ANALYSIS ON CHEMICAL PROPERTIES OF PORTUGUESE WINE

Thomas Simons and Joey Cerquera

PSTAT 131: UC Santa Barbara

12/20/17

Abstract:

We looked a two data sets on white wine and red wine from Portugal. The datasets contained information on the physical/chemical properties of the wine and the wine quality out of 10. Since we know the two types of wine taste very different, one of our questions was what are the differences in composition between red and white wine. We then wanted to see if certain predictors (i.e. pH, sulfate level, etc.) cause the wine to be rated higher or lower. We wanted to see if the quality was based on the actual properties of the wine or just personal preference. Then we wanted to see which predictors are important in predicting the quality of the wine and then compare the important predictors between red and white wines. Finally, we wanted to create a model that best predicts the quality. This is important to Portuguese wine producers, as they must compete with an exorbitant number of competitors, as technology has facilitated the growth of a once small market to expand into an overly complex and competitive one.

We found that there is a clear difference in the composition of white wine compared to red wine. To see if the predictors influenced quality we used Principal Component Analysis (PCA). We found that both wines had clustering of high quality wines around high values of PC1 and had clustering of low quality wines around low values of PC1. The opposite was true for PC2. It was interesting in that predicting quality alcohol percentage was highest for red wine and white wine. However, for red wine the second most important predictor was sulfates, while for white wine it was volatile.acidity. We also found that the best predictive model was Random Forests.

Keywords: Variable selection; Model selection; Support vector machines, Decision Trees, Boosting, RandomForest, Wine, Portugal, Exportation, Macroeconomics

Introduction:

Wine is considered an intricate and chemically associated, alcoholic beverage. Chemical properties frequently mentioned when discussing wines, are acidity, sugar content, and alcoholic content. Yet many more properties exist, and may be important for the evaluation of a wine.

This study will provide a thorough analysis of many chemical properties of wine, and whether a model can be constructed to predict a person's liking of a wine. These models may also provide

insight into whether wine is truly an intricate beverage that depends on the individual, or is just a specific blend of certain properties. These models can also simplify the winemaking process, by providing an ideal range for these properties or whether they are unimportant in the process.

There was a previous study done on this data set by Paulo Cortez et al called "Modeling wine preferences by data mining from physicochemical properties". They wanted to find the best predictive model out of the regression techniques: Multiple Regression, Support Vector Machines, and Neural Networks. Unlike our analysis they wanted to predict each quality score. Whereas in our study we just predicted good or bad quality corresponding to quality greater than or less than 5. They used a SVM using a gaussian kernel while we used a radial kernel. We also wanted to see the differences between red and white wine and find important predictors. Thus, our analysis will build on their work by focusing less on finding a good model and looking at other techniques like PCA.

The data was collected from the vinho verde region of Portugal. The samples were tested by the certification entity CVRVV which is an organization with the goal of improving the quality of this wine. The data was processed to include a distinct wine sample in each row and they used the most common chemical/physical tests. To compute the quality each sample was evaluated using blind tastes and at least 3 sensory assessors which had a scale of 0 to 10. The quality is the median of these evaluations. The data set is large enough to be significant (about 1500 samples for red and 5000 for white). Most of the quality scores are 5 or 6 for each dataset. This is what motivated us to only try to predict good or bad quality wines since there may be a chemical difference between wines that are rated 5 or 6. We will have to analyze both datasets separately since red wine and white wine taste so different.

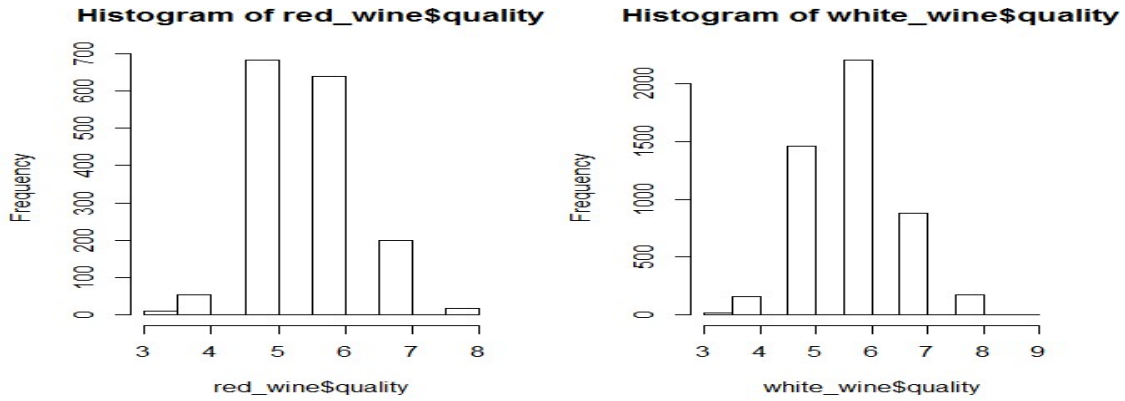
The techniques used for exploratory analysis are histograms, averaging the columns, and Principal Component Analysis. For model selection and variable importance we considered Random Forests, Boosting, and SVM. Once the best model was found we used it to find variable importance. We found that the best model was using random forests. It was interesting that the best model wasn't using SVM like the other study used but this is probably because they used a different kernel. The data was obtained through the UCI machine learning repository and the software used was RStudio.

Data:

The response in the data set is the quality of the wine and the predictors are the properties of the wine. The predictors are: Sulphates, alcohol, residual sugar, citric acid, total sulfur dioxide, free sulfur dioxide, volatile acidity, density, pH, chlorides, and fixed acidity.

We created two new factors to help with the analysis. Class has 2 levels, good if the quality was greater than 5 and bad if the quality was less than 5. This is the factor we will be predicting with the models. To help visualize PCA we created Type which was 0 if the quality was less than 5, 1 if the quality was 5 or 6 and 2 if the quality was greater than 6. This helped us identify clusters in PCA plotting.

Exploratory Analysis:



From the histogram, we see that most of the wines are rated as 5 or 6. Therefore, to investigate the difference between wines rated 5 or 6 we create the factor class with two levels: good or bad.

It is interesting that for both red and white wines there were more observations that had a quality greater than 6. This means that for both red and white wine most of the wines will be classified as good.

To see the difference in chemical composition between red and white wines we look at the column means of each data set.

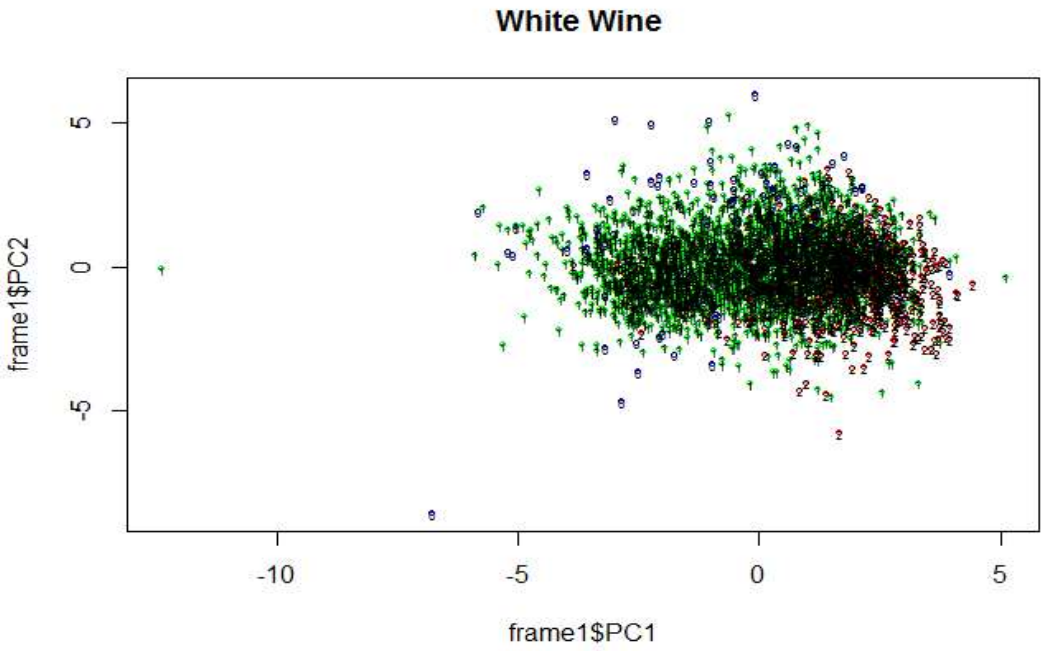
```
##                       red_mean white_mean
## fixed.acidity          8.31964    6.85479
## volatile.acidity       0.52782    0.27824
## citric.acid            0.27098    0.33419
## residual.sugar         2.53881    6.39141
## chlorides              0.08747    0.04577
## free.sulfur.dioxide   15.87492   35.30808
## total.sulfur.dioxide  46.46779  138.36066
## density                0.99675    0.99403
## pH                     3.31111    3.18827
## sulphates              0.65815    0.48985
## alcohol               10.42298   10.51427
## quality                5.63602    5.87791
```

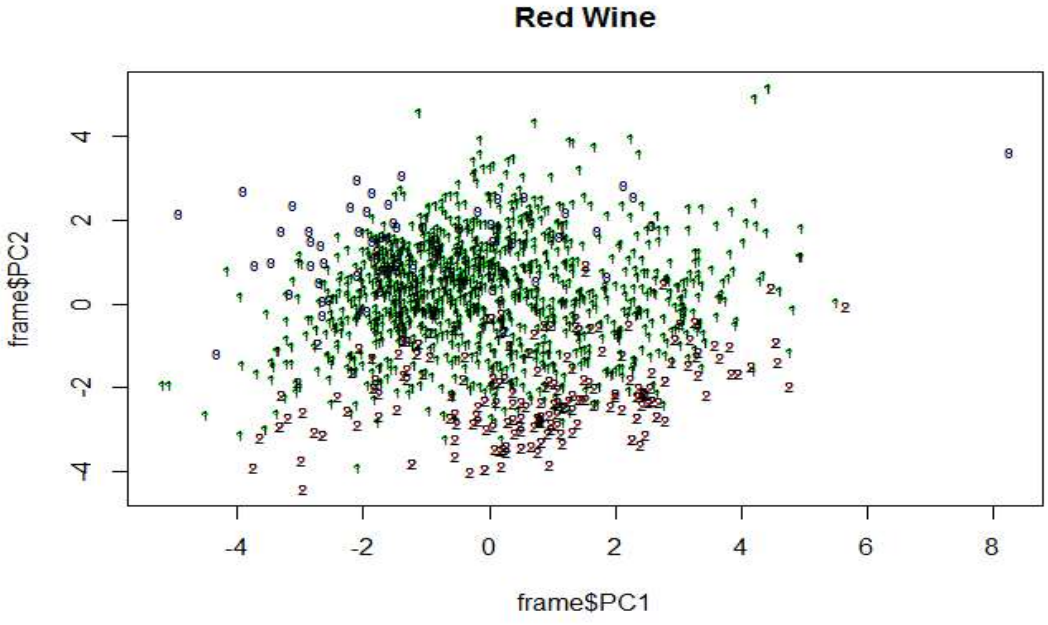
We can see that most of the variable have different mean values for white and red wine. The ones that have similar mean values are density, pH, alcohol and quality. From this we can conclude that the variables without similar means are important for the prediction of wine quality. The means also tell us the different properties of white wine compared to red wine and that there is a difference so it will be significant to compare the two datasets.

Principle Component Analysis

We will perform principal component analysis to see if there is a relationship between the quality of the wine and the properties of the wine. The factor Type is created with level 0 if the wine has quality lower than 5, 1 if the wine has quality 5 or 6 and 2 if the wine is quality greater than 5.

After performing PCA on both data sets we plot the first principal component versus the second to see if there is any clustering of the Types of wine.





For white wine, we see a denser distribution of points than for red wine. For both data sets we see clustering of Type 0 and Type 2 wines. We notice that for red wine the wines that are Type 2 (high quality) tend to have a low PC2 value and the wines that are low quality have a low PC1 value and a high PC2 value.

For white wine, there is a less clear separation than red wine. The plot does look similar but there is more overlap of wines that are high or low quality with wines that are in the middle. From these plots, we can conclude that there is a relationship between the predictors and the quality of the wine. Since the plot for white wine is denser it may be harder to predict the quality of white wine. We will need to do further analysis since the first two principal components only explain about 40% of the variance.

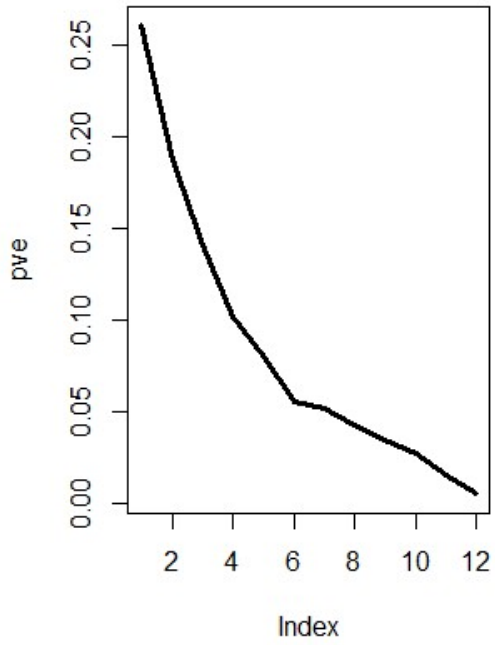
To see what variables, have the most effect on the PC1 direction look at the loadings in decreasing order for red wine.

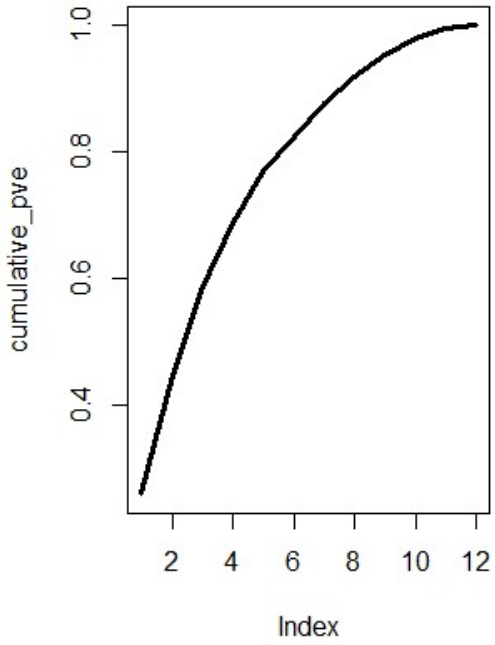
```
##                 rowname      PC1
## 1         fixed.acidity 0.487883
## 2           citric.acid 0.473335
## 3                    pH 0.432721
## 4               density 0.370301
## 5      volatile.acidity 0.265129
## 6             sulphates 0.254535
## 7             chlorides 0.197427
## 8        residual.sugar 0.139154
## 9               quality 0.112489
## 10              alcohol 0.073177
## 11  free.sulfur.dioxide 0.045881
## 12 total.sulfur.dioxide 0.004067
```

Then do the same thing but for the PC2 direction. It is most clear that low PC2 values predict high quality wines. Therefore, the loadings that have low values may be good predictors of high quality red wine.

```
##                 rowname      PC2
## 1               alcohol 0.502709
## 2               quality 0.473166
## 3  total.sulfur.dioxide 0.363971
## 4      volatile.acidity 0.338968
## 5               density 0.330781
## 6   free.sulfur.dioxide 0.259483
## 7             chlorides 0.189788
## 8        residual.sugar 0.167736
## 9           citric.acid 0.137358
## 10            sulphates 0.109334
## 11                   pH 0.065440
## 12        fixed.acidity 0.004173
```

Now look at the proportion of variance explained by each Principal Component to see which PCs are significant.





Model Selection and Analysis:

Our initial decision to use models related to decision trees, Boosting and Random Forests, is due to the fact that decision trees are a very viable manner to tackle a classification type problem. These two decision tree-based models are similar in that they use decision trees, but each has a different use of decision trees. Random Forests uses bags the decision trees with replacement, and considers a random subset of the input fields at each split. The difference between RandomForests and Boosting, is that boosting trees are additive rather than averaged. Each observation improves on the last by trying to fit a gradient.

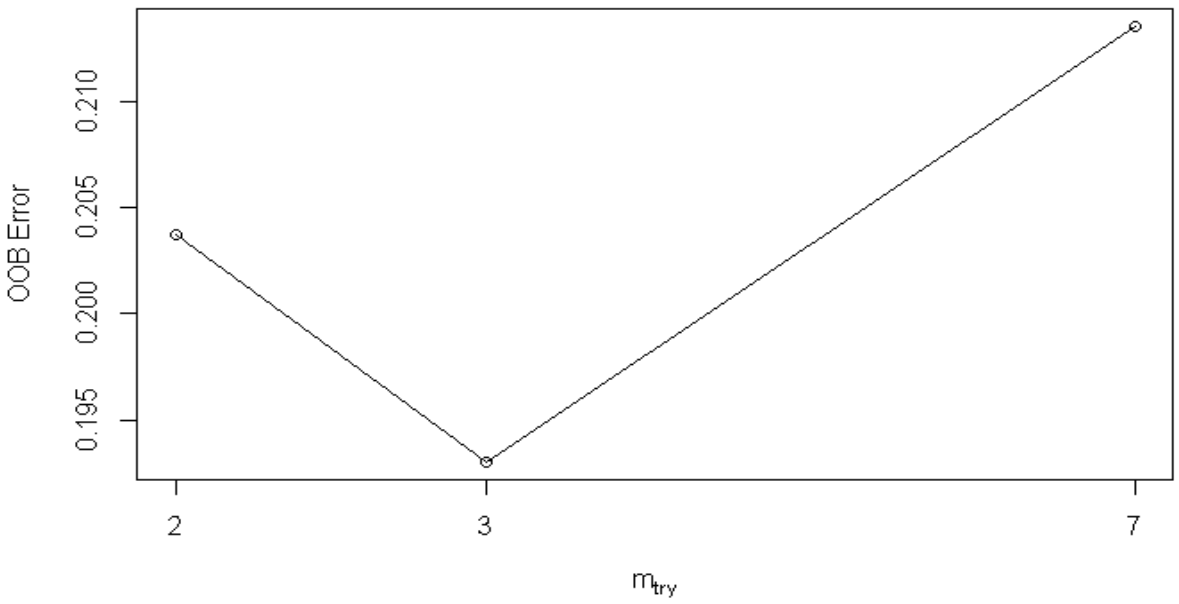
Support Vector Machines, have the particular aim of transforming the original dataset into another dimension, and searching for the best performing separating line or hyperplane. This hyperplane best divides the data, while allotting for some misclassification as to not produce an overly biased model.

We first separate the dataset into a test and training set. The training dataset has 7/10ths of the complete data. This is to test the viability of the data, as it is difficult to find a similar dataset elsewhere. We will look at SVM, Random forests and Boosting to see which is the best and which to use to find variable importance.

I. Random Forest

We decide to tune the Random Forest for the best number of trees to fit, after observing this computation, we note there is not much difference in misclassification error with different tree amounts, this is the case for both wines. In short, we noted that tuning the number of trees has a very small difference on Out of Bag error, same as misclassification. We decide to additionally tune

the number of variables randomly sampled as candidates at each split. By tuning several times on mtry, we decide for 3 for both randomForests.

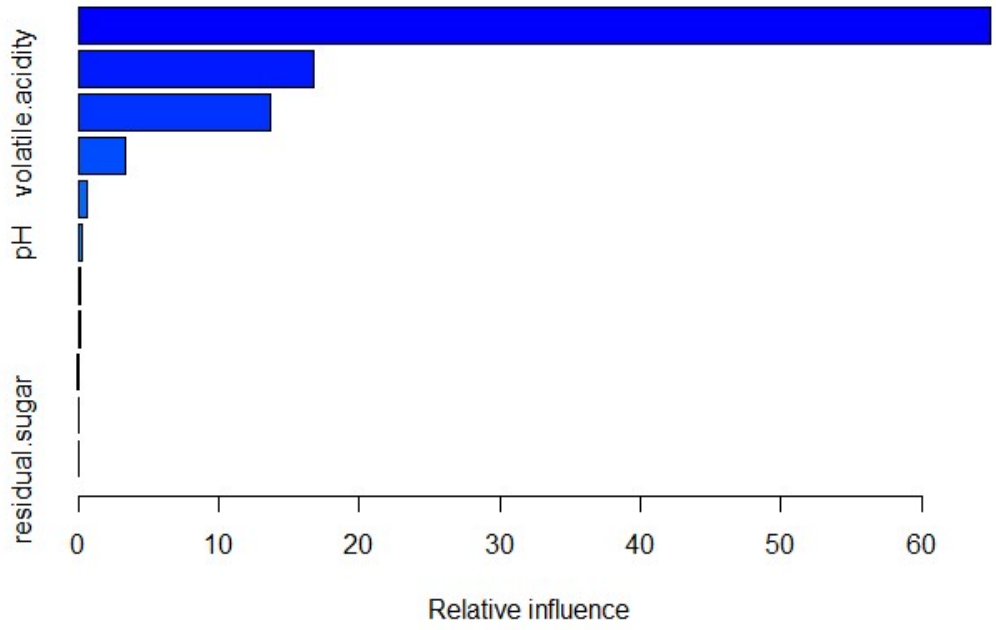



```
## Parameter tuning of 'randomForest':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  ntree
##    750
##
## - best performance: 0.1741
##
## - Detailed performance results:
##   ntree  error dispersion
## 1   150 0.1759    0.02257
## 2   250 0.1788    0.02313
## 3   500 0.1756    0.01851
## 4   750 0.1741    0.02287


Call: randomForest(formula = class ~ ., data = red_w.train2, mtry = 3,ntree = 500, im
portance = TRUE)
                       Type of random forest: classification
                             Number of trees: 500
                    No. of variables tried at each split: 3
                      OOB estimate of  error rate: 18.77%
                                Confusion matrix:
                               good bad class.error
                            good  487 109      0.1829
                            bad   101 422      0.1931
```

```
Call: randomForest(formula = class ~ ., data = white_w.train2, mtry = 3,      ntree =
500, importance = TRUE)
                       Type of random forest: classification
                             Number of trees: 500
                    No. of variables tried at each split: 3

                      OOB estimate of  error rate: 16.66%
                            Confusion matrix:
                           good bad class.error
                        good 2079 220     0.09569
                        bad   351 778     0.31089
```

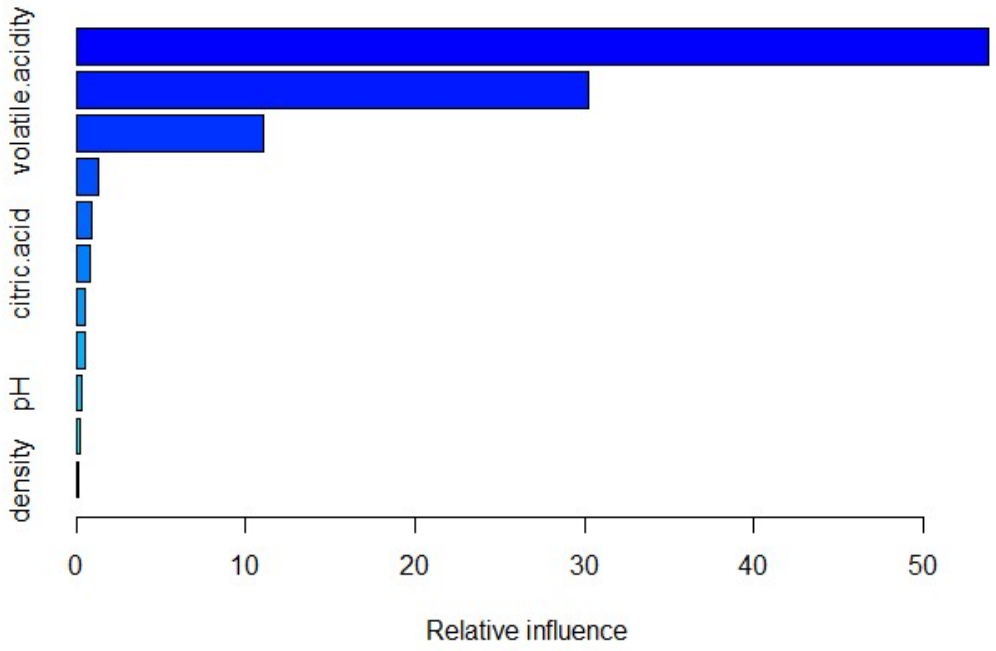
II. Boosting

Boosting for red wine




```
##                                     var  rel.inf
## alcohol                         alcohol 64.93279
## sulphates                     sulphates 16.81591
## volatile.acidity     volatile.acidity 13.69524
## total.sulfur.dioxide total.sulfur.dioxide  3.30770
## fixed.acidity           fixed.acidity  0.64618
## pH                                 pH  0.26889
## free.sulfur.dioxide   free.sulfur.dioxide  0.13032
## chlorides                   chlorides  0.12918
## density                       density  0.07379
## citric.acid                 citric.acid  0.00000
## residual.sugar           residual.sugar  0.00000
```

Boosting for white wine




```
##                                     var rel.inf
## alcohol                         alcohol 53.8690
## volatile.acidity         volatile.acidity 30.1951
## free.sulfur.dioxide   free.sulfur.dioxide 11.0853
## fixed.acidity               fixed.acidity  1.3115
## residual.sugar             residual.sugar  0.9315
## citric.acid                   citric.acid  0.8384
## total.sulfur.dioxide total.sulfur.dioxide  0.5075
## chlorides                     chlorides  0.4852
## pH                                   pH  0.3676
## sulphates                     sulphates  0.2316
## density                         density  0.1773
```

III. SVM

Analysis with SVM

For the Support Vector Machine we use a radial kernel, as we have a binary classifier. That as the only attainable values are "good" and "bad". This allows the model to be able to attain these values which would usually be difficult for most regression model regardless of dimensionality. We create a training and test dataset, to see if replicability is attainable. Next, as the SVM requires a Cost parameter as explained in the Methods section, we use the tune function to find the ideal cost to minimize misclassification error. Below is the summary of the tune function showing that the ideal or "best" Cost is 4.

SVM for white wine

```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  cost
##     4
##
## - best performance: 0.2103
##
## - Detailed performance results:
##   cost  error dispersion
## 1  0.1 0.2378    0.02004
## 2  1.0 0.2156    0.02325
## 3  3.5 0.2121    0.01397
## 4  4.0 0.2103    0.01247
## 5  4.5 0.2106    0.01153
## 6  7.0 0.2112    0.01214
```

SVM for red wine

```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  cost
##     1
##
## - best performance: 0.2377
##
## - Detailed performance results:
##   cost  error dispersion
## 1    1 0.2377    0.04326
## 2    4 0.2440    0.03707
## 3    7 0.2475    0.03909
## 4   10 0.2484    0.03576
## 5   15 0.2404    0.03985
## 6   20 0.2386    0.03608
```

By tuning the SVM we have obtained that the cost of 4.5 is the best model. This is because it reduces the misclassification error compared to several other costs. We should also note it also has the lowest false positive rate, which is desired for this situation.

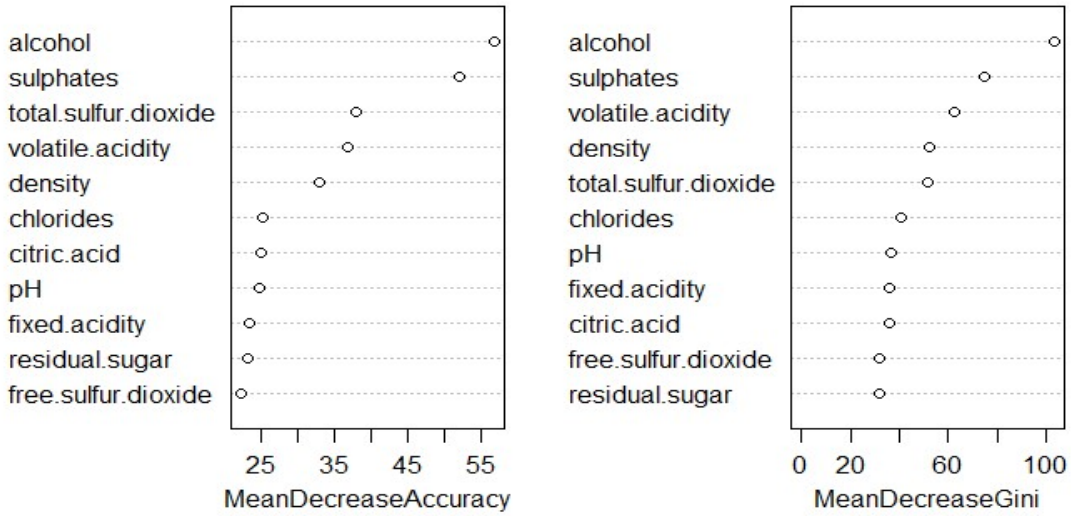
Misclassification Error for all Models:

We will compare misclassification rates for each of the models to see which has the lowest. The misclassification rate is in short, the rate at which the model incorrectly classifies the types of wine. It is the sum of false positives and false negatives over the total number of observations. We will then conclude which model is best, and will look at the variable importance for that model to see which variables are the most important for predicting wine quality.

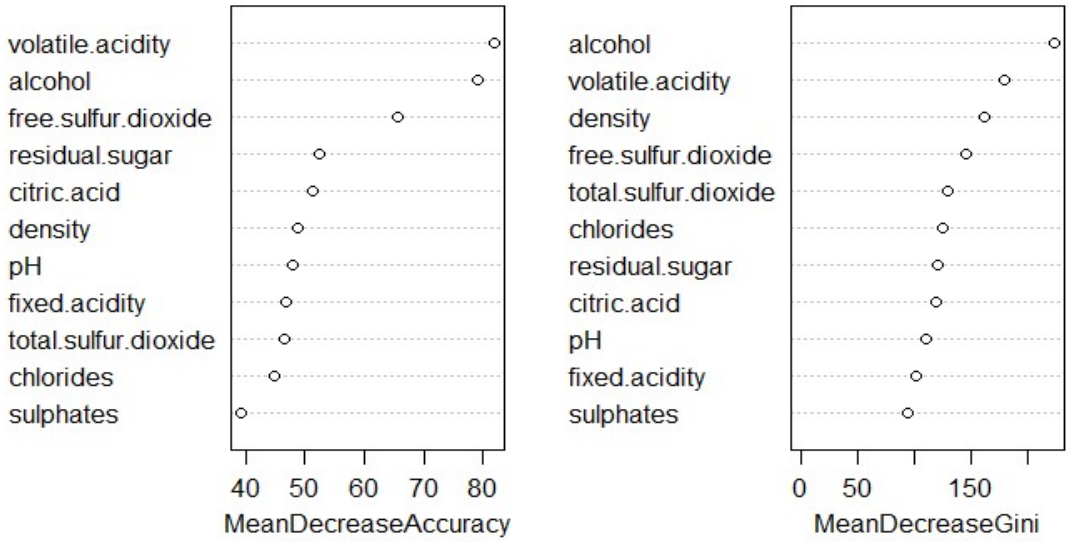
```
"Misclassification for Boost:White"

0.7497

"Misclassification for Boost:Red"

0.4604


"Misclassification for Random Forest:White"

0.1633

"Misclassification for Random Forest:Red"

0.2208


"Misclassification for SVM:White"

0.2177

"Misclassification for SVM:Red"

0.25
```

The model with the lowest misclassification rate was the model using randomForests. It has a misclassification error of .1633 for White Wine and .2208 for Red Wine, which is the lowest for both types of wine compared to the others. Since this is the best model we will use it to determine variable importance and compare between red and white wines.

rf.red



rf.white



Looking at the variable importance plots the most important variable for red wine based to mean decrease accuracy and mean decrease Gini is alcohol followed by sulphates.

The most important predictors for predicting wine quality for white wine are alcohol and volatile acidity. This shows that red and white wine both have alcohol as their most important predictor but the second most important predictor for each is different.

Discussion

Through analyzing multiple models, we found a model with great performance that minimizes classification error. The most important feature is its low false positive rate. This is the most desired quality of the model, as it stabilizes revenue, an explanation follows. Wine was at one time difficult to harvest as it depended on the topography and climate. Since the Age of Discovery, the Portuguese and other Iberians had produced wine for exportation for nearly all countries. Nowadays, technology has dispersed and increased the viability of harvesting wine to many parts previously impossible. This has increased competition, so Portuguese wine producers must do more to compete with an exorbitant number of producers. The desire for low false positive rates is to increase value on a desired wine, and obtain a stable number of consumers. This will allow for a stable cash flow, while the incorrectly classified wine can be sold for a lower value and still gain another stable cash flow, as the wine will probably be very likeable to most.

Another important part of our analysis that had not been expressed in the prior experiment, is the sheer abundance of observation between 5 and 6. Our take of this phenomenon is a cognitive bias, the bias being politeness. Politeness is a key cultural virtue in the relatively small country of Portugal, and it would be improper to harshly criticize a product even if it was awful. This is evident in the PCA as we made three levels to stratify the responses, and found it to be inefficient to observe three classifications when we could have a Bernoulli-type classification, with very little effect to the overall analysis.

Finally, we have achieved our original goal of observing which chemical properties affect wine quality. It seems alcohol volume and sulphates affect red wine's quality greatly, while volatile.acidity and alcohol affect white wine. Some challenges were finding the best way to transform the data and making sure our analysis was rigorous. As suspected there were different important predictors for white wine and red wine. Further questions could be analyzing wines from other parts of the world and looking at different chemical properties besides the one in this study.

References:

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, José Reis, Modeling wine preferences by data mining from physicochemical properties, In Decision Support Systems, Volume 47, Issue 4, 2009, Pages 547-553, ISSN 0167-9236, https://doi.org/10.1016/j.dss.2009.05.016. (http://www.sciencedirect.com/science/article/pii/S0167923609001377)

Amaral, Luciano. "Economic History of Portugal". EH.Net Encyclopedia, edited by Robert Whaples. March 16, 2008. URL http://eh.net/encyclopedia/economic-history-of-portugal/

Keywords: Variable selection; Model selection; Support vector machines, Decision Trees, Boosting, RandomForest, Wine, Portugal, Exportation, Macroeconomics