Analysis of Popular Smartphone Handset Sentiment

Kate Koebbe
Alert! Analytics

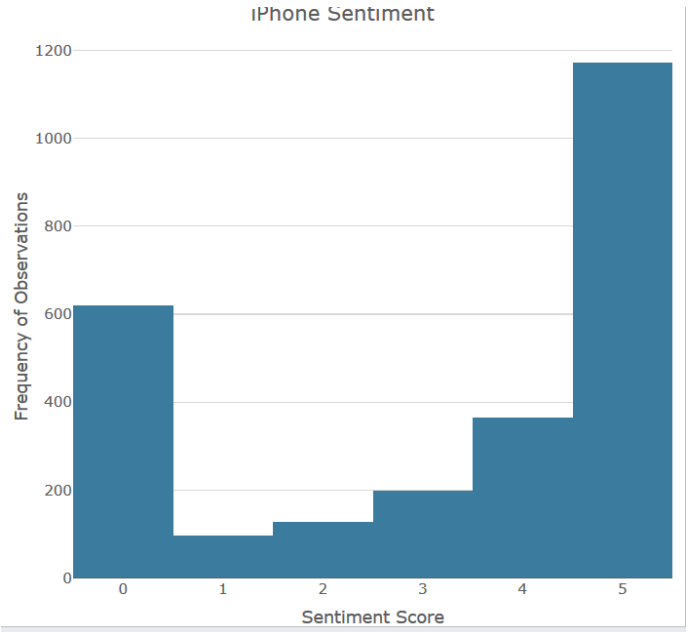
OVERVIEW

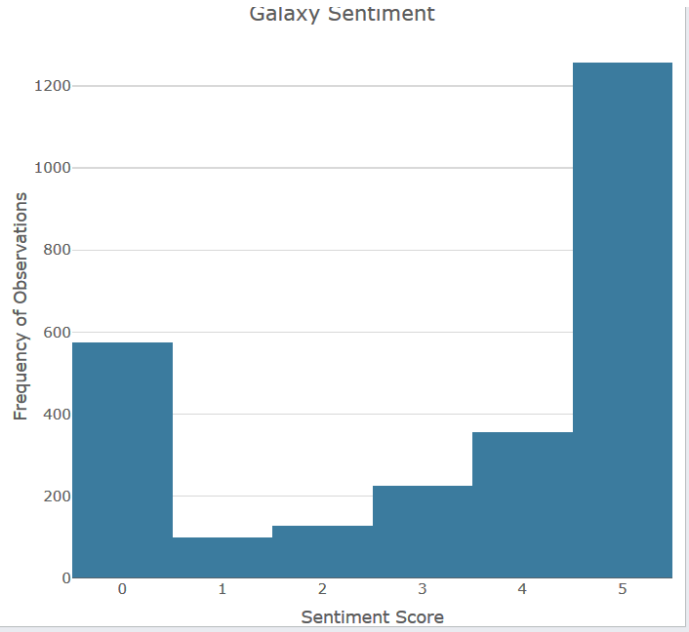
Our client, Helio, is developing of a suite of smartphone medical applications for use by aid workers to manage local health conditions. The applications help facilitate communications between the workers and medical professionals, who may be located elsewhere in the world. The implications of such technologies may have far-reaching impacts on communities in the developing world, potentially enabling specialists in communicable diseases to diagnose conditions by examining images and other patient data collected by local aid workers. To assist in the development of this technology, Helio contracted the Alert! Analytics team to conduct a broad-based sentiment analysis for popular smartphone handsets, as the suite of applications must be bundled with one handset model. Helio's team narrowed our analysis to two handset models capable of running the application software, iPhone and Samsung Galaxy, and it is our task to determine which model generates the most positive sentiment among consumers.

METHODOLOGY

Our methodology follows a typical data science process consisting of the following steps: 1) gather and evaluate data sets, 2) preprocess data for ease of manipulation in visualization and modeling, 3) analyze and visualize data to discover underlying trends, 4) select and train models, and evaluate model performance, and 5) validate the top performing model and make predictions.

Our data sets are scraped from numerous smartphone-related web pages, sourced from an online repository called *Common Crawl*. Alert! Analytics team members manually developed "small" (12,000+ observations) data matrices for each the iPhone and the Samsung Galaxy by reviewing webpages and assigning sentiment scores (0-5 scale). We then assembled data for a "large" (23,701 observations) matrix to make our predictions.

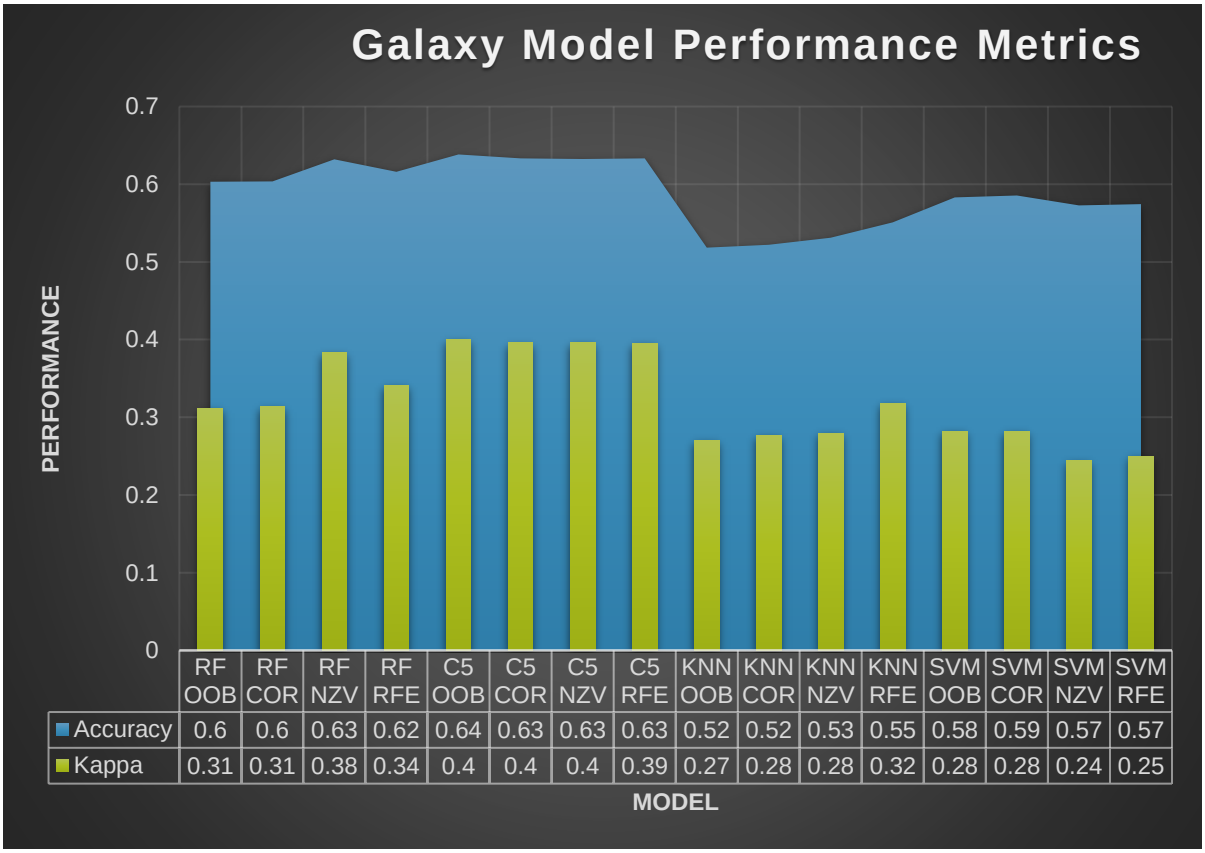




The figures above illustrate the distribution of the sentiment scores, later described as sentiment classes, assigned by Alert! Analytics team members to the instances in the iPhone and Samsung Galaxy matrices; as we can see, the distributions are very similar, with the highest frequency of observations clustered at the highest and lowest scores.

We employed several feature selection techniques to build smaller data sets to see if sets with fewer variables would result in improved accuracy. Fewer variables reduces the amount of "noise" created by redundant data and often reduces the amount of model training time, making our process more efficient. For our small matrices we used correlation analysis to reduce the number of highly correlated variables, resulting in a data set if 50 variables; feature variance analysis to remove variables with zero or near-zero variance, resulting in a data set with 14 variables; and recursive feature elimination, which uses a machine learning algorithm (in this case, Random Forest) to select important features and discard unimportant features, to create a data set with 18 variables. In total, our model training included **eight** data sets: an "out-of-the-box" (OOB) set, a correlation set (COR), a set for near-zero variance (NZV), and a recursive feature elimination set (RFE) for each handset model.

FINDINGS



	RF OOB	RF COR	RF NZV	RF RFE	C5 OOB	C5 COR	C5 NZV	C5 RFE	KNN OOB	KNN COR	KNN NZV	KNN RFE	SVM OOB	SVM COR	SVM NZV	SVM RFE
Accuracy	0.6	0.6	0.63	0.62	0.64	0.63	0.63	0.63	0.52	0.52	0.53	0.55	0.58	0.59	0.57	0.57
Kappa	0.31	0.31	0.38	0.34	0.4	0.4	0.4	0.39	0.27	0.28	0.28	0.32	0.28	0.28	0.24	0.25

We selected four classification algorithms to see which yielded the best fit for our out-of-the-box and feature-selected data sets: Random Forest Classifier (RF), C5.0 (C5), K-Nearest Neighbors (KNN), and Support Vector Machines (SVM). To evaluate performance, we used

accuracy, the number of instances that were classified correctly, and *kappa*, a comparison of observed accuracy and expected accuracy. Generally, the higher the score for each metric, the better fit the model is for our data.

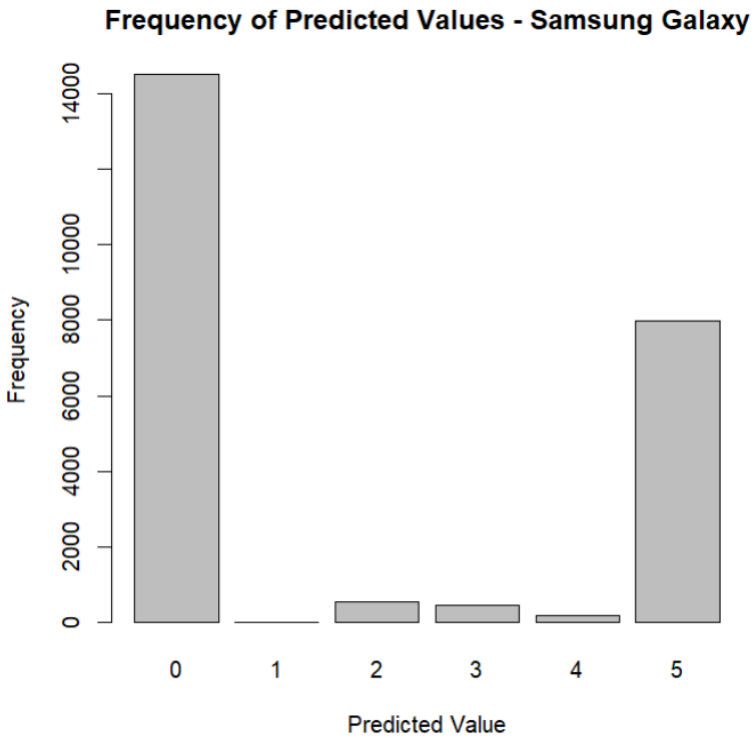
GALAXY

The chart on page 2 illustrates the performance of each model with each data set for our Samsung Galaxy data. We also resampled the data to measure the accuracy of our metrics. For our four top performing models, all using the C5.0 algorithm, the resampling metrics were in-

Samsung Galaxy Resampling Metrics		
Model	**Mean Accuracy**	**Mean Kappa**
C5.0 OOB	0.6415094	0.4001069
C5.0 COR	0.6314363	0.3970662
C5.0 NZV	0.6361186	0.3963900
C5.0 RFE	0.6341463	0.3949042

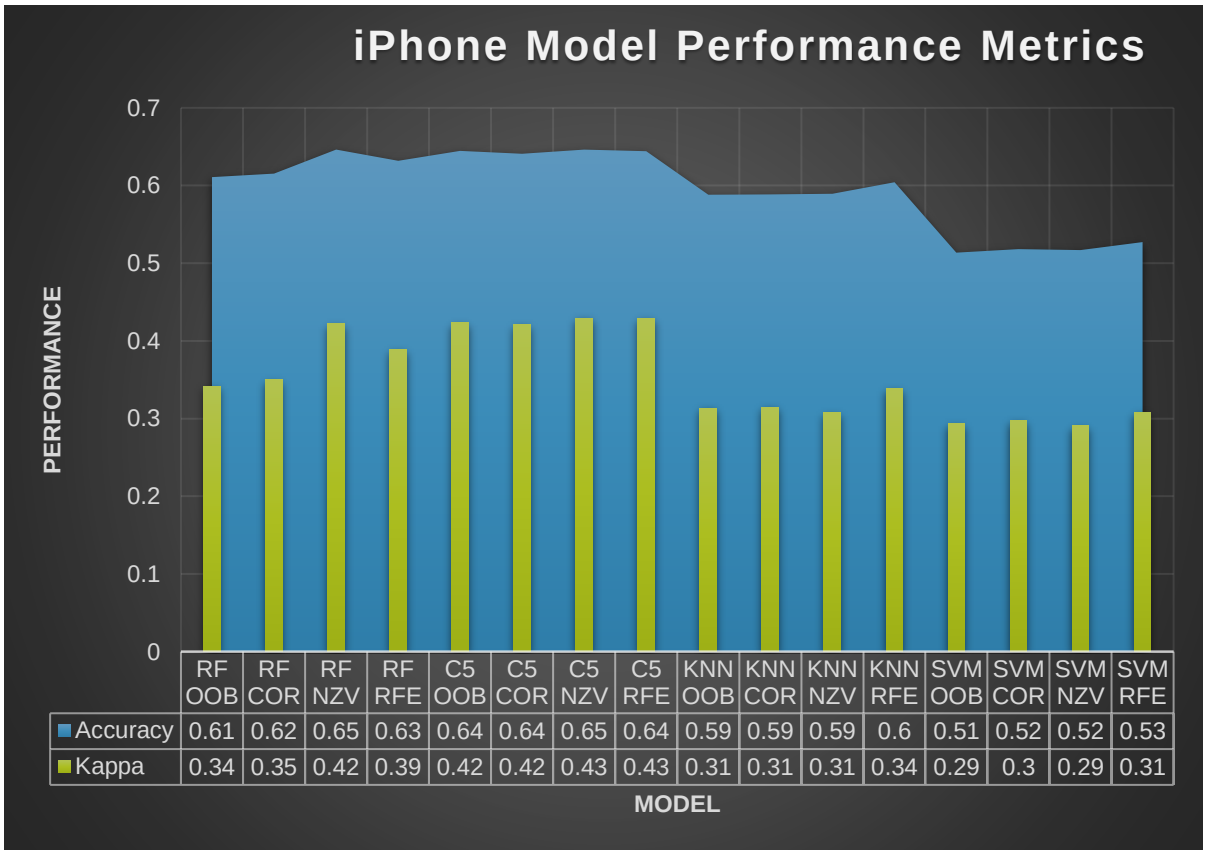
line with our model performance. Thus, we assume we can expect our top performing model, C5.0 OOB, to make predictions with about 63% accuracy for the Galaxy. A more focused look at the metrics, however, reveals some concerns. The low *kappa* statistic is the first indicator of concern – generally, you'd want to see a *kappa* statistic similar to that of the model's accuracy, as that would mean what was expected is what was observed. But that is not what we see. Further analysis of metrics generated by a confusion matrix generated with data from our top performing model, as shown in the table below, illustrates our quandary at the sentiment class level. In the matrix, the rows represent the predicted sentiment class and the columns the actual sentiment class. We see significant discrepancies between the predictions and the outcomes in several classes. For example, C5.0 OOB was not able to accurately predict the "1" sentiment class; instead, all outcomes (29) for this class were "5." The resulting accuracy of the class "1", called the *specificity* or true positives, is 0%. The specificity values for all sentiment classes are included in the matrix. In sum, while the model was table to predict sentiment values of "0" and "5" with higher levels of accuracy, it was not able to predict sentiment values in between with much accuracy at all.

Samsung Galaxy Confusion Matrix						
	0	**1**	**2**	**3**	**4**	**5**
0	123	0	2	4	6	14
1	0	0	0	0	0	0
2	0	0	1	0	0	0
3	0	0	2	13	1	11
4	3	0	3	2	20	8
5	45	29	30	49	80	344
Specificity	0.7193	0.00	0.02632	0.19118	0.18692	0.9125



The final step in our Samsung Galaxy analysis was making predictions with our “large” data set with over 23,000 observations. The chart below illustrates the values predicted by our C5.0 OOB models using the large data matrix. We found the overwhelming majority of the predictions to be skewed toward the “0” sentiment class, the number of observations for this variable being double the number of values with the predicted observations of “5.” Thus, the distribution of our predictions is the reverse of our distribution in the small data matrix.

iPhone



iPhone Resampling Metrics		
Model	**Mean Accuracy**	**Mean Kappa**
C5.0 OOB	0.6456044	0.4235068
C5.0 COR	0.6371191	0.4207564
C5.0 NZV	0.6436464	0.4289801
C5.0 RFE	0.6398892	0.4287880

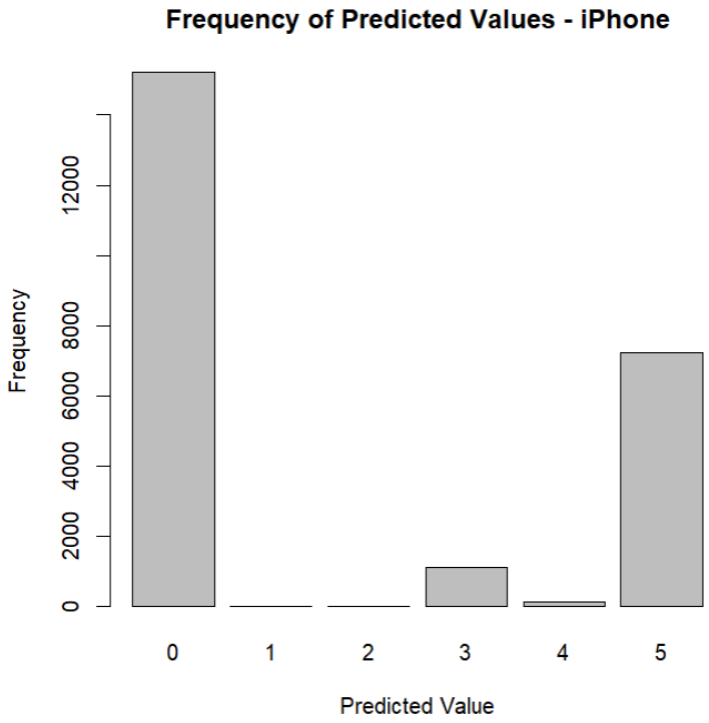
For our analysis of iPhone sentiment, we followed the same methodology and used the same models as those used for the Samsung Galaxy. A visual representation of the iPhone model performance is shown above. The C5.0 models, again, outperformed other algorithms, suggesting that this method is the best fit for data sets like this. Unlike the Samsung Galaxy small matrix, however, a data set employing feature selection, the near-zero variance set (C5.0 NZV), appears to have the highest accuracy and kappa values (0.6462204 and 0.4289801, respectively). Resampled values are largely the same as the model performance metrics and confirm the accuracy of our top model selection.

Also much like our Samsung Galaxy data, the confusion matrix is skewed towards the highest and lowest sentiment classes. However, the C5.0 algorithm applied to the iPhone NZV

IPhone Confusion Matrix						
	0	**1**	**2**	**3**	**4**	**5**
0	134	0	4	1	7	8
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	1	0	0	21	5	8
4	3	0	1	1	31	6
5	48	29	33	36	67	329
Specificity	0.7204	0.00	0.00	0.35593	0.28182	0.9373

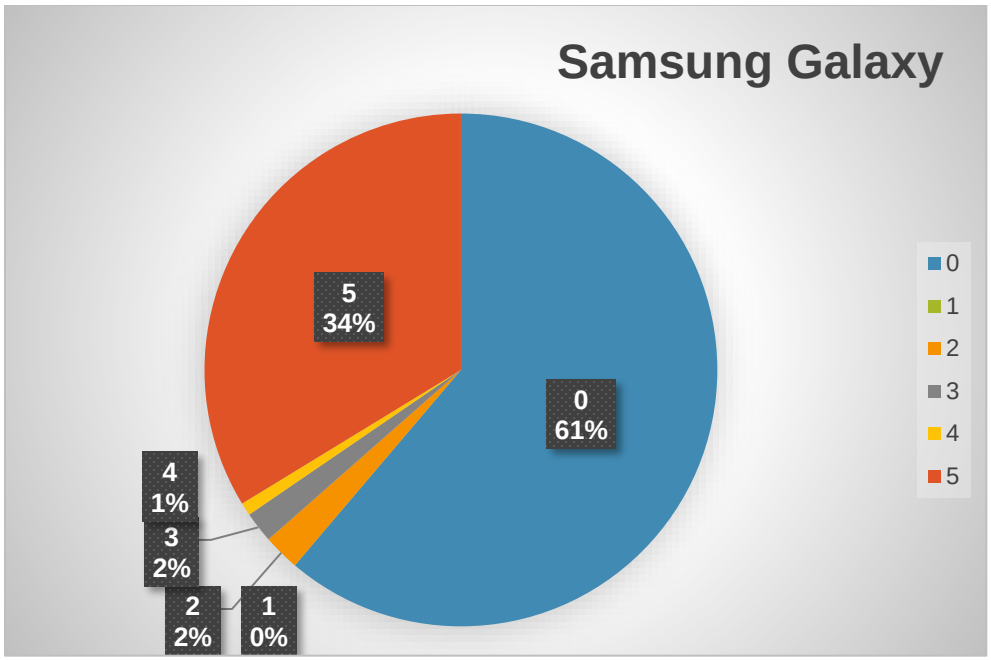
data set reveals important distinctions: the iPhone's C5 model has a lower chance (0%) of predicting of predicting the sentiment class "2," but it has significantly higher rates of predicting true values of "3" and "4."

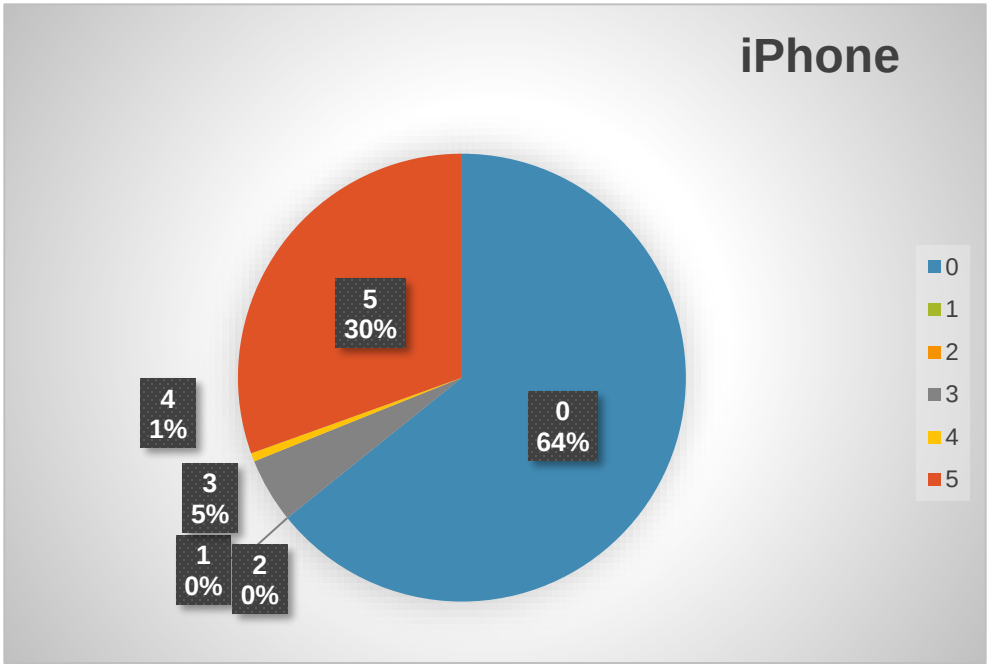
Finally, when considering the predictions made based upon the data with our large matrix containing over 23,000 observations, again, the distribution is largely skewed towards the lowest sentiment class, "0," with the highest sentiment class, "5," having roughly half of that number. Also similar to the Samsung Galaxy predictions, we see a distribution shape that is the reverse of the distribution in the small matrix: the "in between" sentiment classes represent a very small percentage of the predicted values.



IMPLICATIONS

When comparing the metrics of the Samsung Galaxy and iPhone models, we can see many similarities and some slight differences, making the decision of which handset to select rather difficult. Both models have accuracy and kappa metrics that are somewhat underwhelming, largely due to both models' inability to predict the "in between" sentiment classes. To make our handset recommendation, we first have to address an unknown in this analysis – what do the values in the sentiment classes represent? Typically, we would expect lower values to represent lower ratings or negative sentiment, and higher values to indicate higher ratings or positive sentiment. We'll use this assumption and apply percentages to each sentiment class to both the Samsung Galaxy and iPhone predictions. Two charts illustrate the comparison below. Overall, the Samsung Galaxy has a higher percentage of positive sentiment, assuming the sentiment class "5" is positive, and a lower percentage of negative sentiment, assuming the sentiment class "0" is negative, making the Samsung Galaxy our tentative choice for Helio's preferred handset for use with their application. However, it's worthy to note that the differences here in positive and negative sentiment are very slight, so other business considerations, such as the bulk cost of handsets or compatibility with other applications in use by the aid workers, could easily sway a recommendation in favor of the iPhone.





How might we be able to improve our models? First, sentiment, especially when categorized manually by human analysts, is very subjective. Where one individual classifies "OK" as a "2" another person may categorize it as a "3." To obtain a more reliable measure of consumer sentiment, surveys with similar sentiment classes and choices reflecting only these sentiment classes can be conducted. We might also use more standardized approaches to categorizing sentiment if consumer surveys are not a possibility. Second, and as with most projects, more data would help to make clearer the distinctions between Galaxy and iPhone model performance and overall consumer sentiment. Our "large" matrix was roughly twice the size of our smaller matrices; perhaps a matrix three, five, or ten times the smaller matrices would only serve to enhance the distinctions and make the choice clearer.

LESSONS LEARNED

[For Logan]

This was the most enjoyable task I've worked on in the data analytics program, likely because I got to apply what I've learned into one pipeline. In this instance, I completed the pipeline with relative ease, thanks to considerable agonizing and suffering in other tasks, which made writing the report that much easier as well. I was surprised, though, how much clicked into place once I started reviewing and analyzing performance metrics. The information provided in the confusion matrices are a good example. Where I could only vaguely interpret the measures' meanings in previous tasks, I was able to really dig into some of the values in this one. I may not understand it all perfectly at this point, but I do feel more confident in my justifications.

If I could change something about this task, it would be to complete a PowerPoint for non-technical audiences and a report for those who are more technically savvy. PowerPoints are much more focused on visualization, which is better for the layperson in terms of making sense of all the data. Reports lend themselves to text, and I can see those in the non-technical audiences getting lost in the explanations, no matter how simplified they are.

One thing that was particularly challenging in this task was deciding which model was my top performing model, specifically for the iPhone data. I could identify which model has the best accuracy and kappa, but the results of resampling (where the OOB set had better metrics) had me questioning whether the NZV set was truly the best model. I think the distinction was so fine that, ultimately, the choice between the models didn't matter much for this task; however, if a larger data set (like 50,000 instances) might make the distinctions much clearer. So, if I were to complete this task again in the future, I may consider both OOB and NZV data sets. What was pretty clear, though, was that C5.0 was our top algorithm.

Another thing that was challenging was including just enough information in the report to explain my position without being overly "technical." With non-technical audiences, there should be some explanation of the concepts, especially if those same concepts illustrate an important trend in the data. For example, my explanation of the confusion matrices and sensitivity may be a bit too technical for most audiences, but I felt it was important to highlight that the models did not do well at all predicting the "in between" values.

In the future, I would like to employ more techniques to refine my models, like tuning and feature engineering. I am not sure what kind of impact they would have had on these data sets, but even if minimal, if could have helped to make those small distinctions that much clearer.

Overall, I'm glad this task was the last one I had to complete, and not C4T3. I feel like I'm ending the program on a positive note and now have more desire to explore more data. Hopefully my scripts and the report above demonstrate how much I've learned in the past six months!