Data Science Challenge Report (Xuqi LI)

1. Introduction

This report will briefly visualize and analyze the wine data and model result for two tasks, **wine quality prediction** and the **prediction of wine type**.

2.1 Wine quality prediction

First of all, we use histograms to plot the distribution of the wine quality for both white and red wine. Most of the score are distributed in the 5-6. The number of Good quality (>7) and worst quality wine(<4) is extremely less(tail in Figure 1 and 2).

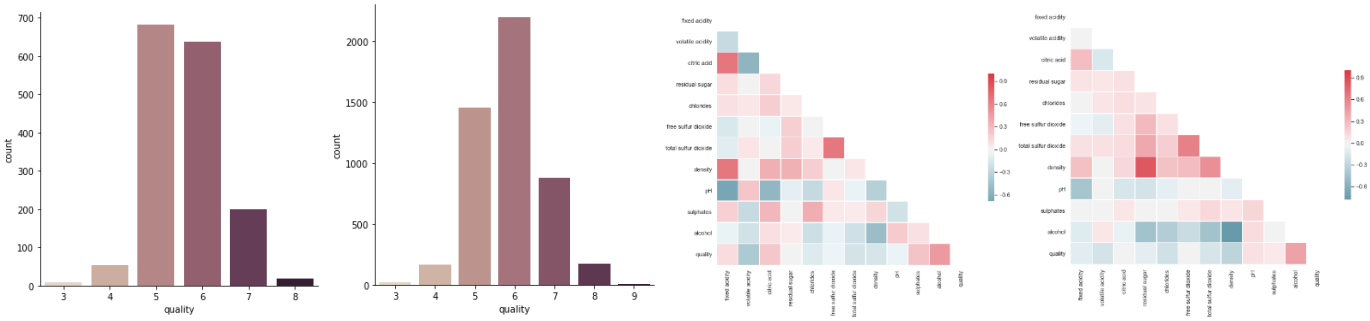


Figure 1&2&3 &4: (left,1) distribution of red wine, (left,2) distribution of white wine: (right,1) Correlation heat map of the red and (right,2) white wine.

Secondly, after visualizing the correlation of each attributes and quality score, it is easily to figure out that in Figure 3 the quality of red wine is much related to alcohol and volatile acidity (last row with deeper color). While in Figure 4 the quality of red wine is much related to alcohol and density (last row with deeper color).

2.2 Model and result

In this prediction task, we choose Support vector regression model to do regression of quality score. And use grid search for model parameter selection. Finally, the parameter of SVC for red wine are:

C=1, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape='ovr', degree=3, gamma=0.5, kernel='rbf', max_iter=-1, probability=False, random_state=8, shrinking=True, tol=0.001, verbose=False.

The training error and test error are listed in Table 1 and the visualization of the prediction and real is shown in Figure4.1:

Train_red:	0.8498659517426274
Test_red:	0.65625

Table 1. Error of the model.

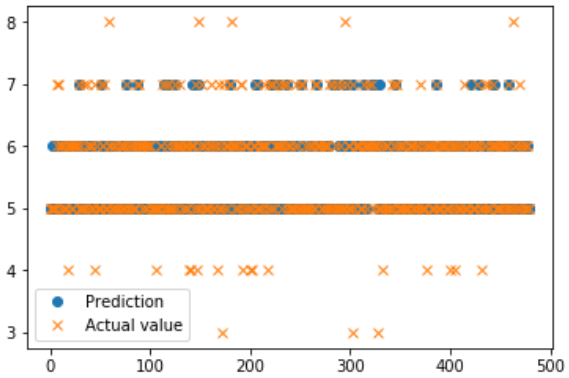


Figure 4.1: Result of model

3.1 Type Prediction

Firstly, we visualized the correlation of each attributes and type, it is easily to figure out that in Figure 5, type is highly related to **total sulfur dioxide**. And we later apply 5 classifiers to predict the type and then use cross-validation to calculate the test error to approximate the real error and the accuracy on test set is **0.993842341678145**, the validation measurement used is accuracy and f1-score. Finally, XGBoost model showed its advantage for this binary prediction task.

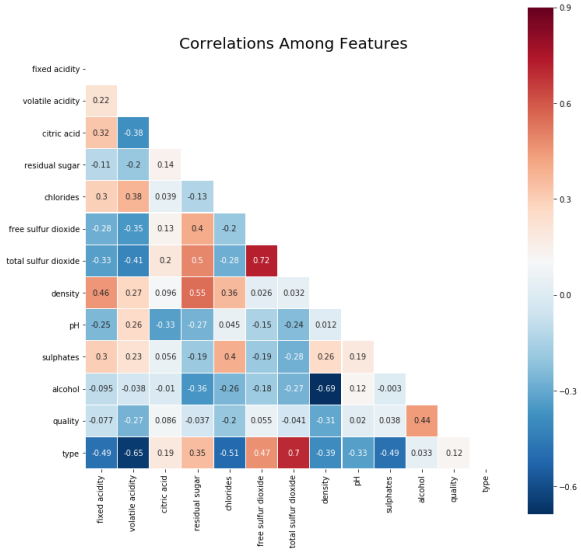


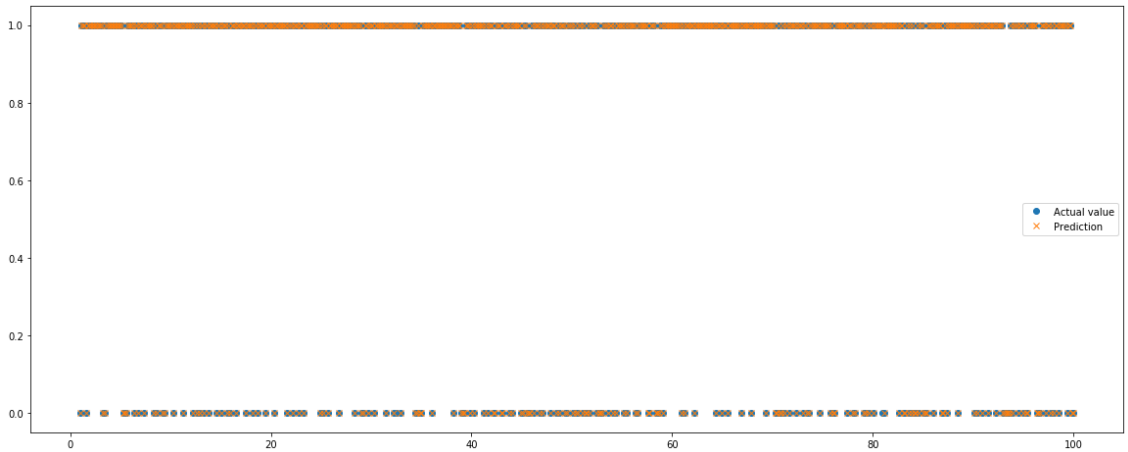
Figure 5: Correlation heat map of two types of wine.

The evaluation of five models is listed below, the XGBoost achieved the highest score on acc and F1, therefore XGBoost is chosen for this task:

	acc:	f1_score:
Decision Tree:	98.92307692307692%	99.26393270241851%
Random Forest:	99.53846153846155%	99.68321013727561%
KNeighbors:	94.61538461538461%	96.31190727081137%
GaussianNB:	98.0%	98.61554845580405%
SVC:	96.0%	97.29166666666667%
XGB:	99.6923076923077%	99.78902953586498%

3.2 The visualization of prediction and real

It is clear to figure out that almost all actual value(o) are covered by prediction(x) on testing set, our model shows the biggest advantage in this task. And f1-score(99.78) told us that this model perfectly balanced the recall and precision dilemma.



Reference:

The code (jupyter notebook file) is attached in the email, please kindly check it, thanks!