PROJECT – 5

ADITYA BHARDWAJ **ABHARDW2**

PART 1

Process:

- Dataset is cleaned of outliers and errors to ensure better fit for the model.
- The neural network calculations take some time, so I added some additional parameters to increase the speed, but it affected the performance of the training of the model. So, I shifted to the original method.

For Hidden Layer = 2, and neurons per layer = 2 and 1:

Hidden Layer: [2, 1]				
Iteration	**MSE**	**SSE**	**RMSE**	**R^2**
1	799.074236658	399537.11832	28.2679011718	0.997023195664
2	925.724801839	462862.400919	30.42572598705	0.99705770650
3	937.544945128	468772.472564	30.6193557268	0.997004448974
4	816.0672765193	408033.6382596	28.5668912645	0.997267397753
5	784.245030659	392122.515329	28.0043752056	0.997504655435
6	1090.997180363	545498.590181	33.03024644721	0.995995397237
7	952.257831416	476128.915708	30.85867514032	0.996778348761
8	856.285008449	428142.504224	29.26234796542	0.997140624791
9	815.944167927	407972.083963	28.56473644071	0.997100043207
10	792.351592831	396175.796415	28.14874051945	0.997461117343
MEAN	877.0492072	438524.6036	29.57489959	0.997033294

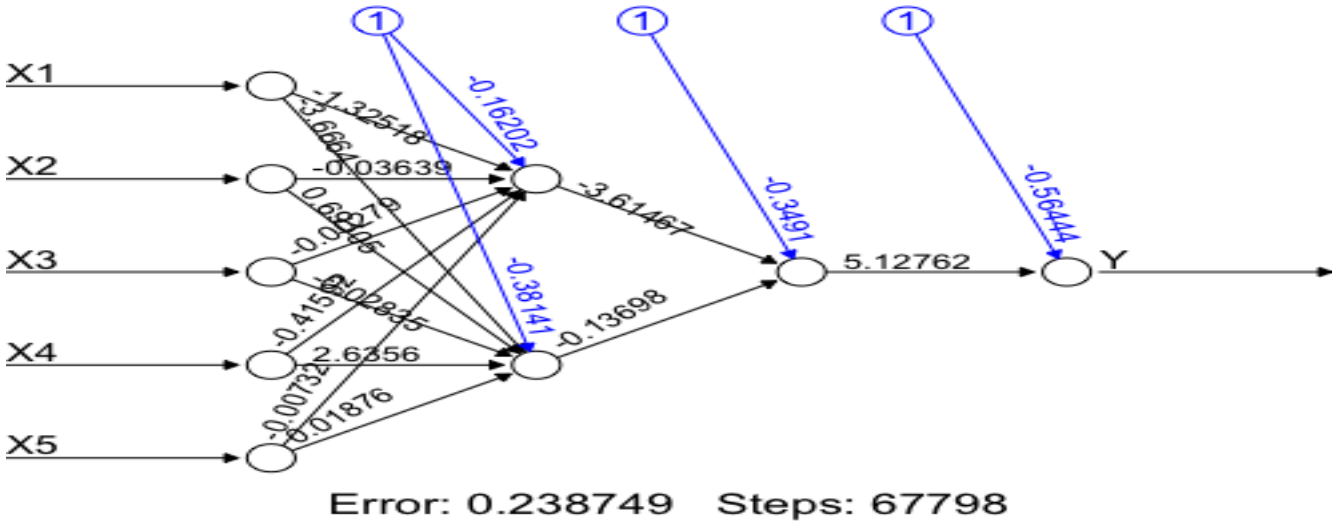

Figure 1Neuron in the layer [2, 1]

Observation Criteria:

1. Higher the value of R^2 (i.e. ~ 1), better the predictions by the model.
2. Lowest the values of SSE, MSE and RMSE, model would be better fit.
3. Given dataset, RMSE value is around 28 – 30.

PART 2

COMPARISON OF HIDDEN LAYERS				
Layers	**MSE**	**SSE**	**RMSE**	**R^2**
[2,1]	877.0492072	438524.6036	29.57489959	0.997033294
[2,2]	844.5414931	422270.7466	29.0379145	0.997144802
[2,1,2]	876.1627688	438081.3844	29.5670376	0.997036158
[3,1]	826.186243121	413093.121560	28.7434556572	0.997352703089
[3,3,1]	855.3272567	427663.6284	29.21620455	0.997111432
[3,2,1]	856.7603361	428380.168	29.24204837	0.997104007
[4,2]	847.0993708	423549.6854	29.08401569	0.997137578

Observations:

1. R^2 value for all the observations is high, i.e., near 1. R^2 value measure ensures a good fit for the regression model, hence, ANN is good for regression on the current dataset. **Also, the R^2 value for all the experiments are similar, but for [3,1] it is highest among all.**
2. The RMSE values for all the experiments are similar, and moves around 29 which shows that increasing or decreasing number of neurons or layers doesn't have much impact on the performance of the model. **The RMSE value for [3,1] is lowest among all the experiments, hence, it should be considered over others when taking RMSE as the measure.**
3. MSE values shows variations between 826 and 877. The lowest recorded MSE value is for [3,1] where 3 neurons are in first layer and 1 is in second layer. **Hence, it should be first choice if we consider lowest MSE.**
4. SSE values also varies between 410000 and 430000, and the lowest value is recorded for [3,1], hence, **[3,1] should be considered as best model.**

The selection criteria for Neurons:

- Total # of neurons = 2/3rd of input/output nodes.
- Neurons need not be added in the hidden layers if adding them doesn't improve the performance, as it would increase the training time for the model.

BEST MODEL: 3 neurons in 1st layer and 1 neuron in 2nd layer.

The detailed results of each iteration for the k-folds method in the program are as below:

Hidden Layer: [2, 2]				
Iteration	**MSE**	**SSE**	**RMSE**	**R^2**
1	808.356884845	404178.442422	28.4316176966	0.996988614863
2	881.321964852	440660.982426	29.6870672996	0.997198835032
3	908.814918607	454407.459303	30.1465573259	0.997096244318
4	825.082027018	412541.013509	28.7242411043	0.997237211850
5	766.561282130	383280.641065	27.6868431232	0.997560922347
6	977.586634795	488793.31739	31.2663818628	0.996411680791
7	915.114173072	457557.08653	30.2508540883	0.996904012115
8	805.521553620	402760.77681	28.3817116048	0.997310138169
9	773.018758753	386509.37937	27.8032148996	0.997252604910
10	784.036733371	392018.36668	28.0006559453	0.997487760128
MEAN	844.5414931	422270.7466	29.0379145	0.997144802

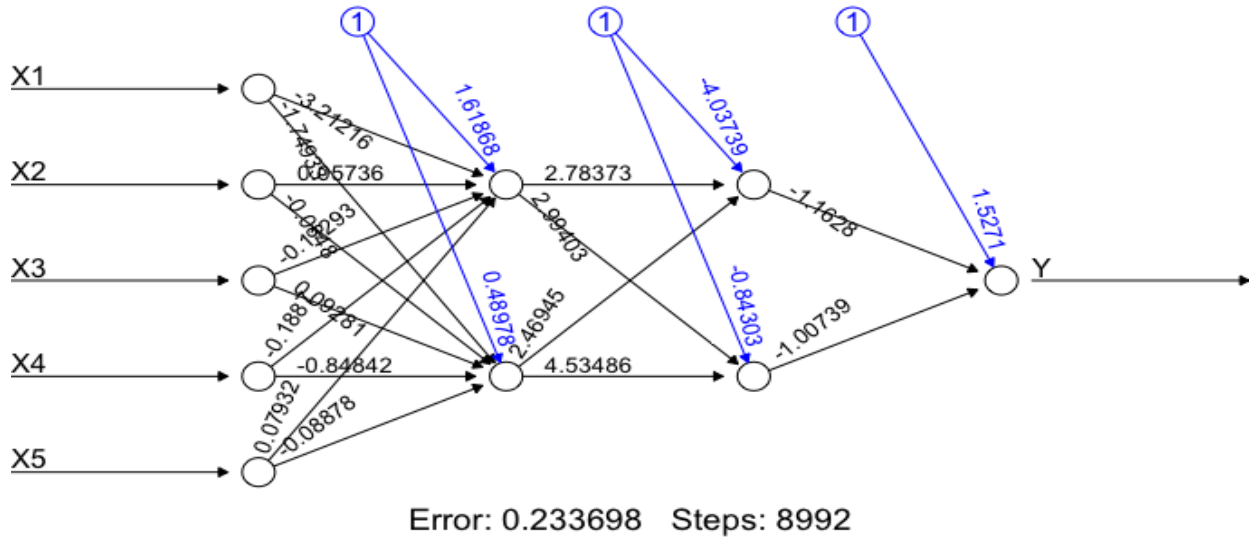

Figure 2Neuron in the layer [2, 2]

Hidden Layer: [2, 1, 2]				
Iteration	**MSE**	**SSE**	**RMSE**	**R^2**
1	816.040773685	408020.386842	28.5664273874	0.996959989946
2	915.116301702	457558.150851	30.2508892712	0.997091424215
3	939.244662631	469622.331315	30.6470987636	0.996999018205
4	829.699091461	414849.545730	28.8044977644	0.997221751605
5	755.844618499	377922.309249	27.4926284392	0.997595021088
6	1049.84165454	524920.827252	32.4012600758	0.996146462277
7	959.988765239	479994.382619	30.9836854689	0.996752193688
8	857.235211702	428617.605851	29.2785794003	0.997137451797
9	832.35489799	416177.448996	28.8505614848	0.997041717637
10	806.261710521	403130.855260	28.3947479390	0.997416546023
MEAN	876.1627688	438081.3844	29.5670376	0.997036158

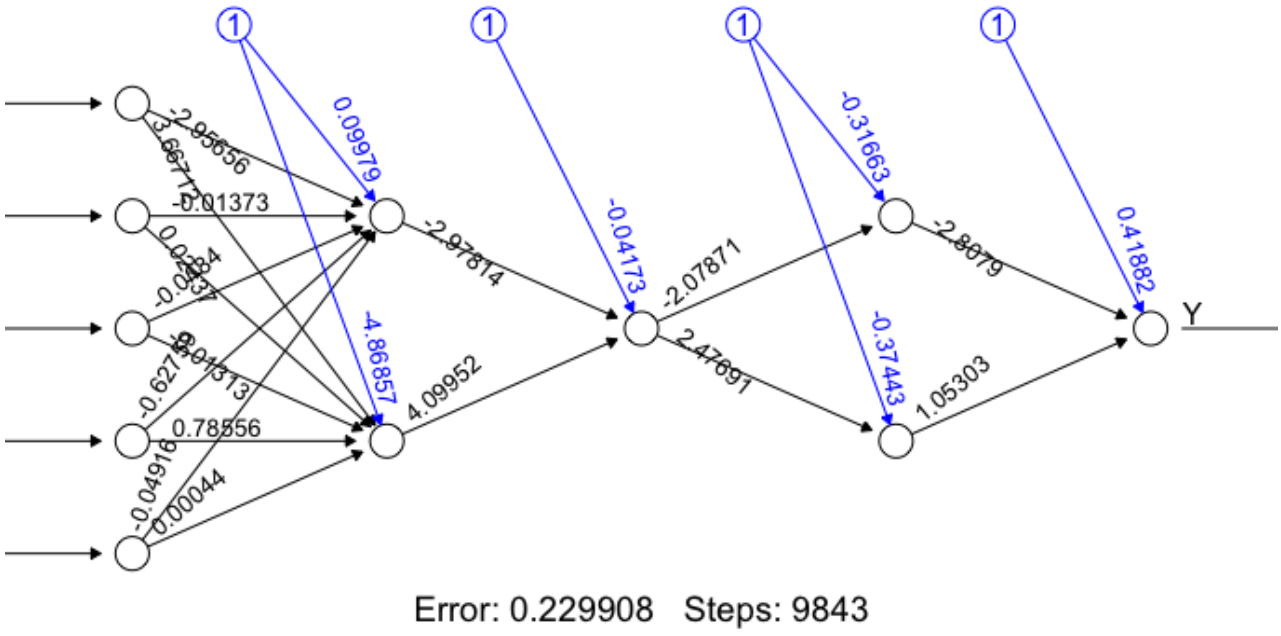

Figure 3Neuron in the layer [2, 1, 2]

Hidden Layer: [3, 1]				
Iteration	**MSE**	**SSE**	**RMSE**	**R^2**
1	802.205364781	401102.682390	28.3232301261	0.997011531221
2	889.378658089	444689.329044	29.8224522480	0.997173227902
3	940.029077700	470014.538850	30.6598936348	0.996996511919
4	807.513000819	403756.500409	28.4167732302	0.997296041756
5	772.197667505	386098.833752	27.7884448558	0.997542988253
6	1033.52010328	516760.051643	32.1484074766	0.996206371991
7	924.469347161	462234.673580	30.4050875210	0.996872361959
8	832.174190033	416087.095016	28.8474295221	0.997221137560
9	776.510441583	388255.220791	27.8659369407	0.997240195079
10	826.186243121	413093.121560	28.7434556572	0.997352703089
MEAN	860.4184094	430209.2047	29.30211112	0.997091307

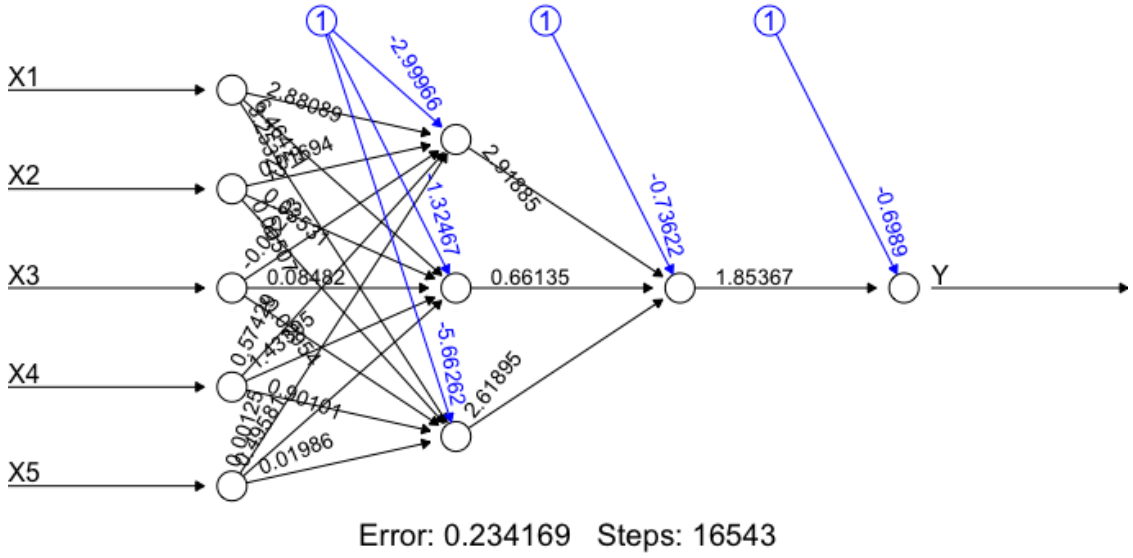

Figure 4Neuron in the layer [3, 1]

Hidden Layer: [3, 3, 1]				
Iteration	**MSE**	**SSE**	**RMSE**	**R^2**
1	763.038943412	381519.471706	27.62315954795	0.997157438532
2	901.444355477	450722.177738	30.02406294087	0.997134878683
3	948.720453411	474360.226705	30.80130603418	0.996968742094
4	830.182347795	415091.173897	28.81288510016	0.997220133420
5	764.438099173	382219.049586	27.64847372231	0.997567677982
6	991.405092360	495702.546180	31.48658591147	0.996360958907
7	917.674546258	458837.273129	30.29314355193	0.996895349934
8	855.599032792	427799.516396	29.25062448550	0.997142915456
9	767.753828536	383876.914268	27.70837109136	0.997271317060
10	813.0158678555	406507.9339277	28.51343311240	0.997394904099
MEAN	855.3272567	427663.6284	29.21620455	0.997111432

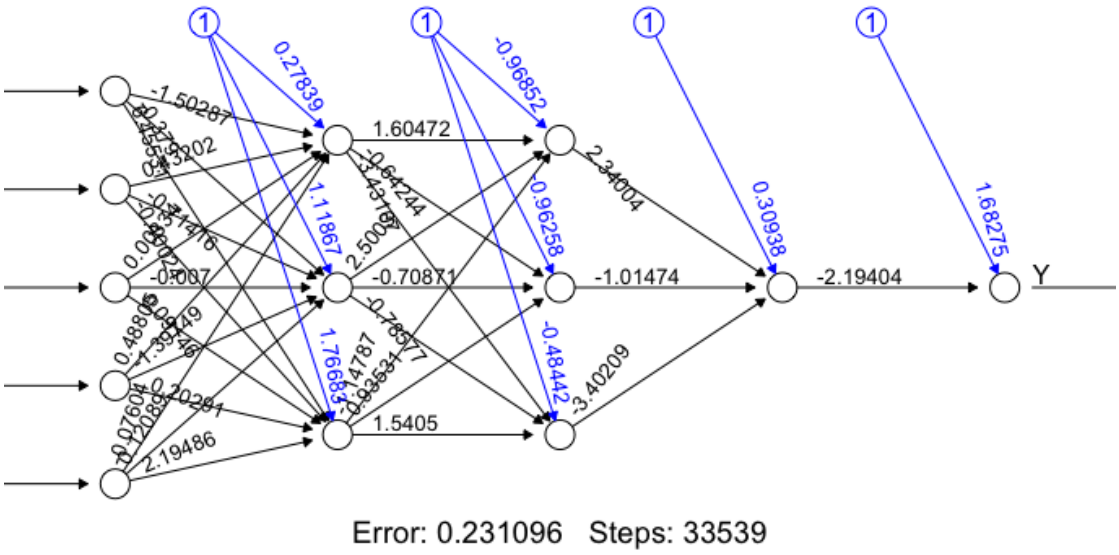

Figure 5Neurons in layer [3, 3, 1]

Hidden Layer: [3, 2, 1]				
Iteration	**MSE**	**SSE**	**RMSE**	**R^2**
1	764.6888638054	382344.4319027	27.65300822343	0.9971512920569
2	936.4970463388	468248.523169	30.6022392373	0.997023468354
3	886.831040159	443415.520079	29.7797085304	0.997166485036
4	822.748838551	411374.419275	28.6835987726	0.997245024535
5	754.938607355	377469.303677	27.4761461518	0.997597903873
6	1005.03752048	502518.760244	31.7023267362	0.996310919859
7	923.586336795	461793.168397	30.3905632852	0.996875349335
8	840.582289747	420291.144873	28.9927972046	0.997193060562
9	834.304958363	417152.479181	28.8843375960	0.997034786905
10	798.387858990	399193.929495	28.2557579793	0.997441775713
MEAN	856.7603361	428380.168	29.24204837	0.997104007

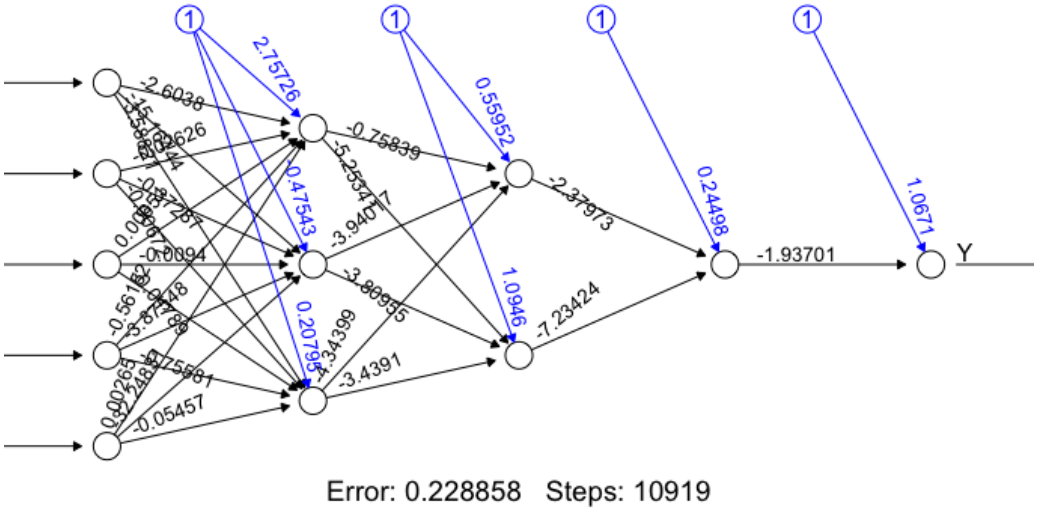

Figure 6Neurons in layer [3, 2, 1]

Hidden Layer: [4, 2]				
Iteration	**MSE**	**SSE**	**RMSE**	**R^2**
1	776.602910566	388301.455283	27.8675960672	0.997106908463

2	875.426814251	437713.407125	29.5876125135	0.997217571987
3	908.142550673	454071.275336	30.1354036089	0.997098392602
4	816.499737438	408249.868719	28.5744595301	0.997265949657
5	782.766695595	391383.347797	27.9779680390	0.997509359265
6	980.045985126	490022.992563	31.3056861468	0.996402653526
7	902.838412758	451419.206379	30.0472696390	0.996945543113
8	850.462700016	425231.350008	29.1626936344	0.997160067108
9	782.310364594	391155.182297	27.9698116653	0.997219581504
10	795.897536926	397948.768463	28.2116560472	0.997449755297
MEAN	847.0993708	423549.6854	29.08401569	0.997137578

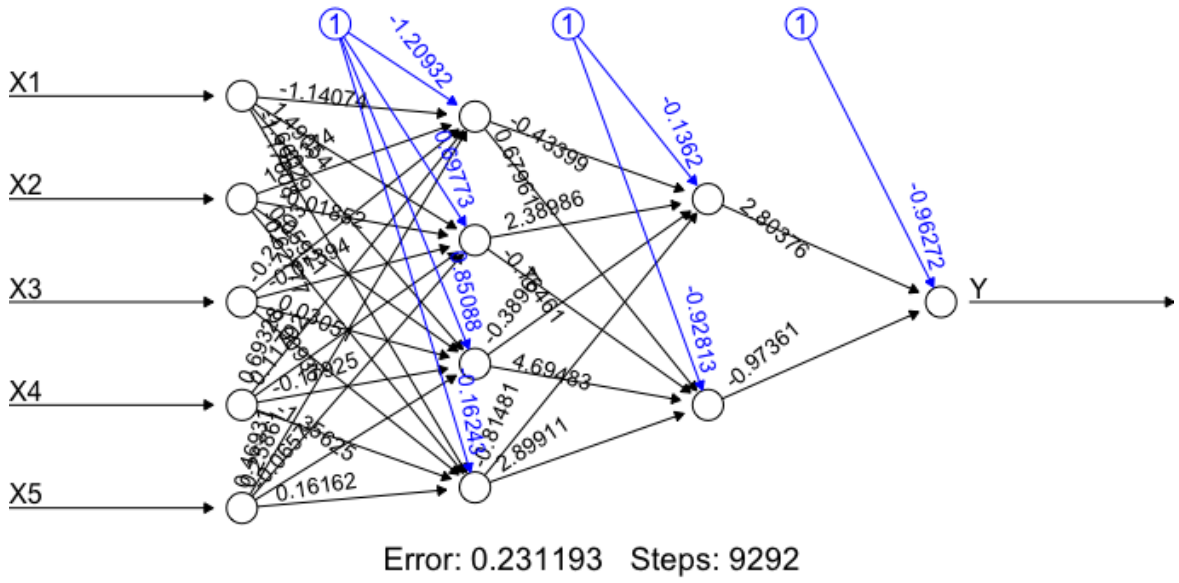

Figure 7Neurons in layer [4, 2]

EXTRA CREDIT

The results of Multivariate regression after running on the dataset are as follows:

Call:
lm(formula = dataset$Y ~ dataset$X1 + dataset$X2 + dataset$X3 +
dataset$X4 + dataset$X5, data = dataset)

Residuals:

Min	1Q	Median	3Q	Max
-0.039423818	-0.006649344	0.000145798	0.006792210	0.037832217

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0080079795	0.0005993085	-13.36203	< 0.00000000000000222 ***
dataset$X1	0.8290802650	0.0007373641	1124.38377	< 0.00000000000000222 ***
dataset$X2	0.0046610054	0.0006656492	7.00220	0.0000000000028536 ***
dataset$X3	0.0017137856	0.0014403641	1.18983	0.2341705
dataset$X4	0.1772739050	0.0006454482	274.65243	< 0.00000000000000222 ***
dataset$X5	0.0055681234	0.0020688335	2.69143	0.0071383 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01000708 on 4994 degrees of freedom
Multiple R-squared: 0.9972313, Adjusted R-squared: 0.9972286
F-statistic: 359754.2 on 5 and 4994 DF, p-value: < 0.00000000000000022204

Conclusions:

The p-value ~ 0 which means it is good model, and also the R^2 value = 0.9972 which is near 1, i.e., if R^2 value is more then, the model is a good fit.

COMPARISON between Regression and ANN				
Layers	**MSE**	**SSE**	**RMSE**	**R^2**
ANN [3,1]	826.186243121	413093.121560	28.7434556572	0.997352703089
Regression	832.690263001	4163451.31500	28.8563730049	0.9972286

OBSERVATIONS:

1. MSE, SSE and RMSE of ANN are low as compared to the Regression model, hence, ANN model is accepted over Multivariate regression.
2. R^2 value for ANN is marginally better than the Regression model, so this factor doesn't play important role in selection of the model.
3. The MSE, SEE and RMSE errors should be less in order for a model to be good fit. Also, higher the R^2 value (i.e. ~ 1), the model is better fit. **Therefore, ANN model is preferable model for the given dataset.**

Hence, ANN model with 2 hidden layers and (3 neurons in 1st layer and 1 neuron in 2nd layer) is a preferred model for the given dataset and can be considered as better option over Multivariate regression model for the given dataset.
Although, there is minimal difference between the value of MSE, SSE, RMSE and R^2 value of Multiple regression and ANN. Hence, any of the model can be used but given the marginal improvement in performance, the ANN model is preferable over Multiple regression.