**C5T3 Lessons Learned**

Jelly P. Edwards

**iPhone**

I chose the RandomForest Classifier over C5.0, kknn, and SVM. This had the best Accuracy and Kappa out of all the classifiers experimented with. After checking the predictions and testing results with confusionmatrix and postResample, it was clear that RF provided the best results with the highest Accuracy and Kappa. For Feature Selection, I preferred the results of iPhone RFE. The RFE function tried every combination of features for the iPhone data set and returned a list of recommended features. It had higher Accuracy and Kappa compared other feature selections. For this method, the features selected were: iPhone, googleandriod, iphonedispos, iphonedisneg, samsunggalaxy, htcphone, iphonedisunc, iphoneperpos, ios, iphoneperneg, sonyxperia, iphoneperunc, iphonecampos, iphonecamneg, iphonecamunc, htcdisunc, htccampos, htcperpos, and htccamneg. I eliminated the other features to help improve computation time and to remove any redundant data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | RF | C5.0 | kknn | SVM |
| Accuracy | 0.78 | 0.78 | 0.34 | 0.70 |
| Kappa | 0.57 | 0.57 | 0.18 | 0.41 |

***Figure 1.*** *PostResample Results for Out of the Box All Feature Tests (iPhone)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | NZV C5.0 | RFE C5.0 | NZV RF | RFE RF |
| Accuracy | 0.76 | 0.77 | 0.76 | 0.81 |
| Kappa | 0.52 | 0.55 | 0.53 | 0.65 |

***Figure 2.*** *Comparing the Confusion Matrix Metrics between NZV and RFE feature selections and RF and C5.0 Classifiers (iPhone)*

**Galaxy**

Once again, I chose the RandomForest Classifier over C5.0, kknn, and SVM. It also had the best Accuracy and Kappa out of all the classifiers. After checking the predictions and testing results with confusionmatrix and postResample, it was clear that RF provided the best results with the highest Accuracy and Kappa. For Feature Selection, I once again preferred the results of RFE for Galaxy. It had higher Accuracy and Kappa compared other feature selections. For this method, it had every feature except: id, nokiadisneg, and sonycamneg. I eliminated the other features to help improve computation time and to remove any redundant data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | RF | C5.0 | kknn | SVM |
| Accuracy | 0.76 | 0.76 | 0.72 | 0.69 |
| Kappa | 0.52 | 0.52 | 0.47 | 0.35 |

***Figure 1.*** *PostResample Results for Out of the Box All Feature Tests (Galaxy)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | NZV C5.0 | RFE C5.0 | NZV RF | RFE RF |
| Accuracy | 0.76 | 0.77 | 0.77 | 0.80 |
| Kappa | 0.60 | 0.55 | 0.53 | 0.60 |

***Figure 2.*** *Comparing the Confusion Matrix Metrics between NZV and RFE feature selections and RF and C5.0 Classifiers (Galaxy)*

**Lessons Learned**

Improving the core usage for RStudio was a major savior for this project. Computation time was an important factor and increase cluster size decreased the amount of time used to run each query. The improvement gave me liberty to experiment with several algorithms without worry that the system would take too long to run through them. Another important factor to this project was running the entire process with the iPhone small data set BEFORE running the same process for Galaxy. Keeping them separated like that eased my mind and I didn’t get the two mixed up. Doing both at the same time would have been more hectic.

In the instructions, we were asked to find correlation for the data sets and to remove any features that had high correlation. Unfortunately, my correlation with the dependent variable never had any high values. This left me without one extra sample data set to test. This wasn’t a large issue, just something to take note of.

When working with the Large Data Matrix, I found that I was unable to repeat the Feature selection processes that I did with the small data matrix. This was due to the fact that the sentiment column (iphonesentiment and galaxysentiment) was empty, which prevented the system from correctly performing the Feature Selection processes. Due to this, I had to manually remove any features I previously deleted in the small matrix from the large matrix. I now know that these require data inside of the dependent variable.

Overall, I found this project to be smoother that others. It was easy to understand, and the coding did not give me as much trouble as others. Part 7 of the POA was a major help with writing the report, and I really liked the provided questions that helped me think about the problem at hand.