High - quality wine Predicition and Analysis

By Ruthy Yao

Summary

- Predictive and Descriptive analysis on red wine quality classification
- Random Forest Model with accuracy score of 93.4% and F1 score 71.2%
- Top five contributing factors and its value range:
  - Alcohol - to control between 10.8 and 12.2;
  - Sulphates - to control between 0.65 and 0.86;
  - Volatile acidity - to control between 0.3 and 0.49;
  - Density - to control between 0.9947 and 0.9974;
  - Total sulfur dioxide - to control between 17 and 43.

Outline

- Business Problem
- The data
- Method
- Results and Insights
- Next Steps

Business Problem

- The wine quality of Riverwood's red wine products become unpredictable in recent years – they lost the "Spectatular Wine" awards for several times.
- The management team needs more insights on the physiochemical properties that contribute to a high-quality wine.
- Using a prediction model to identify the high-quality wines so they can allocate the sales and marketing resources to the best products.

The data

The dataset comes from the red variants of the Portuguese "Vinho Verde" wine [1]

- 1599 data entries
- 11 physiochemical properties
- Wine quality rated from 1 to 8

Note: 1. please refer to [Cortez et al., 2009], http://www3.dsi.uminho.pt/pcortez/wine/

Analytical Method

- Three **machine learning models** are trained
  - Logistic Regression
  - Decision Tree
  - Random Forest
- **Feature Importance** technique to rank the parameters based on the contribution to the wine quality classification from the highest to the lowest
- **Quartile statistics** to compare and contrast the value distribution for each of the top five parameters between “high-quality” and “mediocre” wines.

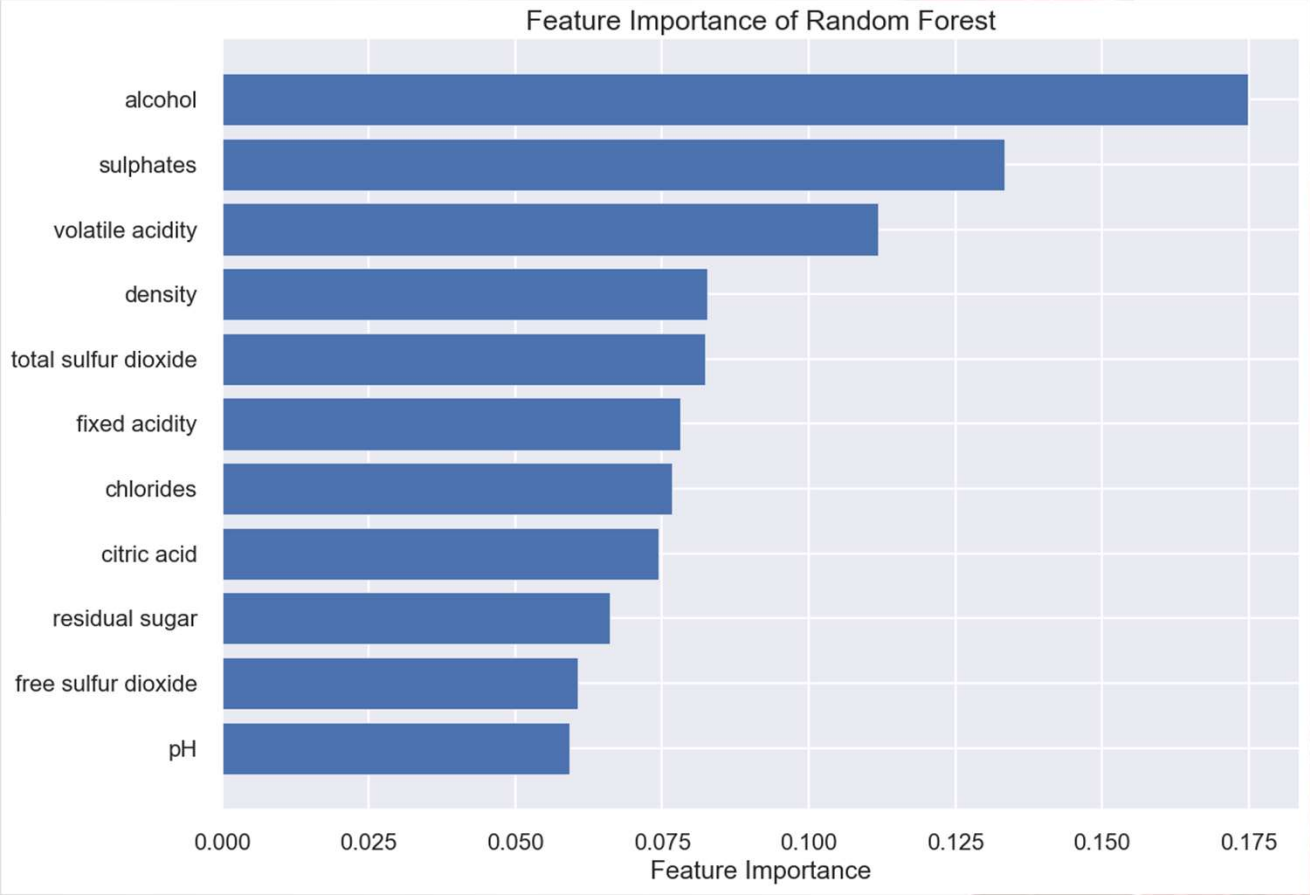
Results and Insights

- Random Forest Model has been selected because it has the highest predictive power

Accuracy score	F1 score[1]
93.4%	**71.2%**

Note: 1. F1 score is a balanced score that combines precision and recall

Results and Insights



- The parameters that contribute the most to a “high-quality“ status are alcohol, sulphates, volatile acidity, density, total sulfur dioxide

Results and Insights

- The desirable value range for each parameter

Physiochemical Properties	Value Range
Alcohol	10.8 - 12.2
Sulphates	0.65 - 0.86
Volatile acidity	0.3 - 0.49
Density	0.9947 - 0.9974
Total sulfur dioxide	17 - 43

Next Steps

- **Enlarge the sample size** - to improve the predicative accuracy.
- **Analyse the interacted features** - Build interaction features to improve the predicative power.
- **Introduce the white wine dataset** - Compare and contrast the results from the two different products' dataset to yield additional insights around the common factors vs the product-specific factors to target the fermentation techniques for wine quality improvement.