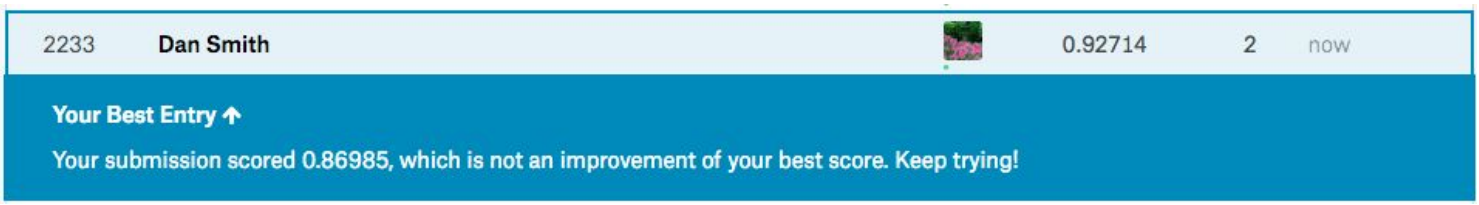
[Colab]



Problem Definition:

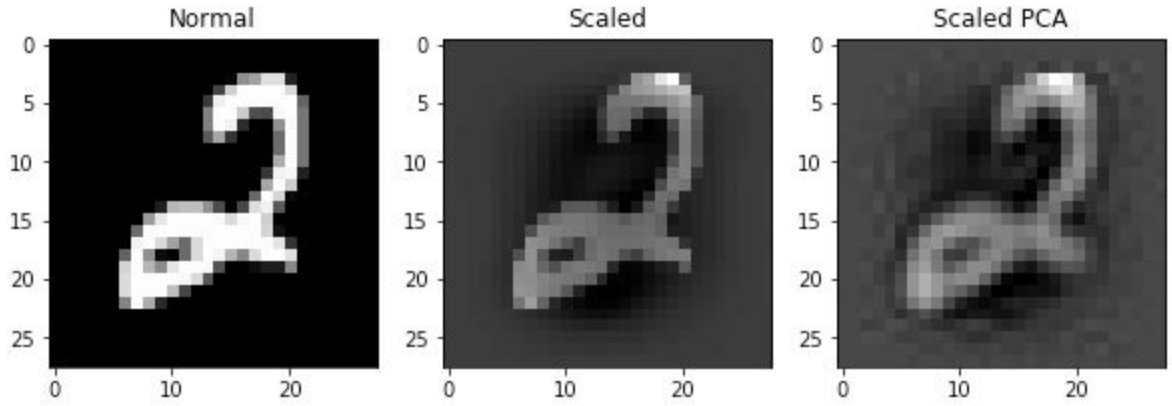
The purpose of this assignment was to conduct multiclass classification on the MNIST dataset, classifying images 0 - 9. Two models were used including RandomForestClassifier and PCA + RandomForestClassifier. Principal Component Analysis (PCA) is an unsupervised learning algorithm used to conduct dimensionality reduction. It is commonly used to simplify visualizations from n to 2-dimensions. It can also be used to decrease the computational resources required to fit a model.

Research Design/Programming:

Training data was split using train_test_split. The subsequent split was standardized to a mean of 0 and standard deviation of 1 using StandardScalar. A model was fit using the RandomForestClassifier algorithm following hyperparameter tuning with Random Search. Next, PCA was conducted such that 95% of the variability was still included in the explanatory variables. This reduced the number of features from 784 to 318.

	Time to Fit (s)	Accuracy Score	# Features
RandomForest	376.086310	92.690476	784
PCA + RandomForest	565.483327	87.070000	318

In theory, the decrease in features should have decreased the time required to fit the model. However time to fit increased using PCA. By decreasing the number of features the RandomForestClassifier took longer to find splits that minimized impurity, resulting in a longer time to fit the model. The accuracy score was better without PCA. This makes intuitive sense when we look at what our model is actually trying to predict. The model using PCA includes includes only a subset of the original 784 pixels. As a result we have an image that is less clear with decreased resolution.



Recommendations for Management:

The RandomForestClassifier without PCA outperformed the model that utilized PCA. PCA has advantages in certain use cases, however in this instance it is not recommended. I suggest that the RandomForestClassifier be used for now. However, going forward, other algorithms should be explored to improve predictive accuracy.

Note: The directions prompt us to perform PCA using the test and training set. PCA should be conducted only on the training set. Out of sample data should be quarantined until final model evaluation. For this reason I did not follow the directions and only used the training set to conduct PCA.