Name Mr. Key
ID# 7777777

ECO 6380
Predictive Analytics for Economists

Prof. Tom Fomby
Spring 2016

MID-TERM EXAM I

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 ~~80~~ 74 points. The points for the separate questions are broken out as follows:

Q1 = 8 (2, 2, 2, 2)
Q2 = 3
Q3 = 6
Q4 = 4
Q5 = 7 (2, 2, 3)
Q6 = 10 (2, 3, 2, 3)
Q7 = 8 (2, 2, 2, 2)
Q8 = 8 (3, 3, 2)
Q9 = 6 (2, 4)
Q10 = 4
Q11 = 10 (5, 3, 2)

1. Briefly define these terms:

a. **Supervised Learning -** Using statistical and machine learning methods to either predict an interval variable or classify a categorical variable. (2)

b. **Unsupervised Learning -** A data mining technique that does not involve prediction or classification e.g. cluster analysis, principal components analysis, (2)

c. **Ordinal versus Nominal Variables -** An ordinal variable is an ordered categorical variable where choice represents ranked choice (bad, adequate, good). A nominal variable is an unordered categorical variable where choice is not ranked. eg choice of transportation mode (train, car, bike, walk) (2)

d. **What is the major difference between a Predictive Analytics Course and a traditional statistics course?** A predictive analytics course is primarily concerned with accurate prediction whereas statistics courses are more interested in hypothesis testing. (2)

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 parts: training, validation, and test data sets. The training data set is used to build competing predictive models. The validation data set is used to choose between competing models. The test data set is used to obtain a measure of unconditional predictive performance (3)

Consider **Computer Output # 1.** Use it to answer questions 3 and 4.

3. In the IBM/SPSS Modeler stream that is depicted in Computer Output # 1, give me the names of the Nodes 1 – 6 and what functions they perform.

what we want to avoid most by data partitioning is the overtraining of models.

(1) Node 1 – Source node. Tells SPSS Modeler where the data is.

(1) Node 2 – Data partition node. Data is partitioned to prevent overtraining of models.

(1) Node 3 – Type Node - where user declares target variable and input variables to be used in the analysis.

(1) Node 4 – Feature selection node. Chooses important explanatory variables.

(1) Node 5 – Regression Node - Backward selection method is used

(1) Node 6 – Artificial Neural Network node. An ANN model is chosen and evaluated.

4. Of the two predictive models considered in the Modeler stream which model is the best? Explain your answer. Backw. Sel. reg. ANN

	Test Val.	Test Val.
MAE	3.657 3.409	2.354 2.504
Corr.	0.816 0.826	0.887 0.886

ANN is better. It has smaller MAEs and larger Corr. in the Test and Validation Data sets. (4)

5. Consider **Computer Output # 2.** Use it to answer the following:

There are two Modeler Streams represented here. What is the purpose of these two Modeler streams? That is, the top stream is to compare two competing prediction methods - A Chaid tree and a K-Nearest Neighbors model (2)

The bottom stream is form an ensemble model based on the Chaid and K-NN models. Use ½ pred. of Chaid plus ½ pred. of K-NN. (2)

From the output of these streams I conclude that Ensemble model is best. (2)

My reasoning for my conclusion is The Test and Val. MAEs and Corr. of Ensemble are better for the ensemble than for either of the individual models - chaid or K-NN. (1)

6. Consider **Computer Output # 3**. Use it to answer the following:

a) What type of model is being represented there? An Artificial Neural Network model (2)

b) Given that the model was derived by SPSS modeler how was it chosen? Thoroughly explain your answer? several competing ANNs were fitted on the Training data set and scored on the Test data set. The ANN with the smallest MAE and highest Corr. was chosen (3)

c) What is the "architecture" of this model? 10-5-1 (2)

d) In the predictor importance diagram, how were the input variables rank-ordered from most important to least important? And how are the lengths of the bars representing each variable's importance determined? Bivariate correlations between each input var. and the target var. were calculated ($c_1, c_2, \ldots, c_{10}$). Then the abs. value of these correlations were summed producing a number let us call Sum. Then the rel. imp. of each var. was calculated as Var i (importance) = $\frac{|c_i|}{Sum}$. (3)

7. Consider **Computer Output # 4.** This is taken from XLMINER. Use it to answer the following questions.

a) What kind of model is being estimated here?

K-Nearest Neighbors model (2)

b) What is the tuning parameter of the model?

The neighborhood size (K) (2)

c) What is the best value of the tuning parameter and how is it determined?

(2) K=4 is best choice. It minimizes the validation RMSE.

d) Is this a parametric model or a non-parametric model? Explain your answer.

(2) Non-parametric model. NO coefficients are estimated in this model.

8. Consider **Computer Output # 5**. Use it to answer the following questions.

a) This is a **CART** regression tree with the target variable being the interval variable "MEDV" of the Boston Housing data set. Thoroughly describe to me how the architecture of this tree is determined.

(3) This CART tree is built using the Binary Recursive method. It therefore was built in a unique order and pruned in the reverse unique order. Each split point of each decision node is chosen so as to guarantee the maximal additional reduction in SSE of the fit. Once the full tree is built it is pruned until the validation RMSE is minimized.

b) In contrast, we could have built a regression tree for MEDV using the **CHAID** method. How does this method differ from the CART method? Explain your answer thoroughly.

(3) The chaid method using in-sample tests of significance when adding a prospective decision node. When the chi-square significance of the next node slips below, say 0.05, the tree building process is stopped and a final tree is chosen.

c) Suppose you had a Boston city district with the following characteristics:

CRIM	ZN	INDUS	CHAS	NOX	RM
0.00632	18	2.31	0	0.538	6.575

(2) What would the predicted MEDV value be for this city district?

MEDV(predicted) = 24.60 (See marked-up tree)

9. Consider **Computer Output # 6**. This is a Neural Network model. Use it to answer the following questions.

a) The name of the method that was used to get the reported coefficient estimates is

(2) Back-propagation.

b) Write out the equations that represent the "architecture" of this model.

(4)

$$Z_1 = -0.417\,CRIM + 0.144\,ZN - 0.11\,INDUS + 0.127\,CHAS + 0.138\,NOX + 1.08\,RM - 1.18$$

$$Z_2 = 0.85\,CRIM + 0.049\,ZN + 0.08\,INDUS - 0.62\,CHAS + 0.35\,NOX - 0.80\,RM + 0.402$$

$$h_1 = \frac{\exp(Z_1)}{1+\exp(Z_1)}, \quad h_2 = \frac{\exp(Z_2)}{1+\exp(Z_2)}$$

$$\phi = 0.0109 + 1.51\,h_1 - 1.60\,h_2$$

10. Regression and Classification trees are often "stabilized" by using the method of **Bagging**. Explain to me what the method of Bagging is.

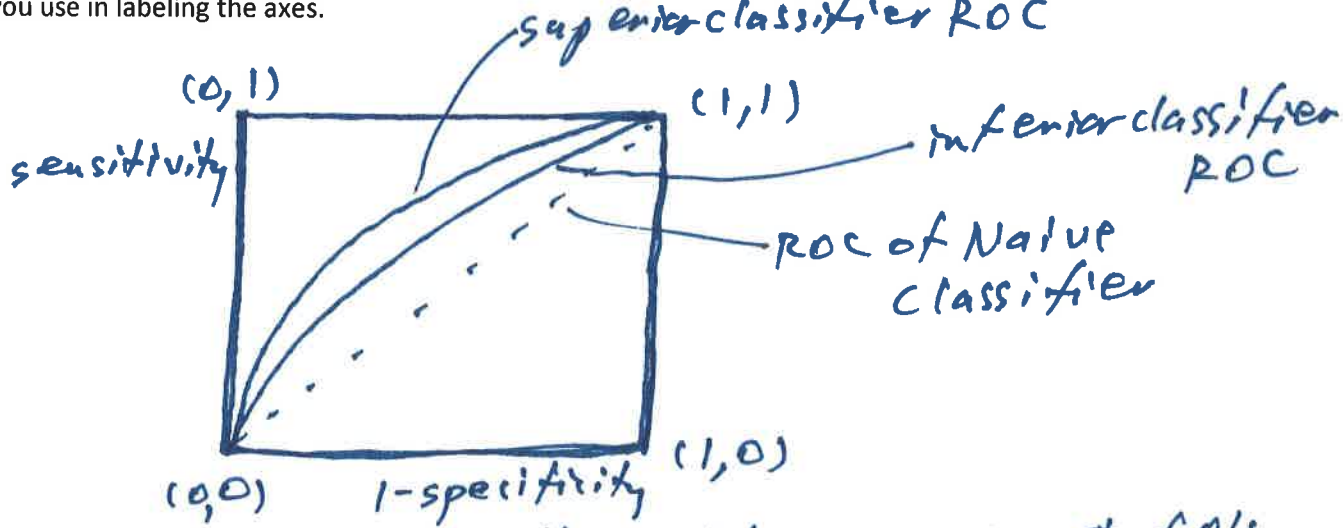
(4)

Several trees are built by k-fold estimation. Then when using these trees we put them into an ensemble and produce a prediction that is the average of the predictions produced by the trees

11. In classification problems we often examine competing classifiers using something called an **ROC curve**.

a) In the below space draw me three ROC curves in one graph. One curve should represent the ROC curve of a naïve classifier. One ROC curve should represent a classifier that is clearly superior to the ROC curve of an inferior competing classifier. Be sure and clearly label the x-axis and y-axis of your diagram. Define the terms that you use in labeling the axes.

(5)



$$\text{sensitivity} = \frac{\#\text{ of 1's correctly predicted}}{\#\text{ of 1's}} \qquad \text{specificity} = \frac{\#\text{ of 0's correctly pred.}}{\#\text{ of 0's}}$$

b) In words, describe to me how the computer generates these ROC curves.

(3)

Say, 99 classification tables are generated using the cutoffs of 0.01, 0.02, ..., 0.98, 0.99. The corresponding (1-specificity, sensitivity) points are plotted. This produces the ROC Curve.

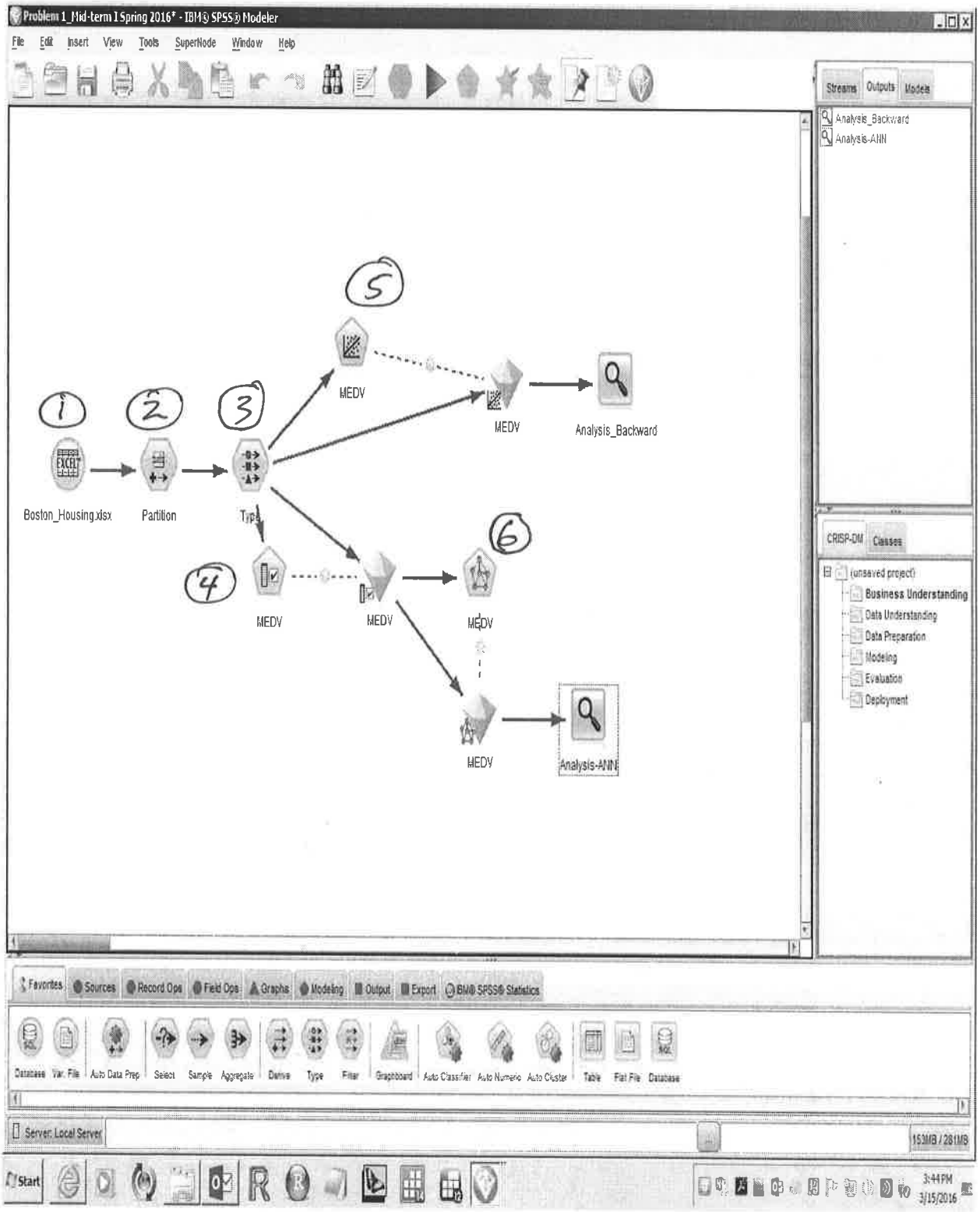
c) Describe to me **two** major uses of the ROC curve in classification analysis.

(2)

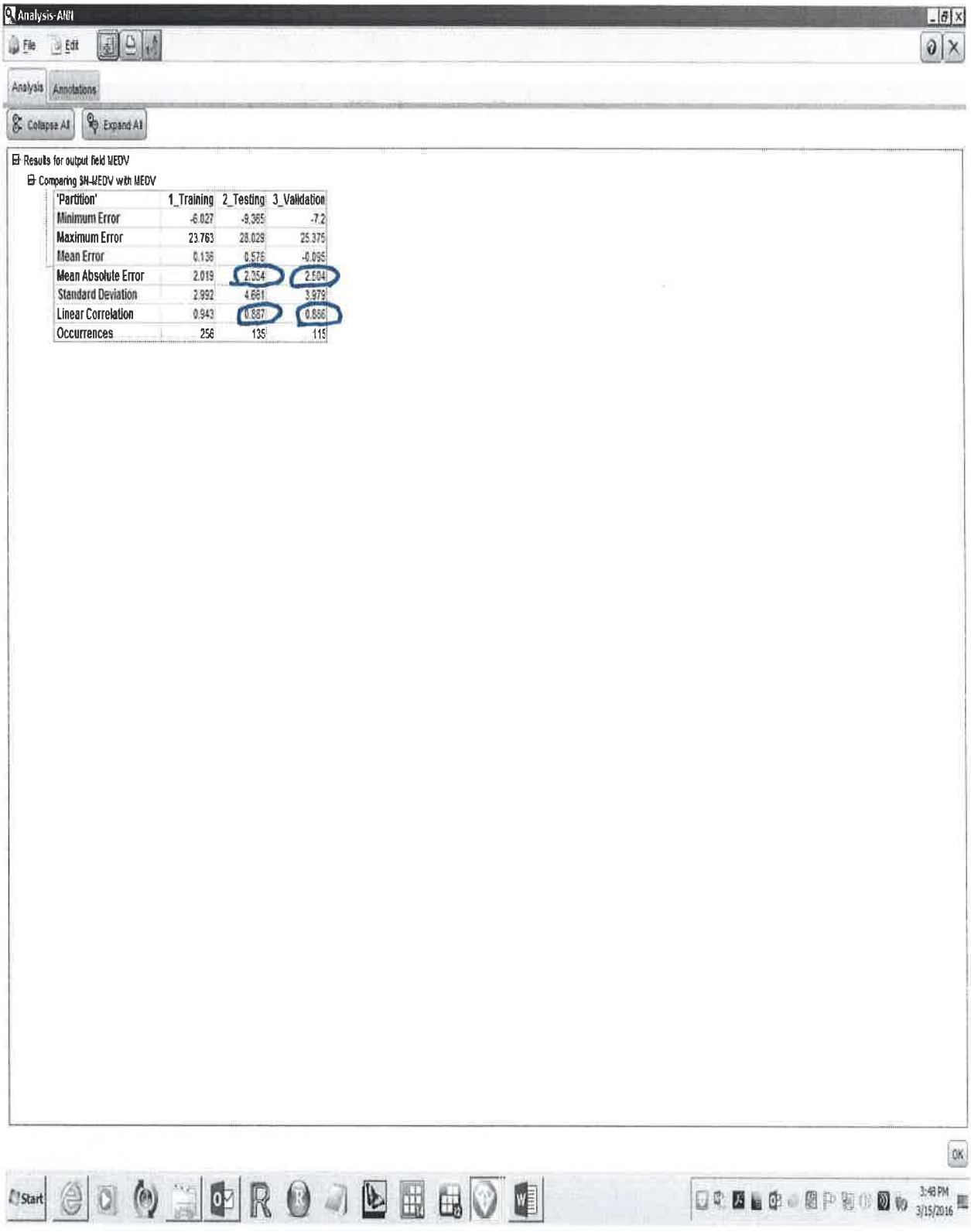
a) To choose the best classifier among a set of competing classifiers. The one with the largest area under the ROC is the best

b) The ROC can be used to choose an optimal cut-off probability. Max. AC_d.

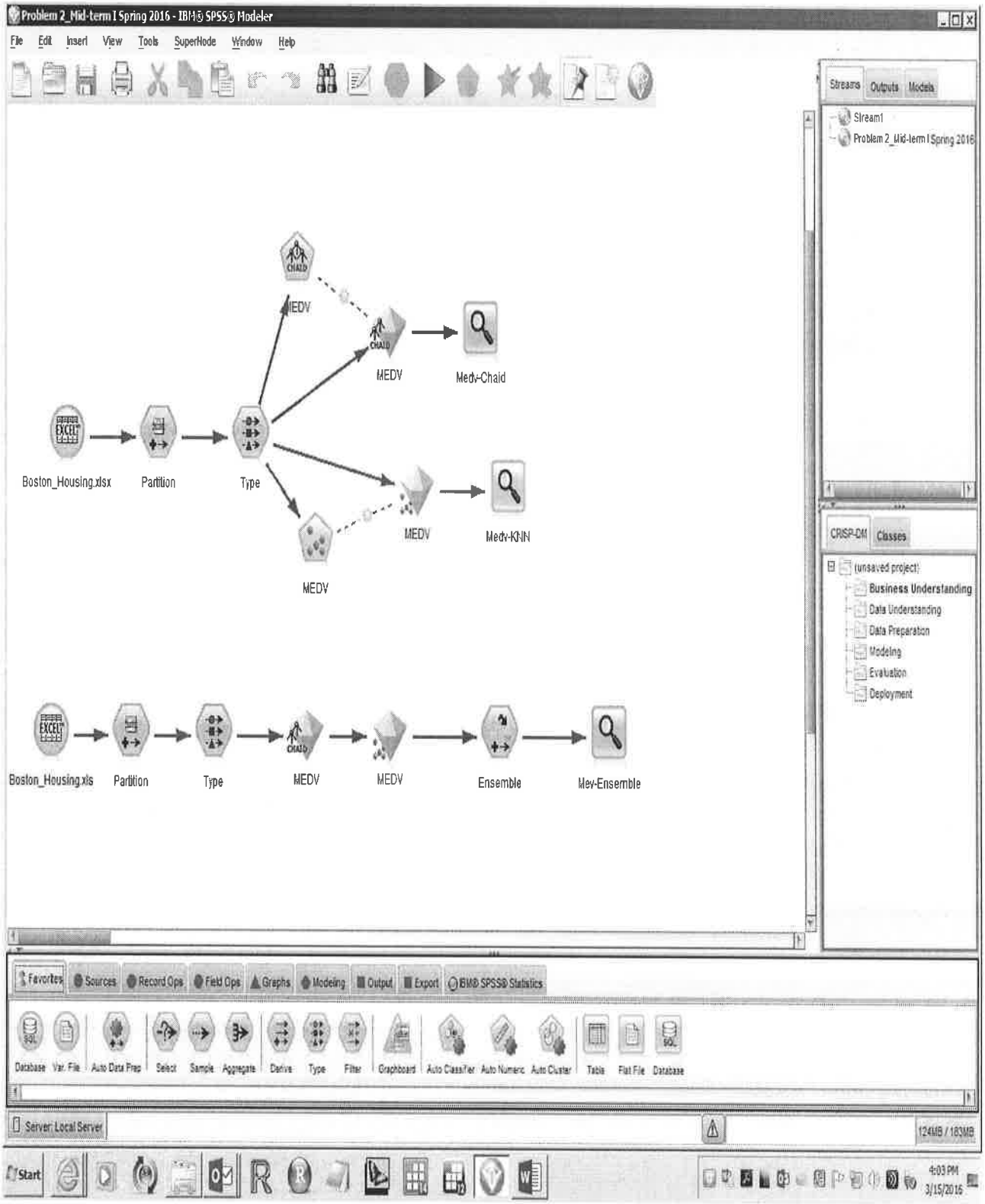
COMPUTER OUTPUT # 1




Analysis_Backward
File
Edit
Analysis
Annotations
Collapse All
Expand All
Results for output field MEDV
Comparing $E-MEDV with MEDV
'Partition' 1_Training 2_Testing 3_Validation
Minimum Error -14.312 -9.024 -10.664
Maximum Error 29.727 31.616 20.316
Mean Error 0.0 0.624 -0.152
Mean Absolute Error 2.983 3.857 3.409
Standard Deviation 4.267 5.832 4.935
Linear Correlation 0.879 0.818 0.826
Occurrences 256 135 115
OK
Start
3:47 PM
3/15/2016


Analysis-Alt
File
Edit
Analysis
Annotations
Collapse All
Expand All
Results for output field MEDV
Comparing $N-MEDV with MEDV
'Partition' 1_Training 2_Testing 3_Validation
Minimum Error -6.027 -8.365 -7.2
Maximum Error 23.763 28.029 25.375
Mean Error 0.136 0.576 -0.095
Mean Absolute Error 2.019 2.354 2.504
Standard Deviation 2.992 4.661 3.979
Linear Correlation 0.943 0.887 0.886
Occurrences 256 135 119
OK
Start
3:48 PM
3/15/2016

COMPUTER OUTPUT # 2



Medv-Chaid

File Edit

Analysis Annotations

Collapse All Expand All

Results for output field MEDV

Comparing $R-MEDV with MEDV

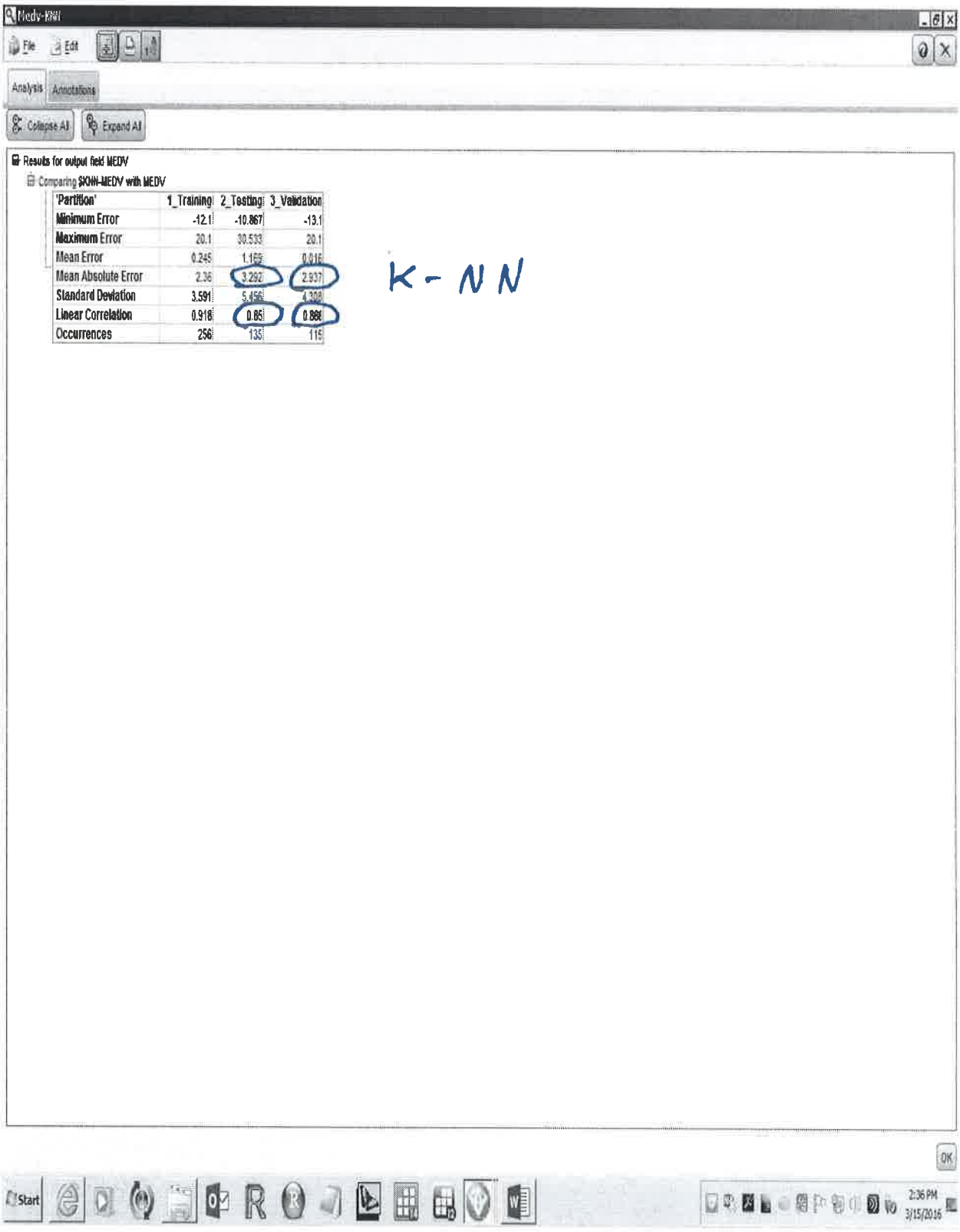
'Partition'	1_Training	2_Testing	3_Validation
Minimum Error	-9.467	-9.867	-16.6
Maximum Error	19.688	28.96	20.888
Mean Error	0.0	0.173	-0.476
Mean Absolute Error	2.458	3.455	3.31
Standard Deviation	3.478	5.458	4.877
Linear Correlation	0.922	0.843	0.83
Occurrences	256	135	115

chaid

OK

Start

2:33 PM
3/15/2016


Medv-KNN
File
Edit
Analysis
Annotations
Collapse All
Expand All
Results for output field MEDV
Comparing $KNN-MEDV with MEDV
'Partition' 1_Training 2_Testing 3_Validation
Minimum Error -12.1 -10.867 -13.1
Maximum Error 20.1 30.533 20.1
Mean Error 0.245 1.165 0.016
Mean Absolute Error 2.36 3.292 2.937
Standard Deviation 3.591 5.456 4.308
Linear Correlation 0.918 0.85 0.866
Occurrences 256 135 115
K-NN
OK
Start
2:36 PM
3/15/2016

Mev-Ensemble

File Edit

Analysis Annotations

Collapse All Expand All

Results for output field MEDV

Individual Models

Comparing $R-MEDV with MEDV

'Partition'	1_Training	2_Testing	3_Validation
Minimum Error	-9.467	-9.867	-16.6
Maximum Error	19.688	28.96	20.888
Mean Error	0.0	0.173	-0.476
Mean Absolute Error	2.458	3.455	3.31
Standard Deviation	3.478	5.458	4.877
Linear Correlation	0.922	0.843	0.82
Occurrences	256	135	115

summary for Chaid

Comparing $KNN-MEDV with MEDV

'Partition'	1_Training	2_Testing	3_Validation
Minimum Error	-12.1	-10.867	-13.1
Maximum Error	20.1	30.533	20.1
Mean Error	0.245	1.169	0.016
Mean Absolute Error	2.36	3.292	2.937
Standard Deviation	3.581	5.456	4.308
Linear Correlation	0.918	0.85	0.866
Occurrences	256	135	115

Summary for K-NN

Comparing $XR-MEDV with MEDV

'Partition'	1_Training	2_Testing	3_Validation
Minimum Error	-7.606	-9.803	-7.107
Maximum Error	18.667	29.747	18.107
Mean Error	0.122	0.671	-0.23
Mean Absolute Error	1.955	2.964	2.692
Standard Deviation	2.942	4.971	3.88
Linear Correlation	0.948	0.874	0.894
Occurrences	256	135	115

Summary for Ensemble

Agreement between $R-MEDV $KNN-MEDV $XR-MEDV

Comparing Agreement with MEDV

'Partition'	1_Training	2_Testing	3_Validation
Minimum Error	-7.606	-9.803	-7.107
Maximum Error	18.667	29.747	18.107
Mean Error	0.122	0.671	-0.23
Mean Absolute Error	1.955	2.964	2.692
Standard Deviation	2.942	4.971	3.88
Linear Correlation	0.948	0.874	0.894
Occurrences	256	135	115

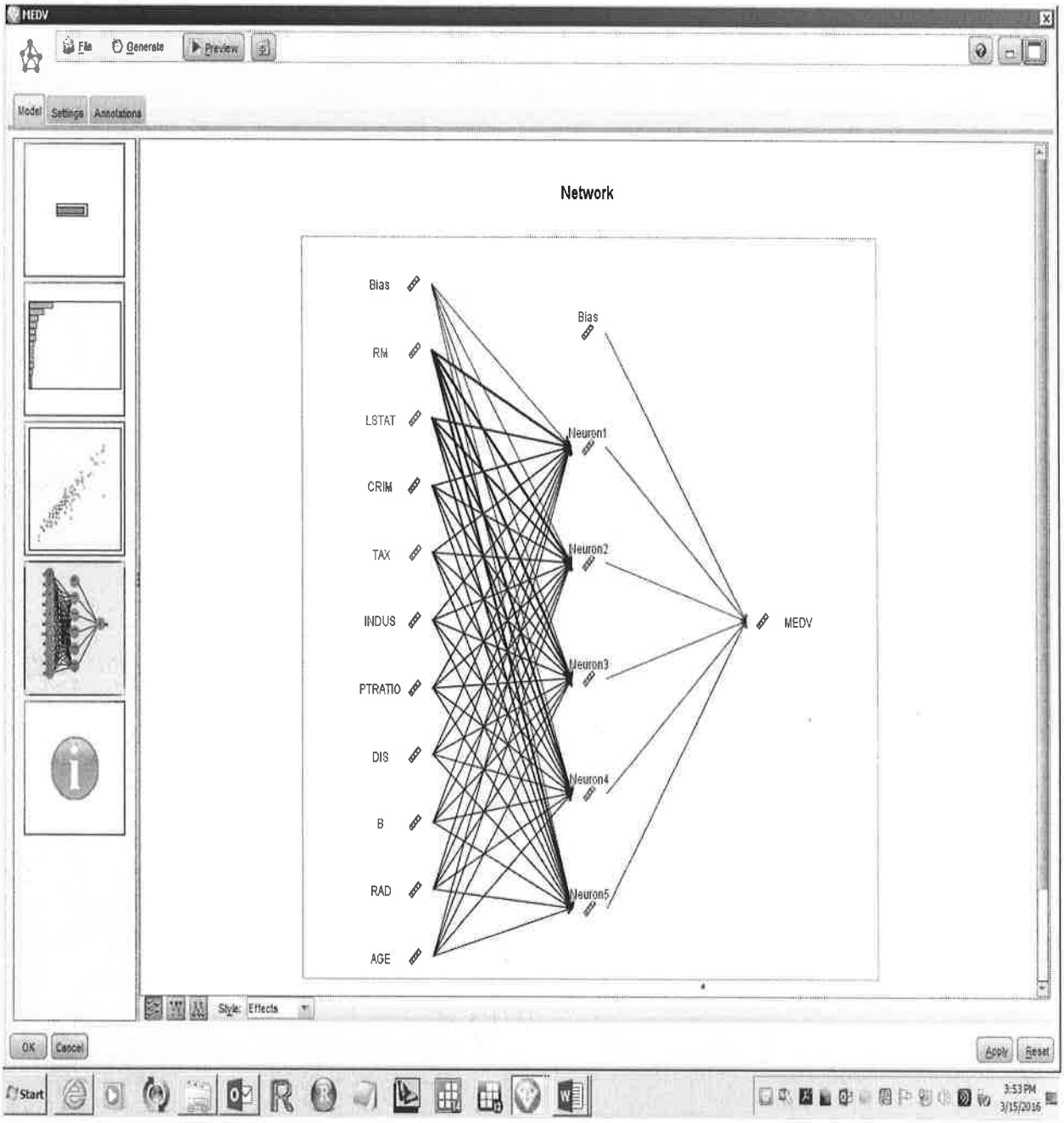
reports best model results
=> Ensemble.

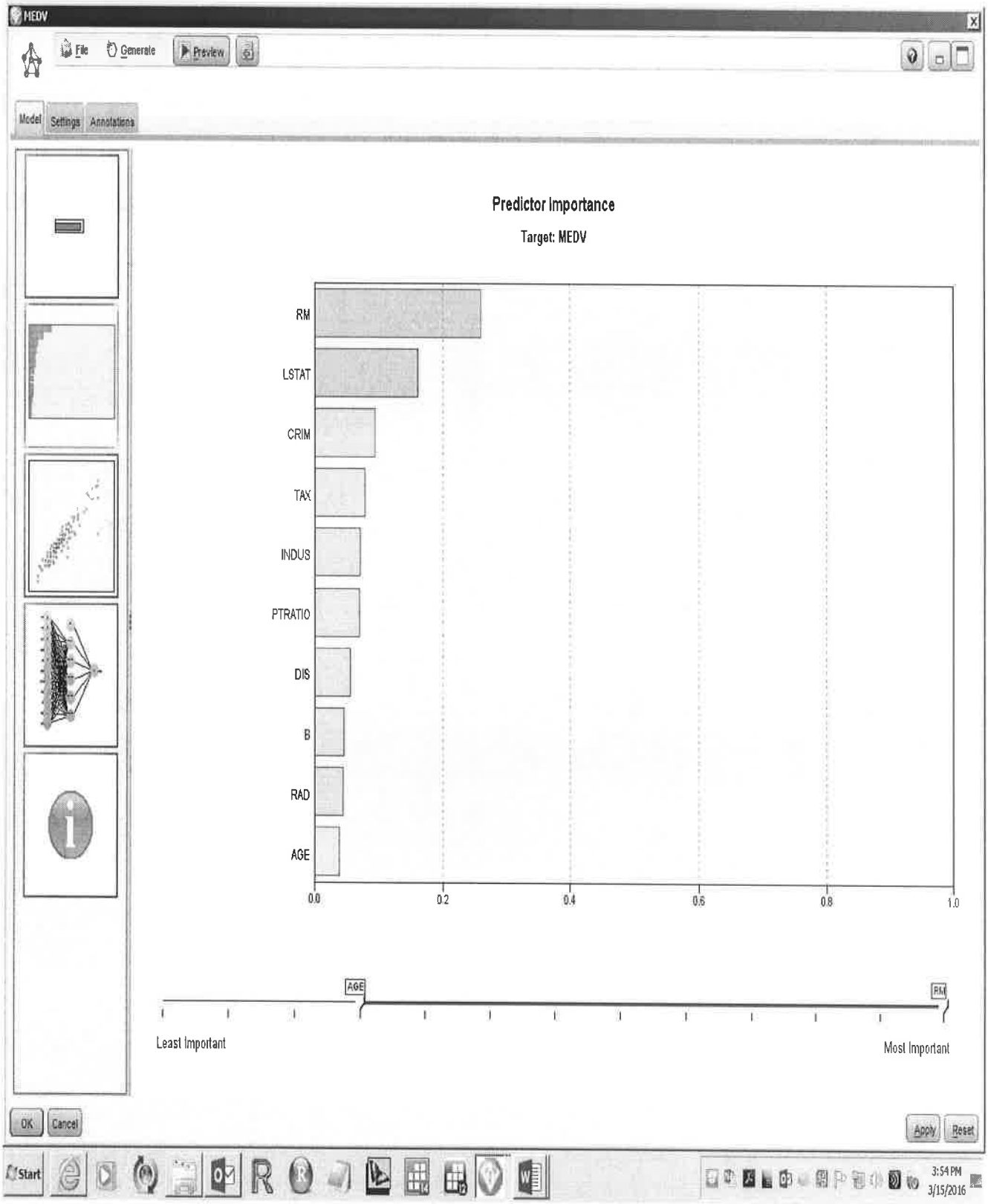
OK

Start

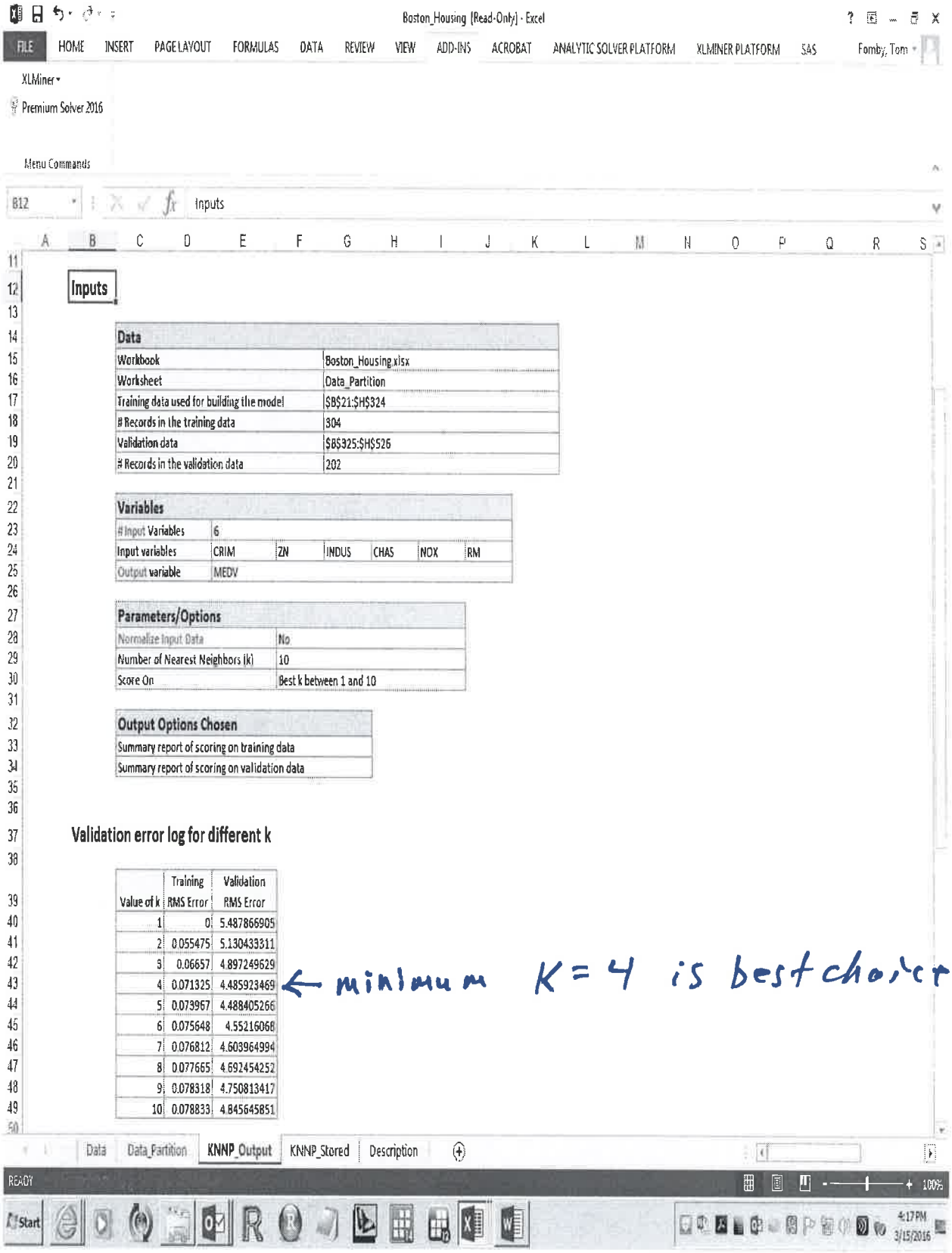
2:39 PM
3/15/2016

COMPUTER OUTPUT # 3




MEDV
File
Generate
Preview
Model
Settings
Annotations
Predictor Importance
Target: MEDV
RM
LSTAT
CRIM
TAX
INDUS
PTRATIO
DIS
B
RAD
AGE
0.0
0.2
0.4
0.6
0.8
1.0
AGE
RM
Least Important
Most Important
OK
Cancel
Apply
Reset
Start
3:54 PM
3/15/2016

COMPUTER OUTPUT # 4



Inputs

Data	
Workbook	Boston_Housing.xlsx
Worksheet	Data_Partition
Training data used for building the model	$B$21:$H$324
# Records in the training data	304
Validation data	$B$325:$H$526
# Records in the validation data	202

Variables						
# Input Variables	6					
Input variables	CRIM	ZN	INDUS	CHAS	NOX	RM
Output variable	MEDV					

Parameters/Options	
Normalize Input Data	No
Number of Nearest Neighbors (k)	10
Score On	Best k between 1 and 10

Output Options Chosen
Summary report of scoring on training data
Summary report of scoring on validation data

Validation error log for different k

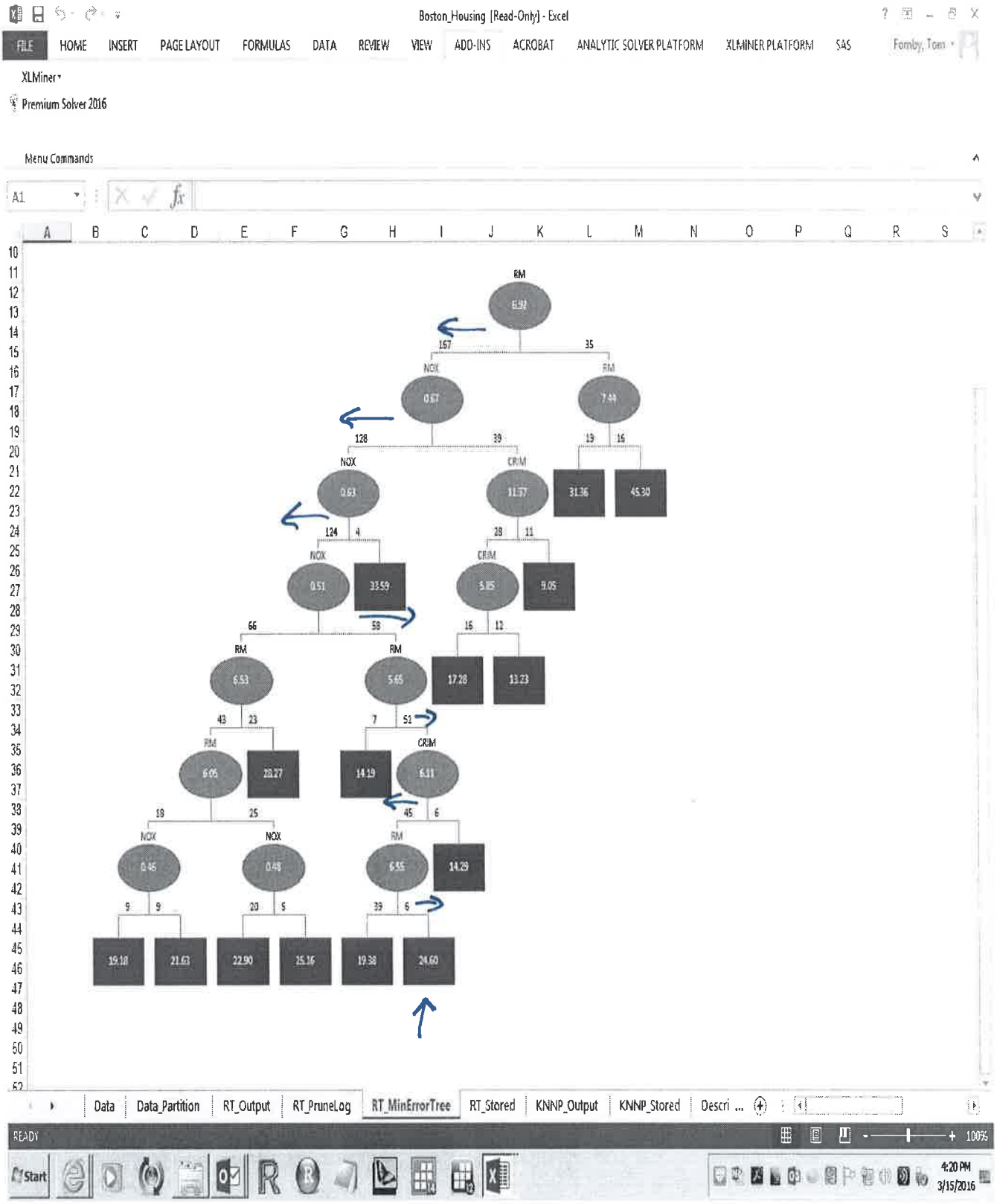
Value of k	Training RMS Error	Validation RMS Error
1	0	5.487866905
2	0.055475	5.130433311
3	0.06657	4.897249629
4	0.071325	4.485923469
5	0.073967	4.488405266
6	0.075648	4.55216068
7	0.076812	4.603964994
8	0.077665	4.692454252
9	0.078318	4.750813417
10	0.078833	4.845645851

← minimum $K = 4$ is best choice

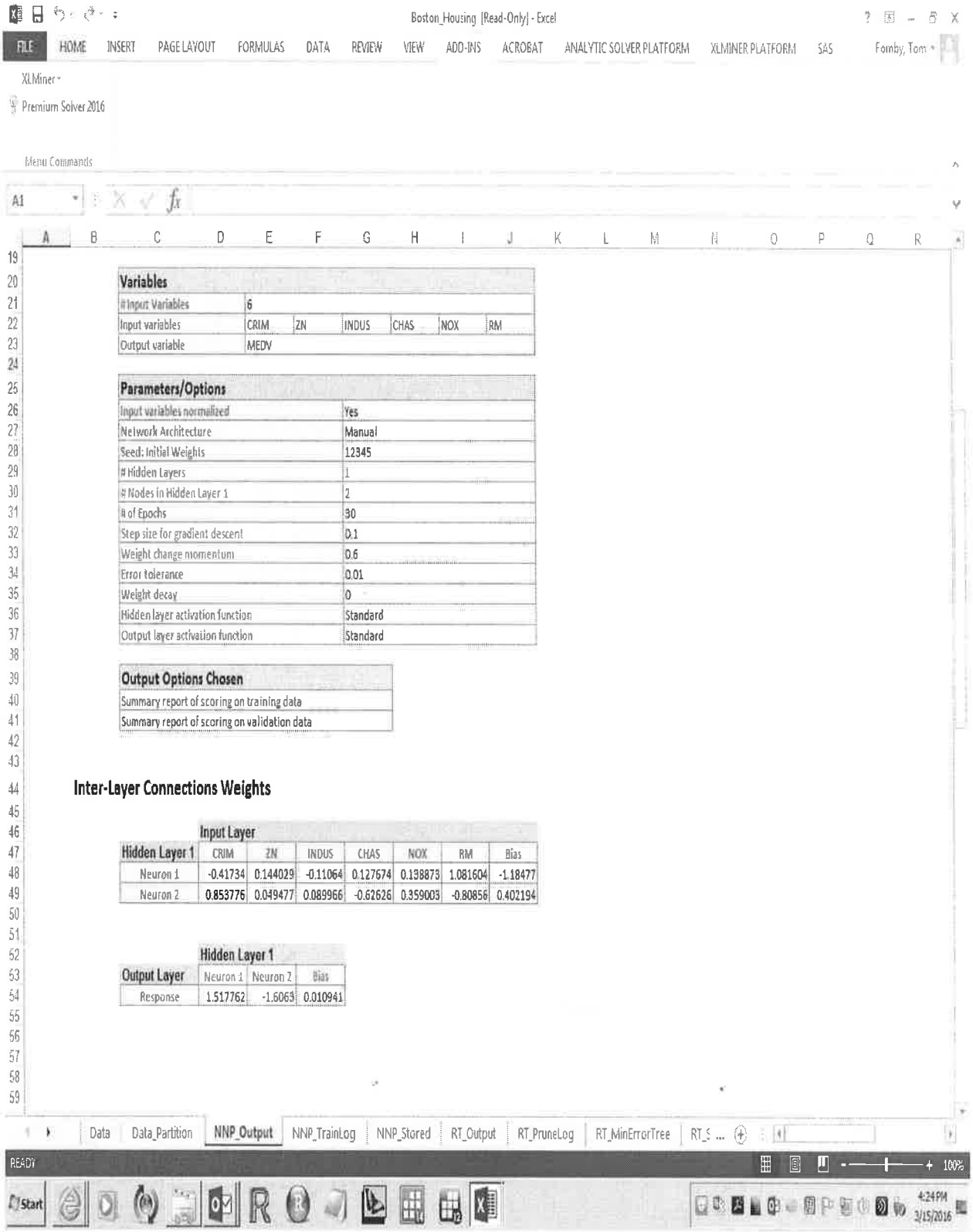
Data | Data_Partition | KNNP_Output | KNNP_Stored | Description

READY

COMPUTER OUTPUT # 5



COMPUTER OUTPUT # 6



Boston_Housing [Read-Only] - Excel

Variables						
# Input Variables	6					
Input variables	CRIM	ZN	INDUS	CHAS	NOX	RM
Output variable	MEDV					

Parameters/Options	
Input variables normalized	Yes
Network Architecture	Manual
Seed: Initial Weights	12345
# Hidden Layers	1
# Nodes in Hidden Layer 1	2
# of Epochs	30
Step size for gradient descent	0.1
Weight change momentum	0.6
Error tolerance	0.01
Weight decay	0
Hidden layer activation function	Standard
Output layer activation function	Standard

Output Options Chosen
Summary report of scoring on training data
Summary report of scoring on validation data

Inter-Layer Connections Weights

	Input Layer						
Hidden Layer 1	CRIM	ZN	INDUS	CHAS	NOX	RM	Bias
Neuron 1	-0.41734	0.144029	-0.11064	0.127674	0.138873	1.081604	-1.18477
Neuron 2	0.853776	0.049477	0.089966	-0.62626	0.359003	-0.80856	0.402194

	Hidden Layer 1		
Output Layer	Neuron 1	Neuron 2	Bias
Response	1.517762	-1.6063	0.010941

Data | Data_Partition | NNP_Output | NNP_TrainLog | NNP_Stored | RT_Output | RT_PruneLog | RT_MinErrorTree | RT_S ...

READY

4:24 PM 3/15/2016