Week 12 Assignment

Vivek Golla
HDS 5230-07

Question 1: Using the data synthesis R script provided by the instructor as part of the week 11 assignment instructions, produce datasets of the following sizes, and fit deep learning models with the configurations shown below. Associated with each model, record the following performance characteristics: training error, validation (i.e., holdout set) error, time of execution. Use an appropriate activation function.

```
 Data Size Layers  Train Error  Validation Error  Execution Time (s)
0      1000  1 x 4       0.0512            0.0750               11.73
1     10000  1 x 4       0.0015            0.0010               37.96
2    100000  1 x 4       0.0008            0.0009              221.84
3      1000  2 x 4       0.1187            0.1400               13.80
4     10000  2 x 4       0.0016            0.0015               42.68
5    100000  2 x 4       0.0008            0.0013              217.98
```

Question 2: Based on the results, which model do you consider as superior, among the deep learning models fit?

I would consider The Model with 1 hidden layer and 4 nodes, using dataset size of 10000, as the best model. This is because it almost has the lowest Validation Error (0.0010 vs. 0.0009), while also being considerably faster than the Models made using 100000 records (37.96s).

Question 3: Next, report the results (for the particular numbers of observations) from applying xgboost (week 11 – provide the relevant results here in a table). Comparing the results from XGBoost and deep learning models fit, which model would you say is superior to others? What is the basis for your judgment?

XGBoost in Python via scikit-learn and 5-fold CV

Dataset Size	Testing-set predictive performance	Testing-set predictive error	Time taken for this model to be fit (s)
1000	0.9480	0.052	0.30
10000	0.9789	0.0211	0.73
100000	0.9867	0.0133	2.04

Comparing the XGBoost models to the Deep Learning models, I would say the Deep Learning model is better. The validation error is significantly lower for the Deep Learning models when the dataset size is higher, although it takes more time to compute than the XGBoost model.