Project 3- Supervised Learning

(a) The use of linear regressor is not sufficient because as highlighted in the graph the data is not linear, hence the sum square error is large with 71387. So, an ANN with at least one hidden layer is needed to fit and predict the data properly.

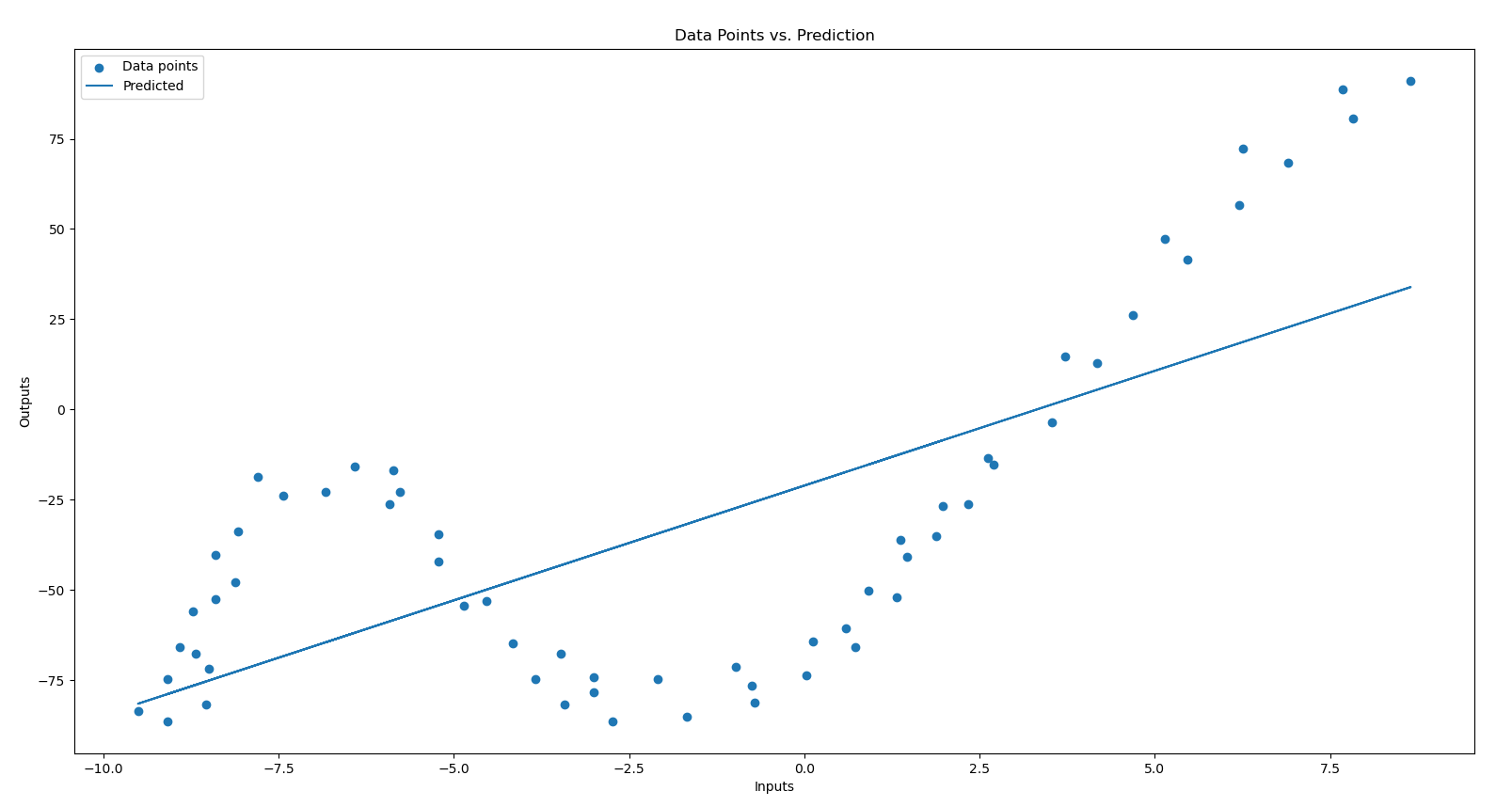


Figure 1- Linear Regression vs. Data Points

Then different hidden layers, learning rates, epochs and their different combinations were experimented with to get the best result. First, two hidden units with learning rate 0.001 and 10,000 epochs are chosen. While selecting the hidden units generally twos multiples are chosen, so 4, 8, 16, 32, and 64 hidden units were experimented with.

|  |  |
| --- | --- |
| Number of Hidden Units | Error |
| 2 | 29349 |
| 4 | 25456 |
| 8 | 21844 |
| 16 | 20431 |
| 32 | 15279 |
| 64 | 13749 |

Then learning rate values 0.01, 0.001, 0.0001, and 0.00001 are tried and 0.01 is found to give the least error.

|  |  |
| --- | --- |
| Learning Rate | Error |
| 0.01 | 55955 |
| 0.001 | 13373 |
| 0.0001 | 31938 |
| 0.00001 | 78891 |

Lastly, the epoch values between 10 and 1000000 increasing with a multiple of 10 were tested, and best result was with 1000000. The downside to choosing the larger epoch value is that training takes more time and it can lead to overfitting. So 100000 was chosen, since training takes a lot of time with 1000000 epochs.

|  |  |
| --- | --- |
| Epochs | Error |
| 10 | 203255 |
| 100 | 88362 |
| 1000 | 31970 |
| 10000 | 13318 |
| 100000 | 4052 |
| 1000000 | 2842 |

Normalization helps with data points having different scales. When inputs have similar scales, it helps with convergence. Also, normalization can help prevent the weights from becoming too large or too small during the learning process. Since the scales of the data points are similar, normalization is not done.

(b)

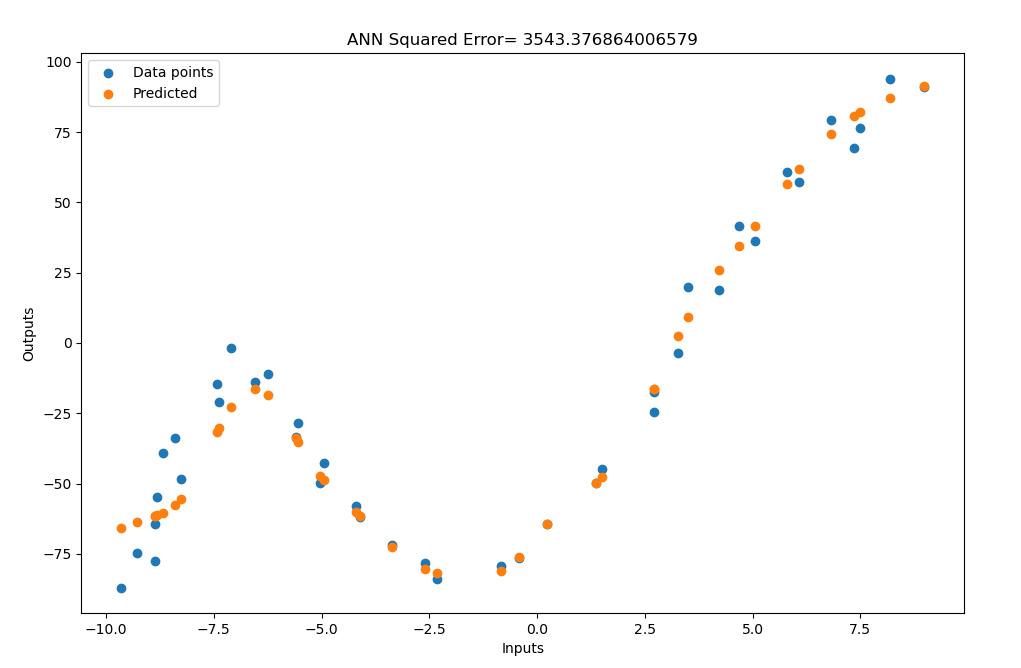
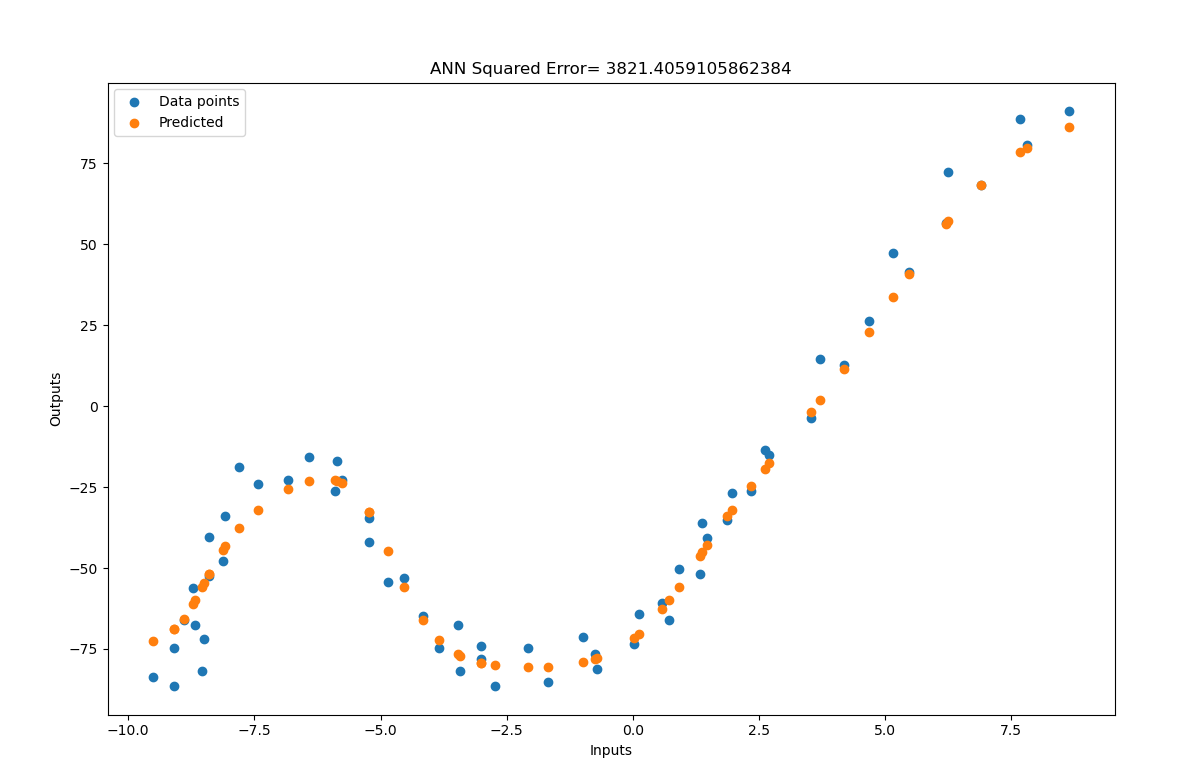


Figure 2- Train Data Points vs. Predicted Figure 3- Test Data Points vs. Predicted

ANN used: 64 When to stop: 100000

Learning rate: 0.001 Is normalization used: No

Range of initial weights: 0.1 Training loss: 11233

Number of epochs: 100000 Test loss: 9038

(c) Finally, for each of the 2, 4, 8, 16, and 32 number of hidden units different learning rates, and epochs are experimented with and the best results are chosen following the steps from previous parts.

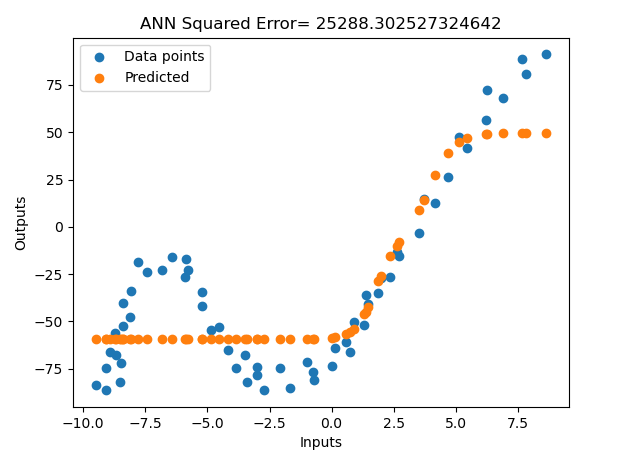
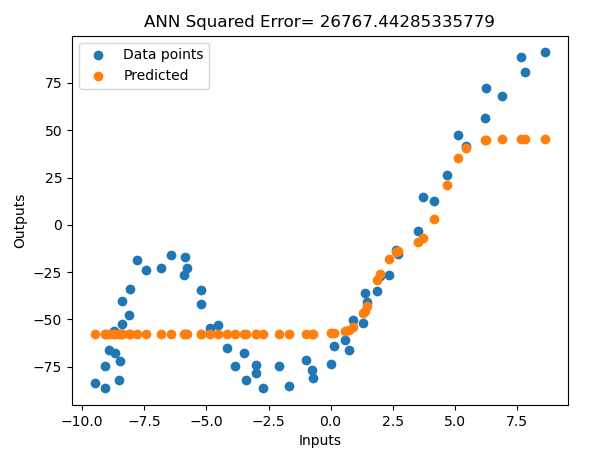


Figure 4- Two Hidden Units Curve Figure 5- Four Hidden Units Curve

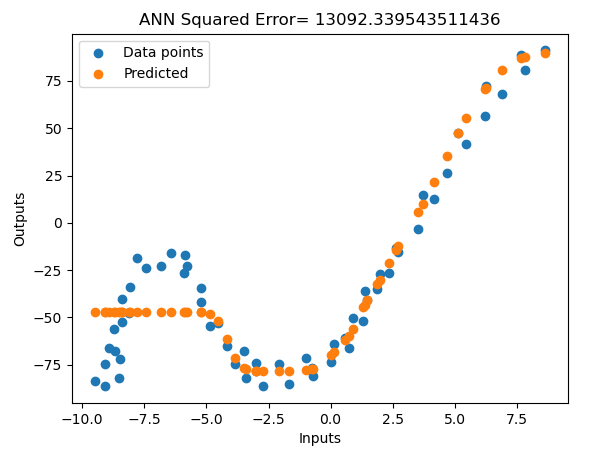
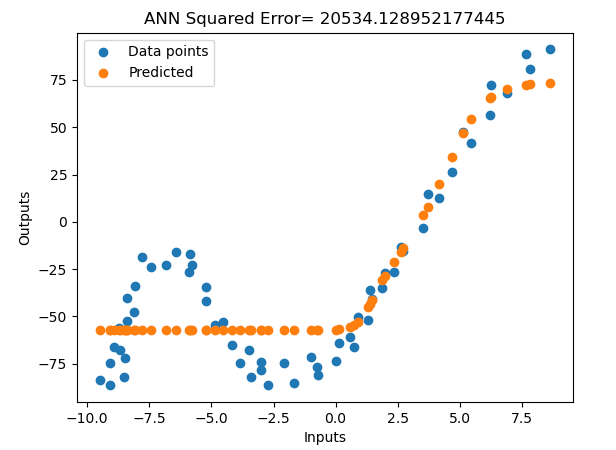


Figure 6- Eight Hidden Units Curve Figure 7- Sixteen Hidden Units Curve

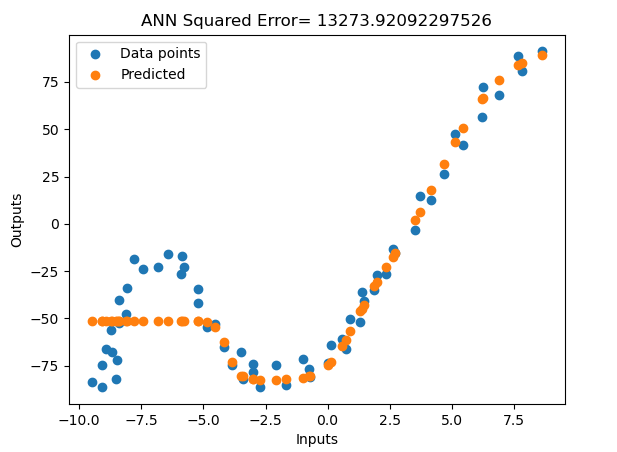


Figure 9- Thirty-Two Hidden Units Curve

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Hidden Units | Learning Rate | Epochs | Training Error | Test Error | Avg. Training Error | Avg. Test Error | Std. Deviation Train | Std. Deviation Test |
| 2 | 0.001 | 100000 | 26660 | 87096 | 444 | 2124 | 450 | 2152 |
| 4 | 0.001 | 100000 | 20208 | 86602 | 336 | 2112 | 341 | 2140 |
| 8 | 0.001 | 100000 | 19869 | 86454 | 331 | 2108 | 335 | 2136 |
| 16 | 0.001 | 100000 | 19918 | 16610 | 331 | 405 | 336 | 411 |
| 32 | 0.001 | 100000 | 12761 | 10629 | 212 | 259 | 216 | 264 |

For all of the hidden unit counts the optimum learning rate and epoch count was the same. Some of the table and graph values are different because outcome of predictions is random at each run. Since graph and table values were taken separately, the values are different. With increasing complexity, the model was better able to capture the shape of the data and consequently error decreased. Also, 32 hidden units were not enough to capture the curve on the lower parts, between -10 and -2.5. But with 64 hidden units it was able to capture the curve of that part as shown in the figure 2 and 3. More complex models are able to learn better unless they over fit.