Wine Quality Prediction Analysis

1. Problem Statement

This wine dataset shows poor wine quality (below 7). Based on the physicochemical features and quality information, we want to make a machine learning model to predict wine quality, and implement corresponding quality control to product good wine (quality >=7).

2. Data Wrangling

Two data sets red wine and white wine was merged into one. No missing values in the data set. The duplicate data was removed before visualize the data. The data type of "quality" was corrected into "category", with the range of 1 to10. Visualization of the target "quality" distribution revealed the dominate wine quality are 5, and 6 with 76%, the good wine (quality >=7) is about 18%, suggesting the data set is imbalanced. The distribution of quality with each numeric feature revealed a promising pattern of quality with alcohol, the good wine seems have relative higher alcohol. The heatmap among numeric features suggests "free_sulfur_dioxide" and "total_sulfur_dioxide" are strongly positive related,density and alcohol are negatively correlated. "free_SO2_ratio" was add into the dataframe for further analysis.

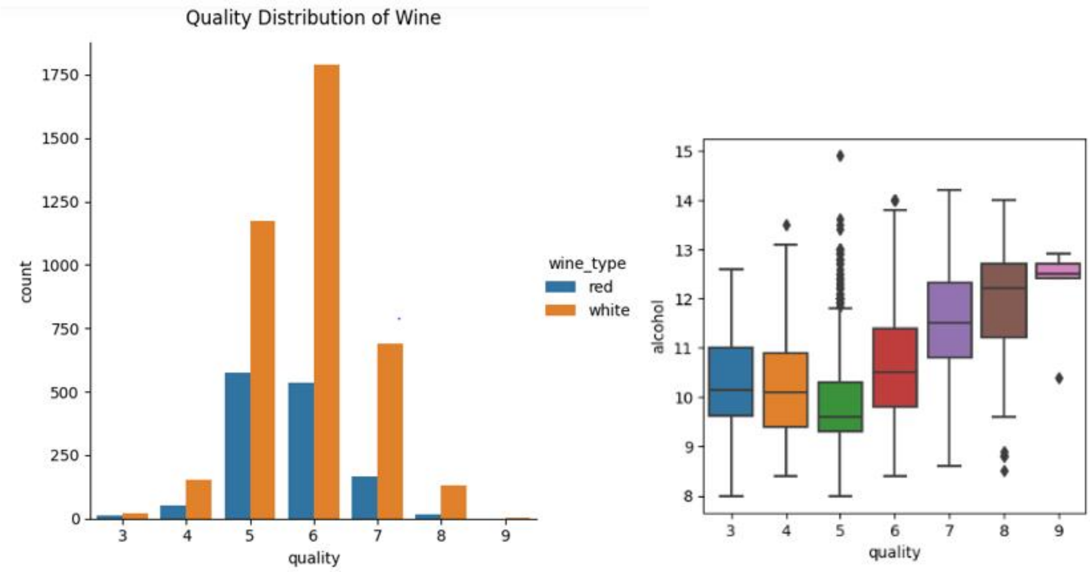


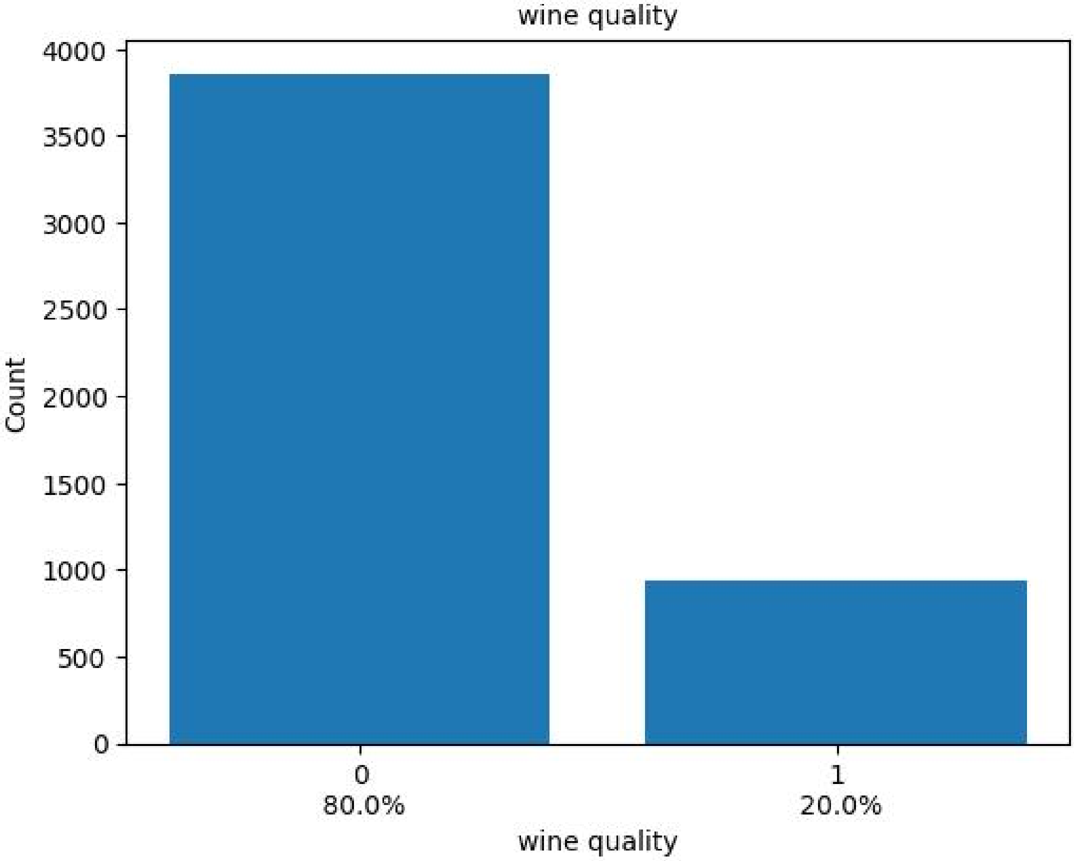
Figure 1 (Left) wine quality distribution, (right) alcohol vs wine quality

3. Data Exploratory

The scatter plot among numeric features reveals the strong positive correlation of density vs residual_sugar,density vs chlorides, density vs fixed_acidity, total_sulfur_dioxide vs residual_sugar, free_sulfur_dioxide vs residual_sugar. Density vs alcohol, citric_acid vs voltalie_acidity, total_sulfur_dioxide vs volatile_acidity shows negative correlation. The outliner of each column was defined as values larger than 3*std. For outliners with the quality >=7 were kept while all others were removed from the data set. New features alcohol/sugar ratio, sugar/acidity ratio, fixed_acidity percentage added for future modeling.

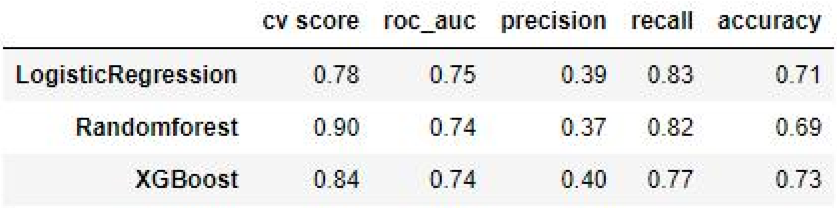
4. Data Preprocessing

Before modeling, the categorical feature wine type was transformed to int type through panda.get_dummies(). The target feature “quality” was bucket into binary classes: 0 (poor quality wine) and 1 (good quality wine). The features were scaled using MinMaxScaler(). The dataset was split into train and test with test_size of 0.3.The target of modeling is to predict good quality wine. To addressing the imbalance of the dataset (3859 class 0 and 940 class 1),SMOTE was used to resampling more class1. , SMOTE was used to resampling more class1 in training data set before modeling.

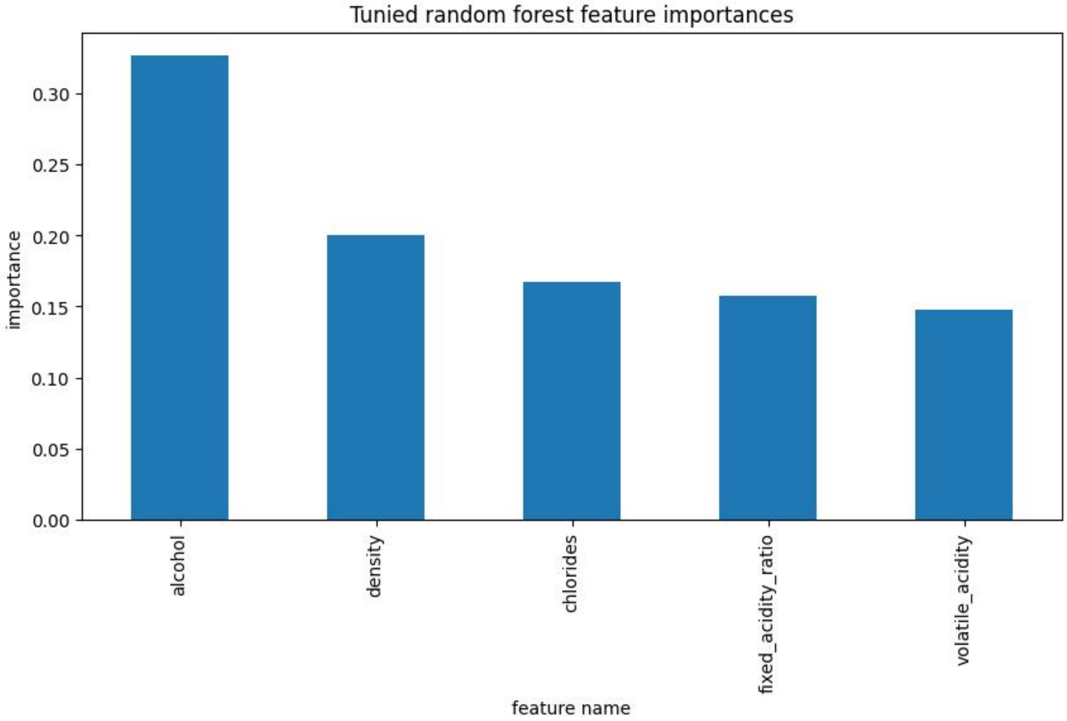


5. Modeling

Based on the purpose of prediction of class1 wine quality, the recall score was chose as the main model performance metrics. Other metrics of each model such as precision score, F1 score, accuracy and roc_auc score were also reported. The DummyClassifier is used for the baseline model, which had a recall of 0.31. Three modeling: Logistic Regression, Random Forest, XGBoost were applied through a pipe line with SelectKBest (k=5). For each model, the modeling parameters was tuned and cross validated with kfold = 5. The random forest model was selected as the best model based on the overall metrics. The features alcohol in all models revealed to be the most important feature for good wine. density, fixed_acidity_ratio Chlorides, and the volatile_acidity were also the top rank feature in wine quality. The logistics regression model revealed the positive importance of alcohol and density, while strong negative importance of chlorides and volatile_acidity in good wine quality.



	cv score	roc_auc	precision	recall	accuracy
LogisticRegression	0.78	0.75	0.39	0.83	0.71
Randomforest	0.90	0.74	0.37	0.82	0.69
XGBoost	0.84	0.74	0.40	0.77	0.73



6. Future Work

- only five feature were used in current modeling, may try more features to feed a model in future.
- in the data preprocessing, the distribution of each numeric features was not fully explored, some of features are not normally distributed, without further process, it maybe the cause of low precision in the modeling. I'd like to do more data exploratory if time allowed in future.
- The modeling reveals alcohol play important role in producing good wine quality. If more data provided, I'd like to