Predicting White Wine Quality in R

IST 687 Introduction to Data Science

Brian Baker
Charlie Vanleuvan

Project Summary

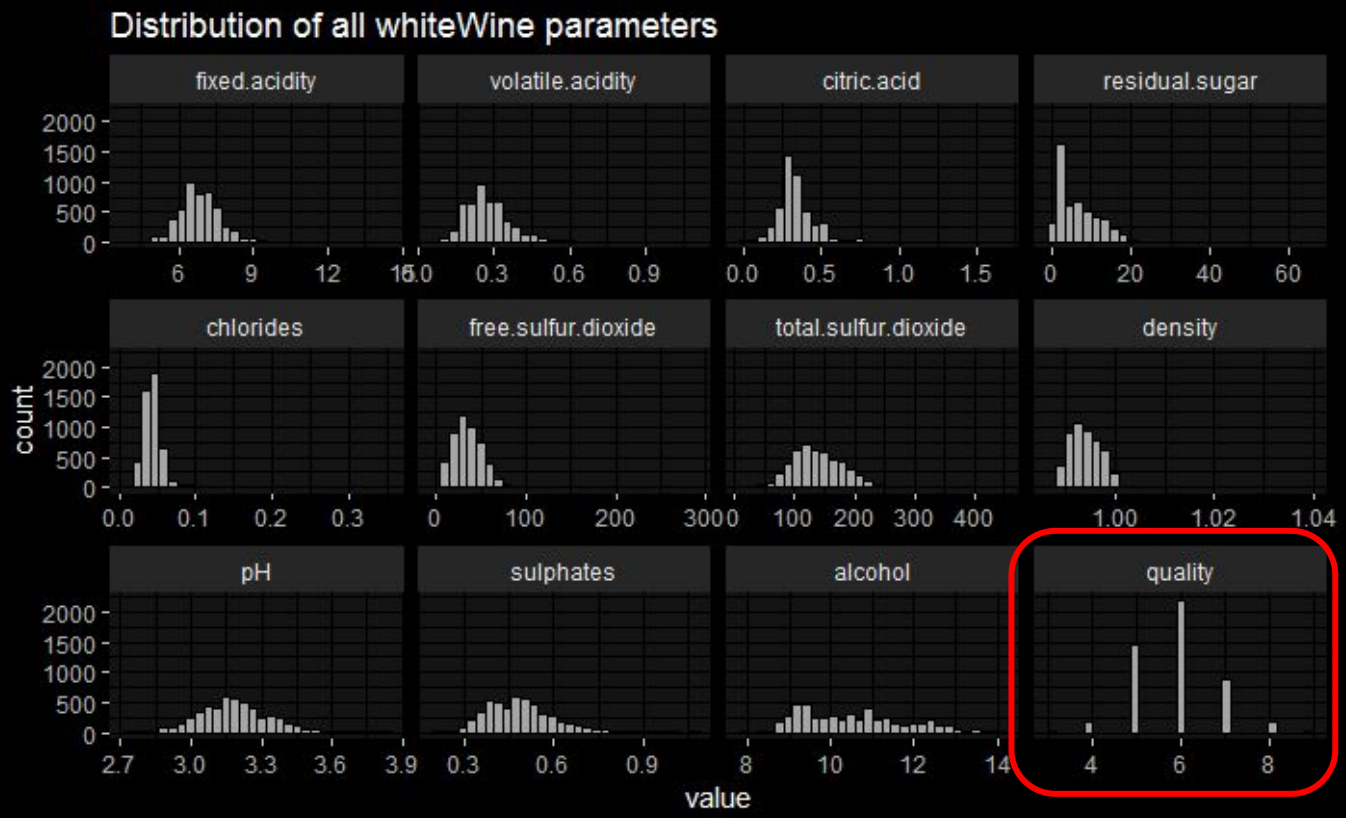
- Summary of Project
- Data Cleaning
- Predictions / Hypothesis
- Exploratory Data Analysis
- Models & Results
  - randomForest
  - SVM
- Next steps / Conclusion

Data Summary / Cleaning

- 4,898 rows of 12 variables of white wine from UC Irvine ML Repository
- Variables 1:11 represent chemical properties of the wine; Variable 12 is a “quality” rating
  - No NAs to scrub
  - Possibly similar or correlated properties:
    - Acidity properties and Sulfur properties
  - Each observation appears to be a separate wine, but are each of the quality ratings from the same individual’s scale, or from multiple individuals?
  - “str” command reveals chemical variables (1:11) are Numeric while “quality” is an Integer (whole numbers only)

```
'data.frame':   4898 obs. of  12 variables:
 $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
 $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
 $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
 $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
 $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
 $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
 $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
 $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
 $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
 $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
 $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
 $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

Variable Summary for White Wine:
Histograms

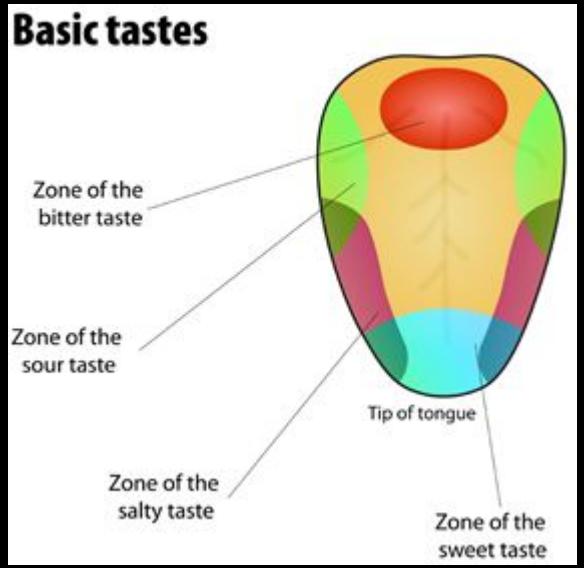


Predictions / Hypothesis

- Can we predict the quality of white wine based on the chemical component variables?
- Expect “Taste” variables to be the most important in quality detection
- Higher Alcohol % could be influential

“Taste” variables could include:

- Fixed Acidity - higher levels produce “tart” taste
- Volatile Acidity - increased levels have “vinegar” taste
  - Could indicate spoilage as this is result of microbes
- Chlorides - “salty” taste
- Residual Sugar - “sweeter”, byproduct of fermentation
- Citric Acid - freshens the wine



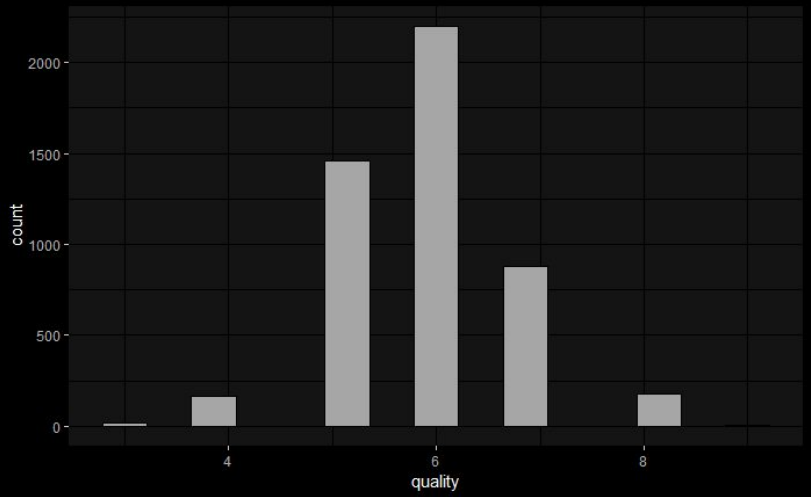
Initial randomForest
Can the raw data prove our hypothesis?

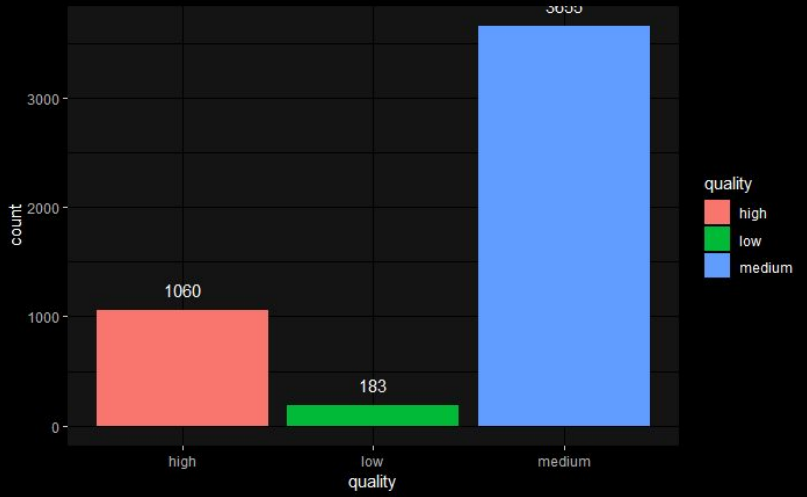
```
Call:
 randomForest(x = whiteWine[, -12], y = (whiteWine[, 12]))
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 0.3435573
                    % Var explained: 56.19
```

- Only ~56% of the variance is explained, no where close to acceptable to move forward
- Hoping to achieve something in the 80%+ range
- Very difficult for the model to predict differences between the Integer labels
- Convert “quality” to Factor to assist with classification & modeling algorithms

Distribution of Quality for White Wine

- Heavy Quality cluster around 5, 6 & 7
- Quality metric is likely based off individual preference and would likely be blurred between integer lines
  - “Fine-Tune” into 3 buckets:
    - “Low” = 3-4
    - “Medium” = 5-6
    - “High” = 7, 8, 9



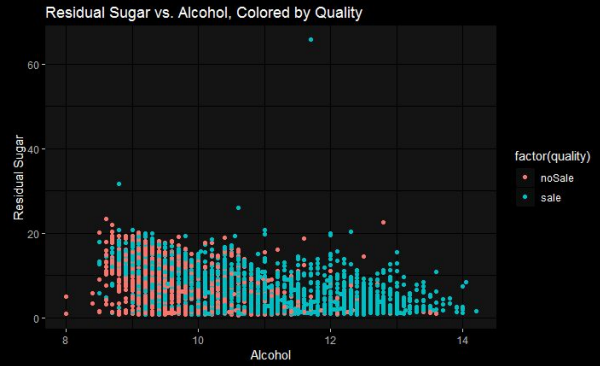


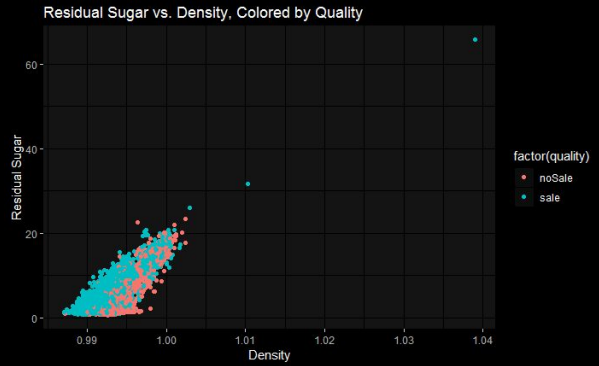
randomForest
How does classification change the model?

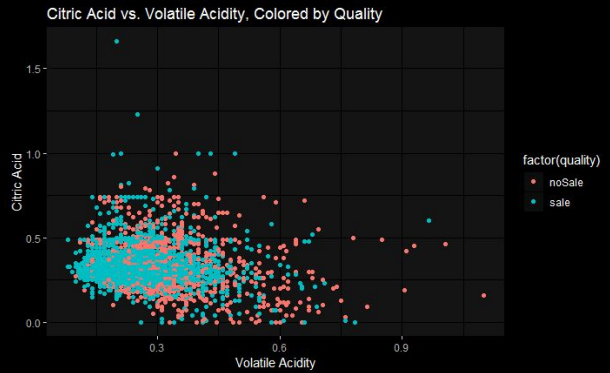
```
Call:
 randomForest(x = whiteWine1.0[, -12], y = (whiteWine1.0[, 12]))
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 13.96%
Confusion matrix:
       high low medium class.error
high    673   0    387  0.36509434
low       1  35    147  0.80874317
medium  142   7   3506  0.04076607
```

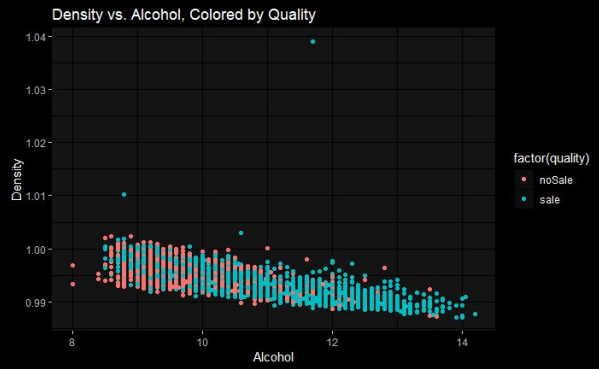
- Error rate of ~14%, provides 86% accuracy!
- Even though the overall results are predictive, there is noise underneath the hood
  - Predicting “high” has an error rate of ~37%
  - Predicting “low” has error rate of nearly 81%
  - Most of the results are “medium”, and predicted as “medium”
- Can we make the model tighter by changing quality to binary (saleable vs. non-saleable)

Exploratory Data Analysis
Scatter Plots



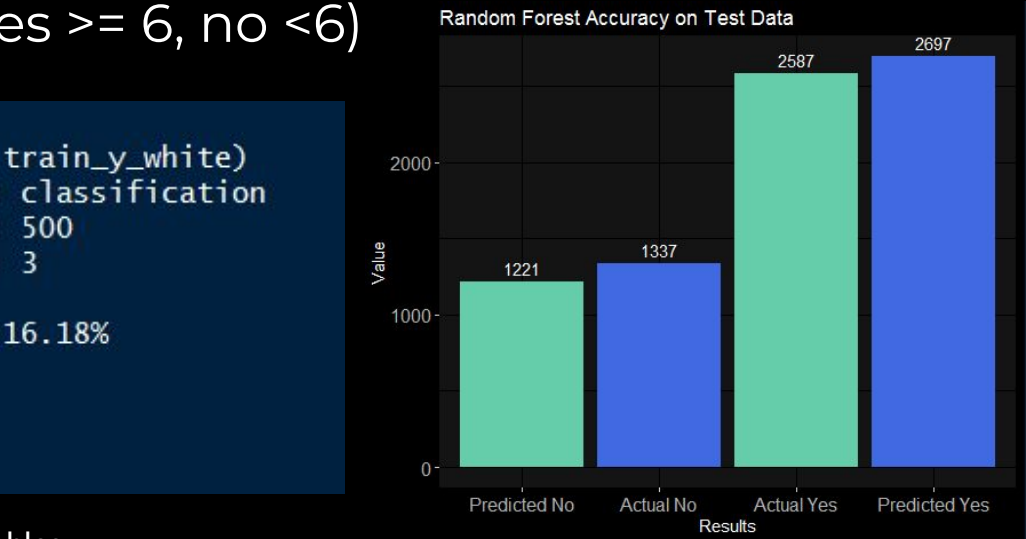






randomForest
Binary Classification (yes >= 6, no <6)

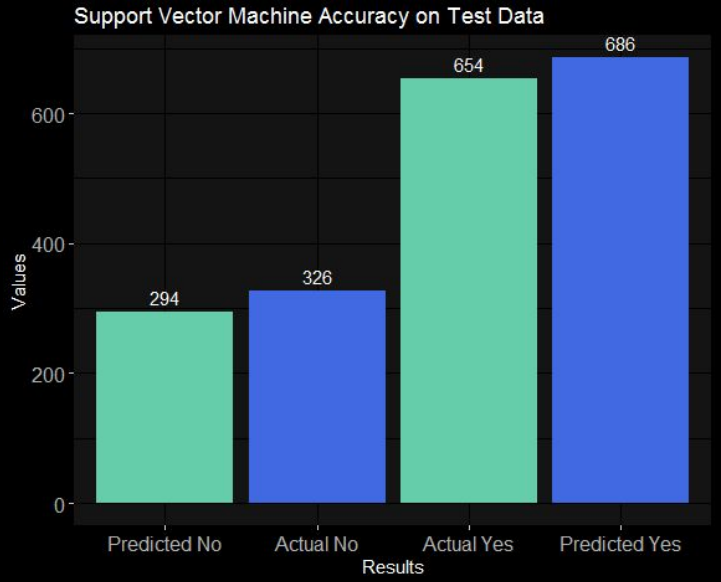
```
Call:
 randomForest(x = train_x_white, y = train_y_white)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 16.18%
Confusion matrix:
     no  yes class.error
no  959  372   0.2794891
yes 262 2325   0.1012756
```



- Predictions of quality based on all other variables
- 80% of white wine data set used for training, 20% used for cross validation
- OOB estimate of error rate ~ 16% (nearly the same as “low”, “medium”, “high” but better under the hood)
  - Nearly 85% correct predictions in training
  - Predictions of “Yes” are correct 90%+
  - Predictions of “No” are correct ~75%
- ~83% cross-validation accuracy in classification
- Feature Selection → Alcohol, Volatile Acidity, and Density were most important features

SVM

```
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 5

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.0778143571663147

Number of Support Vectors : 2036

Objective Function Value : -8017.703
Training error : 0.158244
```



- 80% training / 20% test data split
- With default kernel formula and Cost of Constraints = 5, training error rate ~15%
- Class accuracy:
  - ~63 % for no
  - ~86% for yes
- ~78% cross validation accuracy on test data, 22% prediction error → ksvm overfit our training data
- Model is having difficulty achieving “good” linear separation between all 11 features

Conclusion / Next Steps

- Initial data set is likely better suited for developing outlier detection algorithms as opposed to classification
  - Most of the quality observations lie at 5 or 6 rating
  - Chemical features of wine should be able to predict whether or not the wine will be saleable
  - We were correct in estimating Volatile Acidity and Alcohol would be best predictors
- Tuning SVM for better cross validation accuracy
  - Tune for better “no” classification accuracy
- Feature selection / tuning across both models
- Work with SME’s in the wine business domain to define a clear quality cutoff point for selling wine
  - Also age of wine might be a crucial feature to include