**Question #1 –** Run 9 models (three models with three feature sets) and report the precision and recall by class as well as the corresponding confusion matrix

*Please note: in Python, confusion matrices for binary classification (M) are output as follows:*

*the count of true negatives is M[0,0], false negatives is M[1,0], true positives is M[1,1], and false positives is M[0,1].*

Naive Bayes with Bag of Words Features:

This is the Confusion Matrix for the Naive Bayes model with Bag of Words features:

[[1164 6]

[ 130 439]]

The positive precision for the Naive Bayes model with Bag of Words features is 0.99

The negative precision for the Naive Bayes model with Bag of Words features is 0.9

The positive recall for the Naive Bayes model with Bag of Words features is 0.77

The negative recall for the Naive Bayes model with Bag of Words features is 0.99

Decision Tree with Bag of Words Features:

This is the Confusion Matrix for the Decision Tree model with Bag of Words features:

[[1125 45]

[ 40 529]]

The positive precision for the Decision Tree model with Bag of Words features is 0.92

The negative precision for the Decision Tree model with Bag of Words features is 0.97

The positive recall for the Decision Tree model with Bag of Words features is 0.93

The negative recall for the Decision Tree model with Bag of Words features is 0.96

Random Forest with Bag of Words Features:

This is the Confusion Matrix for the Random Forest model with Bag of Words features:

[[1163 7]

[ 42 527]]

The positive precision for the Random Forest model with Bag of Words features is 0.99

The negative precision for the Random Forest model with Bag of Words features is 0.97

The positive recall for the Random Forest model with Bag of Words features is 0.93

The negative recall for the Random Forest model with Bag of Words features is 0.99

Naive Bayes with TF-IDF Features:

This is the Confusion Matrix for the Naive Bayes model with TF-IDF features:

[[1166 4]

[ 203 366]]

The positive precision for the Naive Bayes model with TF-IDF features is 0.99

The negative precision for the Naive Bayes model with TF-IDF features is 0.85

The positive recall for the Naive Bayes model with TF-IDF features is 0.64

The negative recall for the Naive Bayes model with TF-IDF features is 1.0

Decision Tree with TF-IDF Features:

This is the Confusion Matrix for the Decision Tree model with TF-IDF features:

[[1133 37]

[ 42 527]]

The positive precision for the Decision Tree model with TF-IDF features is 0.93

The negative precision for the Decision Tree model with TF-IDF features is 0.96

The positive recall for the Decision Tree model with TF-IDF features is 0.93

The negative recall for the Decision Tree model with TF-IDF features is 0.97

Random Forest with TF-IDF Features:

This is the Confusion Matrix for the Random Forest model with TF-IDF features:

[[1164 6]

[ 46 523]]

The positive precision for the Random Forest model with TF-IDF features is 0.99

The negative precision for the Random Forest model with TF-IDF features is 0.96

The positive recall for the Random Forest model with TF-IDF features is 0.92

The negative recall for the Random Forest model with TF-IDF features is 0.99

(Word2Vec features on next page)

Naive Bayes with Word2Vec Features:

This is the Confusion Matrix for the Naive Bayes model with Word2Vec features:

[[1125 45]

[ 102 467]]

The positive precision for the Naive Bayes model with Word2Vec features is 0.91

The negative precision for the Naive Bayes model with Word2Vec features is 0.92

The positive recall for the Naive Bayes model with Word2Vec features is 0.82

The negative recall for the Naive Bayes model with Word2Vec features is 0.96

Decision Tree with Word2Vec Features:

This is the Confusion Matrix for the Decision Tree model with Word2Vec features:

[[1152 18]

[ 25 544]]

The positive precision for the Decision Tree model with Word2Vec features is 0.97

The negative precision for the Decision Tree model with Word2Vec features is 0.98

The positive recall for the Decision Tree model with Word2Vec features is 0.96

The negative recall for the Decision Tree model with Word2Vec features is 0.98

Random Forest with Word2Vec Features:

This is the Confusion Matrix for the Random Forest model with Word2Vec features:

[[1161 9]

[ 15 554]]

The positive precision for the Random Forest model with Word2Vec features is 0.98

The negative precision for the Random Forest model with Word2Vec features is 0.99

The positive recall for the Random Forest model with Word2Vec features is 0.97

The negative recall for the Random Forest model with Word2Vec features is 0.99

(Please see next page for Question #2)

**Question #2**: Calculating the business cost of each model

*Per the rules laid out in the assignment, mis-classifying spam to non-spam (a false negative) costs 5 and mis-classifying non-spam to spam (a false positive) costs 100.*

Below is a model-by-model breakdown of the cost of using each model with each of the feature sets, followed by an overall summary table for ease of comprehension.

Bag of Words Features:

The Naive Bayes model with Bag of Words features has 130 false negatives and 6 false positives. At a cost of 5 units per false negative and 100 units per false positive, this model costs the business 1250 units.

The Decision Tree model with Bag of Words features has 40 false negatives and 45 false positives. At a cost of 5 units per false negative and 100 units per false positive, this model costs the business 4700 units.

The Random Forest model with Bag of Words features has 42 false negatives and 7 false positives. At a cost of 5 units per false negative and 100 units per false positive, this model costs the business 910 units.

TF-IDF Features:

The Naive Bayes model with TF-IDF features has 203 false negatives and 4 false positives. At a cost of 5 units per false negative and 100 units per false positive, this model costs the business 1415 units.

The Decision Tree model with TF-IDF features has 42 false negatives and 37 false positives. At a cost of 5 units per false negative and 100 units per false positive, this model costs the business 3910 units.

The Random Forest model with TF-IDF features has 46 false negatives and 6 false positives. At a cost of 5 units per false negative and 100 units per false positive, this model costs the business 830 units.

Word2Vec Features:

The Naive Bayes model with Word2Vec features has 102 false negatives and 45 false positives. At a cost of 5 units per false negative and 100 units per false positive, this model costs the business 5010 units.

The Decision Tree model with Word2Vec features has 25 false negatives and 18 false positives. At a cost of 5 units per false negative and 100 units per false positive, this model costs the business 1925 units.

The Random Forest model with Word2Vec features has 15 false negatives and 9 false positives. At a cost of 5 units per false negative and 100 units per false positive, this model costs the business 975 units.

Summary Table of Costs from Each Model:

| **Model Names** | **Costs to Business: Bag of Words Features** | **Costs to Business: TF-IDF Features** | **Costs to Business: Word2Vec Features** |
| --- | --- | --- | --- |
| Naive Bayes | 1250 | 1415 | 5010 |
| Decision Tree | 4700 | 3910 | 1925 |
| Random Forest | 910 | 830 | 975 |

Under the current rule set of 5 units for a false negative and 100 units for a false positive, the Random Forest model easily wins the day; specifically the Random Forest model using TF-IDF features. Unlike the Decision Tree and Naive Bayes models, the Random Forest model calculates its estimates based on the results of several other models. That is, Random Forest models have several decision trees running “inside” of them, which allows for the estimation of results based on a greater amount of trials and outcomes. This “repeated learning” is likely the reason behind the model’s superiority over the other two in this analysis, both of which only make one attempt at analyzing the data before generating predictions.