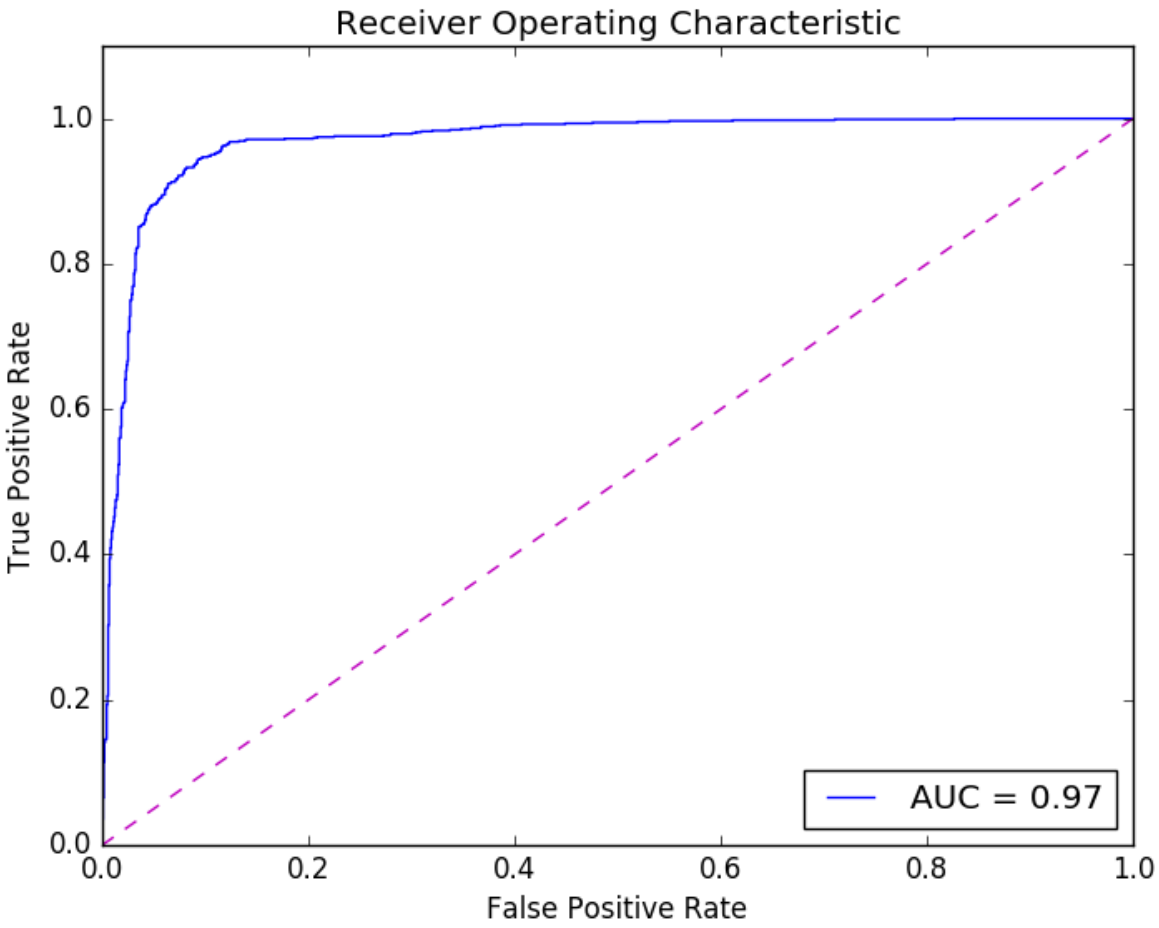
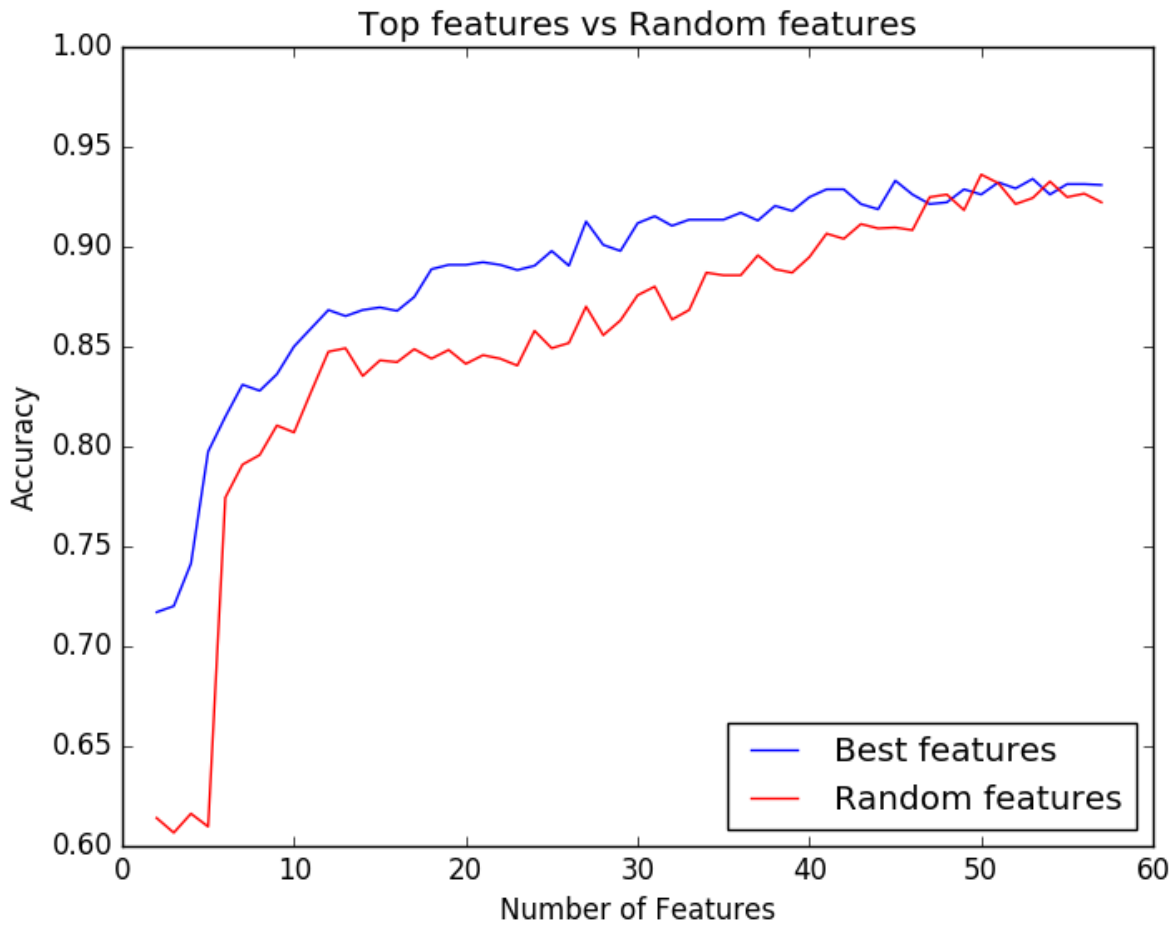
Machine Learning
HW3
Rubin Stricklin

Experiment 1



I used sklearn's SVM package for all of these experiments. Experiment 1 had a final accuracy of 93% and a final precision of 96%.



Experiment 2 & 3

In experiment 2 it looks like the top 5 features were:

1. Total number of capital letters in the e-mail
2. Length of longest uninterrupted sequence of capital letters
3. Average length of uninterrupted sequences of capital letters
4. Frequency of a word / total words
5. Frequency of a word / total words

I hypothesize that these features have the greatist weight because these are all stratigies that the spammers use to get there message across even when that message is trying to be hidden inside of junk text. Having a work show up in CAPS will make that word jump out from the rest of the text and having a word repeated several times will make it more likely that the person will see it.

Selecting the top features gives a pretty good accuracy with only a few features. To get ~70% with only two features seems pretty amazing. As more features are added the accuracy does go up, but the amout of accuracy gained diminishes with each feature. I think that this is because most of the features have actually very little weight and each feature added has less weight than the one before it.

Selecting random features does not work well when few features are selected. The above graph shows a huge spike around 5 features and that must be because one of the heavy weighted features was added. Overall as more features are added the plots start to converge, although it is puzzling that they didn't end with exactly the same accuracy.