Name Mr. Key
ID# 7777777

ECO 5385
Predictive Analytics for Economists

Prof. Tom Fomby
Summer I, 2014

MID-TERM EXAM

Instructions: Fill in your name and student ID above. You have **1 and ½ hours** to complete this exam. This in-class exam is worth a total of 94 points. The points for the separate questions are broken out as follows:

Q1 = 15 (3,3,3,3,3)
Q2 = 4
Q3 = 4
Q4 = 4
Q5 = 9 (3,3,3)
Q6 = 10 (4,3,3)
Q7 = 4
Q8 = 3
Q9 = 3
Q10 = 2
Q11 = 4
Q12 = 4
Q13 = 3
Q14 = 3
Q15 = 3
Q16 = 3
Q17 = 3
Q18 = 3
Q19 = 10 (8,2)

1. Briefly define these terms:

a. **Supervised Learning -** Data mining tasks that involve either prediction or classification. (3)

b. **Unsupervised Learning -** Data mining tasks that do not involve either prediction or classification. Examples include anomoly detection, graphical representation of data, clustering, Association Rules. (3)

c. **Leo Brieman's Concept of "two cultures" among practicing statisticians -** The two cultures refer to the "hypothesis testing" orientation of some data scientists versus the "prediction" orientation of others. (3)

d. **Ordinal versus Nominal Variables -** There are types of classification variables. An ordinal classification variable has a numerical ordering of the categories like high, medium, and low where nominal classification variables do not have a numerical ordering. Red, Blue, Green Cars. (3)

e. **Tuning Parameter -** Many Data Mining techniques have parameters that determine the complexity or simplicity of the technique. For example, in MLR, SLSTAY and SLENTRY are the tuning parameters in SAS. (3)

2. Briefly define for me the **three** parts of a **partitioned data set** (as defined in XLMINER and SAS Enterprise Miner) and what their roles are in the data mining process.

3 Partitions: (1) Training Data set is used to "fit" a model to the Data (2) The validation Data set is used to scored the Trained Model to see if it has or has not been overtrained. (3) The Test data set is used is get an unconditional measure of the accuracy of a model (4)

3. In multiple linear regressions we can engage in finding a "best" regression by pursuing either an "exhaustive" search approach or a "directed" search approach. Tell me the methods that we have studied that fall under the first category and then the methods that we have studied which fall into the second category.

Exhaustive search involves inspection all possible subset regressions. Methods include C_p and Adjusted R^2. Directed search involves looking at only a fraction of all possible subset regressions - Forward, Backward, and Stepwise MLR. (4)

4. In **Directed Search** in Multiple Linear Regression via the SAS procedure REG we have the parameters called SLSTAY and SLENTRY. Suppose SLSTAY = 0.05 and SLENTRY = 0.10. What are the meanings of these values? If we were to increase both of these values by 0.10, what would this imply in terms of models we would select as compared to the models we would have selected with the first choices of SLSTAY and SLENTRY?

In Forward search SLENTRY is the p-value that allows a variable to enter the regression while SLSTAY is the p-value that keeps a variable already in the regression. If we increase both of these values by 0.10 we will produce more complex models. (4)

constant, CRIM, NOX, RM, AGE, PTRATIO, LSTAT.

5. See **Computer Output # 1** that I have given to you. It involves the Boston Housing data with MEDV as the target variable. Use it to answer the below parts of this question.

a) In class I talked about two criteria that you can use to select between these models. What are these two criteria? Which input variables would you select for your first criteria? Explain your answer. Which input variables would you select for your second criteria? Explain your answer.

(3) The C_p Mallows criterion and the Adjusted R^2 criterion. One could use either the minimum C_p value with the resulting model having the variables constant, CRIM, RM, AGE, PTRATIO, and LSTAT in it. Or one uses the maximum Adjusted R^2 the resulting model would have

b) Suppose you wanted to talk about the **unconditional** predictive accuracy of the model with 7 coefficients. Which measure would you report? I am looking for a numerical value. Explain your answer.

(3) the Test Data set RMSE = 4.766. The test data set provides the unconditional measure of RMSE because the test data set has not been used to train models or choose the best ones among them.

c) Suppose someone told you that they had a better model than your 7 coefficient model because their Training RMSE = 5.40. What would you say?

(3) The Training set RMSE might be for an overtrained model and when it is applied to an independent data set like the Test data set it could wind up performing worse than the T-coefficient one.

6. Consider the **Computer Output # 2**. It again involves the Boston Housing data set with MEDV being the target variable. Use this Computer Output to answer the following parts of this question.

a) Write out the functional form of the ANN that is implied by this output.

(4)
$$z_1 = -0.21 ZN - 0.90 CHAS - 0.46 AGE + 0.48 DIS + 0.13 TAX + 3.63 LSTAT - 0.40$$
$$z_2 = -0.43 ZN + 0.50 CHAS - 0.198 AGE + 1.13 DIS + 0.48 TAX + 2.06 LSTAT - 1.08$$
$$z_3 = 0.61 ZN - 0.90 CHAS - 0.06 AGE - 0.96 DIS + 0.27 TAX - 2.15 LSTAT + 0.12$$
$$h_1 = \frac{1}{1+e^{-z_1}}, \quad h_2 = \frac{1}{1+e^{-z_2}}, \quad h_3 = \frac{1}{1+e^{-z_3}}, \quad \Phi = 1.25 - 2.81 h_1 - 1.41 h_2 + 1.97 h_3$$

b) What is the **architecture of this model**? Use the notation I used in class. How many **hidden layers** are there in this model? What is the **activation function** for the hidden nodes of this model? What is the **activation function** for the output layer of this model?

Architecture = 6 - 3 - 1

(3) There is one hidden layer with 3 hidden nodes within that layer. The activation functions for the hidden nodes are logistic in nature. The activation function of the output layer is the linear (sometimes called identity) activation function.

models used the validation Data in refining themselves or not

c) Comparing this ANN and the 7-coefficient MLR reported in **Computer Output # 1**, which is the better model? Explain your answer.

(3) The 7-coefficient MLR has a Test Data RMSE = 4.766
The 6-3-1 ANN has a Test Data RMSE = 5.84. The 7-coefficient MLR is better since its Test RMSE is Less.
We use the Test Data set because we don't know whether either of these models used the Validation Data in refining themselves or not

7. Consider the **Computer Output # 3** generated by XLMINER on the Boston Housing data set with MEDV as the target variable. What are the architectures (using the M-N-R or M-N-R-S notation, whichever is appropriate) of the 4 ANN models presented in Computer Output # 3?

Choose Model 2 (4)

Model 1 = 5-3-2-1 . Val. RMSE = 9.25 Test RMSE = 10.70
Model 2 = 5-3-1 . Val. RMSE = 4.85 Test RMSE = 6.31
Model 3 = 5-2-1 . Val. RMSE = 4.88 Test RMSE = 6.36
Model 4 = 5-1-1 . Val. RMSE = 4.87 Test RMSE = 6.35

In this case we would usually use Validation RMSE as choice. Here Model 2 has smallest Validation RMSE.

8. Of the 4 ANN models of Question 7, which one do you prefer? Explain your answer.

(3) Model 2 = 5-3-1 has smallest validation RMSE.
Incidently, Model 2 also has the smallest Test RMSE as well.

9. Consider the following information on four vectors in the training data set:

Training Data Input	Training Data Output(y)
$(x_{11}, x_{12}, x_{13}) = x_1$	11
$(x_{21}, x_{22}, x_{23}) = x_2$	12
$(x_{31}, x_{32}, x_{33}) = x_3$	18
$(x_{41}, x_{42}, x_{43}) = x_4$	10

Further suppose that we are interested in scoring the vector x_0 in the validation data set. Consider the following information. $d(x_0, x_1) = 11$, $d(x_0, x_2) = 8$, $d(x_0, x_3) = 9$ and $d(x_0, x_4) = 7$. Then the K=2 Nearest Neighbor $\hat{y}$ value associated with x_0 is (show the math you used to get your answer)

(3) $\hat{y}_0 =$ 11 .

$\hat{y}_0 = \frac{12 + 10}{2} = 11$

(2) 10. (True / False). The K-nearest neighbor method is a non-parametric method.

11. Consider **Computer Output #4**. This is the Boston Housing data with MEDV as the target variable and an unnamed set of input variables. What is the optimal neighborhood size for this K-NN model? Explain your answer.

(4) neighborhood size = 14 produces the smallest validation RMSE. It is the "optimal" neighborhood size.
11
4.448.

The 7-coefficient MLR performed the best of the reported models because it had the smallest Test RMSE = 4.76.

12. **Of all of the models considered in Computer Outputs # 1, # 2, # 3, and # 4,** which one do you prefer? (All of these models used the same data partition of the Boston Housing data.) Explain your answer.

(4) As we may have some of the models using the Validation Data set to refine themselves we focus on the Test Data set RMSEs.
7-Coeff. MLR = Test RMSE = 4.76, G-3-1 ANN = Test RMSE = 5.84,
5-3-1 ANN = Test RMSE = 6.39, K = 14 KNN Test RMSE = 5.33.

13. Briefly describe to me why you would ever want to **stratify** your data before proceeding to build a predictive analytics model for a given target variable. By the way what do we mean when we say we have "stratified" our data?

(3) Data are often stratified before modelling proceeds because it may be strongly felt by the analyst that the prediction or classification models might be quite different across the strata and thus prediction strata by strata might be more accurate

14. Briefly explain to me what we mean by **"binning"** a continuous input variable and why we might want to do it.

(3) The binning of highly collinear data may lead to variables that lead to more accurate predictions than using models that have the collinear continuous variables.

15. What is an **ensemble model**? What is the rationale for building an ensemble model? What is meant by the term "trimmed ensemble?" How do you go about building a **"trimmed"** ensemble?

(3) An ensemble model is a combination of prediction or classification models. Ensembles often produce more accurate predictions than the individual methods that make up the ensemble. A trimmed ensemble is a parsimonious ensemble that only uses a small number of the best models to form an ensemble.

16. Briefly describe to me how you can use the technique of **"Bagging"** to build a more accurate prediction model using Multiple Linear Regression.

(3) Bagging = Bootstrap aggregation. The amounts to building B number of MLR based on B bootstrap samples of the Training Data set and then averaging the predictions of the B MLRs to get more accurate predictions.

17. Consider **Computer Output # 5.** Using this SAS output, I want you to write out the formula of the **Granger-Ramanathan combination** (ensemble) forecast that you would use. For simplicity let us call the dependent variable Y. The two competing forecasting methods of Y are ARIMA (a Box-Jenkins model) and ECON (an econometric model). Write your combination forecasting formula in the below space.

(3) $\hat{y}_{GR} = 1.55917 + 0.42\,ECON + 0.51104\,Arima$

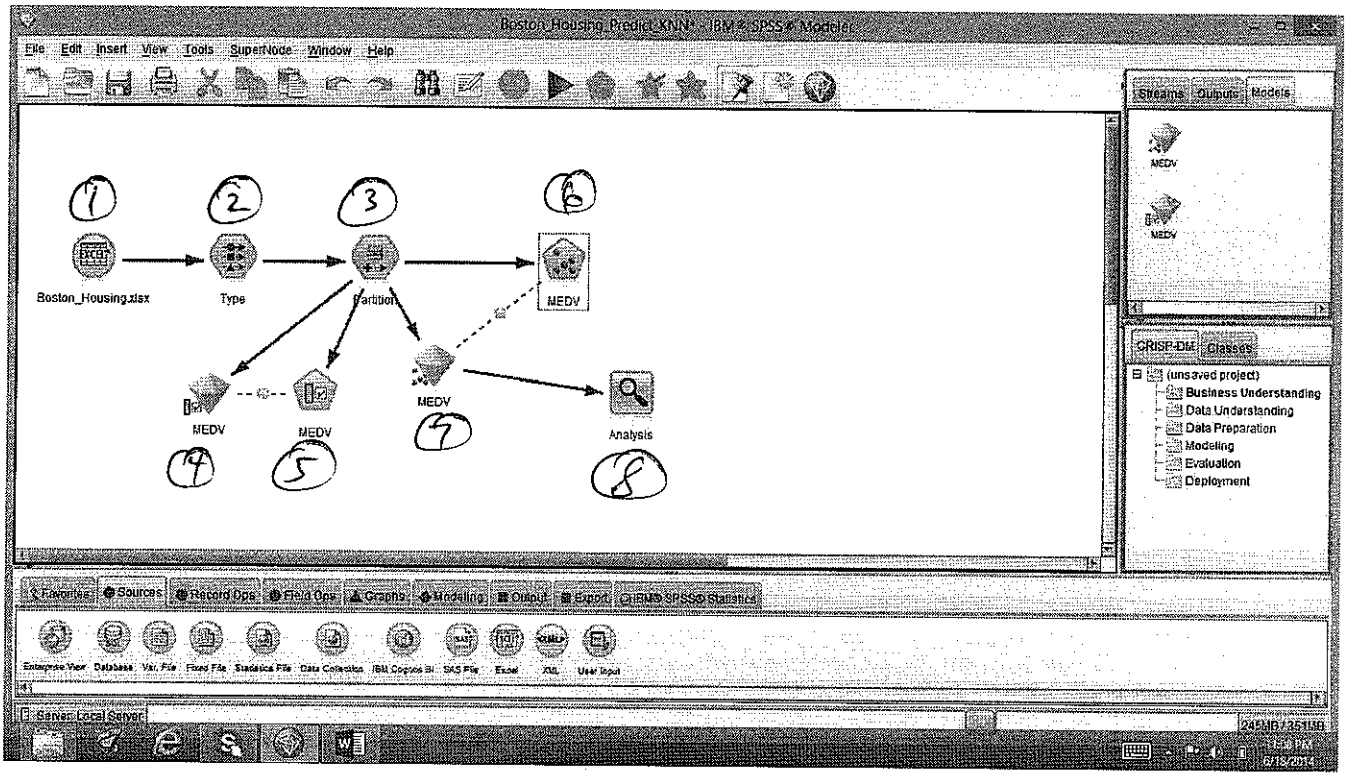
18. Consider **Computer Output # 5** again. On the last page of Computer Output # 5 you have the MSE and MAE performances of the Econometric Model, the ARIMA Model, the Average Ensemble, the Granger-Ramanathan Ensemble and the so-called Nelson Ensemble. Did the ensemble methods perform better than the separate forecasting methods (ARIMA and Econometric) in the hold-out data? Which ensemble method performed best? Explain your answer.

(3) GR performed best overall. In Terms of MAE, all ensemble methods improved on the best of the individual methods, namely, the ARIMA model. Among the combination (ensemble) methods, the G-R performed best.

In terms of MSE, the Nelson and G-R ensembles performed better than the best individual (method (ARIMA). Among the combination methods the G-R performed the best.

19. Consider the below two screen shoots, the first being an SPSS Modeler stream that is analyzing the Boston Housing data and the second being an output produced by the stream. MEDV is the target variable. Explain to me what each node in this stream does. Number the nodes as you explain the purpose of each node so that I will know clearly which node you are talking about. Also, given the partitioned output, how would you judge the worthiness of your model as compared to other competing models?

(10)



(1) = Source Node. Data is in Excel spreadsheet.

(2) = Type Node. Here the input variables to be used and the target variable to be used is declared.

(3) = Partition Node. Partitions the data (either two-way or three-way partition). Since a k-NN model is being fit and validated we would use a 3-way partition.

(4) = The Nugget produced by the

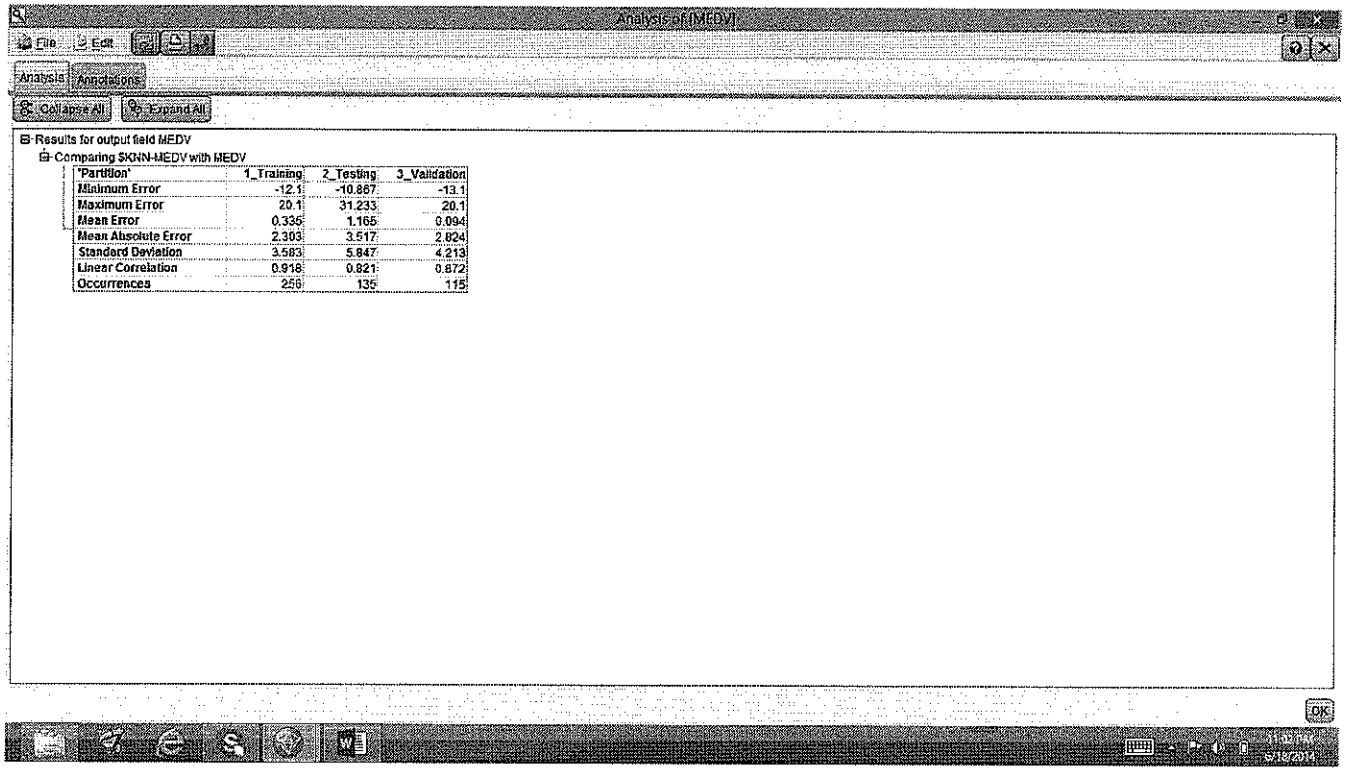
(5) = Feature selection Node. (The nugget produces a list of important variables.)

(6) = k-NN Node. Builds a k-NN model for the data.

(7) = Nugget providing details of k-NN model, in particular, it reports the optimal neighborhood size for model.

(8) = Provides accuracy measures for the optimal k-NN model.

(2)



'Partition'	1_Training	2_Testing	3_Validation
Minimum Error	-12.1	-10.887	-13.1
Maximum Error	20.1	31.233	20.1
Mean Error	0.335	1.165	0.094
Mean Absolute Error	2.303	3.517	2.824
Standard Deviation	3.583	5.847	4.213
Linear Correlation	0.918	0.821	0.872
Occurrences	256	136	115

The worthiness of this K-NN model would be judged either by the 3rd-partition MAE (the smaller the better) or the 3rd partition linear correlation between the predicted and actual values in the 3-partition data set (the larger the better).

Competing models using the same data partition then would be judge by how their validation (3rd partition) MAEs or linear correlation of predictions vs. actual values measure up against ~~[illegible]~~ each other.

Computer Output # 1

#Coeffs	RSS	Cp	R-Squared	Adj. R-Squared	Probability	Model (Constant present in all models)						
							2	3	4	5	6	7
2	10322.42285	82.36524963	0.521502	0.519595633	0	Constant	LSTAT	*	*	*	*	*
3	8828.72168	36.41519928	0.59074282	0.587468762	0.00000046	Constant	RM	LSTAT	*	*	*	*
4	7896.22168	8.48055172	0.633969046	0.629559034	0.06058468	Constant	RM	PTRATIO	LSTAT	*	*	*
5	7795.35498	7.24257994	0.638644742	0.632816432	0.12204693	Constant	CRIM	RM	PTRATIO	LSTAT	*	*
6	7708.384277	6.45068979	0.642676287	0.635443013	0.22957009	Constant	CRIM	RM	AGE	PTRATIO	LSTAT	*
7	7663.193359	6.9999938	0.644771121	0.636107002	1	Constant	CRIM	NOX	RM	AGE	PTRATIO	LSTAT

Based on MLR with 7 coefficients:

Training Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
7663.193427	5.503571757	-7.11304E-07

Validation Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
3610.840983	4.873964791	-0.718388769

Test Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
2294.313059	4.766127451	-0.34600535

Computer Output # 2

Inter-layer connections weights

	Input Layer						
Hidden Layer # 1	ZN	CHAS	AGE	DIS	TAX	LSTAT	Bias Node
Node # 1	-0.2107214	-0.90575301	-0.4693863	0.487191359	0.139588878	3.635431472	-0.4022626
Node # 2	-0.43972437	0.504404415	-0.19801267	1.133580956	0.483516225	2.063354092	-1.08550315
Node # 3	0.615081802	-0.40801715	-0.06615399	-0.96558255	0.277040508	-2.15617176	0.123364117

	Hidden Layer # 1			
Output Layer	Node # 1	Node # 2	Node # 3	Bias Node
Output Node	-2.81112214	-1.41381309	1.974137995	1.259780859

Training Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
7365.531622	5.395625377	0.728618893

Validation Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
3086.808481	4.506434348	-0.08118634

Test Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
3448.4205	5.843182115	0.659419416

Computer Output # 3

First ANN Model: 5-3-2-1

Inter-layer connections weights

	Input Layer					
Hidden Layer # 1	CRIM	AGE	DIS	TAX	LSTAT	Bias Node
Node # 1	-0.96840484	-0.70890647	-0.15780758	-0.35261682	0.307820449	0.790610074
Node # 2	-0.78478448	0.748203221	0.019932003	0.844505297	0.59568019	0.730463718
Node # 3	0.908437155	-0.84638557	-0.19768082	-0.40600408	0.586920465	-0.32106862

	Hidden Layer # 1			
Hidden Layer # 2	Node # 1	Node # 2	Node # 3	Bias Node
Node # 1	-0.45455434	0.167066284	-0.08505813	-0.02591335
Node # 2	0.134240383	-0.23112394	0.469692997	-1.34045127

	Hidden Layer # 2		
Output Layer	Node # 1	Node # 2	Bias Node
Output Node	-0.67099589	-0.13282	-0.54239522

Training Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
27387.82909	10.40443602	4.848790174

Validation Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
13020.20349	9.255227359	3.115052086

Test Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
11574.64451	10.70515954	4.890507752

Second ANN Model: 5-3-1

Inter-layer connections weights

	Input Layer					
Hidden Layer # 1	CRIM	AGE	DIS	TAX	LSTAT	Bias Node
Node # 1	0.269252665	-0.85338312	0.232368415	-0.3685157	3.665482858	-0.11789287
Node # 2	-0.24338609	0.002029157	0.066599946	0.458022531	2.014611514	-0.60542561
Node # 3	0.492046485	-0.91772408	-0.37346356	-0.79244388	-0.93476034	-0.07112736

	Hidden Layer # 1			
Output Layer	Node # 1	Node # 2	Node # 3	Bias Node
Output Node	-2.94341466	-1.4842496	1.283095366	1.751947583

Training Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
9045.194204	5.979277193	1.146430779

Validation Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
3580.102706	4.853174969	0.099421862

Test Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
4033.416493	6.319400032	1.035043109

Third ANN Model: 5-2-1

Inter-layer connections weights

	Input Layer					
Hidden Layer # 1	CRIM	AGE	DIS	TAX	LSTAT	Bias Node
Node # 1	0.211692578	-0.68494444	0.205332791	-0.30127173	3.616309386	-0.23756256
Node # 2	-0.13557757	0.045752642	0.169278125	0.456973282	2.405950652	-0.81759037

	Hidden Layer # 1		
Output Layer	Node # 1	Node # 2	Bias Node
Output Node	-2.80583189	-1.91042335	2.142467219

Training Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
9231.881053	6.040666224	1.297840569

Validation Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
3623.289452	4.882359123	0.229936197

Test Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
4091.630784	6.364840602	1.165241693

Fourth ANN Model: 5 – 1 – 1

Inter-layer connections weights

	Input Layer					
Hidden Layer # 1	CRIM	AGE	DIS	TAX	LSTAT	Bias Node
Node # 1	0.312346639	-0.55881054	0.270748286	-0.04850161	4.288585145	-0.620774

	Hidden Layer # 1	
Output Layer	Node # 1	Bias Node
Output Node	-3.64918092	1.59488442

Training Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
9183.948531	6.024964052	1.42300083

Validation Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
3618.357936	4.879035399	0.305485178

Test Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
4073.556848	6.350767357	1.268280099

Computer Output # 4

Validation error log for different k

Value of k	Training RMS Error	Validation RMS Error
1	0	6.345869979
2	0	5.314981085
3	0	5.071470964
4	0	4.818336403
5	0	4.653424611
6	0	4.588099094
7	0	4.522481009
8	0	4.527442111
9	0	4.475512144
10	0	4.464699631
11	0	4.488631927
12	0	4.475123046
13	0	4.46382566
14	0	4.448444877 ← $k_{opt.}$
15	0	4.464709377

Training Data scoring - Summary Report (for best k)

Total sum of squared errors	RMS Error	Average Error
0	0	0

Validation Data scoring - Summary Report (for best k)

Total sum of squared errors	RMS Error	Average Error
3029.911733	4.464709377	-0.74702222

Test Data scoring - Summary Report (for best k)

Total sum of squared errors	RMS Error	Average Error
2869.713119	5.330384712	-0.61692063

Computer Output # 5

SAS Code:

```
Proc reg data = work;
   model actual = econ arima;

 run;
```

SAS Output:

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: Actual

Number of Observations Read	16
Number of Observations Used	16

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	256.08926	128.04463	10.48	0.0019
Error	13	158.88608	12.22201		
Corrected Total	15	414.97534			

Root MSE	3.49600	**R-Square**	0.6171
Dependent Mean	44.01688	**Adj R-Sq**	0.5582
Coeff Var	7.94241		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.55917	9.47275	0.30	0.7652
Econ	1	0.42473	0.34868	0.73	0.4780
Arima	1	0.51104	0.32617	2.03	0.0637

The Mean Absolute Errors of the Competing Methods

Obs	maeecon	maearima	maenelson	maegr	maeave
1	5.78	4.49125	4.34595	4.31942	4.76937

The Mean Square Errors of the Competing Methods

Obs	mseecon	msearima	msenelson	msegr	mseave
1	40.2572	30.6437	29.8051	29.1645	31.0209