**C5T3 – Michelle Giniewicz – Findings Report**

**iPhone & Galaxy Sentiment Analysis**

For each of the iPhone and Galaxy sentiment datasets, I created four different versions of the datasets using feature selection: 1) the out of the box (OOB) dataset with no features changed, 2) the OOB dataset with highly correlated features removed, 3) the OOB dataset with Near Zero Variance features removed, and 4) the OOB dataset with Recursive Feature Elimination, an automated feature selection.

I started each of the iPhone and Galaxy analyses on the corresponding OOB dataset and used it to train and test the C5.0, Random Forest, SVM, and kknn models. The models that performed the best were Random Forest and C5.0, so I then used my three remaining datasets to train and test those two models. After performing this analysis, my best model for iPhone was Random Forest for the OOB dataset at 77% accuracy and 0.56 Kappa score, and my best model for Galaxy was C5.0 for the OOB dataset at 77% accuracy and 0.54 Kappa score. All of the other datasets that had features removed were less accurate.

I then performed feature engineering on both the iPhone and Galaxy OOB datasets, starting with engineering the dependent variable by reducing the sentiment values from six (0-5) to only four values (1-4). Using this engineered dataset with Random Forest, I was able to get my iPhone model to perform even better – 85% accuracy and 0.62 Kappa Score. Using the same feature engineering for the Galaxy dataset with C5.0, I was able to get my Galaxy model to perform better as well – 85% accuracy and 0.62 Kappa Score. I also used feature engineering to do Principle Component Analysis; however, that actually reduced my model accuracy on both the iPhone and Galaxy datasets.

**What Worked Well**

Following the pipeline made it easy to keep things very organized for this analysis. I was a little overwhelmed at first since we had to run four different datasets through up to four different models. However, following the pipeline I made sure to do all four models on the OOB dataset first, then worked through each of the other datasets using the best model. I also found that creating the histograms and pie charts made it easy to compare the iPhone and Galaxy sentiment.

**What Didn’t Work Well / What was Difficult**

I had a hard time with the feature engineering where you reduce the sentiment values from six to four. I was unable to complete it based on the recode() example given in the POA. I had to research other ways to perform this change and ended up having success with a case\_when statement. I also had to be really careful when re-running the modeling with the Galaxy dataset after completing the iPhone dataset. There were a few places I forgot to change “iPhone” to “Galaxy” in my analysis and ended up with iPhone sentiment in my Galaxy dataset. Thankfully, this was pretty obvious using the str() and summary() functions, and I was able to quickly resolve the issue and get the correct sentiment data added back into the dataset.

**How the Process Should be Changed**

I honestly cannot think of any ways the process should be changed. Everything worked pretty seamlessly for me by following our pipeline, and I really enjoyed this as the final project of the course. I also found this project really interesting to me personally, since I have started looking at employee sentiment in my job, managing our first ever internal IT Net Promoter Score survey at my company.