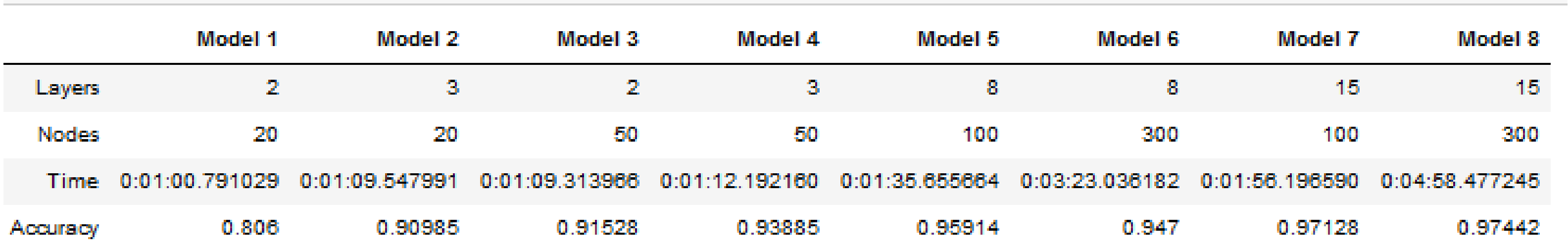
MSDS 422 Assignment 5

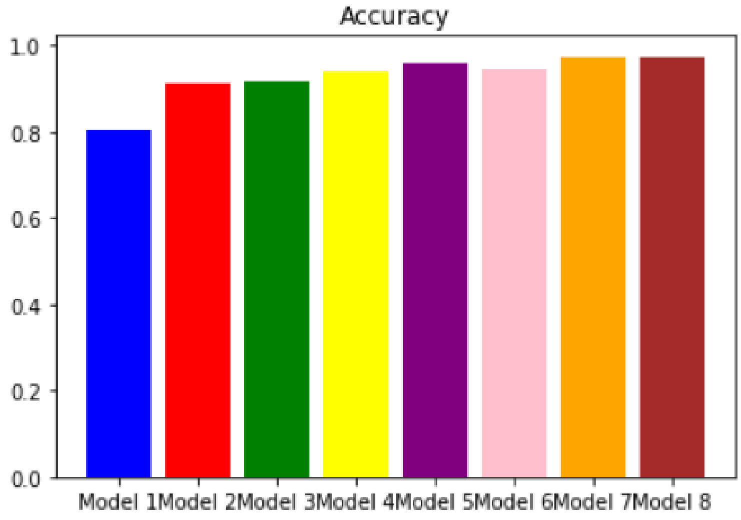
In order to determine the cost and effectiveness of neural networks, one will use the MNIST dataset. One will compare neural networks models by altering the number of nodes and layers in each model. By examining the duration and accuracy, one can determine which type of neural network should optical character recognition. This dataset used for this study 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:

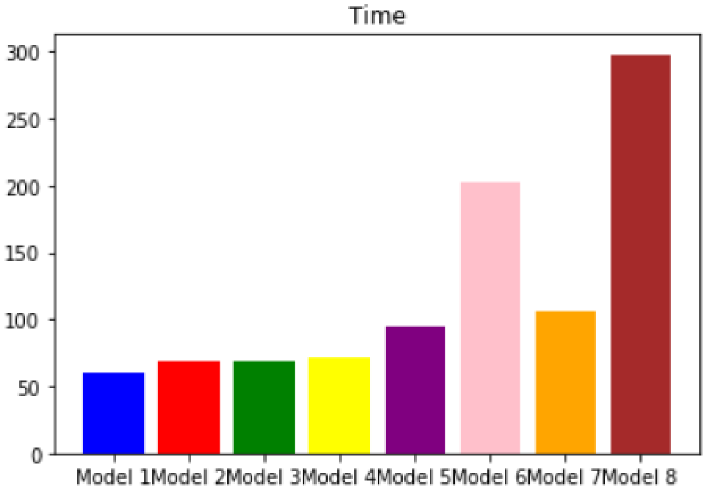
For our determination of the cost and effectiveness of neural networks, training and testing datasets were obtained from kaggle.com. The training dataset included a label column, which detailed which digit was written, and also the binary data for the 784 pixels. Also, it included 42,000 hand-written 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.

A DNN classifier was used in order to build models. In order to analyze these neural network models, the amount of nodes and layers were adjusted. By having varying nodes and layers one can determine which model produces the most accurate predictions and spends the least amount of time. Eight models were produced, and the results can be seen below.



	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Layers	2	3	2	3	8	8	15	15
Nodes	20	20	50	50	100	300	100	300
Time	0:01:00.791029	0:01:09.547991	0:01:09.313966	0:01:12.192160	0:01:35.655664	0:03:23.036182	0:01:56.196590	0:04:58.477245
Accuracy	0.806	0.90985	0.91528	0.93885	0.95914	0.947	0.97128	0.97442





One can recognize that as more nodes and layers are added the accuracy increases. However, this also increases the amount of time it takes to produce the model. It appears an increase in layers has a greater impact on accuracy than nodes, which seen when nodes are held constant compared to when layers are held constant. Furthermore, one can notice that nodes have a larger impact on the amount of time, as seen by the large increases in time when nodes are increased.

Based on the study, I would choose model 7, which features 100 nodes and 15 layers. This model produces one of the most accurate results at 97.1% and a time just short of 2 minutes. Although, it is not the best time, it is not terribly long compared to model 6 and model 8 which have 300 nodes. In general, I would increase the amount of layers before increasing the number of nodes in order to preserve time and increase accuracy. Furthemore, I would not use the DNN classifier with the parameters of models 1-4. This is due to the fact that the Random Classifier from the previous analysis produced an accuracy of 94.3% in only 12 seconds.

I used 2 accounts for this assignment, since I ran out of submissions

First 7 models:

Kaggle User name: jcasey2

Last Model:

Kaggle User name: jcasey18