MSDS 422 Assignment 5

In order to determine the cost and effectiveness of principal components analysis (PCA), a preliminary to machine learning classification, one will employ the MNIST dataset. This dataset contains data for hand-written digits. Each digit includes 784 pixels, and each of these pixels is represented in the dataset as a binary data point. By using binary classification methods one can predict what the number (represented as label in the dataset) is. To get an understanding of what these digits look like, here are a few examples of some hand-written digits, obtained from the dataset:

For our determination of the cost and effectiveness of PCA, training and testing datasets were obtained from kaggle.com. The training dataset included a label column, which detailed which digit was written, and also the binary data for the 784 pixels. Also, it included 42,000 hand-written digits. In contrast the testing dataset only included the 784 digits, and included 28,000 digits. Therefore, when making a predictive model the label column will be used as the explanatory variable. Predictions for the labels of the test data will be submitted to kaggle.com for an accuracy score.

The first method we employed was a Random Forest Classifier. The model was timed and then submitted for an accuracy score. The duration to build the model was about 12.7s and received an accuracy score of 94.3%

Next, PCA was used as preliminary step for classification. The principle components were identified in this process, and it took about 32.3 to do this. The principle components were then used for a Random Forest Classifier. The duration to build the model was 31.2s and received an accuracy score of 88.2%

Here is a summary of the data:

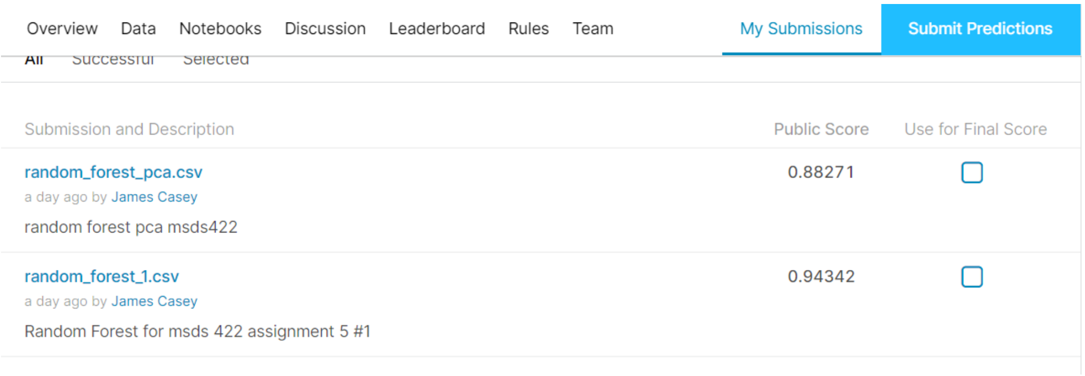
	Random Forest	PCA	2nd Random Forest	PCA + 2nd Random Forest
Time	12.7 s	32.3s	31.2s	63.5s

	Random Forest	PCA + 2nd Random Forest
Score	94.3%	88.2%

When it comes to making a management decision for which method to use, I would make use of the first method of using the random forest classifier without the PCA beforehand. The first method produced the most accurate results and had the lowest duration to complete the model. The second method which included PCA produced a lower accuracy and cost nearly 5X the amount of time. If this method had produced more accurate results, this would have been a more difficult decision. However, in this specific case there is no benefit to PCA.

Kaggle Display name: James Casey

Kaggle User name: jcasey2



Submission and Description	Public Score	Use for Final Score
random_forest_pca.csv a day ago by James Casey random forest pca msds422	0.88271	□
random_forest_1.csv a day ago by James Casey Random Forest for msds 422 assignment 5 #1	0.94342	□