Wine Quality Prediction

In Depth Analysis ML

OVERVIEW

After EDA and Statistical testing, we are finally on the ML section of the project. Here I try to build a classifier which accurately predicts the quality of wines as good (rated 7 or above) and poor (rated < 7).

MACHINE LEARNING

In the wines dataset, there were 1382 red wines of “poor quality”, rated below and 7 and 217 of high quality. 3838 white wines were “poor quality” and 1060 of high quality.

We used sklearn’s train_test_split to split the already standardised dataset into train and test sets.

The models were trained using sklearn’s Logistic Regression, XGBoost, Decision Tree and Random Forest.

All models were tuned for hyper parameters to get the best possible performance. GridSearchCV was used for hyper parameter tuning. AUC score was used to evaluate the model performance.

Summary

A summary of all the models is displayed below.

Red wines

Model Type	Parameters with Grid Search	AUC score
Logistic Regression	'C: 5.179474679231202, penalty: l2	0.8033
XGBoost	learning_rate=0.05, max_depth=4, n_estimators=180	0.9089
Decision Tree	max_depth=4, max_leaf_nodes=6, min_samples_leaf=1, min_samples_split=2	0.7652
Random Forest	‘max_depth : 20, 'max_features': 0.25, 'min_samples_split': 2, 'n_estimators': 150	0.9182

White wines

Model Type	Parameters	AUC score
Logistic Regression	‘C': 268.2695795279727, 'penalty': 'l2'	0.8033
XGBoost	learning_rate=0.1, max_depth=3, n_estimators=100	0.9071
Decision Tree	max_depth=13, max_leaf_nodes=41, min_samples_leaf=1, min_samples_split=2	0.7434
Random Forest	'max_features': 0.25, 'min_samples_split': 2, 'n_estimators': 250	0.9190

The best performing model for red and white wines, was the Random Forest classifier. For both the wine types, Random Forest is able to correctly classify high and poor quality wines, ~92% of the time!

Classification report and confusion matrix of random forest model for red wines:

```
              precision    recall  f1-score   support

         0.0       0.93      0.98      0.96       420
         1.0       0.82      0.52      0.63        60

    accuracy                           0.93       480
   macro avg       0.88      0.75      0.80       480
weighted avg       0.92      0.93      0.92       480

[[413   7]
 [ 29  31]]
```

The random forest model for red wines has a precision of 0.82 for the positive class, it means when the model predicts a wine to be of high quality, it is correct 82% of the time! A recall of 0.52 for the positive class means, the model correctly identifies 52% of all high quality red wines in the dataset. As can be seen from the confusion matrix, out of 480 wines of the test set, the model predicts 29 red wines as FNs and 7 as FP. It means the model tends to classify high quality red wines (wines rated < 7) as low quality.

Classification report and confusion matrix of random forest model for white wines:

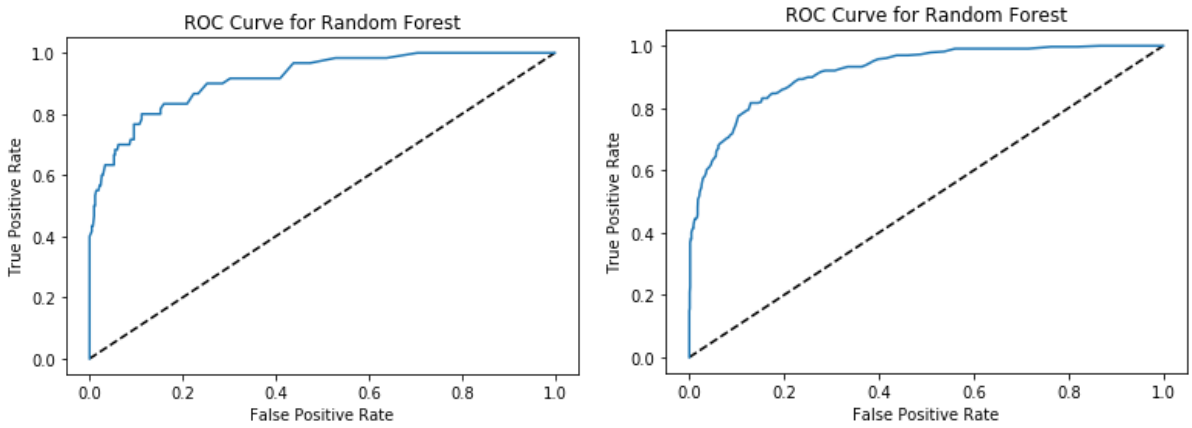
```
              precision    recall  f1-score   support

         0.0       0.88      0.97      0.93      1143
         1.0       0.86      0.55      0.67       327

    accuracy                           0.88      1470
   macro avg       0.87      0.76      0.80      1470
weighted avg       0.88      0.88      0.87      1470

[[1114   29]
 [ 147  180]]
```

The hyper parameter tuned, random forest model for white wines has a precision of 0.86 for the positive class, it means when the model predicts a wine to be of high quality, it is correct 86% of the time! A recall of 0.55 for the positive class means, the model correctly identifies 55% of all high quality white wines in the dataset. As can be seen from the confusion matrix, the model predicts more FNs (147), compared to FP (29). In this case it could indeed be better to get a high quality wine predicted as low quality, compared to a low quality wine being predicted as high quality!



Red and White wine ROC Curves for Random Forest classifier model.

Custom Thresholding on random forest model

The default threshold for all ML models is 0.5. I tried to change it in order to get better model performance.

Results for red wine thresholding

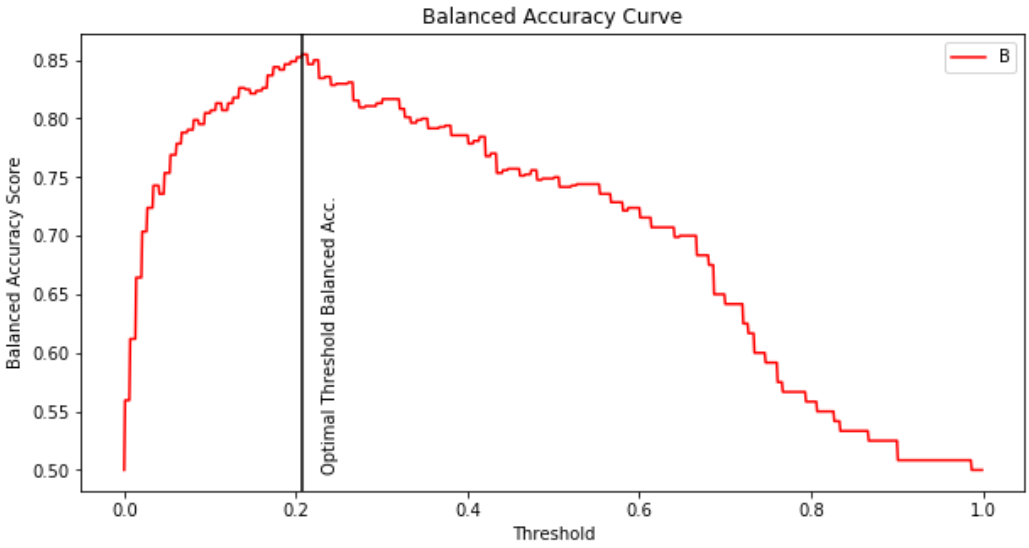
```
Threshold=0.207, Balanced Accuracy Score=0.85476
              precision    recall  f1-score   support

         0.0       0.97      0.88      0.92       420
         1.0       0.49      0.83      0.62        60

    accuracy                           0.87       480
   macro avg       0.73      0.85      0.77       480
weighted avg       0.91      0.87      0.88       480

[[368  52]
 [ 10  50]]
```

After checking the balanced accuracy score for thresholds ranging from 0 to 1, a threshold value of 0.207 gave the maximum balanced accuracy score of 0.85476. The recall for positive class now increased to 0.83, from 0.52 and precision decreased from 0.82 to 0.49. The model now misclassified 52 low quality red wines as high quality and 10 high quality red wines as low quality.



Results for white wine thresholding

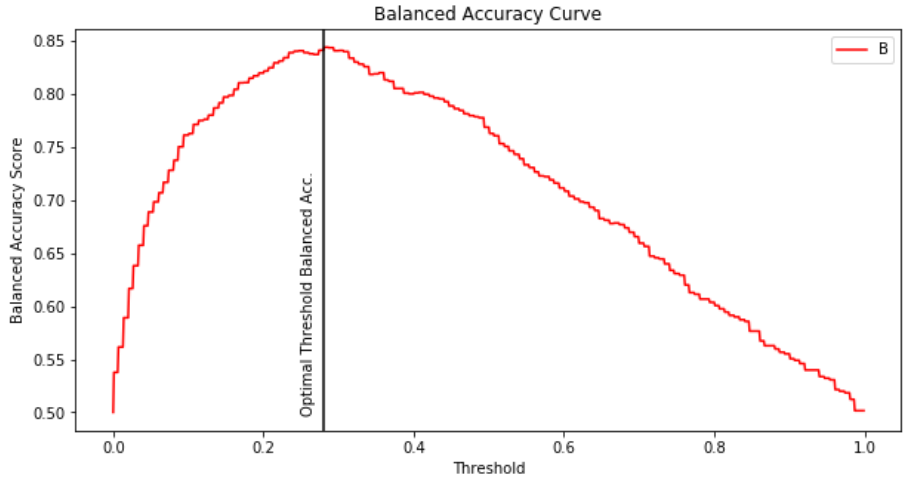
```
Threshold=0.281, Balanced Accuracy Score=0.84372
              precision    recall  f1-score   support

         0.0       0.95      0.86      0.90      1143
         1.0       0.62      0.83      0.71       327

    accuracy                           0.85      1470
   macro avg       0.78      0.84      0.81      1470
weighted avg       0.87      0.85      0.86      1470

[[978 165]
 [ 55 272]]
```

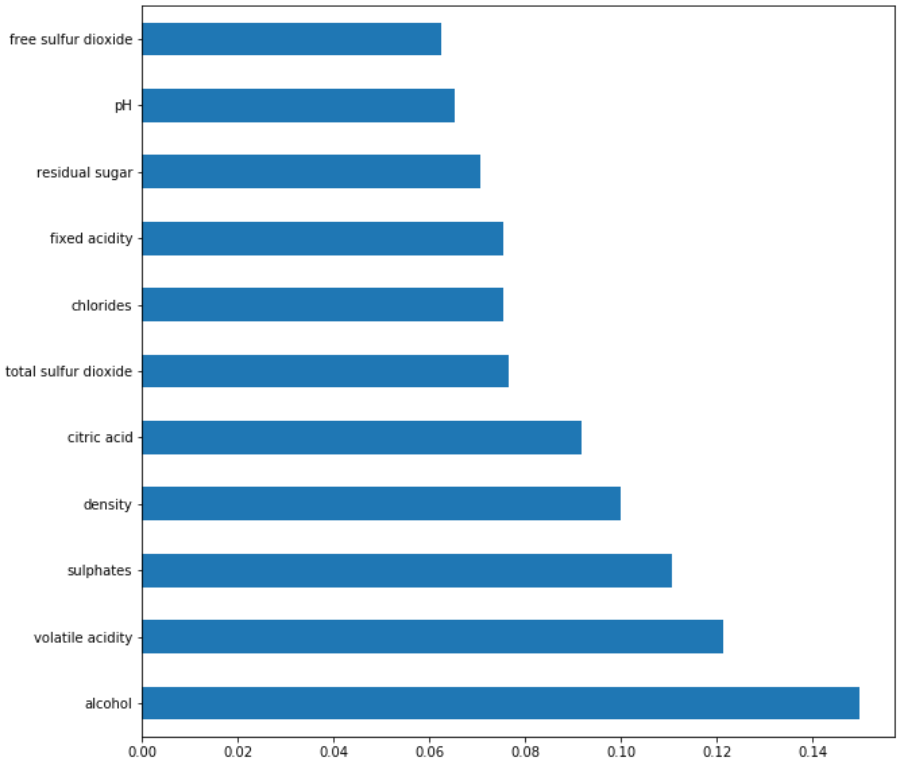
A maximum balanced accuracy score of 0.8472 was obtained for a threshold of 0.281. Precision reduced to 062 but recall increased to 0.83. The F1 score improved to 0.71 from 0.67. The model misclassified 165 white wines as high quality, when they were low quality and 55 white wines as low quality when they were of high quality.



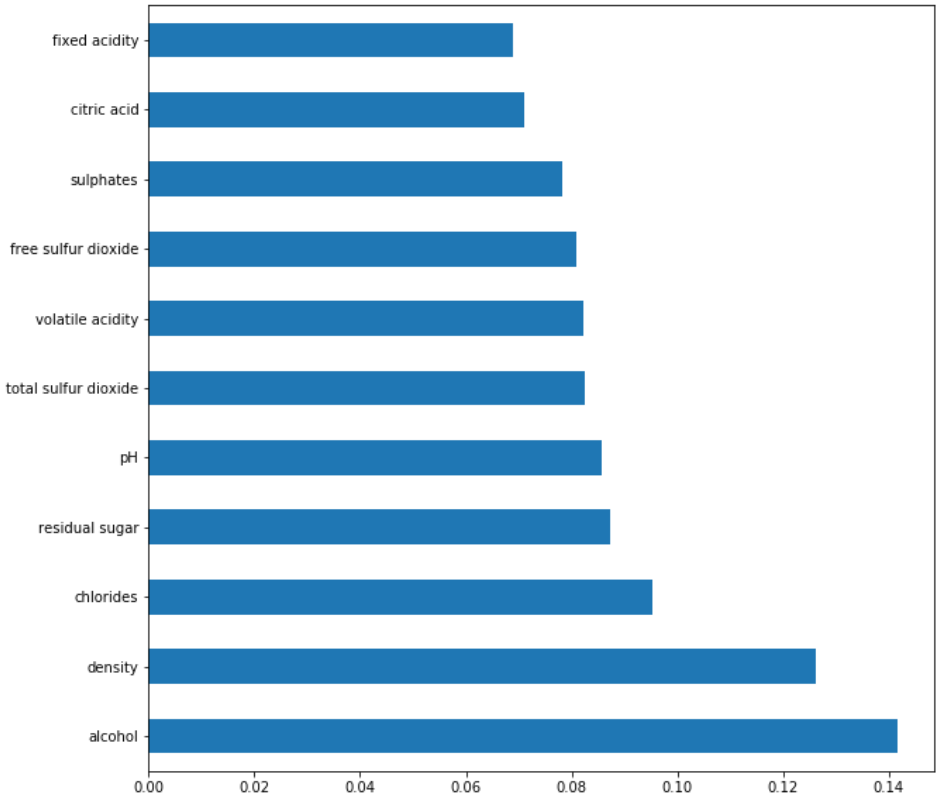
After thresholding, the model for red and white wines had an increase in recall value and a slight decrease in precision, but the accuracy did not see any improvement. It also had more misclassifications when compared to the earlier default threshold models.

Feature Importance using Random Forest

On the basis of the model built, I wanted to understand which wine features have the most influence on its quality. The feature_importances_ attribute of Random Forest Classifier was used for this purpose.



As in the image seen above, alcohol and volatile acidity have the most influence on the quality of red wines.



Alcohol and density have the most effect on the quality of white wines!

CONCLUSION

From the correlation matrix, box plots, Tukey test, logistic regression and the machine learning analysis, it is clear that alcohol content has the largest effect on the quality of red and white wines! Sulphate content is another important deciding factor for red wines whereas chlorides and residual sugars could play an important role in white wine quality. The odds ratio was high for pH and sulphates but the feature_importances_ attribute of random forest differed from that.