Wine quality predictions

Andrew Hercules

For this exercise, I opted to use a classification models. When I looked at the distribution of scores, I was concerned about the uneven distribution and variation, both within the datasets and between the datasets and felt that regression would not be suitable. To support the use of classification models, I transformed the numeric quality value into a categorical bin and then assigned each category a label:

- Wines with a quality value of 1-4 are categorised as “poor” and labelled 0
- Wines with a quality value of 5-6 are categorised as “average” and labelled 1
- Wines with a quality value of 7-10 are categorised as “good” and labelled 2

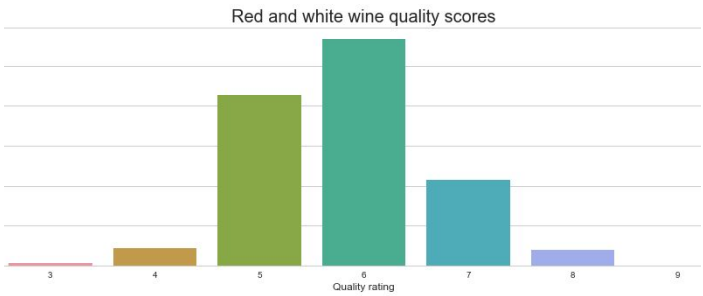
Limitations:

- When the two datasets are combined, there are more white wine records than red wine records and because the important features for both sets is different, it changes the performance of the model and could lead to lower accuracy
- The dataset does not include information about price and brand, which do impact the perceived quality of wine (or any item for that matter)
- I took a random sample of 5 classification models and there may be models that are more accurate
- Ideally, I would like to use more samples to build more robust models
- I did not fully tune the parameters to improve the accuracy of the Random Forest model due to limits in my laptop’s processing power

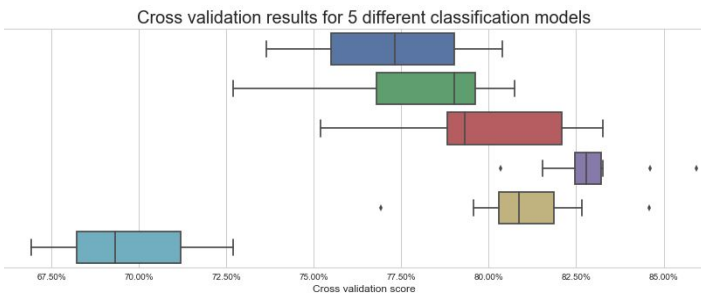
---

Red & white wine

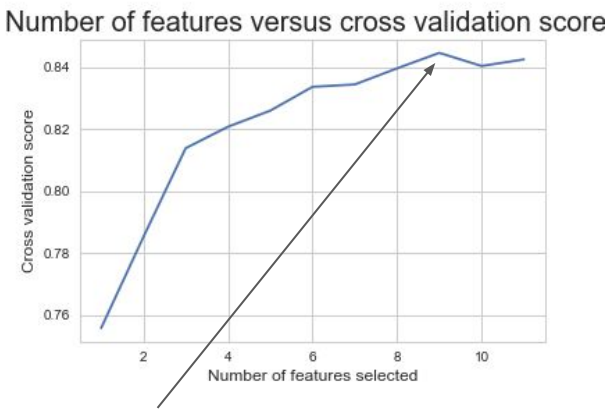
Number of records: 6497
Average quality rating: 5.81



4974 out of 6497 wines were ranked either 5 or 6 and no wines ranked higher than 9 or lower than 3



Compared to 5 other classification models, **Random Forest** had the highest average accuracy at **83%**



Volatile acidity, Residual sugar, Chlorides, Free sulfur dioxide, Total sulfur dioxide, Density, pH, Sulphates, Alcohol

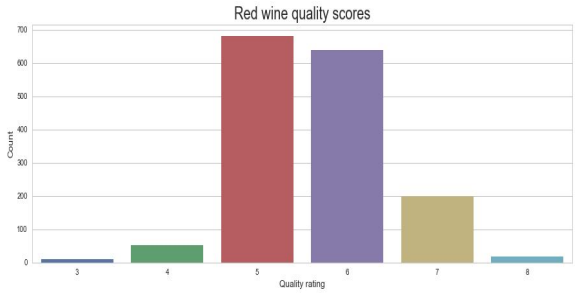
The 9 features identified by the Random Forest classifier as resulting in the most optimal prediction model

84.5%

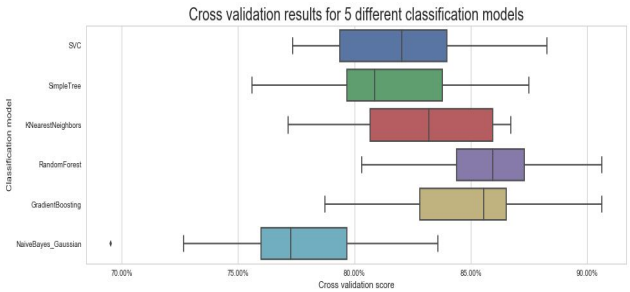
Accuracy when using a Random Forest classifier and the 9 optimal features

Red wine

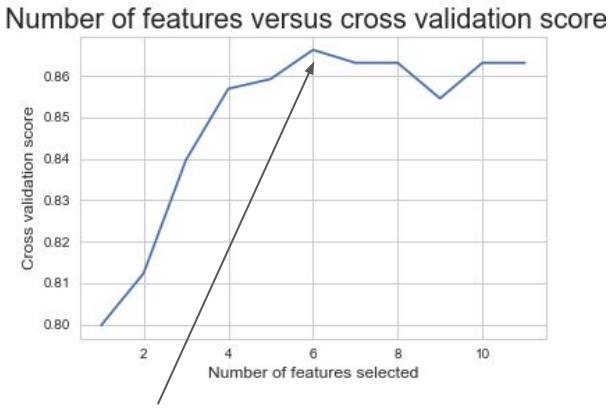
Number of records: 1599
Average quality rating: 5.64



1319 out of 1599 wines were ranked either 5 or 6 and no wines ranked higher than 8 or lower than 3



Compared to 5 other classification models, **Random Forest** had the highest average accuracy at **86%**



Fixed acidity, Volatile acidity, Total sulfur dioxide, Density, Sulphates, Alcohol

The 6 features identified by the Random Forest classifier as resulting in the most optimal prediction model

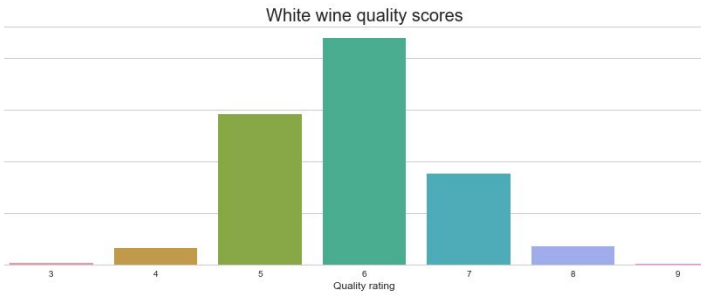
86.5%

Accuracy when using a Random Forest classifier and the 6 optimal features

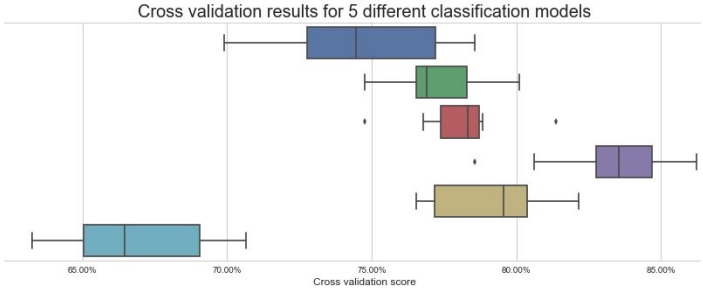
---

White wine

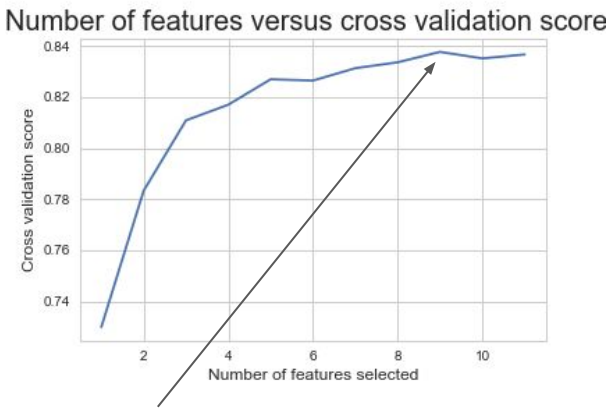
Number of records: 4898
Average quality rating: 5.88



2198 out of 4898 wines were ranked 6 and no wines ranked higher than 9 or lower than 3



Compared to 5 other classification models, **Random Forest** had the highest average accuracy at **83%**



Volatile acidity, Residual sugar, Chlorides, Free sulfur dioxide, Total sulfur dioxide, Density, pH, Sulphates, Alcohol

The 9 features identified by the Random Forest classifier as resulting in the most optimal prediction model

86.2%

Accuracy when using a Random Forest classifier and the 9 optimal features