Sahil Rangwala
MSDS 422

Module 6 Assignment 1: Digit Recognizer

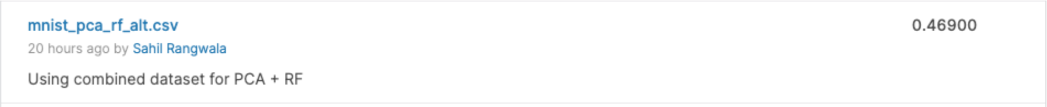
This week's assignment we utilized a combination of PCA, Random Forest, and KNN. The initial dataset was clean and was void of empty values thus there was not much data cleaning needed. Some EDA showcased the optimal PCA component values as well as the overall trends of the data. The incorporation of PCA is useful for training a model to represent a multivariate data table as a smaller set of explanatory variables in order to observe trends, jumps, clusters and outliers. This overview may uncover the relationships between observations and variables, and showcase insights not understood before. After identifying the principal components, we can utilize that to help tweak or tune or model to become more refined.

The initial Random Forest model was not tweaked or tuned in any way. It used the default parameters as well as the default data passed in (not tuned or scaled data). This model offered a greatly accurate score. Additionally, this model still offered a great score within the Kaggle hidden test cases. The score is below.

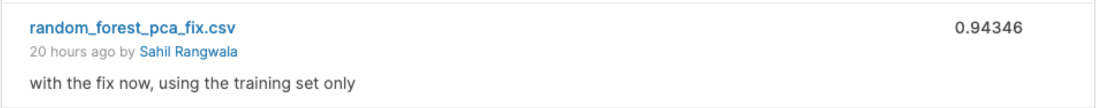
Now after conducting PCA, we are able to identify the number of components that are relevant and can scale the data we pass into the model in this configured way. With this, we tried to preserve 95% variability thus it equated to about 154 components. Basically, PCA can make the process of finding the perfect decision boundary much easier by aligning your training set along the directions with highest variance.

K-means clustering is an unsupervised machine learning method; consequently, the labels assigned by our KMeans algorithm refer to the cluster each array was assigned to, not the actual target integer. To fix this, we defined a few functions that will predict which integer corresponds to each cluster.

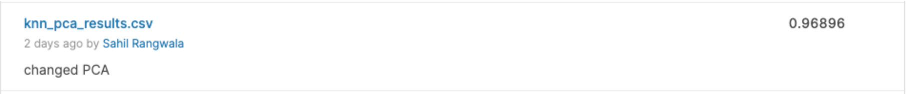
The flaw that was introduced was that we conducted PCA on the whole dataset. It is not needed and actually brought down the overall score of the random forest model. Once we used PCA on the subset of the data set (the training set) it resulted in a better score on Kaggle as well as a quicker time it took to conduct the analysis. Surprisingly, the default random forest score fared really well in comparison to the other iterations of the model.



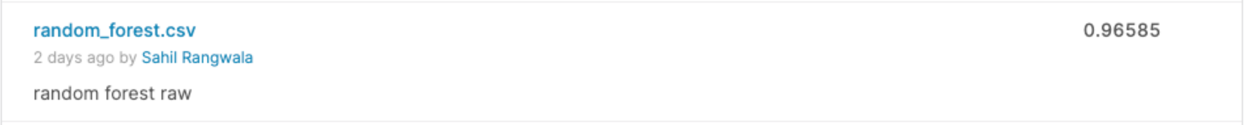
mnist_pca_rf_alt.csv 0.46900
20 hours ago by Sahil Rangwala
Using combined dataset for PCA + RF



random_forest_pca_fix.csv 0.94346
20 hours ago by Sahil Rangwala
with the fix now, using the training set only



knn_pca_results.csv 0.96896
2 days ago by Sahil Rangwala
changed PCA



random_forest.csv 0.96585
2 days ago by Sahil Rangwala
random forest raw