### [30 points] Train a classifier on the train dataset that is able to predict spam.

Choose stratified sampling training ratio

| Average accuracy | MLP | SVC | KNeighbors | GaussianNB | RandomForest |
|---|---|---|---|---|---|
| **Tf-idf Setting** | Default | Default | Default | Default | Default |
| **Training Ratio** | | | | | |
| **0.5** | 0.77459207459 | 0.810955710 | 0.78444055594 | 0.735780885780 | 0.82004662004662 |
| **0.8** | 0.79353146853 | 0.802622377 | 0.78444055594 | 0.681759906759 | 0.82004662004662 |
| **0.9** | 0.77051282051 | 0.793589743 | 0.76282051282 | 0.701282051282 | 0.816666666666667 |

*Average Accuracy: The average of accuracy_score and mean of 5-fold cross_val_score

At training ratio = 0.8, almost all classifiers have the best accuracy score. So I will use training ratio = 0.8 for later experiment.

| Classifier | Tf-idf Setting | Accuracy | CV_score |
|---|---|---|---|
| **MLP** | Default | 0.769230769230769 | 0.73939393939394 |
| | sublinear_tf=True | 0.769230769230769 | 0.774242424242424 |
| | ngram_range=(2, 2), sublinear_tf=True | 0.923076923076923 | 0.724242424242424 |
| **SVC** | Default | 0.769230769230769 | 0.775757575757576 |
| | sublinear_tf=True | 0.846153846153846 | 0.757575757575758 |
| | ngram_range=(2, 2), sublinear_tf=True | 0.846153846153846 | 0.775757575757576 |
| **KNeighbors** | Default | 0.615384615384615 | 0.689393939393939 |
| | sublinear_tf=True | 0.692307692307692 | 0.725757575757576 |
| | ngram_range=(2, 2), sublinear_tf=True | 0.846153846153846 | 0.672727272727273 |
| **GaussianNB** | Default | 0.538461538461538 | 0.56969696969697 |
| | sublinear_tf=True | 0.538461538461538 | 0.692424242424243 |
| | ngram_range=(2, 2), sublinear_tf=True | 0.846153846153846 | 0.504545454545455 |
| **RandomForest** | Default | 0.846153846153846 | 0.793939393939394 |
| | sublinear_tf=True | 0.846153846153846 | 0.793939393939394 |

| Classifier | Tf-idf Setting | Accuracy | CV_score |
|---|---|---|---|
|  | ngram_range=(2, 2),<br>sublinear_tf=True | 0.846153846153846 | 0.793939393939394 |

*CV_score: The mean of 5-fold cross validation score

According to the spreadsheet above

1. When apply tf-idf setting: ngram_range=(2, 2), sublinear_tf=True, the overall accuracy of all classifiers is the highest.

2. RandomForest Classifier has the highest CV_score, SVC classifier is second high

[10 points] Using the best and second-best classifiers you got from above, apply them once each on the entire test dataset.

If you run into the out-of-vocabulary problem i.e. there are words in your test data that are not in your training data,

you can delete the word, although in practice we would be taking a more sophisticated approach.

What are the accuracy measures that you are getting for both? Is the difference greater, smaller or equal compared to performance

difference on validation set/cross-validation? What would be the accuracy for a classifier that labeled the data 'randomly'?

(Hint: for the last question, use a bernoulli distribution i.e. toss a coin, with spam coming up with probability p and non-spam

with probability 1-p. Use the number of spam/legit samples in the training data to estimate the ideal value for p.)