[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	0.774592074592	0.810955710	0.7844405594	0.735780885780	0.82004662004662
0.8	0.793531468531	0.802622377	0.7844405594	0.681759906759	0.82004662004662
0.9	0.770512820512	0.793589743	0.7628205128	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	0.686363636363636
	sublinear_tf=True	0.692307692307692	0.739393939393939
	ngram_range=(2, 2), sublinear_tf=True	0.846153846153846	0.721212121212121
SVC	Default	0.769230769230769	0.775757575757576
	sublinear_tf=True	0.846153846153846	0.757575757575758
	ngram_range=(2, 2), sublinear_tf=True	0.846153846153846	0.793939393939394
KNeighbors	Default	0.769230769230769	0.706060606060606
	sublinear_tf=True	0.692307692307692	0.793939393939394
	ngram_range=(2, 2), sublinear_tf=True	0.846153846153846	0.742424242424242
GaussianNB	Default	0.769230769230769	0.686363636363636
	sublinear_tf=True	0.615384615384615	0.739393939393939
	ngram_range=(2, 2), sublinear_tf=True	0.846153846153846	0.721212121212121
RandomForest	Default	0.846153846153846	0.793939393939394
	sublinear_tf=True	0.846153846153846	0.793939393939394

Classifier	Tf-idf Setting	Accuracy	CV_score
	ngram_range=(2, 2), 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.

1. What are the accuracy measures that you are getting for both?
2. Is the difference greater, smaller or equal compared to performance difference on validation set/cross-validation?
3. 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.)

Result Using Test Dataset

Classifier	Tf-idf Setting	Accuracy	CV_score
RandomForest	ngram_range=(2, 2), sublinear_tf=True	0.846153846153846	0.793939393939394
SVC	ngram_range=(2, 2), sublinear_tf=True	0.846153846153846	0.793939393939394

Conclusion: When apply my best and second-best classifier to the entire test dataset, the performance(accuracy and cross-validation score) is **almost equal**.

Result When Labeled the Data Randomly

Classifier	Tf-idf Setting	Accuracy	CV_score
RandomForest	ngram_range=(2, 2), sublinear_tf=True	0.615384615384615	0.759090909090909
SVC	ngram_range=(2, 2), sublinear_tf=True	0.846153846153846	0.793939393939394

Conclusion: During all 10 trails that I had tried with random labeling, 8 of them generate **smaller** accuracy and cross-validation score than original method.

[20 points Extra Credit] Using the subject tf-idf rather than main-body, re-train your classifier/model of choice above and apply it to the test set. Does the performance improve?

Subject Analysis on Test Dataset

Classifier	Accuracy	CV_score
MLP	0.692307692307692	0.759090909090909
SVC	0.769230769230769	0.777272727272727
KNeighbors	0.846153846153846	0.793939393939394
GaussianNB	0.230769230769231	0.501515151515152
RandomForest	0.846153846153846	0.777272727272727

Conclusion: Using only the subject tf-idf will **not improve** the performance