DA Assignment 7

Morgan Ford

April 2022

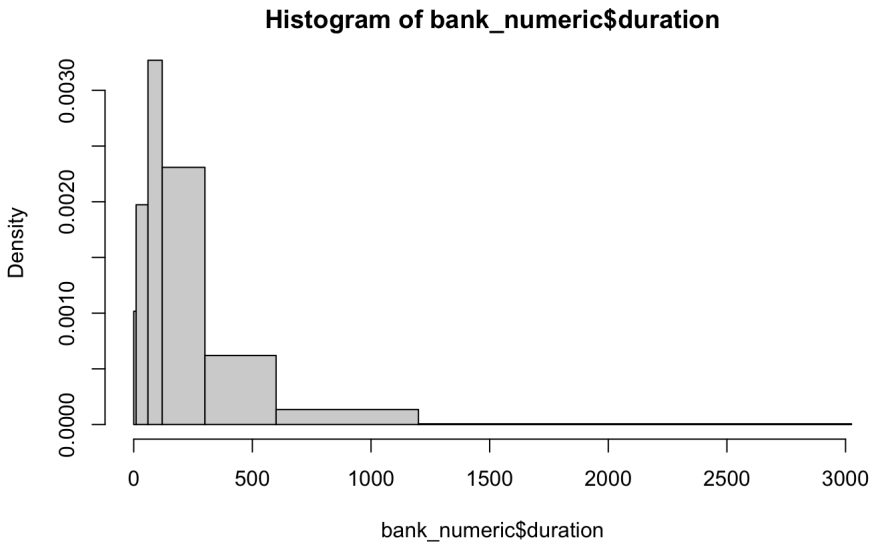
1 Exploratory Data Analysis

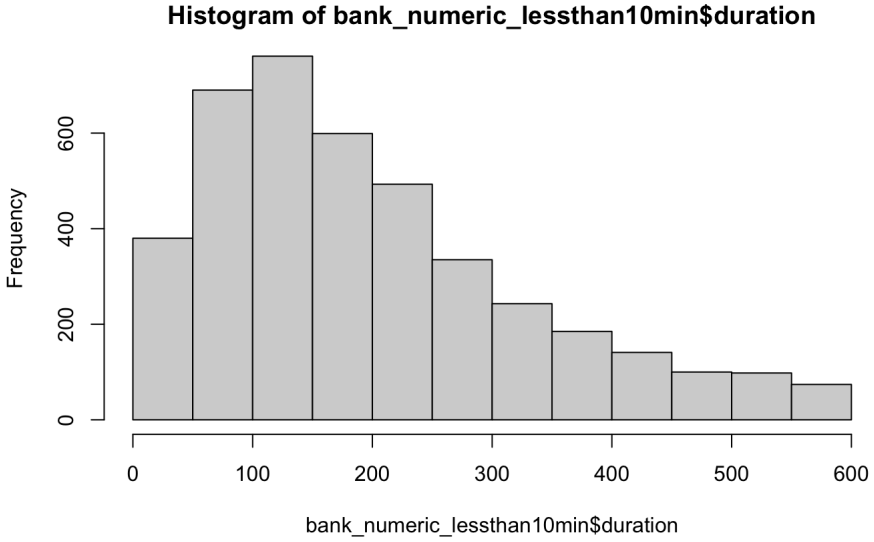
1.1 Bank EDA

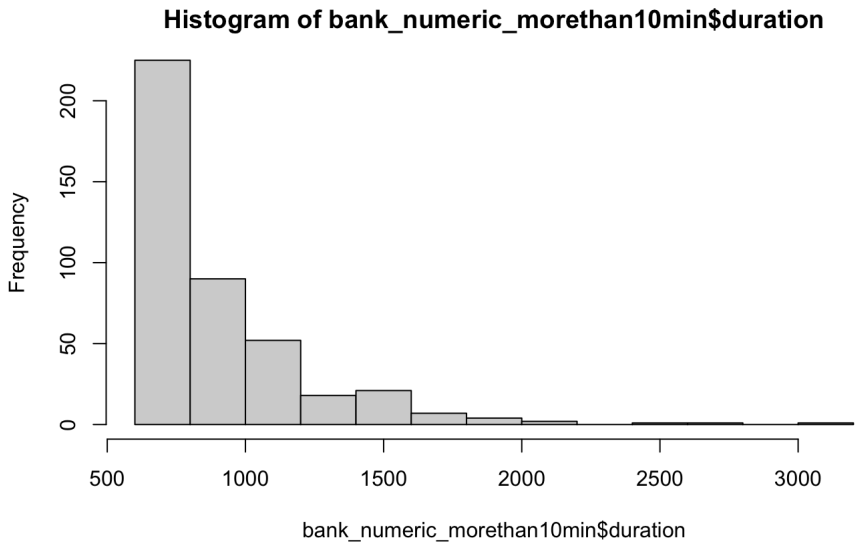
The columns job, martial, education, default, housing, loan, contact, month, poutcome, and y needed to be converted from characters. The column marital was converted with "married" becoming 1, "single" becoming 0, and "divorced" becoming -1. The column education was converted with "primary" becoming 1, "secondary". becoming 2, "tertiary" becoming 3, and unknown becoming NA. the contact column was converted with "cellular" becoming 1, "telephone" becoming 2, and "unknown" becoming NA. The pcoutcome column was converted with "success" becoming 1, the "other" column becoming 0, the "failure" column becoming -1, and "unknown" becoming 0. The default column, housing column, loan column, and y column were converted with "yes" becoming 1 and "no" becoming 0.

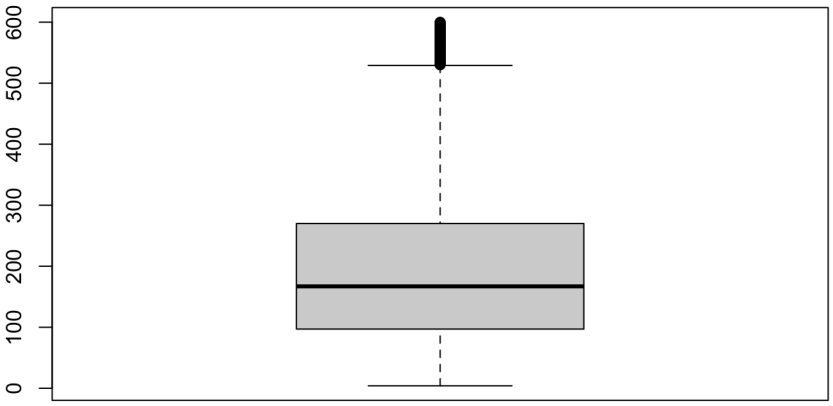
Next, the month column was fixed. It was first converted to Title case then using the months data, it was converted to numeric form. The job column was converted to numeric by converted it to a factor and then unclassing. The new dataframe with all of the columns converted was saved as a separate dataframe.

I wanted to examine the duration column to see how useful it would be. The following three histograms show the histogram of the entire duration column as well as a histogram of the calls where the duration was less than 10 minutes and more than 10 minutes. There is also a box plot of the calls that were less than 10 minutes.


Histogram of bank_numeric$duration
Density
0.0000
0.0010
0.0020
0.0030
0
500
1000
1500
2000
2500
3000
bank_numeric$duration


Histogram of bank_numeric_lessthan10min$duration
Frequency
0
200
400
600
0
100
200
300
400
500
600
bank_numeric_lessthan10min$duration

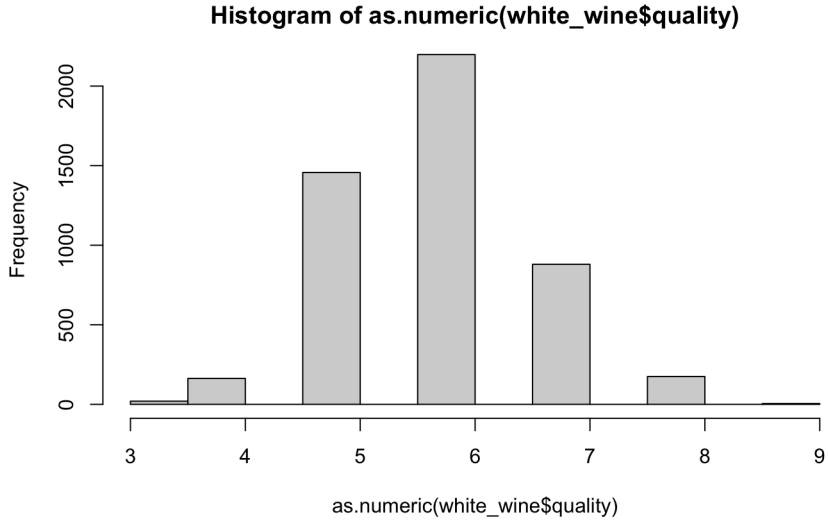




1.2 Wine EDA

There wasn't a much to do with the wine datasets. After looking at a histogram of the quality column of the white wine data set, I determined that instead of trying to classify the quality was the specific numbers, it would be better to classify them as "poor", "normal", and "excellent". "Poor" represented quality between 3 and 4, "normal" represented quality between 5 and 7, and "excellent" represented quality between 8 and 9. "Poor" became -1, "normal" became 0,

and "excellent" became 1. The vast majority of the data was normal quality, with 138 poor quality, 4535 normal quality, and 180 being excellent quality for the white wine data set. The same was done for the red wine data set, which, in general, was much smaller than the white wine data set.



2 Model Development, Optimization, and Tuning

Testing and training sets were creating with a testing sample size of 70%.

2.1 Bank Models

The first model was K Nearest Neighbor.

K	Accuracy %
5	88.50
7	89.09
9	89.01
11	89.01

I chose to go with $k = 9$ neighbors since it had the highest accuracy at 89.09%.

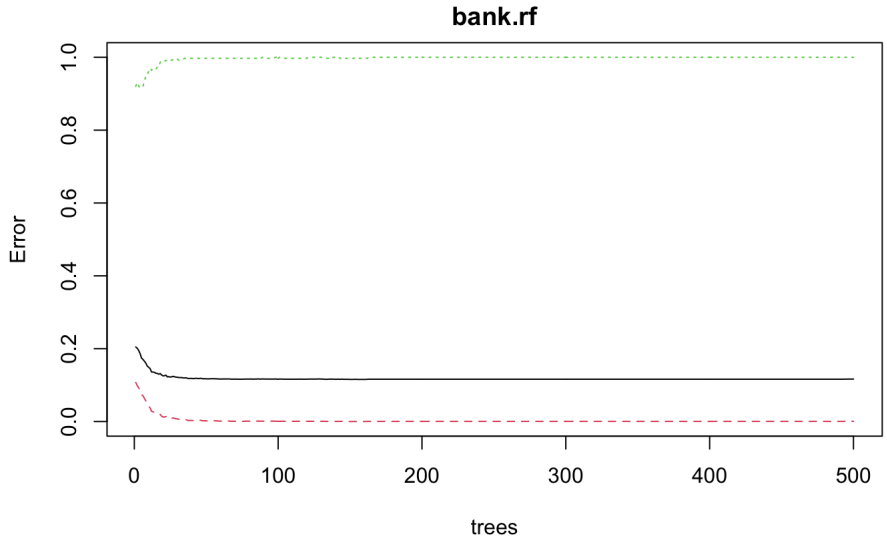
	0	1
0	1207	1
1	148	0

The confusion matrix is shown above.

The next model that was used was kernlab's ksvp. Type "C-svc" was used since I was classifying the data set. It performed with 89.0%, slightly worse than the knn model. The confusion matrix is below.

	0	1
0	1193	163
1	0	0

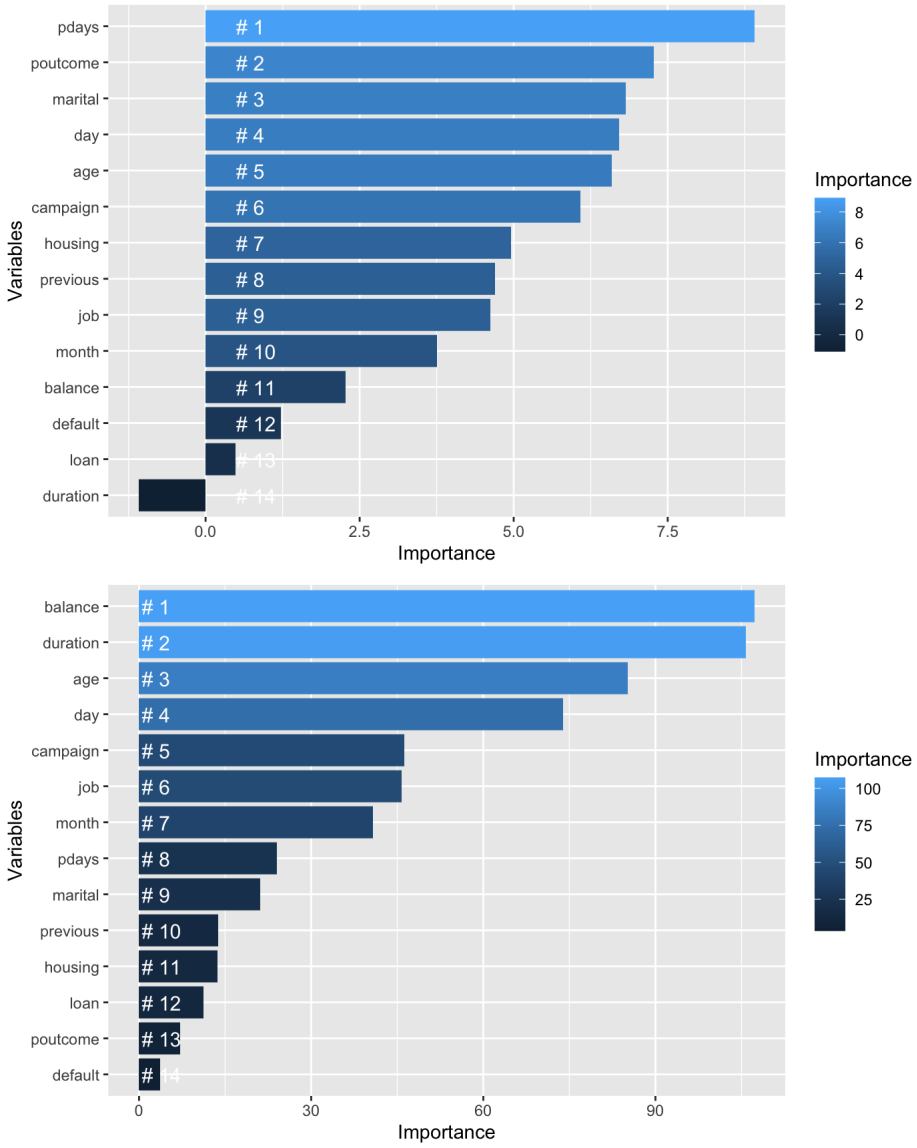
Last, a Random Forest model was created. This model performed the best. 45 trees were chosen with 3 variables tried at each split. A plot showing the error based on the number of trees is shown below.



Using 45 trees performed with an accuracy score of 83.0%. The confusion matrix is shown below.

	0	1
0	1111	146
1	85	14

The importance of the factors used to create the random forest model are shown below using the accuracy score and gini respectively.



2.2 Wine Models

2.2.1 White Wine

The first model was K Nearest Neighbor.

K	Accuracy %
5	92.65
7	92.78
9	92.78
11	92.78

I chose to go with $k = 7$ neighbors since it had the highest accuracy at 92.78% and the accuracy did not meaningfully change going to 9 and 11 neighbors.

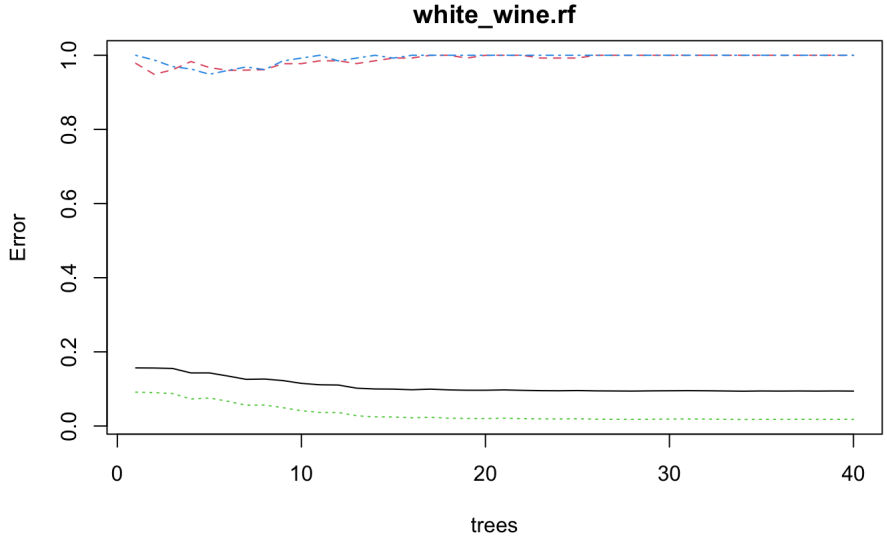
	−1	0	1
−1	0	59	1
0	0	1363	0
1	0	47	0

The confusion matrix is shown above.

Next, the ksvm model was created, using the C-svc type. It performed with 92.78% accuracy, similar to the knn model. Again, it only classified the wines as being normal quality. The confusion matrix is shown below.

	−1	0	1
−1	0	0	0
0	59	1363	47
1	0	0	0

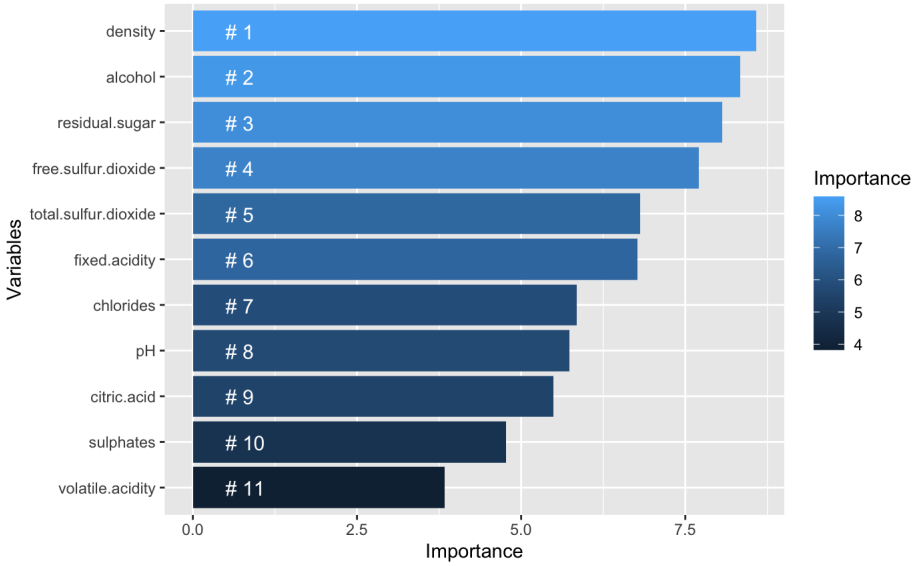
Last, a Random Forest model was made with 40 trees. It performed with 88.97% accuracy. The following plot shows the error based on how many trees.

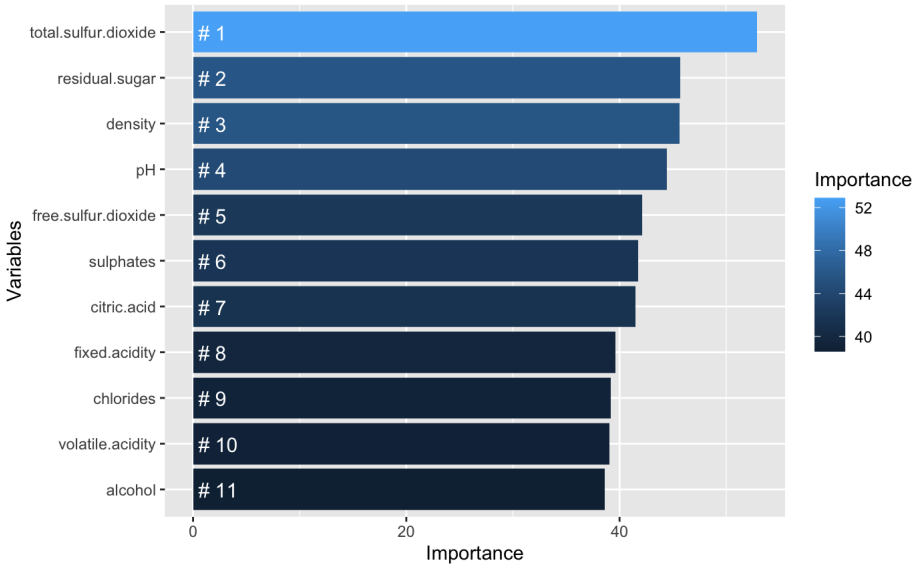


The confusion matrix is shown below.

	−1	0	1
−1	2	28	1
0	55	1305	46
1	2	30	0

We can also examine the importance of factors in creating this model. The following plots show the importance of the factors based on accuracy and gini.I don't really know anything about wine so it's hard to guage





2.2.2 Red Wine

The first model was K Nearest Neighbor.

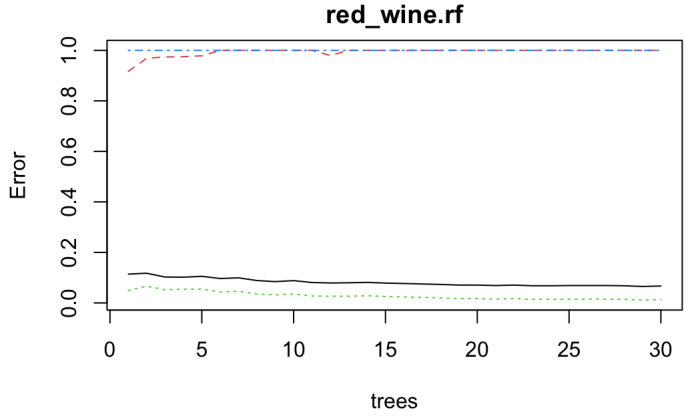
K	Accuracy %
5	93.74
7	93.53
9	93.74
11	93.74

I chose to go with $k = 9$ neighbors since it had the highest accuracy at 93.74% and the accuracy did not meaningfully change going to 11 neighbors.

The confusion matrix below shows similar results to the results from the white wine knn model.

	−1	0	1
−1	0	21	0
0	0	449	0
1	0	9	0

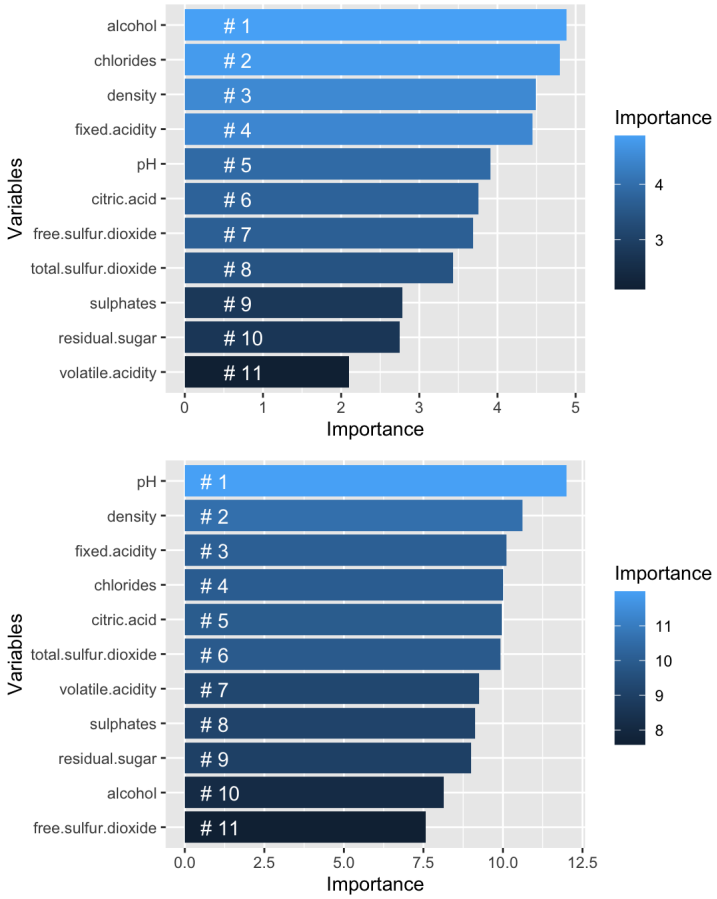
The next model was the ksvm model. It was created with 30 trees and 3 variables tried at each split. The following plot shows the error vs the number of trees.



This model performed with 91.85% error. The confusion matrix is below.

	−1	0	1
−1	1	10	0
0	20	439	9
1	0	0	0

The following plots show the factor importance. Interestingly, these are actually different than the white wine.



3 Decisions

3.1 Bank

As you can see from the confusion matrix, for whatever reason the knn model did hardly classified any consumers as positive outcomes, and none correctly. This is likely due to the fact that the original training data set had very few positive outcome results. As such, this model is good in moderately good at predicting the outcome of the phone call, but not useful because it cannot correctly predict which phone calls will have a positive outcome.

Similar to the knn model, this svm was unable to correctly classify any phone calls as positive outcomes. However, it also did not classify any phone calls as false negatives either. In my opinion, this makes this model a little more useful, but still not very useful because no positive outcome phone calls were correctly classified.

Next, the random forest model. Looking at the confusion matrix, this model

is much more useful. It does correctly classify some positive outcome phone calls. It also has a slightly higher accuracy score in general. Some other interesting things can be gained from this model, such as the most important variables affecting the accuracy and gini. The following plots show these results, respectively. Interesting, variables relating to the date the phone call was made had more weight than I was expecting them to.

In the scenario of marketing calls, it is more harmful to falsely classify positive outcomes as failures than to falsely classify negative outcomes as successes because of potential revenue loss. As such, the random forest model is the best model for classifying marketing calls.

Something that could have improved all of these models would have been to chose to use the larger data sets provided. However, I chose the smaller sets so I could run the svm model. I believe all of these models would perform much better with the larger data set, because that would include more successful outcome calls for which to train the models.

3.2 Wines

Starting with the knn model, looking at the confusion matrix, this model only classified wines as being normal quality. So while it is pretty accurate, it isn't useful at all. The svm model performed similarly. The random forest model, while this is the lowest accuracy of all the models, was able to predict wines as qualities other than normal. Notably, for the red wine data, the training data set had no excellent quality wines, and as such none where classified as excellent quality in the model. This is shown in the confusion matrix. This is arguably a more useful model since it is able to actually classify the wines, instead of just classifying them all as the same quality.

Something to note about this data set is that is was rather small. In the future, better models could be made by including more data, espeically more poor and excellent quality wines.

4 References

http://archive.ics.uci.edu/ml/datasets/Bank+Marketing
http://archive.ics.uci.edu/ml/datasets/Wine+Quality