INTRODUCTION

Project focuses on utilizing unsupervised learning method for multi-class prediction of handwritten digits in the MNIST dataset. The challenge is to build a model that maps digit images to proper labels.

SUMMARY AND PROBLEM DEFINITION FOR MANAGEMENT

Approach uses combination of Random Forest (RF) learning and Principal Component Analysis (PCA) methods in order to classify images. The challenge is to build an accurate model well as to build a model that runs in reasonable time. The prediction is sent Kaggle.com Digit Recognizer competition for scoring https://www.kaggle.com/c/digit-recognizer/ with following KEGGEL ID: YGIZHITSA , YURIY G, **ygizhitsa@hotmail.com**)

MEASUREMENT AND STATISTICAL METHODS

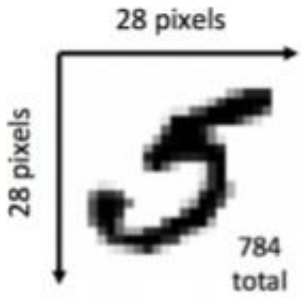
The steps for the analysis are the following:

- Begin by fitting a random forest classifier using the full set of 784 explanatory variables and the model training set (train.csv). Record the time it takes to fit the model and then evaluate the model on the test.csv data by submitting to Kaggle.com.
- (Execute principal components analysis (PCA) on the combined training and test set data together, generating principal components that represent 95 percent of the variability in the explanatory variables. The number of principal components in the solution should be substantially fewer than the 784 explanatory variables.
- Using the identified principal components from step (2), use the train.csv to build another random forest classifier. Record the time it takes to fit the model and to evaluate the model

EXPLORATORY DATA ANALYSIS METHODS

There are 70,000 handwritten digits. Each row represents one of these digits. There are 785 columns of data. 784 of them are the integer grey scale values of each pixel in a 28 x 28-pixel square. The first column is the 'response' variable, which is the actual value to test the predicted estimate against. An example of a plotted row of data (784 pixels). The plot on the below is a binary plot showing a row of

data that has a y value of '5'. The first column of data is the actual value – for training & testing. The rest 784 columns of data are the greyscale values for each of the 28x28 pixels representing the digit.



An example of a plotted row of data (784 pixels)

Overview of Programming Work

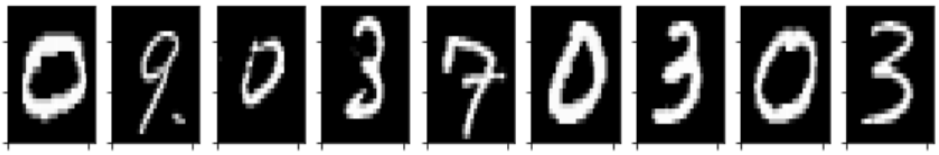
Models

	RF 784	RF Search Grid	PCA-RF-Scaled	PCA-RF (Not Scaled)
Classifier	Random Forest	Random Forest	semi-supervised learning: PCA and Random Forest.	semi-supervised learning: PCA and Random Forest. Data is scaled prior to PCA fitting
Hyperparameters	All 784 enumerators used	200 estimators, criterion = gini	PCA- FULL (70000) RF -42000/28000	PCA- FULL (70000) RF -42000/28000
Data	Train	Train	FULL	FULL
Transformation			Data is scaled prior to PCA fitting and converted to integer	

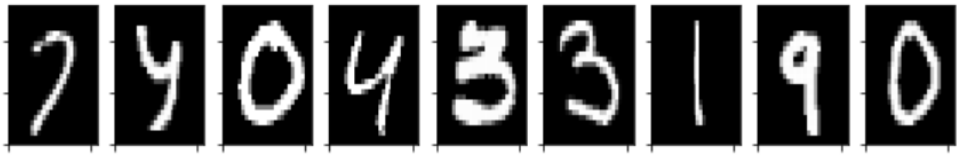
RESULT AND CONCLUSIONS

Visualization of the result

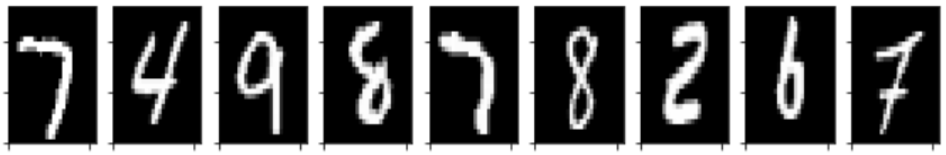
For the visualization of the pixel data was rendered and compare with four model prediction in five 10 digit batches.



RF-784 [0 9 9 3 7 0 3 0 3]
RF-SG [0 9 9 3 7 0 3 0 3]
PCA-RF-S [0 9 4 3 7 0 3 0 3]
PCA-RF [0 9 4 3 7 0 3 0 3]



RF-784 [7 4 0 4 3 3 1 9 0]
RF-SG [7 4 0 4 3 3 1 9 0]
PCA-RF-S [7 4 0 4 3 3 1 9 0]
PCA-RF [7 4 0 4 3 3 1 9 0]



RF-784 [7 4 9 8 7 8 2 6 7]
RF-SG [7 4 9 8 7 8 2 6 7]
PCA-RF-S [7 4 9 8 7 8 8 6 7]
PCA-RF [7 4 9 8 7 8 8 6 7]

RF-784 [1 1 5 7 4 2 7 4 7]
RF-SG [1 1 5 7 4 2 7 4 7]
PCA-RF-S [1 1 5 7 4 2 7 7 7]
PCA-RF [1 1 5 7 4 2 7 4 7]

RF-784 [5 4 2 6 2 5 5 1 6]
RF-SG [5 4 2 6 2 5 5 1 6]
PCA-RF-S [5 4 2 6 2 5 5 1 6]
PCA-RF [5 4 2 6 2 5 5 1 6]

KEGGEL SCORING AND TIME

KEGGEL ID: YGIZHITSA , YURIY G, ygizhitsa@hotmail.com

	RF 784	RF SEARCH GRID	PCA-RF-SCALED	PCA-RF (NOT SCALED)
KEGGEL SCORE	0.96614	0.96600	0.93842	0.94885
EXECUTION TIME	108.8296866	57.4337041378	PCA- 13.01689 RF - 48.202514	PCA- 9.9489 RF - 149.73435354

CONCLUSIONS:

Random Forest model with performance tuning recommend by Search Grid provides better balance for accuracy (0.966) and speed (57s). More Over while PCA might be used for the semi-supervised learning, in general PCA has very sensitive flow for predictive analysis – different data produces different eigen vectors and eigen values and need to be retrained. Another potential problem might raise from the scaling naturally binary data (pixels). It is easily impact original digit representation.