**Summary**: The RandomForest (“rf”) predictive model was chosen for the Galaxy and iPhone sentiment datasets. The rf model had the highest accuracy and kappa scores, but did not show signs of overfitting. For both datasets, 46 out of 60 attributes were used for sentiment prediction for each dataset after running feature selection code. See subsequent pages for the attributes and the accuracy metrics for each dataset.

**Lessons learned:**

**Items that worked well**

Sampling the dataset will reduce the runtime for training predictive models, but make sure to get a sufficient sample size. For this project, 4000 of 20K observations were sampled. Utilizing parallel processing further reduced runtime for training. Do not assign all of your CPU cores to the parallel clusters as you will need 1 or 2 cores to run background processes on your PC. Using the “findCorrelation” function for Feature selection helps prevent overfit by removing items with high covariance (0.80 or more). Note that a correlation matrix needs to be generated prior to running this function. By removing attributes, training predictive model runtime is reduced.

The discretize function allowed us to change the sentiment attribute from a numeric to factor data type. This changed predicting sentiment into a classification problem instead of a regression problem. This simplified the data model and helped us determine whether user sentiment was negative, neutral, or positive in a more effective manner. Also, it was easier to graph the distribution of user sentiment for a phone when you assign factors. To quickly summarize the accuracy metrics for the predictive models, use the “dotplot” function on the resample R list object. This allows you to easily describe your results to a non-technical audience.

**Future recommendations**

Remove unique identifiers like the “id” attributes, which have no predictive power. This will help the model run more effectively. Learn how to run RStudio sessions in AWS EC2. All models could be run at once across multiple EC2 instances to reduce model optimization selection runtime. Please note the setup for the process is nuanced, so it will take time to learn. By learning how to run RStudio on AWS now, it will reduce your runtime for all future projects that allow for parallel processing. In addition, this allows you to run scripts effectively even if you have an old PC.

Finally, encasing comments like so (#### feature selection ####) allows users to create collapsible sections to make the R script more organized and easier to review.

**Galaxy sentiment modeling**

**Classifier Selected**: RandomForest(“rf”)

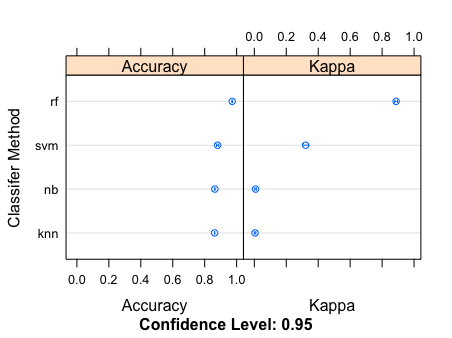
46 Attributes used: I removed all attributes that had a covariance over 0.80 to prevent overfitting of data, which did not remove our dependent attribute (GalaxySentiment). In addition, I should have removed the “id” attribute because it is a unique identifier

Count

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | id | iphone | samsunggalaxy | sonyxperia | nokialumina | htcphone |
| 7 | iphonecampos | sonycampos | nokiacampos | iphonecamneg | nokiacamneg | sonycamunc |
| 13 | samsungdispos | sonydispos | htcdispos | iphonedisneg | samsungdisneg | sonydisneg |
| 19 | iphonedisunc | samsungdisunc | sonydisunc | htcdisunc | iphoneperpos | samsungperpos |
| 25 | iphoneperneg | samsungperneg | sonyperneg | htcperneg | iphoneperunc | samsungperunc |
| 31 | iosperpos | googleperpos | iosperneg | googleperneg | iosperunc | googleperunc |
| 37 | ios | googleandroid | nokiacamunc | iphonedispos | nokiadisneg | htcdisneg |
| 43 | sonyperpos | htcperpos | nokiaperunc | htcperunc |  |  |

**Classifier metric summary**:

Classifiers used: rf, Support Vector Machines (“svm”), K Nearest Neighbor (“knn”), and Naïve Bayes(“NB”)



**IPhone sentiment modeling**

**Classifier Selected**: RandomForest(“rf”)

46 Attributes used: (although I should have removed “id” because it is a unique identifier). I removed all attributes that had a covariance over 0.80 to prevent overfitting of data, which did not remove our dependent attribute (IPhoneSentiment).

Count

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | id | iphone | samsunggalaxy | sonyxperia | nokialumina | | htcphone |
| 7 | iphonecampos | sonycampos | iphonecamneg | nokiacamneg | sonycamunc | | nokiacamunc |
| 13 | sonydispos | htcdispos | iphonedisneg | samsungdisneg | sonydisneg | | nokiadisneg |
| 19 | samsungdisunc | sonydisunc | htcdisunc | iphoneperpos | samsungperpos | | sonyperpos |
| 25 | iphoneperneg | samsungperneg | sonyperneg | htcperneg | iphoneperunc | | samsungperunc |
| 31 | iosperpos | googleperpos | iosperneg | googleperneg | iosperunc | | googleperunc |
| 37 | ios | googleandroid | iphonedispos | samsungdispos | htcdisneg | | iphonedisunc |
| 43 | nokiaperpos | htcperpos | nokiaperunc | htcperunc |  |

**Classifier metric summary**:

Classifiers used: rf, Support Vector Machines (“svm”), K Nearest Neighbor (“knn”), and Naïve Bayes(“NB”)

