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

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

Sources:

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

Past Usage: P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.

Modeling wine preferences by data mining from physicochemical properties.

In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.

In the above reference, two datasets were created, using red and white wine samples. The inputs include objective tests (e.g. PH values) and the output is 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 plot the relative importance of the input variables (as measured by a sensitivity analysis procedure).

Relevant Information:

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 Instances: red wine - 1599; white wine - 4898.

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

Output variable (based on sensory data):

12 - quality (score between 0 and 10)

13 – type (red / white)

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

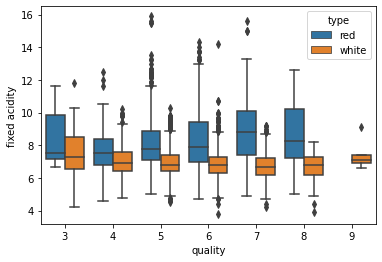
4 216

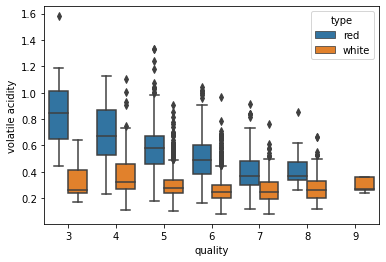
8 193

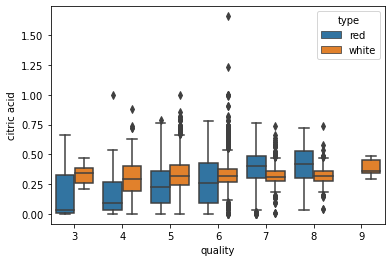
3 30

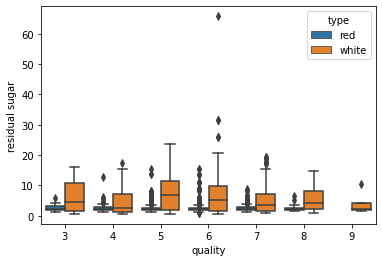
9 5

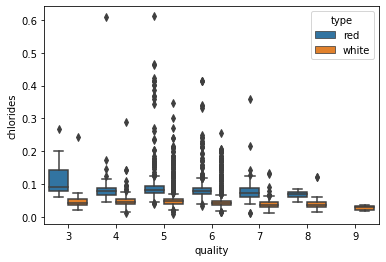
These will be split into a binary feature in the preprocessing step, with 7 and higher being high quality and everything else being lower quality. The plots for the output variable vs each of the other features are below.

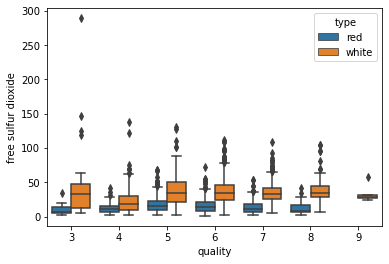


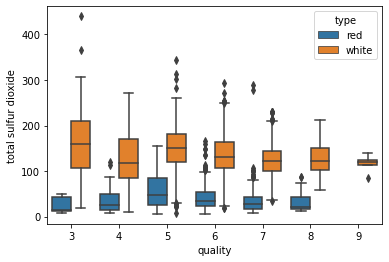


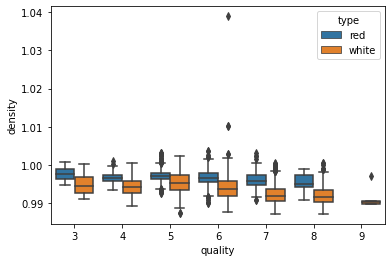


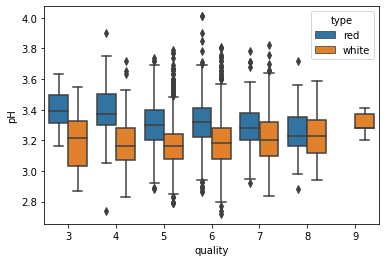


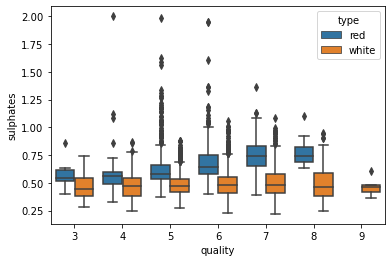


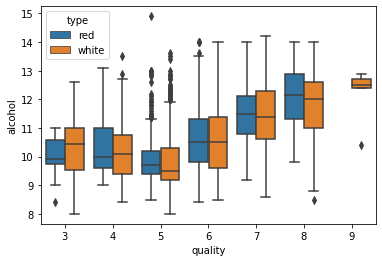




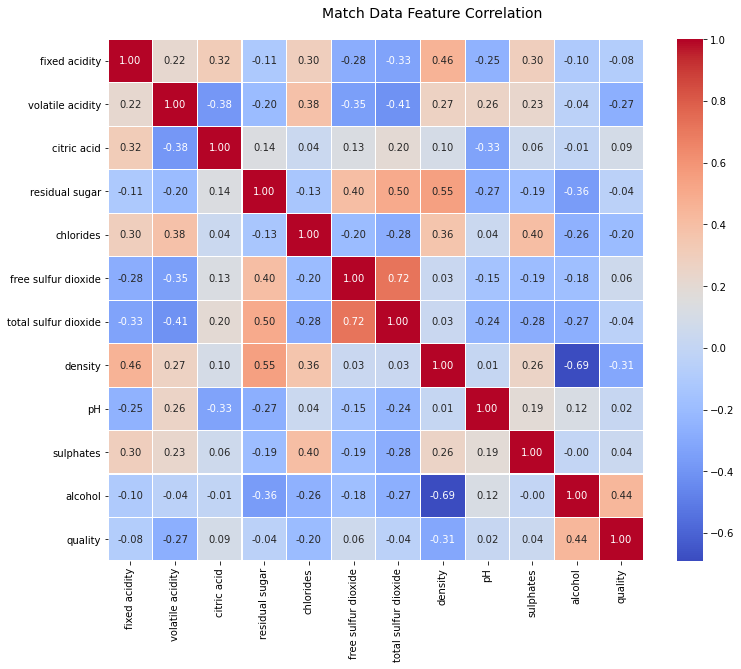








The Correlation Heatmap is as follows, with alcohol and density having a high correlation as well as free & total sulfur dioxide:



**Preprocessing and Modeling:**

I started this section by splitting the data into high and low quality wines, as well as red and white, for a total of two subsets of data: Red and White. 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. I didn’t think any of the correlations were high enough to leave out any features, but I did find that the following features ranked significantly higher than the others in importance: alcohol, chlorides and density.

Feature: 0, Score: 0.02320

Feature: 1, Score: 0.06012

Feature: 2, Score: 0.03225

Feature: 3, Score: 0.04622

Feature: 4, Score: 0.11633

Feature: 5, Score: 0.02718

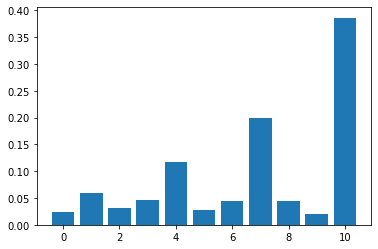
Feature: 6, Score: 0.04506

Feature: 7, Score: 0.19901

Feature: 8, Score: 0.04355

Feature: 9, Score: 0.02057

Feature: 10, Score: 0.38652



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: 0.825

White: 0.7993197278911565

Random Forest Classifier

Red: 0.875

White: 0.8755102040816326

Gradient Boosting Classifier

Red: 0.8520833333333333

White: 0.8333333333333334

Logistic Regression

Red: 0.8791666666666667

White: 0.7938775510204081

KNeighbors Classifier

Red: 0.8708333333333333

White: 0.8387755102040816

GaussianNB Classifier

Red: 0.84375

White: 0.7034013605442176

SVC Classifier

Red: 0.88125

White: 0.8251700680272109

I then 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)

0.8280810805172524

White:

RandomForestClassifier(max\_depth=4, min\_samples\_leaf=2, n\_estimators=72)

0.8280810805172524

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

Red:

0.8666666666666667

0.8676174545407577

White:

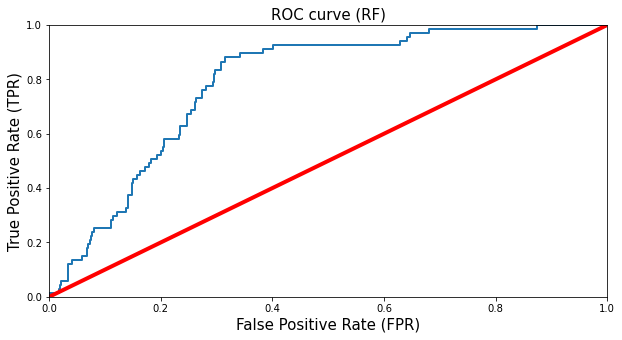
0.7993197278911565

0.8254003121039435

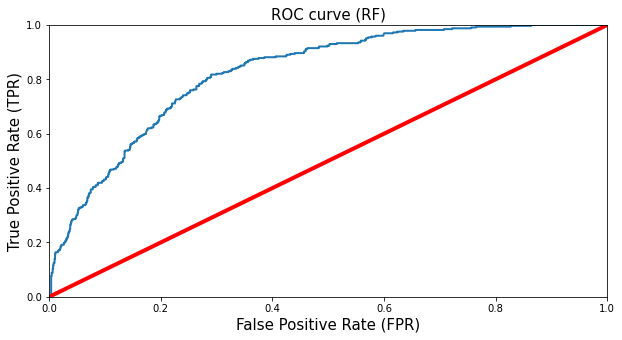
Red ROC-AUC Score: 0.7865996892053052

White ROC-AUC Score: 0.8248550170623006

Red



White



**Citations:**

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.

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