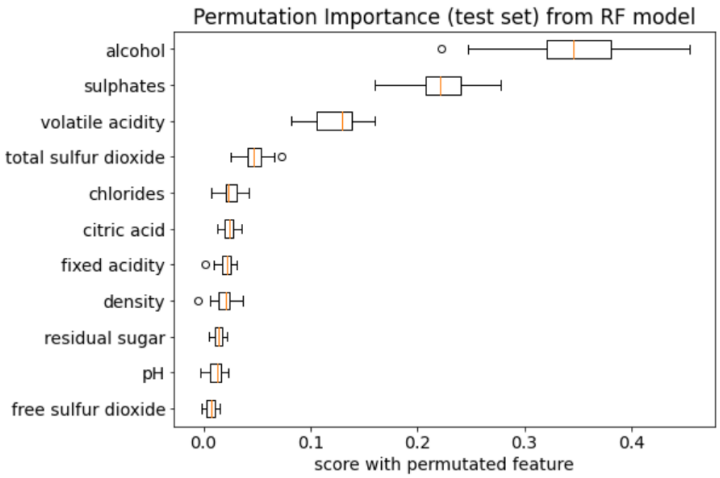


Figure 5 Permutation importance on red wine with RF

From figure 6, we can see the most important features of model prediction are “alcohol”, “volatile acidity” and “free sulfur dioxide” on white wine dataset.

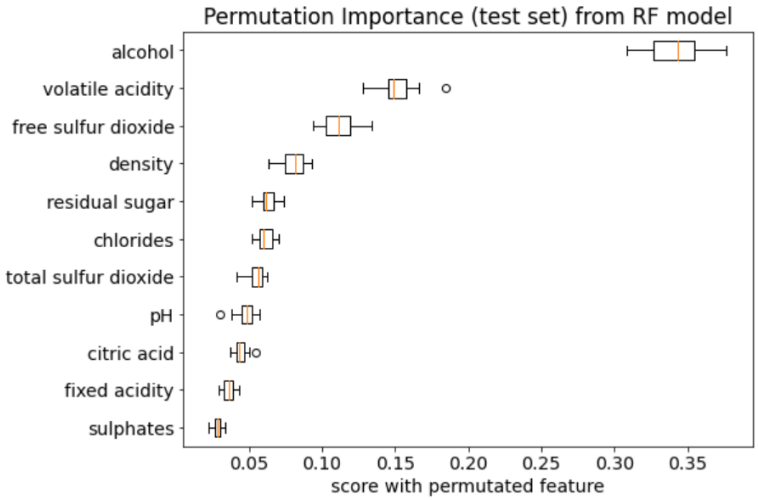


Figure 6 Permutation importance on white wine with RF

From the EDA session, we find that alcohol may have a dominant influence on wine quality, which has been approved, as no matter for white or red wine, alcohol level is the most importance feature influence the wine quality. We are guessing from EDA that density may have a significant influence on quality, however, it turns out that density does not have a significant influence on wine quality, especially on red wine.

Besides, from EDA session, we are guessing another potential candidate is volatile acidity. We could see the clear pattern that when volatile acidity level is low, it tends be have better quality of wine, and it turns out that volatile acidity is ranking high in both datasets towards how importance it is. It can be observed more clearly in shap values that decreasing volatile acidity increases wine quality.

Shap value

From figure 7**,** we can see for red wine**,** having a high “alcohol” is associated with high and positive values on wine quality, and same for “sulphates” . Although having high “volatile acidity”, “total sulfur dioxide”, “density”, “chlorides’ and “pH” is associated with high and negative values on target, which means increasing these feature values results in lower quality red wine.

Similarly, for white wine, from figure 8, having high “alcohol” level result in better quality of wine, though decreasing “volatile acidity” increases white quality. Also, increasing “free sulfur dioxide” does not improve quality much, but decreasing it really make it worse.

Compare the features influence each dataset from figure 7 and 8, we found that red wine and white wine does not have exact same important features. “Free sulfur dioxide” is important to white wine, though it does not matter for red wine. “Density” is important to white wine, though not crucial to red wine, etc.

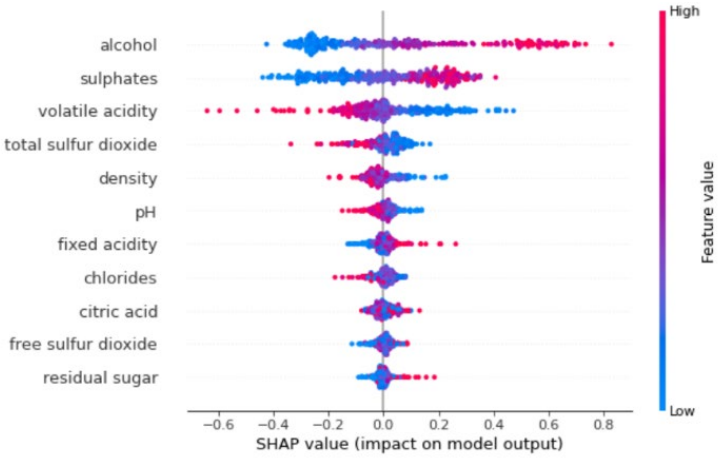


Figure 7 Global feature importance for red wine

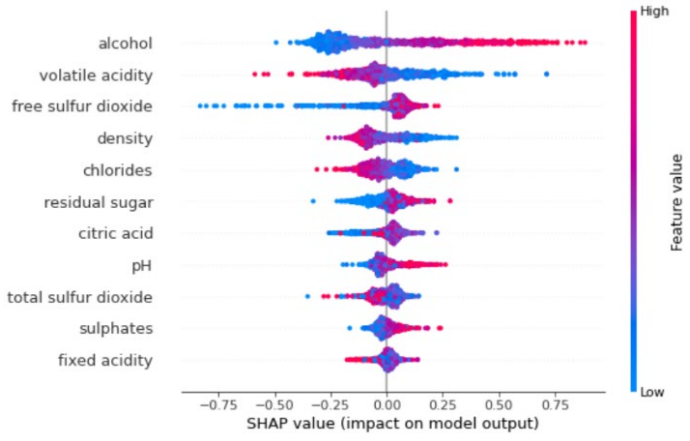


Figure 8 Global feature importance for white wine

Local feature importance

From figure 9, we could see generally increase “free sulfur dioxide” level will result better quality of white, though when “free sulfur dioxide” is between 0 and 2.5, higher “alcohol” level bring better quality of wine.

From figure 10, we see a clearer pattern of how “volatile acidity” and “alcohol” affect quality. When “volatile acidity” is below -0.5, lower “alcohol” level results in better quality wine, however, when “volatile acidity” becomes larger than -0.5, higher alcohol is associated with better quality. This is an interesting result, because though individually increasing “alcohol” level or decreasing “volatile acidity” result in better quality, by combining them, a lower level of “volatile” with a lower level of “ alcohol” actually result in better quality of white wine.

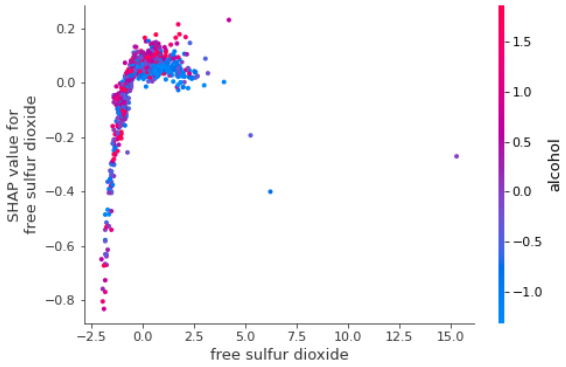


Figure 9 Free sulfur dioxide vs. alcohol for white

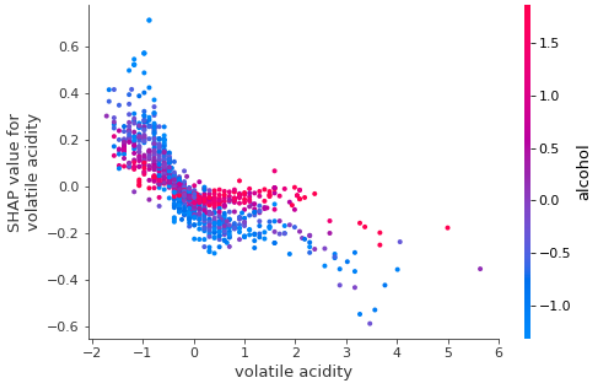


Figure 10 volatile acidity vs. alcohol relationship for white

Since for red wine the data points are much less, we barely see clear pattern in their local feature interactions. But from figure 11 we can still see the interaction between “volatile acidity” and “sulphates”. When “volatile acidity” is below -0.5, increasing “sulphates” brings about better quality. When “volatile acidity” goes up, lower sulphates level results in better quality, but generally decreasing the volatile acidity will have better quality of red.

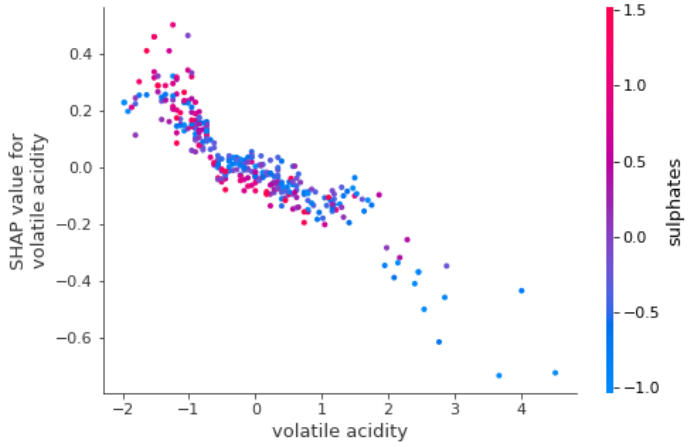


Figure 11 volatile acidity vs. sulphates for red wine

Classification results

I also implement Random forest classifier, baseline accuracy for red wine is 0.426, white wine 0.449. After tuning parameters and choosing best model, I got model accuracy result for red wine is 0.712, white wine 0.683. Accuracy has improved by model as 67.13% for red wine and 52.1%

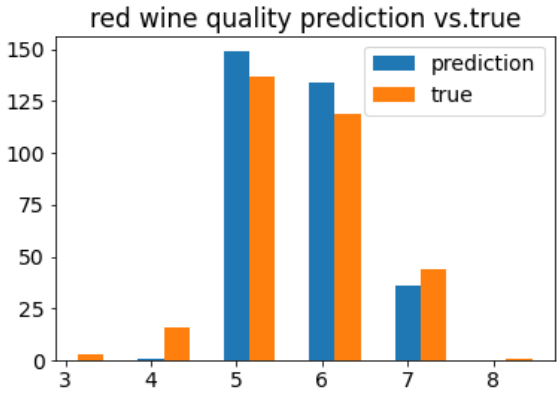


Figure 12 RF classifier prediction on red wine

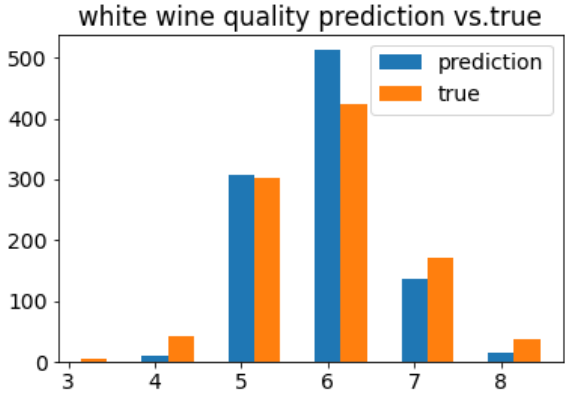


Figure 13 RF classifier prediction on white wine

◆ Outlook

From the prediction comparison figure 12 and 13, we can see the models fail to predict extrme quality wine. As the target variable is already rounded up to integer, it may be more accurate using classification method to predict quality, because we do not have original float data here, predicting numerical value may not be accurate. During this project, I have implemented Random Forest classifier as RF performs best above all others. But it worth to try other models, I may find other surprising findings, I will put it into future work.

In addition, it also worth to try XGBoost. It gives good result and we can take advantage of that it has more build-in functions for studying feature importance to better understand models, though it takes too long to train. If there is more time, I would like to study on the parameter tuning of XGBoost, may find a better performance model.

◆ references

Cortez, Paulo, et al. "Modeling wine preferences by data mining from physicochemical properties." *Decision Support Systems* 47.4 (2009): 547-553.

Data source: Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez

A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal @2009