Helio sentiment analysis project

by

Qing Wu

Alert! Analytics

September 10th, 2018

**Overview and Business goal:**

Our client is Helio, a smart phone and tablet app developer. Helio is now working with a government health agency to create a suite of smart phone medical apps for use by aiding workers in developing countries. The government agency requires that the app suite be bundled with one model of smart phone. Helio has created a short list of five devices that are all capable of executing the app suite’s functions. To help Helio narrow their list down to one device, we have been asked to examine the prevalence of positive and negative attitudes toward these devices on the web.

**Our Approach to the Project:**

Our general approach to this project is to count words associated with sentiment toward these devices within relevant documents on the web. We look for positive, negative, and neutral words that are within five words of a feature word (camera, display, performance) or an alternate of a feature word (e.g., iSight). We then leverage this data and machine learning methods to look for patterns in the documents that enable us to label each of these documents with a value that represents the level of positive or negative sentiment toward each of these devices. We then analyze and compare the frequency and distribution of the sentiment for each of these devices.

**Description of data sources:**

In order to really examine the sentiment toward these devices, we must do this on a very large scale. We use the cloud computing platform provided by Amazon Web Services (AWS) to conduct the analysis. The data sets we analyze will come from Common Crawl. Common Crawl is an open repository of web crawl data (over 5 billion pages so far) that is stored on Amazon’s Public Data Sets. We take advantage of Python mapper, reducer, and output aggregator programs to efficiently collect and compile this data across the billions of documents on the Common Crawl. She then put this data into a matrix that contains approximately 12,000 entries (we call it the small data matrix). This small data matrix is manually labeled with a sentiment rating that we will use later in the process to develop machine learning models capable of determining web page sentiment automatically. We compile data from 400 to 500 archive files from the Common Crawl. This give us around 20,000 web pages that contain references to smart phones. We develop a data matrix in the range of 20 thousand instances (called the large data matrix) through AWS including Elastic Compute Cloud (EC2), Elastic MapReduce (EMR), and Simple Storage Service (S3).  We apply our models to the large matrices to predict their sentiment scores.

**Attribute Information:**

Each row in the csv file is a webpage from the Common Crawl that is relevant to the analysis.

1. Attributes that collect information about the relevancy of the webpage toward each device (columns A-E).
2. Attributes that collect information about the sentiment toward the operating system used on the phone.  (columns F-G)
3. Attributes that collect information about the sentiment toward a phone’s camera (columns H-V).
4. Attributes that collect information about the sentiment toward a phone’s display (columns W-AK)
5. Attributes that collect information about the sentiment toward a phone’s performance (columns AL-BF).

**Model building, tuning and feature engineering for iPhone**

**1.out of the box modeling for iPhone**

Six classifiers-C5.0, Random Forest, SVM, SVM Radial, KKNN and GBM are tried and all features are used to train the model. Small matrix is partitioned into training (.70) and test dataset (.30).

For classifier C5.0, 10-fold cross validation is applied and model is built by using an Automatic Tuning Grid with a tune Length of 10. Model with trial=1, model=tree and winnow=FALSE has highest accuracy and kappa. The accuracy and kappa of cross-validation are 0.7723770, 0.5585957 respectively. This model is used to predict test dataset, and the accuracy and kappa of test dataset are 0.7666, 0.5461 respectively.

For classifier Random Forest, 10-fold cross validation is applied and model is built by using an Automatic Tuning Grid with a tune Length of 10. The model with mtry=20 had the highest kappa. The accuracy and kappa of cross-validation are 0.7760432, 0.5664790 respectively. This model is used to predict test dataset, and the accuracy and kappa of test dataset are 0.7751, 0.5632 respectively.

For classifier SVM Linear, 10-fold cross validation is applied and model is built by using an Automatic Tuning Grid with a tune Length of 5. Model with C=1 was the best model. The accuracy and kappa of that model were 0.7111868, 0.4191533 respectively, which is lower than C5.0 and Random Forest.

For classifier SVM Radial, 10-fold cross validation is applied and model is built by using an Automatic Tuning Grid with a tune Length of 5. Model with sigma = 3.343925 and C = 1 was the best model. The accuracy and kappa of that model were 0.7381494, 0.4701697 respectively, which is lower than C5.0 and Random Forest.

For classifier KKNN, 10-fold cross validation is applied and model is built by using an Automatic Tuning Grid with a tune Length of 5. Model with kmax = 13, distance = 2 and kernel = optimal was the best model. The accuracy and kappa of that model were 0.3460640, 0.1706907 respectively, which was quite low.

For classifier GBM, 10-fold cross validation is applied and model is built by using an Automatic Tuning Grid with a tune Length of 3. Model with n.trees = 150, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10 was the best model. The accuracy and kappa of that model were 0.7731814, 0.5619847 respectively. This model is used to predict test dataset, and the accuracy and kappa of test dataset are 0.771, 0.5555 respectively.

**2. Improving the performance of model with feature selection for iPhone**

Three feature selection methods including Correlation method, near zero variance method, Recursive Feature Elimination method are used to improve the performance of the model.

Using Correlation method, correlation of variables of the small matrix for iPhone sentiments is examined. Some features that had collinearity were detected and some highly-related features are removed. New dataset that contains 34 predictors are created to develop model. Classifier Random Forest are used to develop the model with 10-fold cross validation and an Automatic Tuning Grid with a tune Length of 5. Model with mtry=10 had the highest kappa. The accuracy and kappa of cross-validation are 0.7548172, 0.5209051respectively. This model is used to predict test dataset, and the accuracy and kappa of test dataset are 0.754, 0.5162 respectively.

Using near zero variance method, features variances is examined, and indexes of near zero variance features is created. Features for removal are identified and new dataset that contains 11 predictors are created to develop model. Classifier Random Forest are used to develop the model with 10-fold cross validation and an Automatic Tuning Grid with a tune Length of 5. Model with mtry=4 had the highest kappa. The accuracy and kappa of cross-validation are 0.7612230, 0.5324151 respectively. This model is used to predict test dataset, and the accuracy and kappa of test dataset are 0.7581, 0.5261 respectively.

Using Recursive Feature Elimination method, a Random Forest algorithm was used on each iteration to evaluate the model. The algorithm is configured to explore all possible subsets of the attributes. Subset size of 27 attributes giving best accuracy among different attribute subset sizes is selected, and the top 5 important variables were iPhone, samsunggalaxy, googleandroid, iphonedisunc and iphonedispos. Classifier Random Forest is used to develop the model with 10-fold cross validation and an Automatic Tuning Grid with a tune Length of 5. Model with mtry=8 had the highest kappa. The accuracy and kappa of cross-validation are 0.7767257, 0.5675275 respectively. I used this model to predict test dataset, and the accuracy and kappa of test dataset are 0.7728, 0.5581 respectively.

**3. Improving the performance metrics with feature engineering for iPhone**

The dependent variable is engineered and some of these levels of this variable are combined, which would help increase accuracy and kappa. New dataset was created. Two methods were tried to develop the model.

First, Random forest model with Recursive Feature Elimination is used to develop model with 10-fold cross validation and an Automatic Tuning Grid with a tune Length of 5. Model with mtry=14 had the highest kappa. The accuracy and kappa of cross-validation are 0.8483874, 0.6232784 respectively. This model is used to predict test dataset, and the accuracy and kappa of test dataset are 0.8509, 0.6337 respectively.

Second, Principal Component Analysis (PCA) is used to extract 25 important variables (in form of components) from a large set of variables available in a data set to capture 95% variance. Classifier Random Forest is used to develop the model with 10-fold cross validation and an Automatic Tuning Grid with a tune Length of 5. Model with mtry=7 had the highest kappa. The accuracy and kappa of cross-validation are 0.8428179, 0.6108781 respectively. I used this model to predict test dataset, and the accuracy and kappa of test dataset are 0.8383, 0.598 respectively.

**Model evaluation for iPhone**

Accuracy and kappa are used to evaluate the model. When some models had very similar accuracy and kappa, confusion matrix is a good idea to explore additional metrics.

**1.out of the box modeling evaluation for iPhone**

Resamples function is used to compared metrics of these six models and both accuracy and kappa of random forest is higher than other models. I selected random forest as best model. The table below showed the details accuracy and kappa of these six models.

Comparisons of accuracy from six classifier

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA's |
| C5.0 | 0.7535754 | 0.7656121 | 0.7719008 | 0.7723770 | 0.7779919 | 0.8019802 | 0 |
| RF | 0.7568757 | 0.7689133 | 0.7754540 | 0.7760432 | 0.7822137 | 0.8063806 | 0 |
| SVM | 0.6864686 | 0.7037445 | 0.7114534 | 0.7111868 | 0.7181392 | 0.7403740 | 0 |
| KKNN | 0.3142227 | 0.3361808 | 0.3472755 | 0.3460640 | 0.3558249 | 0.3817382 | 0 |
| SVMR | 0.7120879 | 0.7305573 | 0.7370737 | 0.7381494 | 0.7442821 | 0.7665198 | 0 |
| GBM | 0.7524752 | 0.7673458 | 0.7733773 | 0.7731814 | 0.7795535 | 0.8030803 | 0 |

Comparisons of Kappa from six classifier

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA's |
| C5.0 | 0.5173203 | 0.5446247 | 0.5578421 | 0.5585957 | 0.5715318 | 0.6240833 | 0 |
| RF | 0.5245796 | 0.5505577 | 0.5653320 | 0.5664790 | 0.5787580 | 0.6346380 | 0 |
| SVM | 0.3633234 | 0.4036823 | 0.4194541 | 0.4191533 | 0.4353608 | 0.4852688 | 0 |
| KKNN | 0.1354306 | 0.1596399 | 0.1720956 | 0.1706907 | 0.1788448 | 0.2086101 | 0 |
| SVMR | 0.4055091 | 0.4533903 | 0.4674321 | 0.4701697 | 0.4847049 | 0.5375705 | 0 |
| GBM | 0.5171666 | 0.5480796 | 0.5622067 | 0.5619847 | 0.5754679 | 0.6280857 | 0 |

Since C5.0, Random Forest and GBM had very similar accuracy and kappa, confusion matrix was calculated for these three models.

This table below listed the overall statistics of testing datasets for these three models.

|  |  |  |  |
| --- | --- | --- | --- |
| Overall Statistics | C5.0 | Random  Forest | GBM |
| Accuracy | 0.7666 | 0.7751 | 0.771 |
| Kappa | 0.5461 | 0.5632 | 0.5555 |

The table below listed the error metrics of testing datasets such as sensitivity, specificity, precision, recall, F1 score of these three models for each level. These metrics were also very similar.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | very negative | negative | somewhat negative | somewhat positive | positive | very positive |
| C5.0 | Sensitivity | 0.62755 | 0.00000 | 0.147059 | 0.63202 | 0.32947 | 0.9841 |
| Specificity | 0.99515 | 1.00000 | 0.999201 | 0.99802 | 0.99162 | 0.4760 |
| Precision | 0.95844 | NA | 0.869565 | 0.96983 | 0.83041 | 0.7230 |
| Recall | 0.62755 | 0.00000 | 0.147059 | 0.63202 | 0.32947 | 0.9841 |
| F1 | 0.75848 | NA | 0.251572 | 0.76531 | 0.47176 | 0.8336 |
| Random  Forest | Sensitivity | 0.63605 | 0.00000 | 0.147059 | 0.65449 | 0.34571 | 0.9898 |
| Specificity | 0.99697 | 0.9994699 | 0.998934 | 0.99745 | 0.99566 | 0.4871 |
| Precision | 0.97396 | 0.0000000 | 0.833333 | 0.96281 | 0.90854 | 0.7284 |
| Recall | 0.63605 | 0.00000 | 0.147059 | 0.65449 | 0.34571 | 0.9898 |
| F1 | 0.76955 | NaN | 0.250000 | 0.77926 | 0.50084 | 0.8392 |
| GBM | Sensitivity | 0.62585 | 0.0000000 | 0.147059 | 0.65169 | 0.33875 | 0.9872 |
| Specificity | 0.99727 | 0.9997350 | 0.999734 | 0.99377 | 0.99451 | 0.4846 |
| Precision | 0.97613 | 0.0000000 | 0.952381 | 0.91339 | 0.88485 | 0.7269 |
| Recall | 0.62585 | 0.0000000 | 0.147059 | 0.65169 | 0.33875 | 0.9872 |
| F1 | 0.76269 | NaN | 0.254777 | 0.76066 | 0.48993 | 0.8373 |

Overall, random forest has a little bit higher accuracy and kappa than other models.

**2.Evaluation on models with three feature selections methods for iPhone**

Accuracy and kappa are used to evaluate the models. This table below listed the overall statistics of accuracy and kappa of testing datasets for these three models using Correlation, near zero variance, Recursive Feature Elimination feature selection method respectively.

|  |  |  |  |
| --- | --- | --- | --- |
| Overall Statistics | Correlation | near zero variance | Recursive Feature Elimination |
| Accuracy | 0.754 | 0.7581 | 0.7728 |
| Kappa | 0.5162 | 0.5261 | 0.5581 |

The table below listed the error metrics such as sensitivity, specificity, precision, recall, F1 score of testing datasets of these three models.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | very negative | negative | somewhat negative | somewhat positive | positive | very positive |
| Correlation | Sensitivity | 0.62925 | 0.0000000 | 0.147059 | 0.48596 | 0.33875 | 0.9832 |
| Specificity | 0.99667 | 0.9997350 | 1.000000 | 0.99264 | 0.99480 | 0.4466 |
| Precision | 0.97113 | 0.0000000 | 1.000000 | 0.86935 | 0.89024 | 0.7117 |
| Recall | 0.62925 | 0.0000000 | 0.147059 | 0.48596 | 0.33875 | 0.9832 |
| F1 | 0.76367 | NaN | 0.256410 | 0.62342 | 0.49076 | 0.8257 |
| near zero variance | Sensitivity | 0.6378 | 0.0000000 | 0.000000 | 0.55337 | 0.34107 | 0.9859 |
| Specificity | 0.9891 | 0.9994699 | 0.9992009 | 0.99717 | 0.99537 | 0.4631 |
| Precision | 0.9124 | 0.0000000 | 0.0000000 | 0.95169 | 0.90184 | 0.7184 |
| Recall | 0.6378 | 0.0000000 | 0.000000 | 0.55337 | 0.34107 | 0.9859 |
| F1 | 0.7508 | NaN | NaN | 0.69982 | 0.49495 | 0.8312 |
| Recursive Feature Elimination | Sensitivity | 0.63265 | 0.0000000 | 0.147059 | 0.64888 | 0.34107 | 0.9885 |
| Specificity | 0.99546 | 0.9994699 | 0.999201 | 0.99774 | 0.99624 | 0.4822 |
| Precision | 0.96124 | 0.0000000 | 0.869565 | 0.96653 | 0.91875 | 0.7262 |
| Recall | 0.63265 | 0.0000000 | 0.147059 | 0.64888 | 0.34107 | 0.9885 |
| F1 | 0.76308 | NaN | 0.251572 | 0.77647 | 0.49746 | 0.8373 |

Overall, model with Recursive Feature Elimination had the highest accuracy and kappa and error metrics of RFE is a little better than other methods. I selected model with Recursive Feature Elimination as the best one.

3.Evaluation on models with two feature engineering methods for iPhone

Accuracy and kappa are used to evaluate the models. This table below listed the overall statistics of accuracy and kappa of testing datasets for these two models using dependent variable engineering. These two models were developed using Random forest with Recursive Feature Elimination feature engineering, and Random forest with Principal Component Analysis respectively.

|  |  |  |
| --- | --- | --- |
| Overall Statistics | RFE | PCA |
| Accuracy | 0.8509 | 0.8383 |
| Kappa | 0.6337 | 0.598 |

The table below listed the error metrics such as sensitivity, specificity, precision, recall, F1 score of testing datasets of these two models.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | negative | somewhat negative | somewhat positive | positive |
| RFE | Sensitivity | 0.5688 | 0.147059 | 0.66573 | 0.9848 |
| Specificity | 0.9918 | 0.997869 | 0.99660 | 0.5539 |
| Precision | 0.9391 | 0.714286 | 0.95181 | 0.8324 |
| Recall | 0.5688 | 0.147059 | 0.66573 | 0.9848 |
| F1 | 0.7085 | 0.243902 | 0.78347 | 0.9022 |
| PCA | Sensitivity | 0.53050 | 0.147059 | 0.62360 | 0.9822 |
| Specificity | 0.98807 | 0.997603 | 0.99774 | 0.5205 |
| Precision | 0.90777 | 0.689655 | 0.96522 | 0.8217 |
| Recall | 0.53050 | 0.147059 | 0.62360 | 0.9822 |
| F1 | 0.66965 | 0.242424 | 0.75768 | 0.8948 |

Overall, model with Recursive Feature Elimination and recoding dependent variable had the highest accuracy and kappa and error metrics of RFE is a little better than PCA. I selected model with Recursive Feature Elimination and recoding dependent variable to predict large matrix of iPhone sentiment.

**Model building, tuning and feature engineering for Galaxy**

**1.out of the box modeling for Galaxy**

Six classifiers-C5.0, Random Forest, SVM, SVM Radial, KKNN and GBM are tried and all features are used to train the model. Small matrix is partitioned into training (.70) and test dataset (.30).

For classifier C5.0, 5-fold cross validation is applied and model is built by using an Automatic Tuning Grid with a tune Length of 5. Model with trial=1, model=tree and winnow=FALSE has highest accuracy and kappa. The accuracy and kappa of cross-validation are 0.7631191, 0.5259074 respectively. This model is used to predict test dataset, and the accuracy and kappa of test dataset are 0.7675, 0.5315 respectively.

For classifier Random Forest, 5-fold cross validation is applied and model is built by using an Automatic Tuning Grid with a tune Length of 5. The model with mtry=16 had the highest kappa. The accuracy and kappa of cross-validation are 0.7670127, 0.5338944 respectively. This model is used to predict test dataset, and the accuracy and kappa of test dataset are 0.7696, 0.5378 respectively.

For classifier SVM Linear, 10-fold cross validation is applied and model is built by using an Automatic Tuning Grid with a tune Length of 5. Model with C=1 was the best model. The accuracy and kappa of that model were 0.7121678, 0.3953766 respectively, which is lower than C5.0 and Random Forest.

For classifier KKNN, 10-fold cross validation is applied and model is built by using an Automatic Tuning Grid with a tune Length of 5. Model with kmax = 13, distance = 2 and kernel = optimal was the best model. The accuracy and kappa of that model were 0.7519685, 0.5093890 respectively.

For classifier GBM, 5-fold cross validation is applied and model is built by using an Automatic Tuning Grid with a tune Length of 3. Model with n.trees = 150, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10 was the best model. The accuracy and kappa of that model were 0.7651324, 0.5291339 respectively. This model is used to predict test dataset, and the accuracy and kappa of test dataset are 0.7701, 0.5373 respectively.

**2. Improving the performance of model with feature selection for Galaxy**

Three feature selection methods including Correlation method, near zero variance method, Recursive Feature Elimination method are used to improve the performance of the model.

Using Correlation method, correlation of variables of the small matrix for galaxy sentiments is examined. Some features that had collinearity were detected and some highly-related features are removed. New dataset that contains 33 predictors are created to develop model. Classifier Random Forest are used to develop the model with 10-fold cross validation and an Automatic Tuning Grid with a tune Length of 5. Model with mtry=9 had the highest kappa. The accuracy and kappa of cross-validation are 0.7514819, 0.4974913 respectively. This model is used to predict test dataset, and the accuracy and kappa of test dataset are 0.7476, 0.4897 respectively.

Using near zero variance method, features variances is examined, and indexes of near zero variance features is created. Features for removal are identified and new dataset that contains 11 predictors are created to develop model. Classifier Random Forest are used to develop the model with 10-fold cross validation and an Automatic Tuning Grid with a tune Length of 5. Model with mtry=4 had the highest kappa. The accuracy and kappa of cross-validation are 0.7527644, 0.5004511 respectively. This model is used to predict test dataset, and the accuracy and kappa of test dataset are 0.7587, 0.5126 respectively.

Using Recursive Feature Elimination method, a Random Forest algorithm was used on each iteration to evaluate the model. The algorithm is configured to explore all possible subsets of the attributes. Subset size of 25 attributes giving best accuracy among different attribute subset sizes is selected, and the top 5 important variables were iPhone, samsunggalaxy, googleandroid, iphonedisunc and htcphone. Classifier Random Forest is used to develop the model with 10-fold cross validation and an Automatic Tuning Grid with a tune Length of 5. Model with mtry=7 had the highest kappa. The accuracy and kappa of cross-validation are 0.7666145, 0.5336587 respectively. I used this model to predict test dataset, and the accuracy and kappa of test dataset are 0.7711, 0.5412 respectively.

**3. Improving the performance metrics with feature engineering for Galaxy**

The dependent variable is engineered and some of these levels of this variable are combined, which would help increase accuracy and kappa. New dataset was created. Two methods were tried to develop the model.

First, Random forest model with Recursive Feature Elimination is used to develop model with 5-fold cross validation and an Automatic Tuning Grid with a tune Length of 5. Model with mtry=7 had the highest kappa. The accuracy and kappa of cross-validation are 0.8460444, 0.6014976 respectively. This model is used to predict test dataset, and the accuracy and kappa of test dataset are 0.8419, 0.589 respectively.

Second, Principal Component Analysis (PCA) is used to extract 24 important variables (in form of components) from a large set of variables available in a data set to capture 95% variance. Classifier Random Forest is used to develop the model with 5-fold cross validation and an Automatic Tuning Grid with a tune Length of 5. Model with mtry=2 had the highest kappa. The accuracy and kappa of cross-validation are 0.8352668, 0.5757603 respectively. I used this model to predict test dataset, and the accuracy and kappa of test dataset are 0.8362, 0.5787 respectively.

**Model evaluation for Galaxy**

Accuracy and kappa are used to evaluate the model. When some models had very similar accuracy and kappa, confusion matrix is a good idea to explore additional metrics.

**1.out of the box modeling evaluation for Galaxy**

Resamples function is used to compared metrics of these six models and both accuracy and kappa of random forest is higher than other models. I selected random forest as best model. The table below showed the details accuracy and kappa of these five models.

Comparisons of accuracy from five classifier

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA's |
| C5.0 | 0.7483407 | 0.7599558 | 0.7628524 | 0.7631191 | 0.7665929 | 0.7809735 | 0 |
| RF | 0.7553957 | 0.7631433 | 0.7667219 | 0.7670127 | 0.7693584 | 0.7826327 | 0 |
| SVM | 0.6950747 | 0.7072496 | 0.7140487 | 0.7121678 | 0.7182320 | 0.7367257 | 0 |
| KKNN | 0.7293857 | 0.7470946 | 0.7538717 | 0.7519685 | 0.7588496 | 0.7672747 | 0 |
| GBM | 0.7515219 | 0.7625899 | 0.7648035 | 0.7651324 | 0.7678275 | 0.7798673 | 0 |

Comparisons of Kappa from five classifier

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA's |
| C5.0 | 0.4980952 | 0.5188530 | 0.5255550 | 0.5259074 | 0.5314367 | 0.5651669 | 0 |
| RF | 0.5103113 | 0.5248984 | 0.5339376 | 0.5338944 | 0.5389400 | 0.5691963 | 0 |
| SVM | 0.3521243 | 0.3867967 | 0.3970922 | 0.3953766 | 0.4084361 | 0.4532139 | 0 |
| KKNN | 0.4755206 | 0.4986195 | 0.5117041 | 0.5093890 | 0.5172126 | 0.5378172 | 0 |
| GBM | 0.5009611 | 0.5237389 | 0.5286343 | 0.5291339 | 0.5362801 | 0.5608608 | 0 |

Since in this step, I found that C5.0, Random Forest and GBM had very similar accuracy and kappa, confusion matrix is calculated for these three models.

This table below listed the overall statistics of accuracy and kappa of testing datasets for these three models.

|  |  |  |  |
| --- | --- | --- | --- |
| Overall Statistics | C5.0 | Random  Forest | GBM |
| Accuracy | 0.7675 | 0.7696 | 0.7701 |
| Kappa | 0.5315 | 0.5378 | 0.5373 |

The table below listed the error metrics of testing datasets such as sensitivity, specificity, precision, recall, F1 score of these three models for each level of galaxy sentiment. These metrics were also very similar.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | very negative | negative | somewhat negative | somewhat positive | positive | very positive |
| C5.0 | Sensitivity | 0.68898 | 0.00000 | 0.111111 | 0.58523 | 0.28941 | 0.9743 |
| Specificity | 0.98662 | 1.00000 | 0.998929 | 0.99176 | 0.99507 | 0.4752 |
| Precision | 0.88608 | NA | 0.789474 | 0.87660 | 0.87857 | 0.7388 |
| Recall | 0.68898 | 0.00000 | 0.111111 | 0.58523 | 0.28941 | 0.9743 |
| F1 | 0.77519 | NA | 0.194805 | 0.70187 | 0.43540 | 0.8404 |
| Random  Forest | Sensitivity | 0.69094 | 0.0000000 | 0.111111 | 0.59943 | 0.31059 | 0.9713 |
| Specificity | 0.98543 | 0.9994677 | 0.998929 | 0.99318 | 0.99449 | 0.4824 |
| Precision | 0.87750 | 0.0000000 | 0.789474 | 0.89787 | 0.87417 | 0.7409 |
| Recall | 0.69094 | 0.0000000 | 0.111111 | 0.59943 | 0.31059 | 0.9713 |
| F1 | 0.77313 | NaN | 0.194805 | 0.71891 | 0.45833 | 0.8406 |
| GBM | Sensitivity | 0.68898 | 0.00000 | 0.111111 | 0.60795 | 0.28941 | 0.9752 |
| Specificity | 0.98811 | 1.00000 | 0.998394 | 0.99119 | 0.99594 | 0.4791 |
| Precision | 0.89744 | NA | 0.714286 | 0.87347 | 0.89781 | 0.7404 |
| Recall | 0.68898 | 0.00000 | 0.111111 | 0.60795 | 0.28941 | 0.9752 |
| F1 | 0.77951 | NA | 0.192308 | 0.71692 | 0.43772 | 0.8417 |

Overall, random forest has a little bit higher accuracy and kappa than other models.

**2.Evaluation on models with three feature selections methods** for **Galaxy**

Accuracy and kappa are used to evaluate the models. This table below listed the overall statistics of accuracy and kappa of testing datasets for these three models using Correlation, near zero variance, Recursive Feature Elimination feature selection method respectively.

This table below listed the overall statistics of accuracy and kappa of testing datasets for these three models.

|  |  |  |  |
| --- | --- | --- | --- |
| Overall Statistics | Correlation | near zero variance | Recursive Feature Elimination |
| Accuracy | 0.7476 | 0.7587 | 0.7711 |
| Kappa | 0.4897 | 0.5126 | 0.5412 |

The table below listed the error metrics such as sensitivity, specificity, precision, recall, F1 score of testing datasets of these three models.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | very negative | negative | somewhat negative | somewhat positive | positive | very positive |
| Correlation | Sensitivity | 0.6850 | 0.0000000 | 0.111111 | 0.45455 | 0.29176 | 0.9615 |
| Specificity | 0.9854 | 0.9994677 | 0.998662 | 0.98494 | 0.99420 | 0.4472 |
| Precision | 0.8766 | 0.0000000 | 0.750000 | 0.75117 | 0.86111 | 0.7260 |
| Recall | 0.6850 | 0.0000000 | 0.111111 | 0.45455 | 0.29176 | 0.9615 |
| F1 | 0.7691 | NaN | 0.193548 | 0.56637 | 0.43585 | 0.8273 |
| near zero variance | Sensitivity | 0.69488 | 0.0000000 | 0.0000000 | 0.51705 | 0.31059 | 0.9713 |
| Specificity | 0.98008 | 0.9994677 | 0.9994647 | 0.99261 | 0.99420 | 0.4674 |
| Precision | 0.84048 | 0.0000000 | 0.0000000 | 0.87500 | 0.86842 | 0.7353 |
| Recall | 0.69488 | 0.0000000 | 0.0000000 | 0.51705 | 0.31059 | 0.9713 |
| F1 | 0.76078 | NaN | NaN | 0.65000 | 0.45754 | 0.8370 |
| Recursive Feature Elimination | Sensitivity | 0.69488 | 0.0000000 | 0.111111 | 0.60227 | 0.31294 | 0.9722 |
| Specificity | 0.98632 | 0.9994677 | 0.998929 | 0.99261 | 0.99478 | 0.4850 |
| Precision | 0.88471 | 0.0000000 | 0.789474 | 0.89076 | 0.88079 | 0.7420 |
| Recall | 0.69488 | 0.0000000 | 0.111111 | 0.60227 | 0.31294 | 0.9722 |
| F1 | 0.77839 | NaN | 0.194805 | 0.71864 | 0.46181 | 0.8416 |

Overall, model with Recursive Feature Elimination had the highest accuracy and kappa and error metrics of RFE was better than other methods. I selected model with Recursive Feature Elimination as the best one.

**3.Evaluation on models with two feature engineering methods for Galaxy**

Accuracy and kappa are used to evaluate the models. This table below listed the overall statistics of accuracy and kappa of testing datasets for these two models using dependent variable engineering. These two models were developed using Random forest with Recursive Feature Elimination feature engineering, and Random forest with Principal Component Analysis respectively.

This table below listed the overall statistics of accuracy and kappa of testing datasets for these two models.

|  |  |  |
| --- | --- | --- |
| Overall Statistics | RFE | PCA |
| Accuracy | 0.8419 | 0.8362 |
| Kappa | 0.589 | 0.5787 |

The table below listed the error metrics such as sensitivity, specificity, precision, recall, F1 score of testing datasets of these two models.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | negative | somewhat negative | somewhat positive | positive |
| RFE | Sensitivity | 0.56501 | 0.111111 | 0.57955 | 0.9736 |
| Specificity | 0.98584 | 0.998662 | 0.99148 | 0.5216 |
| Precision | 0.88442 | 0.750000 | 0.87179 | 0.8351 |
| Recall | 0.56501 | 0.111111 | 0.57955 | 0.9736 |
| F1 | 0.68952 | 0.193548 | 0.69625 | 0.8990 |
| PCA | Sensitivity | 0.56270 | 0.118519 | 0.58239 | 0.9652 |
| Specificity | 0.97538 | 0.997859 | 0.99346 | 0.5284 |
| Precision | 0.81395 | 0.666667 | 0.89912 | 0.8360 |
| Recall | 0.56270 | 0.118519 | 0.58239 | 0.9652 |
| F1 | 0.66540 | 0.201258 | 0.70690 | 0.8960 |

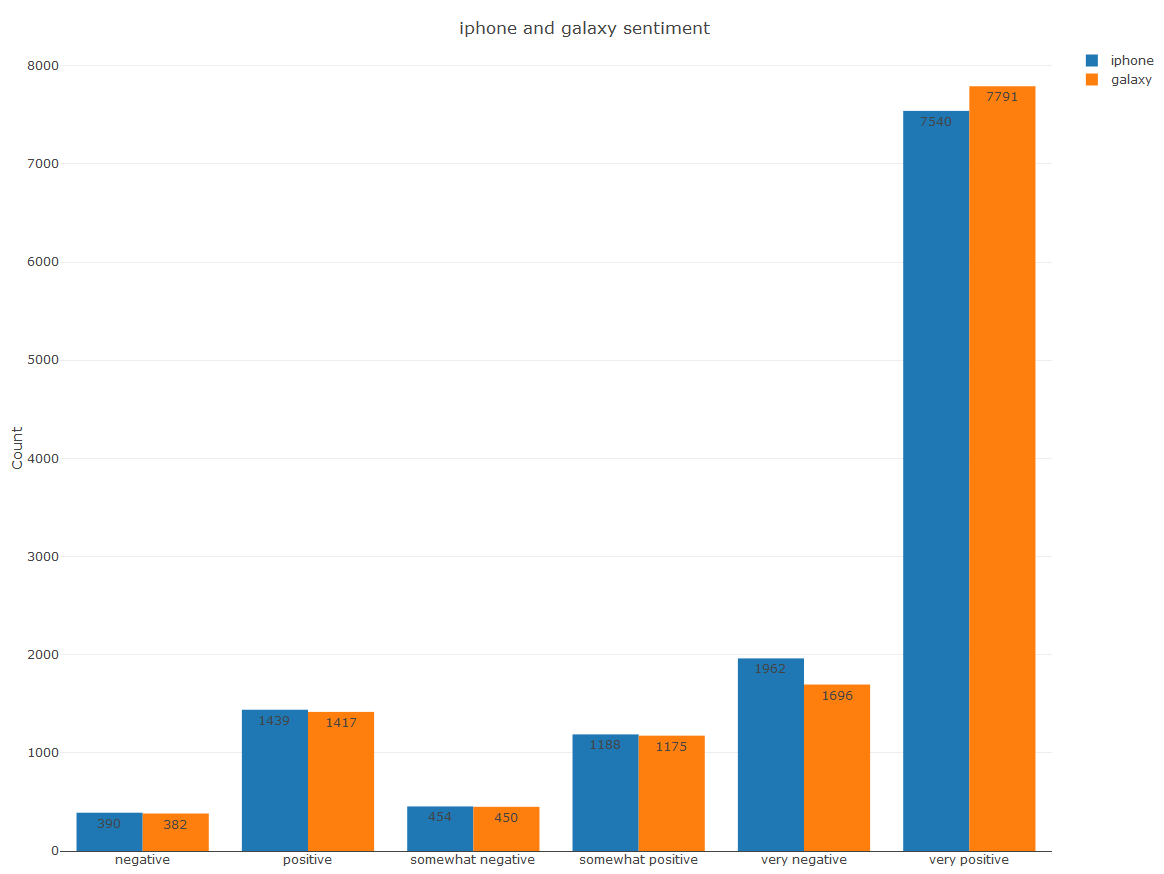
Overall, model with Recursive Feature Elimination and recoding dependent variable had the highest accuracy and kappa and error metrics of RFE is similar with PCA.

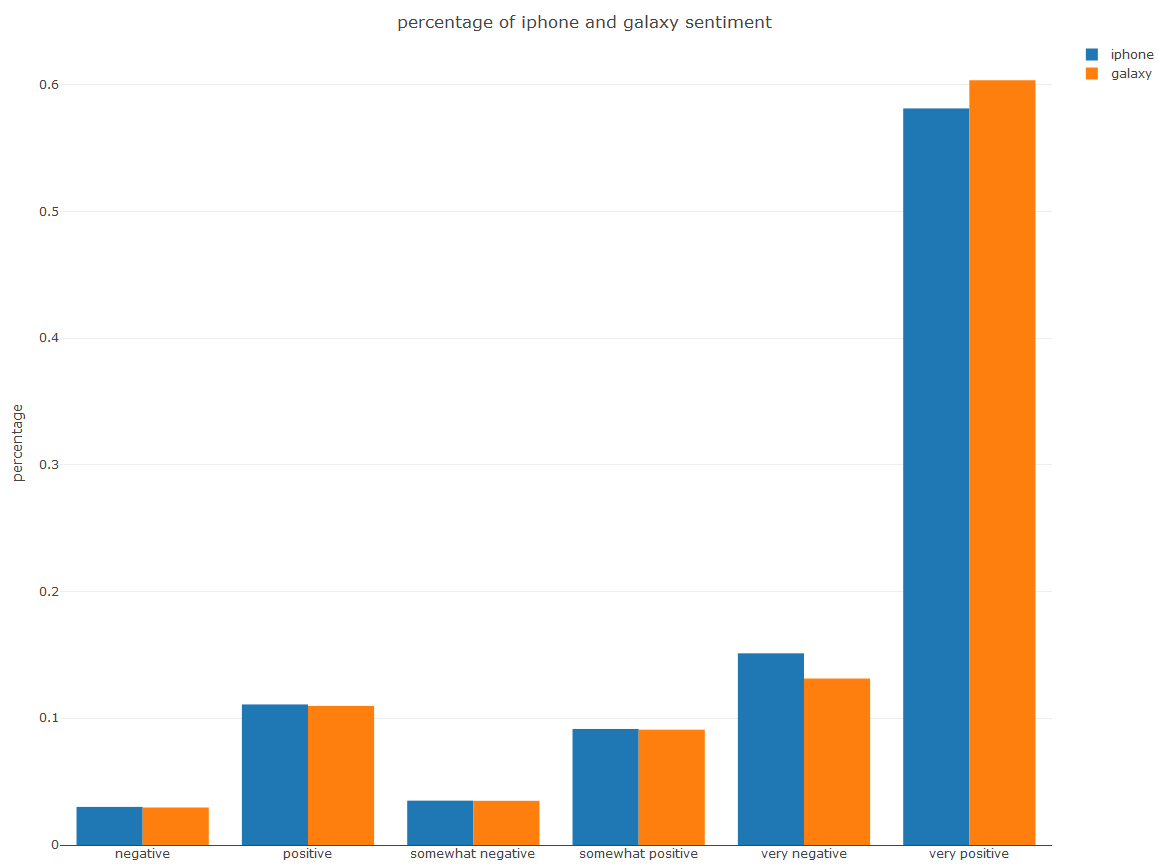
Model with Recursive Feature Elimination and recoding dependent variable is selected to predict large matrix of galaxy sentiment.

**Findings:**

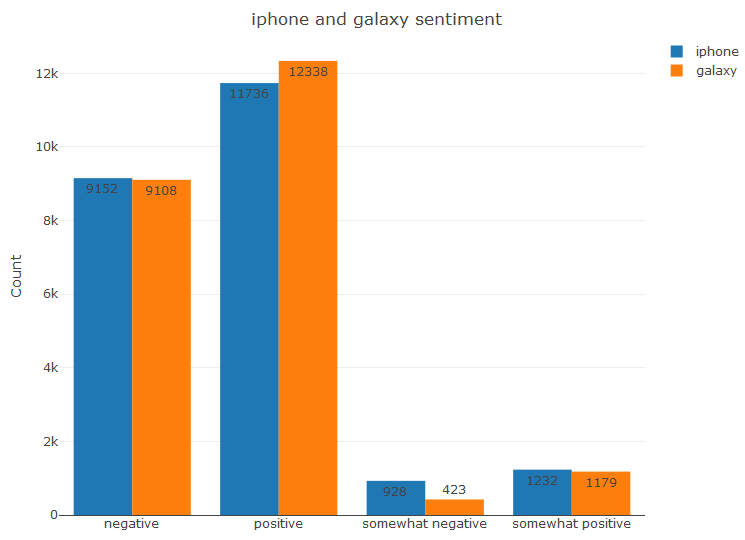
The original sentiment categories include very negative, negative, somewhat negative, somewhat positive, positive, very positive for iPhone and Galaxy in small matrix. The dependent variable was engineered, and some of these levels of this variable were combined, which would help increase accuracy. After that the sentiment categories include negative, somewhat negative, somewhat positive, and positive for iPhone and Galaxy in large matrix. The large matrix we collected from Common Crawl has 23048 observations. Model was optimized to predict the sentiment of iPhone and galaxy of large matrix. The distributions of prediction of iPhone and galaxy sentiment of large matrix were not evenly. 50.9% of documents collected from Common Crawl had positive reviews toward iPhone, while 53.5% of documents collected from Common Crawl had positive reviews toward galaxy. 39.7% of documents collected from Common Crawl had negative reviews toward iPhone, while 39.5% of documents collected from Common Crawl showed positive reviews toward galaxy. 5.35% of documents collected from Common Crawl had somewhat positive reviews toward iPhone, while 5.12% of documents collected from Common Crawl had somewhat positive reviews toward galaxy. 4.03% of documents collected from Common Crawl had somewhat negative reviews toward iPhone, while 1.88% of documents collected from Common Crawl had somewhat negative reviews toward galaxy.

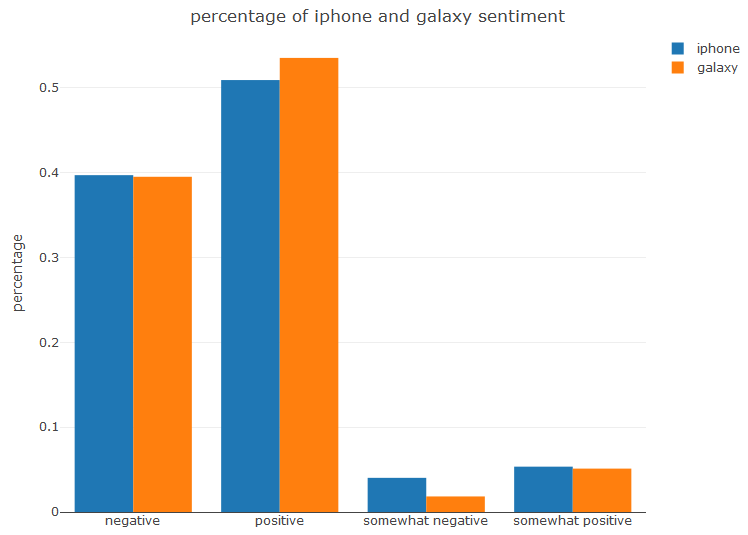
The two charts below are stacked bar comparing two different sets of counts and percentages for iPhone and Galaxy by using small matrix data.





The two charts below are stacked bar comparing two different sets of counts and percentages for iPhone and Galaxy by using large matrix data.





Confidence:

1) The reported error metrics from R

I selected the best model to predict iPhone sentiment, and the overall statistics of accuracy and kappa of testing datasets were 0.8509, 0.6337 respectively, which were quite good.

The table below listed the error metrics such as sensitivity, specificity, precision, recall, F1 score of testing datasets of this best model for iPhone sentiment. The higher of all the values in the error metrics, the better.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | negative | somewhat negative | somewhat positive | positive |
| RFE | Sensitivity | 0.5688 | 0.147059 | 0.66573 | 0.9848 |
| Specificity | 0.9918 | 0.997869 | 0.99660 | 0.5539 |
| Precision | **0.9391** | 0.714286 | 0.95181 | 0.8324 |
| Recall | 0.5688 | 0.147059 | 0.66573 | 0.9848 |
| F1 | 0.7085 | 0.243902 | 0.78347 | 0.9022 |

I selected the best model to predict galaxy sentiment, and the overall statistics of accuracy and kappa of testing datasets were 0.8419, 0.589 respectively, which were moderate.

The table below listed the error metrics such as sensitivity, specificity, precision, recall, F1 score of testing datasets of this best model for galaxy sentiment. The higher of all the values in the error metrics, the better.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | negative | somewhat negative | somewhat positive | positive |
| RFE | Sensitivity | 0.56501 | 0.111111 | 0.57955 | 0.9736 |
| Specificity | 0.98584 | 0.998662 | 0.99148 | 0.5216 |
| Precision | 0.88442 | 0.750000 | 0.87179 | 0.8351 |
| Recall | 0.56501 | 0.111111 | 0.57955 | 0.9736 |
| F1 | 0.68952 | 0.193548 | 0.69625 | 0.8990 |

2) The degree of the attribute captured pages that have relevant sentiment.

IPhone：

The proportion of actual negative attribute of web pages toward IPhone that was predicted as negative is only 0.5688, which was not high. The proportion of actual not negative attribute of web pages that was predicted as not negative is 0.9918. Therefore, the predicted 39.5 percent of documents having negative reviews toward iPhone was lower than actual proportion.

The proportion of actual somewhat negative attribute of web pages toward IPhone that was predicted as somewhat negative is 0.147059, which was quite low. The proportion of actual not somewhat negative attribute of web pages that was predicted as not somewhat negative is 0.997869. Therefore, the predicted 4.03 percent of documents having somewhat negative reviews toward iPhone were much lower than actual proportion.

The proportion of actual somewhat positive attribute of web pages toward IPhone that was predicted as somewhat positive is 0.66573, which was moderate. The proportion of actual not somewhat negative attribute of web pages that was predicted as not somewhat negative is 0.99660. Therefore, the predicted 5.35 percent of documents having somewhat positive reviews toward iPhone were lower than actual proportion.

The proportion of actual positive attribute of web pages toward IPhone that was predicted as positive is 0.9848, which was high. The proportion of actual not positive attribute of web pages that was predicted as not positive is 0.5539. Therefore, the predicted 50.9 percent of documents having positive reviews toward iPhone were higher than actual proportion.

Galaxy：

The proportion of actual negative attribute of web pages toward galaxy that was predicted as negative is only 0.56501, which was not high. The proportion of actual not negative attribute of web pages that was predicted as not negative is 0.98584. Therefore, the predicted 39.7 percent of documents having negative reviews toward galaxy was lower than actual proportion.

The proportion of actual somewhat negative attribute of web pages toward galaxy that was predicted as somewhat negative is 0.111111, which was quite low. The proportion of actual not somewhat negative attribute of web pages that was predicted selected as not somewhat negative is 0.998662. Therefore, the predicted 1.88 percent of documents having somewhat negative reviews toward galaxy was much lower than actual proportion.

The proportion of actual somewhat positive attribute of web pages toward galaxy that was predicted as somewhat positive is 0.57955, which was moderate. The proportion of actual not somewhat negative attribute of web pages that was predicted as not somewhat negative is 0.99148. Therefore, the predicted 5.12 percent of documents having somewhat positive reviews toward galaxy was lower than actual proportion.

The proportion of actual positive attribute of web pages toward galaxy that was predicted as positive is 0.9736, which was high. The proportion of actual not positive attribute of web pages that was predicted as not positive is 0.5216. Therefore, the predicted 53.5percent of documents having positive reviews toward galaxy were higher than actual proportion.

3)The predicted percent of negative, somewhat negative, somewhat positive was lower than actual proportion respectively. The predicted percent of positive was higher than actual proportion with regard to iPhone and galaxy sentiment.

We need to continue to increase the accuracy and specificity of model. We can try ensemble technique such as bagging, boosting and stacking to improve the model. In this project, I tried random forest, GBM to train the model. I have set the tune length to 5 to the random foreset model. I can tune parameter in large wide range in the next round.

Suggestion:

The chart above showed most reviews either hated the given handset or loved it. Few reviewers were neutral. Most customers hold strong preference toward the given handset. I suggest Helio can develop galaxy apps for customers. Because after analyzing the sentiment of large matrix, a few more customers had positive attitudes toward galaxy than iPhone while a few less customers had negative attitudes toward galaxy than iPhone.