Capstone 2: Project Report

Del Wester

2/1/2021

Wine Quality

Many wine brands are seeking new ways to maximize the success of their wines. Before making any decisions, it might be helpful to know which features contribute to a wine's quality. Knowing these features can enable a brand to make more intelligent decisions when making it. But what exactly are these features? Using ML techniques with wine data retrieved from the following website, I plan to answer this question. https://archive.ics.uci.edu/ml/datasets/wine+quality

Data Wrangling

The dataset for this project was wrangled by another party prior to beginning this project.

Using red and white wine samples, inputs include objective tests (PH values) and the output is based on sensory data (wine tasting by experts). Using a median of at least 3 evaluations, each expert graded the wine quality between 0 (very bad) and 10 (very excellent). Several data mining methods were applied to model these datasets under a regression approach to determine wine quality.

Objective: Find if any of the features other than quality can be used to distinguish quality.

Two datasets were created, using red and white wine samples. The input included objective tests (e.g. PH values) and the output was based on sensory data (median of at least 3 evaluations made by wine experts). Each expert graded the wine quality between 0 (very bad) and 10 (very excellent). Several data mining methods were applied to model these datasets under a regression approach. The support vector machine model achieved the best results. Several metrics were computed: MAD, confusion matrix for a fixed error tolerance (T),etc. Also, we plotted the relative importance of the input variables (as measured by a sensitivity analysis procedure).

The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.). These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are munch more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines.

Number of Attributes: 12 + output attribute

Attribute information:

Input variables (based on physicochemical tests):

1 - fixed acidity

2 - volatile acidity

3 - citric acid

4 - residual sugar

5 - chlorides

6 - free sulfur dioxide

7 - total sulfur dioxide

8 - density

9 - pH

10 - sulphates

11 - alcohol

12 - type

Output variable (based on sensory data):

13 - quality (score between 0 and 10)

Missing Attribute Values: None

Exploratory Data Analysis

The Wine Quality data includes mostly continuous data with a few categorical columns. Exploratory data analysis can be used to derive relationships between the wine quality and the various features available from the wine's profile and suggest improvements to the profiles that would increase the wine's quality. This analysis takes into consideration a certain spectrum of wines related to the two specific types – red and white.

The quality feature of this set was the result of wine tasters opinions, ranging from 3 to 9, with the higher numbers being higher quality. The spread is shown below:

"quality" value counts:

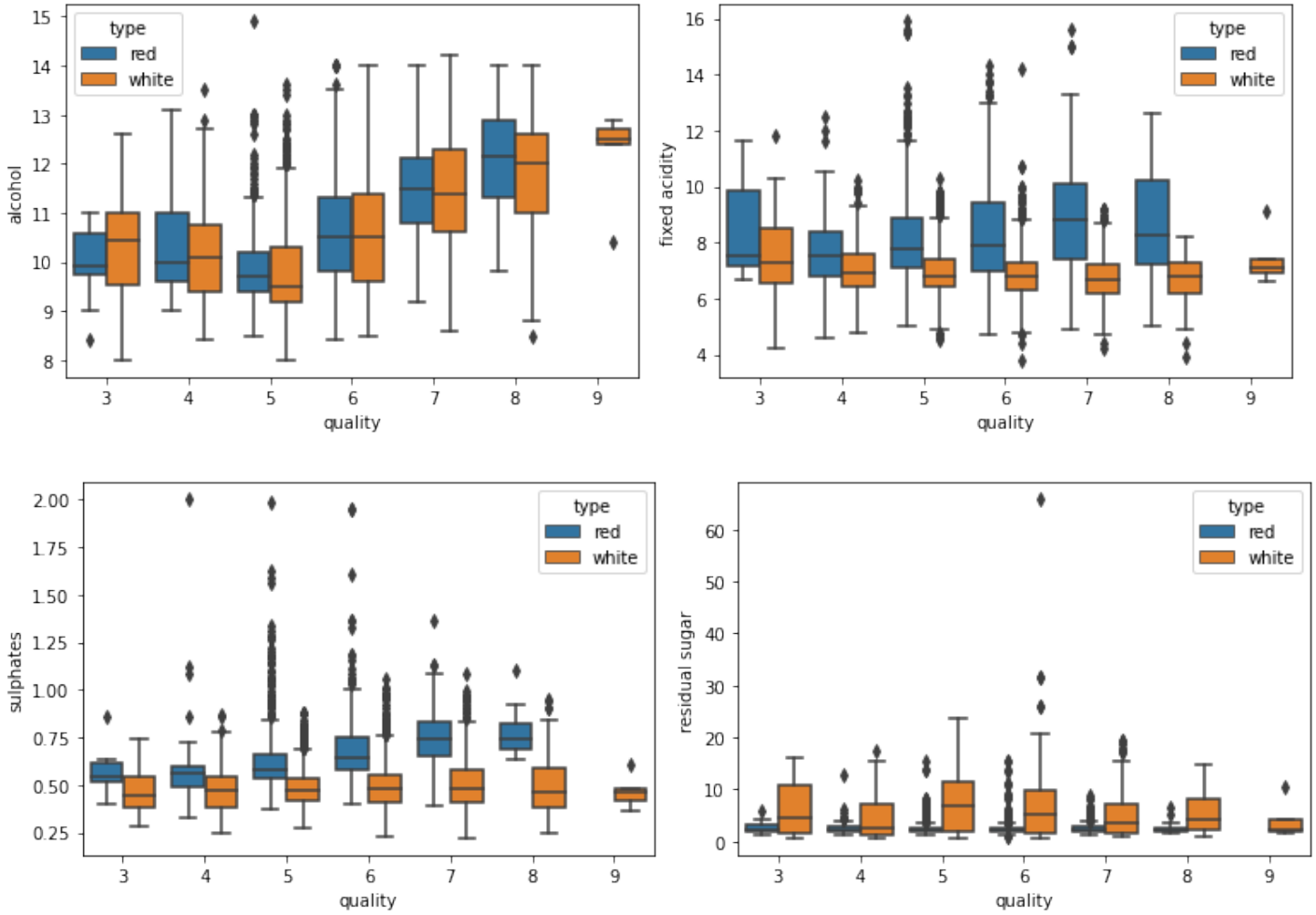
6 2836
5 2138
7 1079
4 216
8 193
3 30
9 5

Type Distribution - Red: 1599 White: 4898

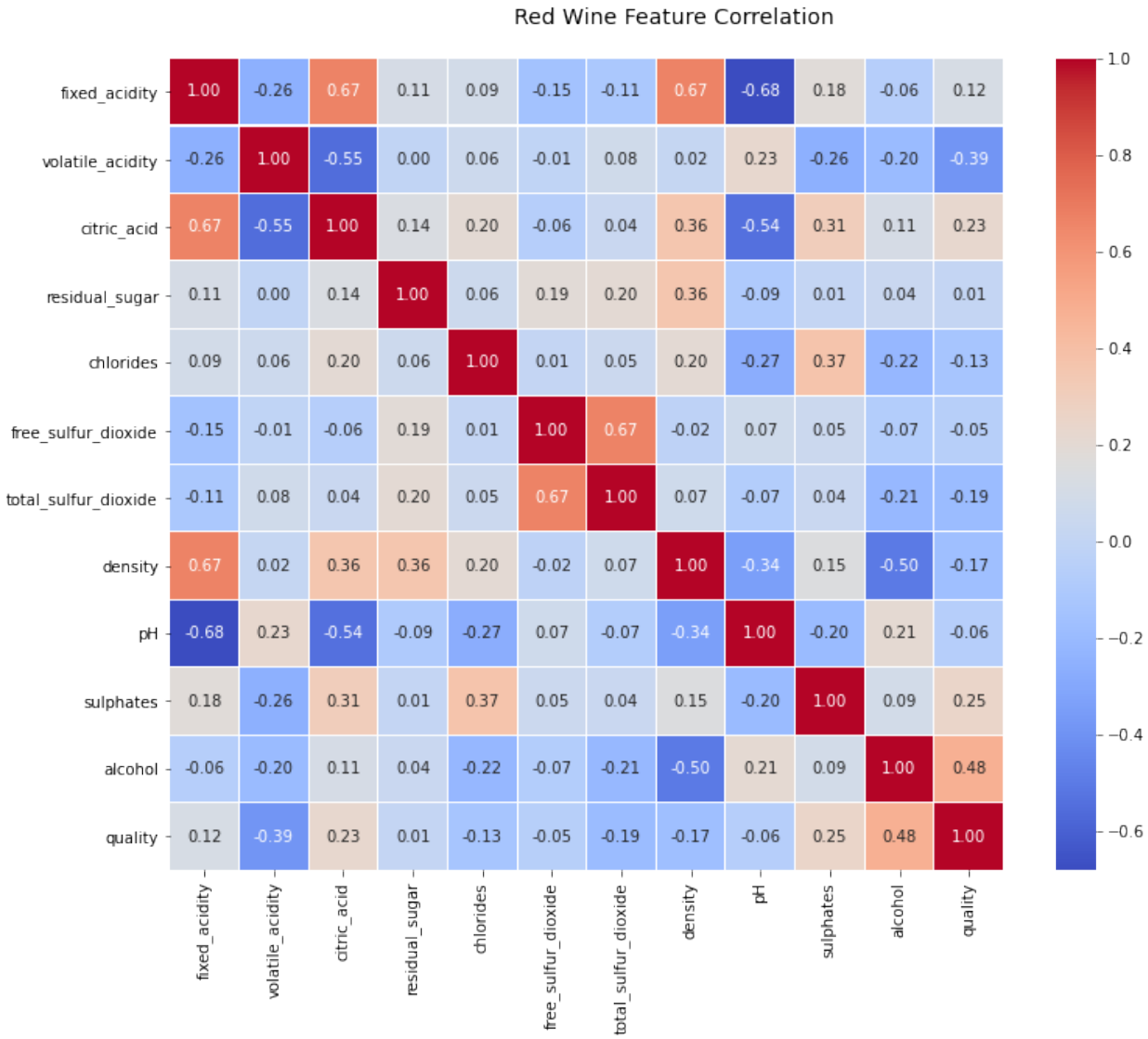
These will be split into a binary feature in the preprocessing step, with 7 and higher being high quality and everything else being lower quality. The plots for the output variable vs each of the top important features are below.

Red – alcohol, fixed acidity, sulphates

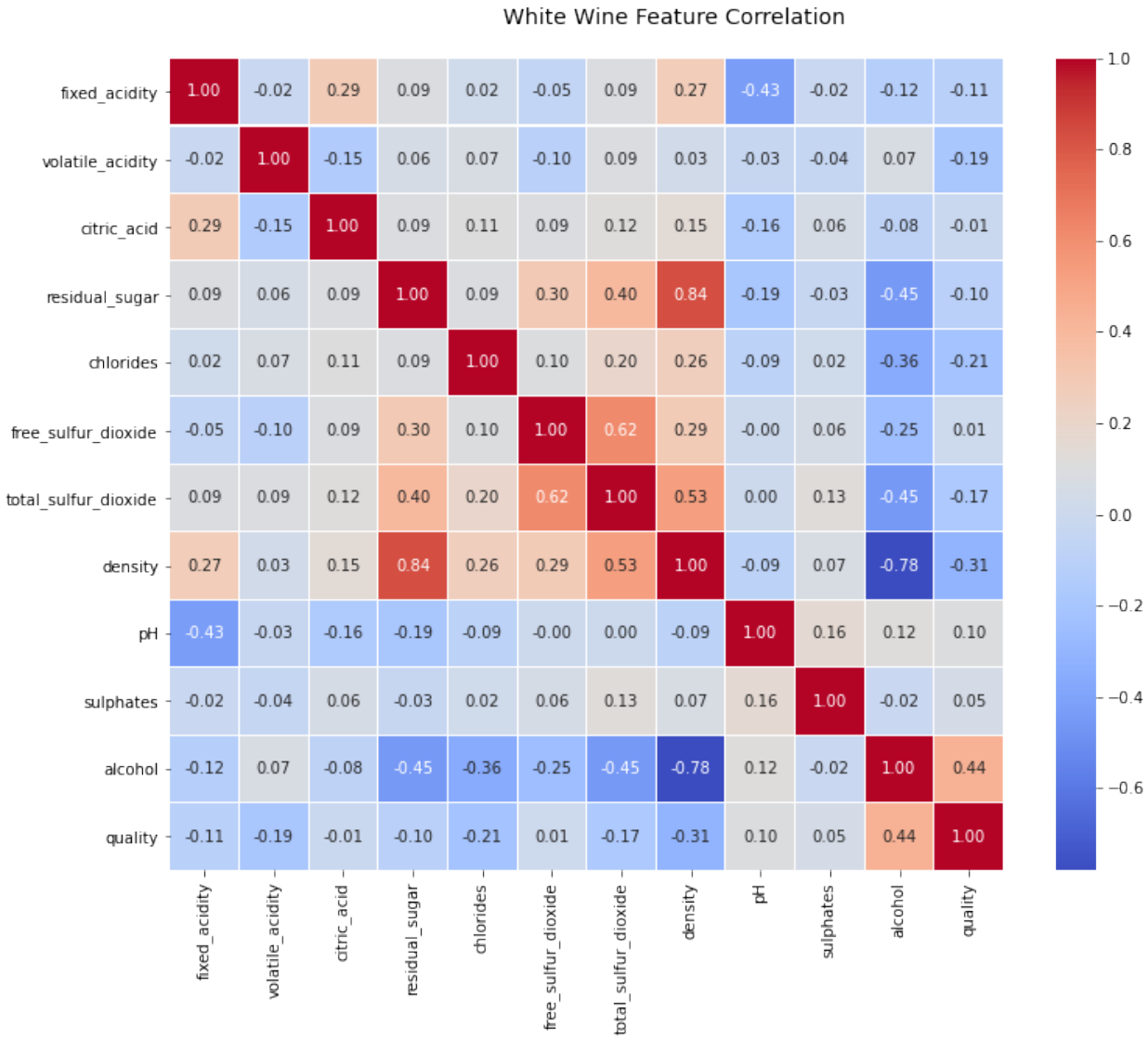
White - alcohol, fixed acidity, residual sugar



Correlation Matrix



Reds appear to have correlations between fixed acidity and pH, fixed acidity and density, fixed acidity and citric acid, and free sulfur dioxides and total sulfur dioxides.



Whites have correlations between residual sugar and density and alcohol and density.

Preprocessing and Training

I started this section by splitting the data into high and low quality wines, with quality 7 and higher being high and the rest low, as well as segregating by type: red and white. 15 percent of the red wines in this data set are high quality and 25 percent of the whites. The data was then scaled using a Robust Scaler and I ran another correlation check to find any features to leave out of the final set due to high correlation as well as which features were of highest importance. Since highly correlated features can skew the data during modeling, I dropped any features that were highly correlated with the most important features for each type.

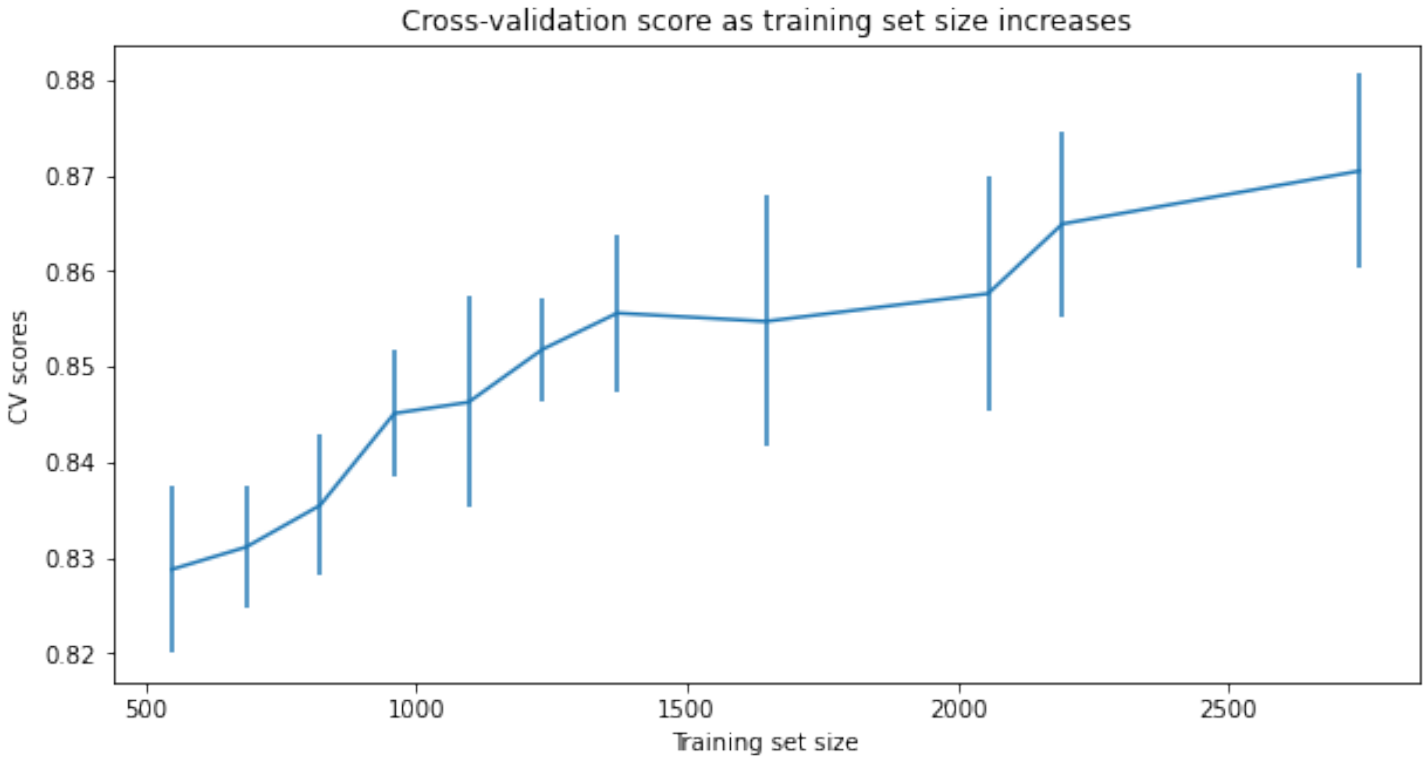
Dropped from Reds: volatile_acidity, citric_acid, total_sulfur_dioxide, density

Dropped from Whites: density

I then tested and trained a model to find the optimal training set size.

Red: 700

White: 2500



Modeling

I then tested and scored the following models on both sets of data: DecisionTreeClassifier, RandomForestClassifier, GradientBoostingClassifier, LogisticRegression, KNeighborsClassifier, GaussianNB and SVC. The Random Forest Classifier scored significantly higher on both datasets than the others and was what I went with for the final model.

Decision Tree Classifier
Red: 0.83
White: 0.80

Random Forest Classifier
Red: 0.89
White: 0.89

Gradient Boosting Classifier
Red: 0.85
White: 0.83

Logistic Regression
Red: 0.88
White: 0.79

KNeighbors Classifier
Red: 0.87
White: 0.84

GaussianNB Classifier
Red: 0.84
White: 0.70

SVC Classifier
Red: 0.88
White: 0.83

I performed a gridsearch hyperparameter tuning for the Random Forest model as well determine the ROC_AUC scores and ROC curve.

Red:
RandomForestClassifier(max_depth=4, min_samples_leaf=2, n_estimators=72)
0.83

White:
RandomForestClassifier(max_depth=4, min_samples_leaf=2, n_estimators=72)
0.83

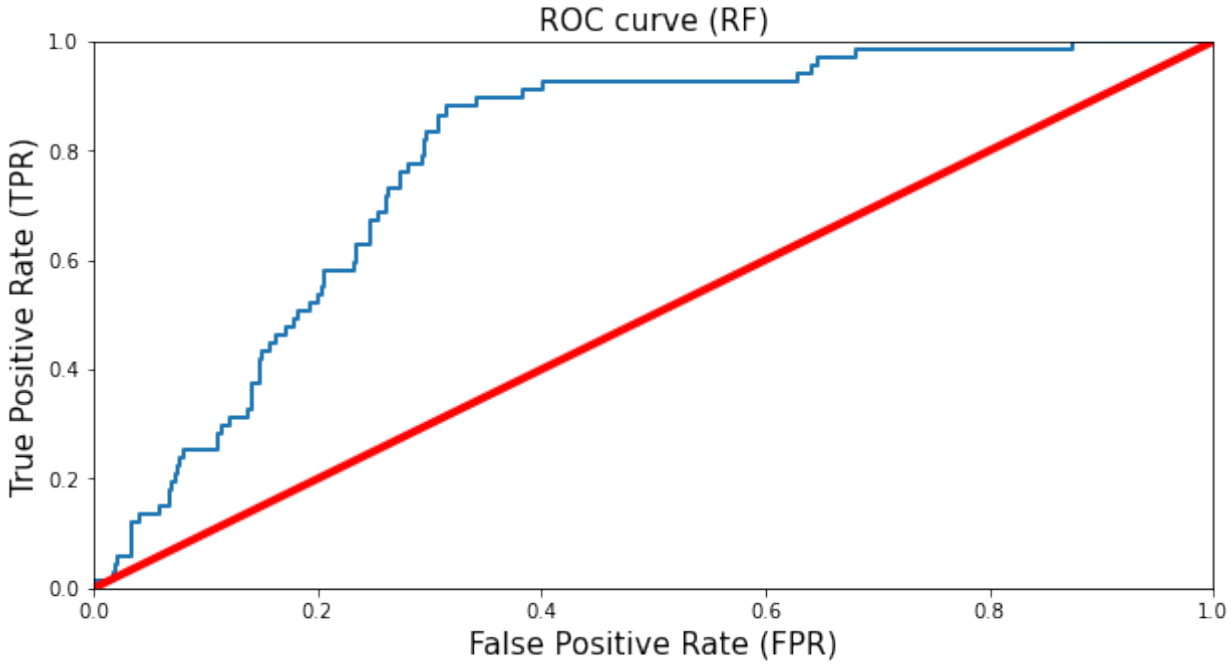
I then scored the Random Forest model with the new hyperparameters and ran the ROC AUC scores and curve.

Red:
0.87
0.87

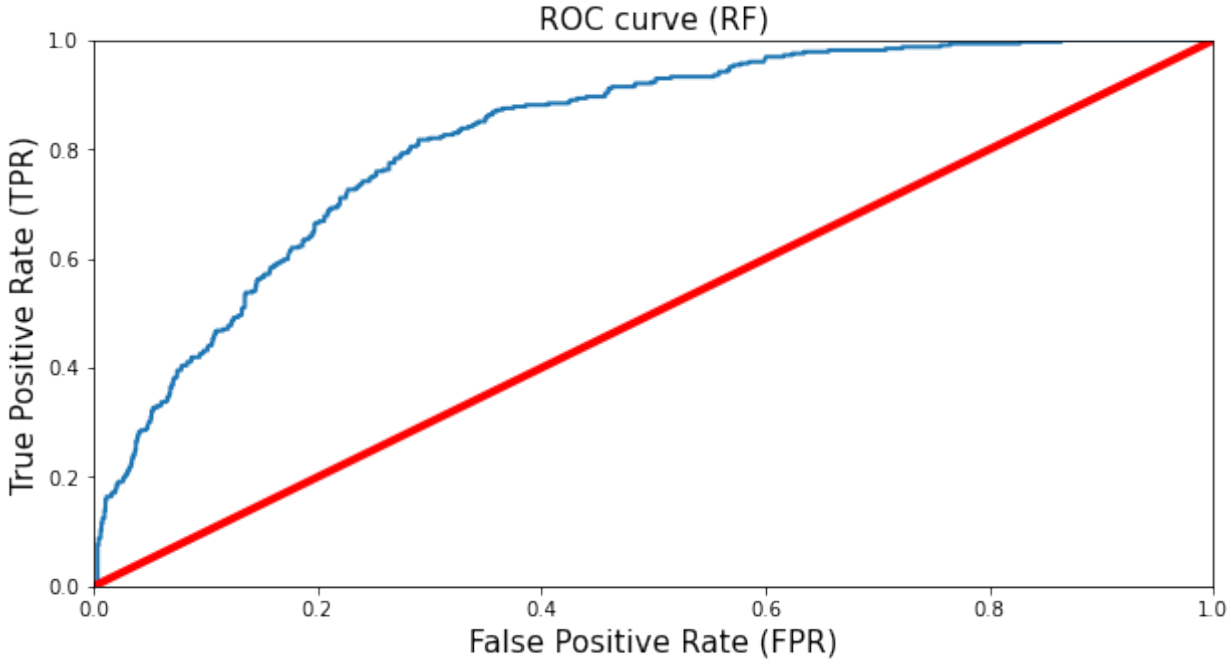
White:
0.80
0.83

Red ROC-AUC Score: 0.79
White ROC-AUC Score: 0.82

Red


ROC curve (RF)
True Positive Rate (TPR)
False Positive Rate (FPR)
0.0
0.2
0.4
0.6
0.8
1.0

White


ROC curve (RF)
True Positive Rate (TPR)
False Positive Rate (FPR)
0.0
0.2
0.4
0.6
0.8
1.0

Red:

	precision	recall	f1-score	support
Lo quality	0.88	0.99	0.93	410
High quality	0.76	0.19	0.30	70
accuracy	0.87	480		
macro avg	0.82	0.59	0.61	480
weighted avg	0.86	0.87	0.84	480

White:

	precision	recall	f1-score	support
Lo quality	0.82	0.99	0.89	1166
High quality	0.72	0.14	0.24	304
accuracy	0.81	1470		
macro avg	0.77	0.57	0.57	1470
weighted avg	0.80	0.81	0.76	1470

Sources & Citations:

Modeling wine preferences by data mining from physicochemical properties.

In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Available at: https://archive.ics.uci.edu/ml/datasets/wine+quality

Paulo Cortez (Univ. Minho), Antonio Cerdeira, Fernando Almeida, Telmo Matos and Jose Reis (CVRVV) @ 2009

Data source: http://www3.dsi.uminho.pt/pcortez/winequality09.pdf.