TUTORIAL LL

Models for Predicting Rapid Intensification of Tropical Storms

Juan S. Lebrón and José F. Nieves

Department of Physics

University of Puerto Rico San Juan, PR 00931

Hurricanes are among the most complex macroscopic systems that are governed, in principle, by the Laws of Physics. One of the purposes of our project is to develop models for making statistical predictions about the intensification of hurricanes without having to deal with the individual physical parameters that characterize such systems. This tutorial gives an example of the kind of *Data Mining* analysis that can be carried out based on the ideas that we have been developing.

I. INTRODUCTION AND MOTIVATION

The main focus of a hurricane forecast is primarily to predict the trajectory and the intensity. While the method for predicting the trajectories have advanced in recent years together with the advances in computational resources, that has not been the case with respect to the intensity. Recent events of *unexpected* rapid intensification (e.g., Hurricane Wilma in 2005) have exacerbated the situation, and have contributed to the question of what could happen if a tropical cyclone goes through a stage of rapid intensification, unexpectedly, just before touching ground. Fortunately, the recent hurricanes that have suffered rapid intensification just before touching ground (e.g., Humberto and Lorenzo in 2007) have been relatively weak and therefore did not cause much damage. Nevertheless, it is clearly just a question of time before such an event occurs involving a stronger system, with potentially catastrophic consequences.

Most of the information about the position and intensity of hurricanes that appear in public information bulletins are estimated on the basis of satellite images of various kinds (e.g., visible and infrared images), such as that shown in Fig. 1. The technique for doing this was proposed by Vern Dvorak in 1975, and has being refined over the years. It is outside the scope of this tutorial to review the details of the technique and the methods involved in estimating such parameters from the images. Suffice it to say that, while Dvorak pretended that his technique be used for estimating not just the value of the intensity at a particular instance but also forecast it in the future, the fact is that his technique is indeed used for the former but no for the latter.

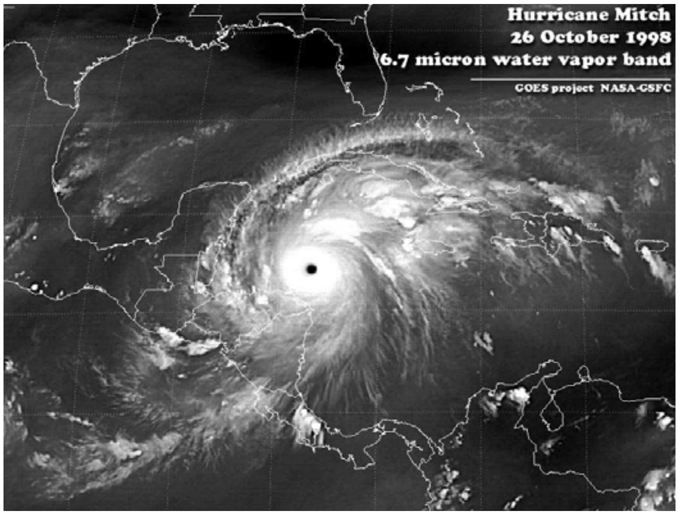


FIGURE 1: Satellite image (water vapor band) of Hurricane Mitch (16 October 1998)

The main purpose of our project is to produce a technique that is similar in spirit to the Dvorak technique for estimating the instantaneous intensity, but which is instead able to forecast the intensity in the future. In particular, the idea is to predict whether a hurricane will undergo rapid intensification during the next 24 hours. Although this can be expressed as a yes/no result, we also hope to generalize the model to give the expected increase in intensity for the next 24h period, regardless of whether it is rapid or not.

II. THE METHOD

The method is based on analyzing a satellite image, and determines from it the values of a set of relevant parameters. The salient feature of the technique is that, instead of attempting to determine the precise value of some physical variables (e.g., wind speed, cloud sizes or heights) from the image, the focus is on determining whether such variables may fall within one range of values or another. Thus, for example, we do not bother in trying to estimate the size or height of the clouds, but instead in determining whether or not there are Cirrus clouds at all and, in addition, whether they persist for a given distance from the center, or whether they cover three quadrants, or two.

There are 8 conditions in the model: the one defined as “inner core” and the other ones labeled as 1a, 1b, 2a, 2b, 3, 4a, and 4b. “Inner core” refers to a particular type of organization of the clouds in a hurricane which allows for faster intensification than otherwise. (This is usually equivalent to the hurricane having an “eye” or not). The other 7 conditions characterize the environmental conditions, in a categorical fashion as we have indicated above. It is outside the scope of this tutorial to explain how those values are determined and what they mean in detail. Here we accept that data as given, and the idea is to fit these 7 conditions to the actual observed changes in intensity in order to derive a useful predictive formula. In the dataset, intensity is expressed as a T value which in theory can range from 1 (weak) to 8 (strong hurricane). The model output is expected to be a ΔT such that today's T plus the model-obtained ΔT produces (predicts) tomorrows T.

Inner core	1a	1b	2a	2b	3	4a	4b	ΔT
0	1	0	0	0	0	1	1	0
0	1	1	1	0	0	1	0	0
0	1	1	1	0	0	1	1	0
0	1	0	0	0	0	0	0	0
1	1	1	0	0	0	1	1	1
1	1	1	1	0	0	1	1	1
0	1	1	0	0	0	1	0	0

TABLE 1: A sample of the data set used in the tutorial.

In the language of statistics, it is clear that we have done is to attempt to characterize such a complex system by a set of categorical variables that presumably carry the most relevant information about physical properties of the system, or at least those properties that are most influential during the periods of rapid intensification.

In this way, we thus end up with a data table that looks like the one shown in Table 1. In short, the problem at hand is to see if the value of ΔT shown in the last column of that table can be predicted on the basis of the values of the eight parameters labeled "inner core" through 4b.

III. MODEL BUILDING

For this tutorial we are using the *Data Mining* models of Statistica. The model building process starts by choosing the item labeled *Data Miner Recipes* from the *Data Mining* menu as shown in Fig. 2. In the next dialog appears asking to open an existing recipe of or a new one, we choose New, which brings us to the window shown in Fig. 3, and there we will click on *Open/Connect data file* to open our data file.

NOTE: Since the DMRecipe is used in many of the other tutorials in this HANDBOOK, we will run through the DMRecipe format rather rapidly in the following example:

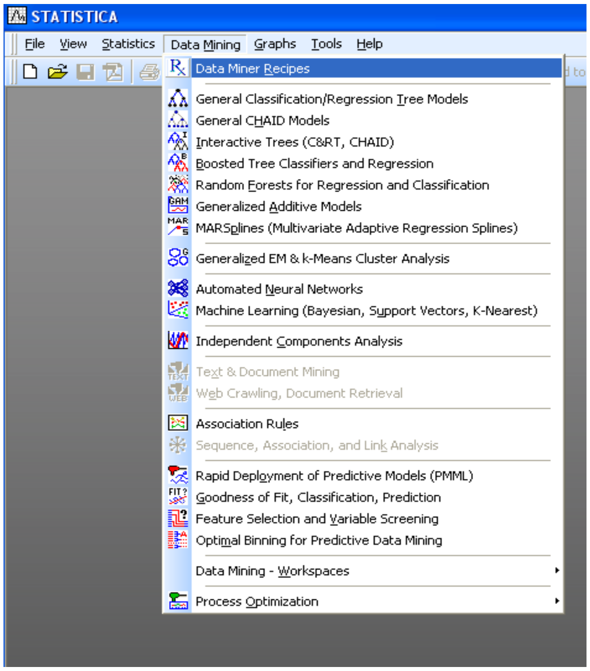
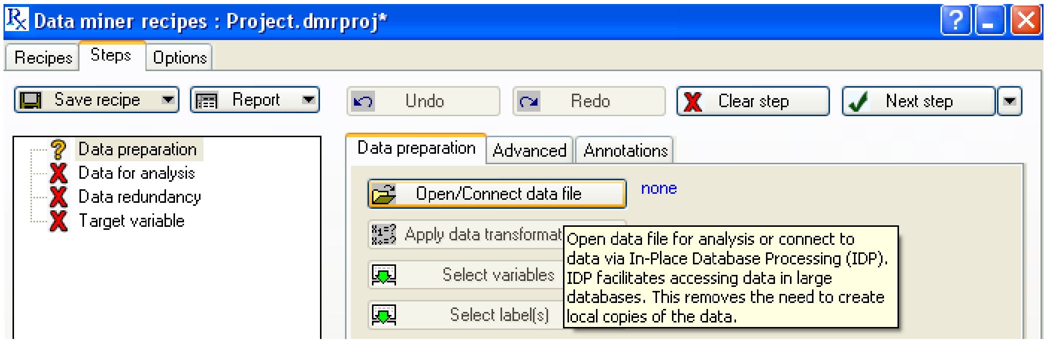

STATISTICA
File View Statistics Data Mining Graphs Tools Help
Data Miner Recipes
General Classification/Regression Tree Models
General CHAID Models
Interactive Trees (C&RT, CHAID)
Boosted Tree Classifiers and Regression
Random Forests for Regression and Classification
Generalized Additive Models
MARSplines (Multivariate Adaptive Regression Splines)
Generalized EM & k-Means Cluster Analysis
Automated Neural Networks
Machine Learning (Bayesian, Support Vectors, K-Nearest)
Independent Components Analysis
Text & Document Mining
Web Crawling, Document Retrieval
Association Rules
Sequence, Association, and Link Analysis
Rapid Deployment of Predictive Models (PMML)
Goodness of Fit, Classification, Prediction
Feature Selection and Variable Screening
Optimal Binning for Predictive Data Mining
Data Mining - Workspaces
Process Optimization

FIGURE 2. Data Mining Menu from STATISTICA Data Miner.

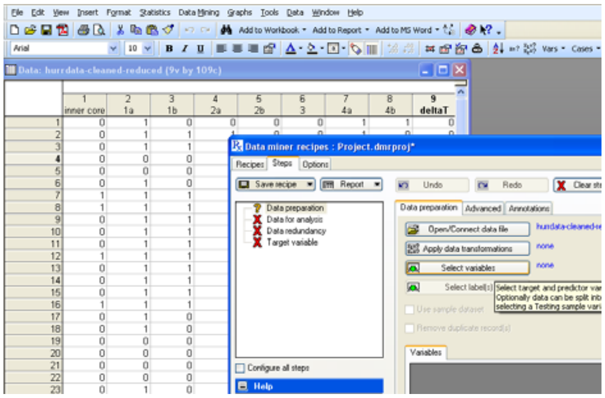


Open/Connect data file: none

Open data file for analysis or connect to data via In-Place Database Processing (IDP). IDP facilitates accessing data in large databases. This removes the need to create local copies of the data.

FIGURE 3. DATA MINING RECIPE initial dialog.

Our original data file was in prepared in a spreadsheet-like comma-separated format. Statistica recognizes the format and asks for confirmation, and after loading the file our workspace looks as shown in Fig. 4, where we will click on *Select variables*. This brings up the window shown in Fig. 5, where we have already selected columns 1-8 as the input categorical parameters, and column 9 as the target categorical variable.



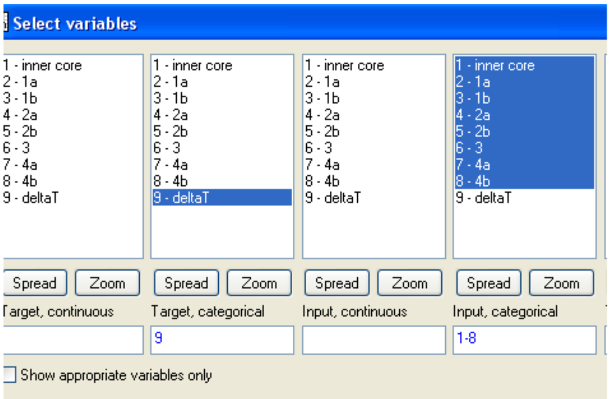
Open/Connect data file: hurrdata-cleaned-re

Apply data transformations: none

Select variables: none

Select target and predictor var
Optionally data can be split into
selecting a Testing sample vari

FIGURE 4 . DMRecipe dialog after loading data



Target, continuous:

Target, categorical: 9

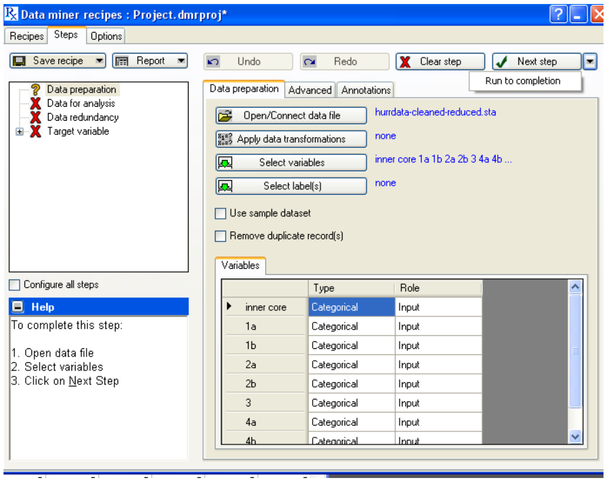
Input, continuous:

Input, categorical: 1-8

☐ Show appropriate variables only

FIGURE 5. Variables selected.

Our workspace then looks as shown in Fig. 6 where, instead of clicking on the item labeled *Next Step*, we will click on the arrow that is to it right side to *Run to completion*. That will instruct the Statistica software to go automatically through a series of steps that would have to be carried manually in order to run anyone of the supported models.



Open/Connect data file: hurrdata-cleaned-reduced.sta

Apply data transformations: none

Select variables: inner core 1a 1b 2a 2b 3 4a 4b ...

Select label(s): none

☐ Use sample dataset

☐ Remove duplicate record(s)

☐ Configure all steps

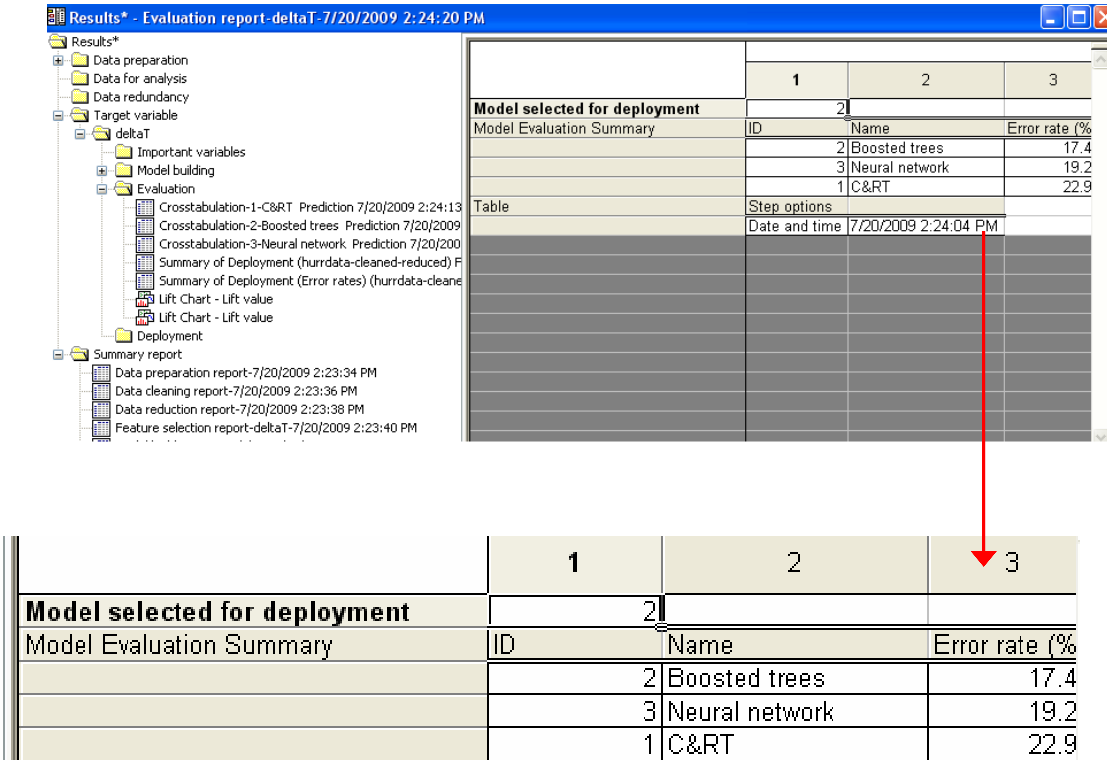
To complete this step:

1. Open data file
2. Select variables
3. Click on Next Step

	Type	Role
inner core	Categorical	Input
1a	Categorical	Input
1b	Categorical	Input
2a	Categorical	Input
2b	Categorical	Input
3	Categorical	Input
4a	Categorical	Input
4b	Categorical	Input

FIGURE 6. Click the RUN TO COMPLETION selection on the upper-right corner to run the complete DMRecipe set of computations.

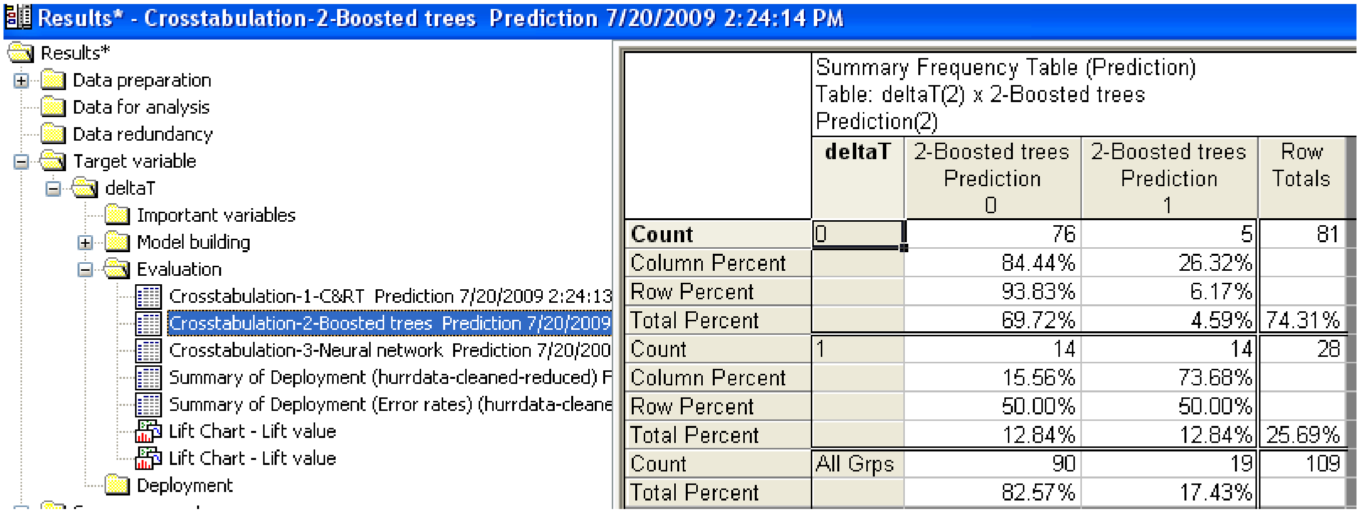
The end result of that process is shown in Fig. 7, which indicates that the best model is the Boosted trees model, and the software has automatically selected it as the model for deployment. The various menu items that are shown on the left hand side in that window, under *Evaluation*, can be chose to display various tables and graphs related to the evaluation of the models as shown, for example, in Figs. 8 and 9.


Results* - Evaluation report-deltaT-7/20/2009 2:24:20 PM
Results*
Data preparation
Data for analysis
Data redundancy
Target variable
deltaT
Important variables
Model building
Evaluation
Crosstabulation-1-C&RT Prediction 7/20/2009 2:24:13
Crosstabulation-2-Boosted trees Prediction 7/20/2009
Crosstabulation-3-Neural network Prediction 7/20/200
Summary of Deployment (hurrdata-cleaned-reduced) F
Summary of Deployment (Error rates) (hurrdata-cleane
Lift Chart - Lift value
Lift Chart - Lift value
Deployment
Summary report
Data preparation report-7/20/2009 2:23:34 PM
Data cleaning report-7/20/2009 2:23:36 PM
Data reduction report-7/20/2009 2:23:38 PM
Feature selection report-deltaT-7/20/2009 2:23:40 PM

	1	2	3
Model selected for deployment	2		
Model Evaluation Summary	ID	Name	Error rate (%
	2	Boosted trees	17.4
	3	Neural network	19.2
	1	C&RT	22.9
Table	Step options		
	Date and time	7/20/2009 2:24:04 PM	

FIGURE 7. Summary Results spreadsheet result of the DMRecipe computations.

Figure 7 results show that the Boosted Trees have the lowest Error Rate at 17.4%, meaning that the ACCURACY for predicting Hurricanes is 100-Error rate, or 82.6% using the BT modeling. If we decide to accept this BT Model, we can deploy it on new data coming in during the hurricane season in future years to get predictions at this 82.6% accuracy.


Results* - Crosstabulation-2-Boosted trees Prediction 7/20/2009 2:24:14 PM
Results*
Data preparation
Data for analysis
Data redundancy
Target variable
deltaT
Important variables
Model building
Evaluation
Crosstabulation-1-C&RT Prediction 7/20/2009 2:24:13
Crosstabulation-2-Boosted trees Prediction 7/20/2009
Crosstabulation-3-Neural network Prediction 7/20/200
Summary of Deployment (hurrdata-cleaned-reduced) F
Summary of Deployment (Error rates) (hurrdata-cleane
Lift Chart - Lift value
Lift Chart - Lift value
Deployment

Summary Frequency Table (Prediction)
Table: deltaT(2) x 2-Boosted trees Prediction(2)

	deltaT	2-Boosted trees Prediction 0	2-Boosted trees Prediction 1	Row Totals
Count	0	76	5	81
Column Percent		84.44%	26.32%	
Row Percent		93.83%	6.17%	
Total Percent		69.72%	4.59%	74.31%
Count	1	14	14	28
Column Percent		15.56%	73.68%	
Row Percent		50.00%	50.00%	
Total Percent		12.84%	12.84%	25.69%
Count	All Grps	90	19	109
Total Percent		82.57%	17.43%	

FIGURE 8. Crosstabulation Table for BOOSTED TREES spreadsheet result.

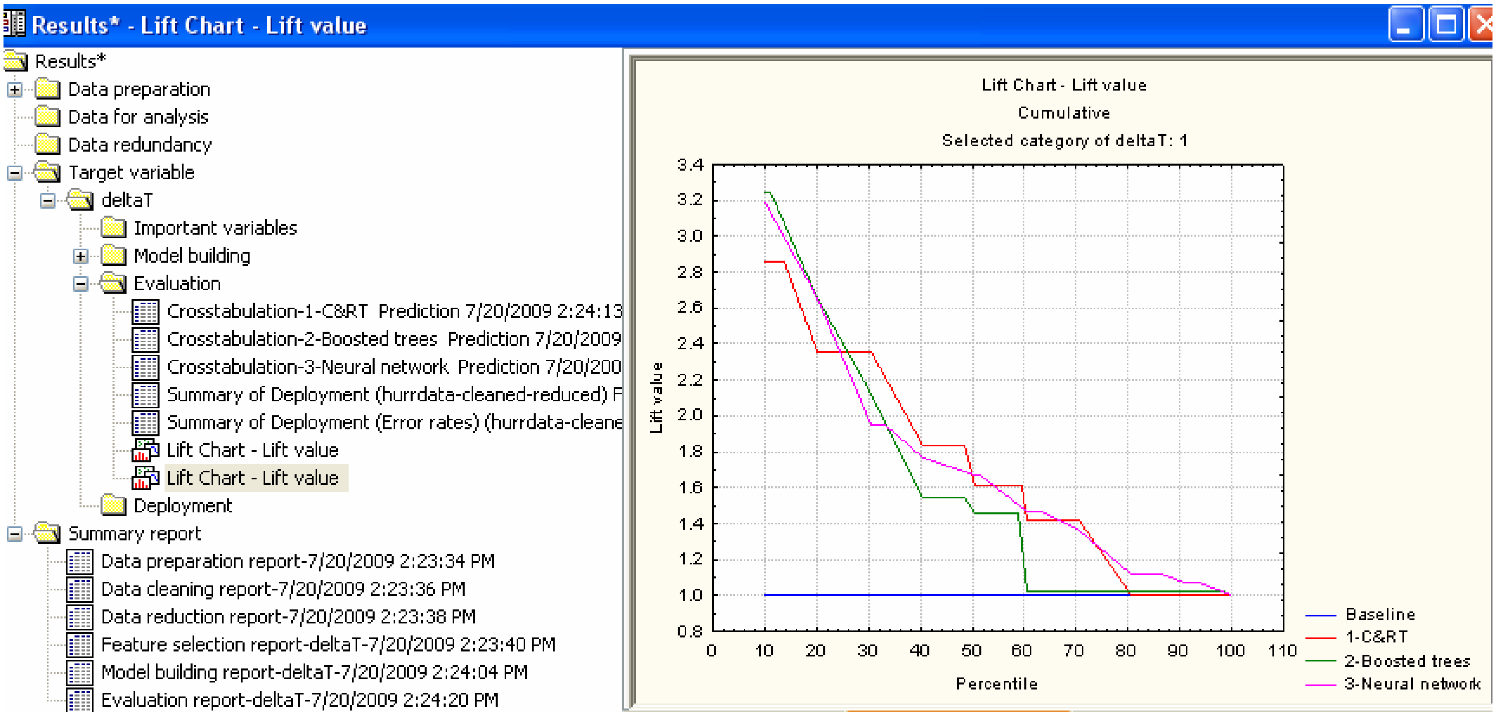
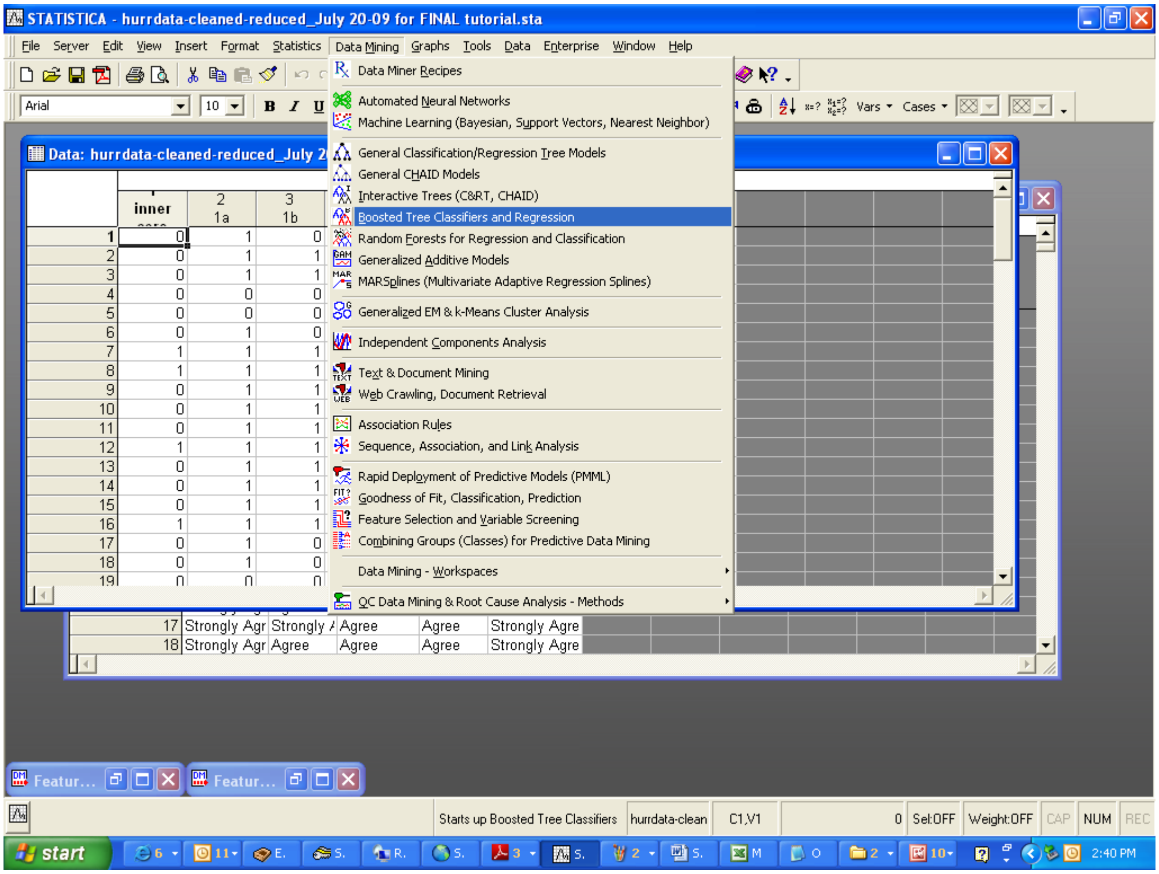

Results* - Lift Chart - Lift value
Results*
Data preparation
Data for analysis
Data redundancy
Target variable
deltaT
Important variables
Model building
Evaluation
Crosstabulation-1-C&RT Prediction 7/20/2009 2:24:13
Crosstabulation-2-Boosted trees Prediction 7/20/2009
Crosstabulation-3-Neural network Prediction 7/20/200
Summary of Deployment (hurrdata-cleaned-reduced) F
Summary of Deployment (Error rates) (hurrdata-cleane
Lift Chart - Lift value
Lift Chart - Lift value
Deployment
Summary report
Data preparation report-7/20/2009 2:23:34 PM
Data cleaning report-7/20/2009 2:23:36 PM
Data reduction report-7/20/2009 2:23:38 PM
Feature selection report-deltaT-7/20/2009 2:23:40 PM
Model building report-deltaT-7/20/2009 2:24:04 PM
Evaluation report-deltaT-7/20/2009 2:24:20 PM
Lift Chart - Lift value
Cumulative
Selected category of deltaT: 1
Lift value
3.4
3.2
3.0
2.8
2.6
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0
10
20
30
40
50
60
70
80
90
100
110
Percentile
Baseline
1-C&RT
2-Boosted trees
3-Neural network

FIGURE 9. Lift Chart Graph comparing the 3 data mining algorithms used in this modeling process. All are quite similar, however the Boosted Trees did have a slight higher value at the left side of the graph, which is the important place to look with Life Charts.

IV. DEPLOYMENT

Since the Boosted Trees gave the highest accuracy, we will use it for deployment. We find this easier to do using the Interactive Boosted Trees (IBT). Thus we will nor re-run the analysis just using the IBT module, as follows.

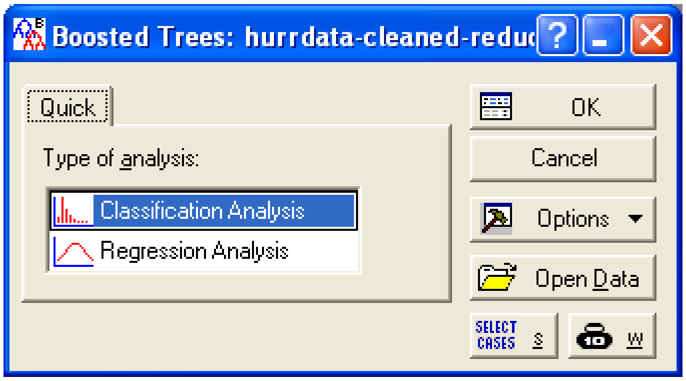
NOTE: Since the INTERACTIVE modules of data mining algorithms have not been extensively used in other tutorials in this HANDBOOK, we will run through this Interactive BOOSTED TREES format thoroughly, step-by-step, in the following example:


STATISTICA - hurrdata-cleaned-reduced_July 20-09 for FINAL tutorial.sta
File Server Edit View Insert Format Statistics Data Mining Graphs Tools Data Enterprise Window Help
Arial
10
Vars
Cases
Data: hurrdata-cleaned-reduced_July 2
Data Miner Recipes
Automated Neural Networks
Machine Learning (Bayesian, Support Vectors, Nearest Neighbor)
General Classification/Regression Tree Models
General CHAID Models
Interactive Trees (C&RT, CHAID)
Boosted Tree Classifiers and Regression
Random Forests for Regression and Classification
Generalized Additive Models
MARSplines (Multivariate Adaptive Regression Splines)
Generalized EM & k-Means Cluster Analysis
Independent Components Analysis
Text & Document Mining
Web Crawling, Document Retrieval
Association Rules
Sequence, Association, and Link Analysis
Rapid Deployment of Predictive Models (PMML)
Goodness of Fit, Classification, Prediction
Feature Selection and Variable Screening
Combining Groups (Classes) for Predictive Data Mining
Data Mining - Workspaces
QC Data Mining & Root Cause Analysis - Methods
Featur...
Featur...
Starts up Boosted Tree Classifiers
hurrdata-clean
C1,V1
0
Sel:OFF
Weight:OFF
CAP
NUM
REC
start
2:40 PM

	inner	2 1a	3 1b
1	0	1	0
2	0	1	1
3	0	1	1
4	0	0	0
5	0	0	0
6	0	1	0
7	1	1	1
8	1	1	1
9	0	1	1
10	0	1	1
11	0	1	1
12	1	1	1
13	0	1	1
14	0	1	1
15	0	1	1
16	1	1	1
17	0	1	0
18	0	1	0
19	0	0	0

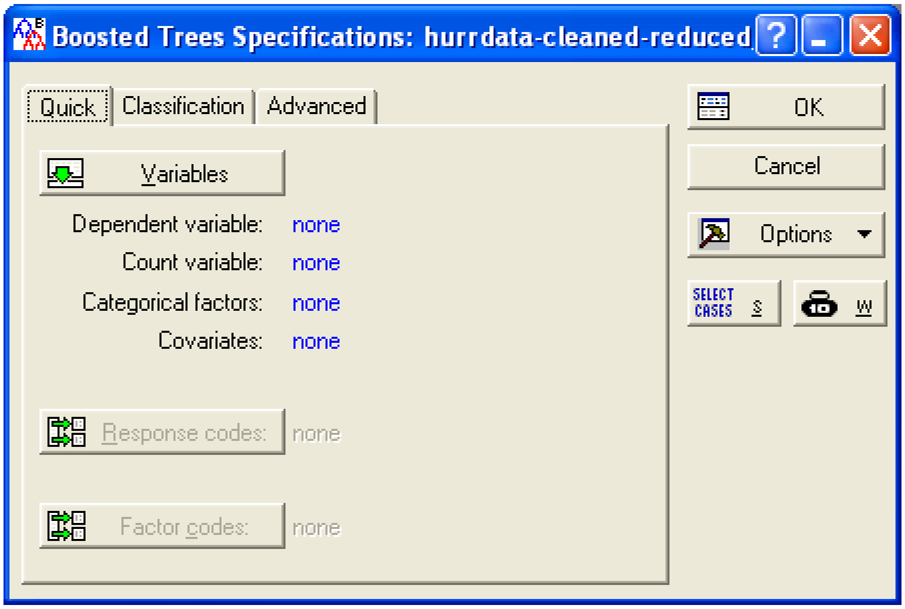
17	Strongly Agr	Strongly /	Agree	Agree	Strongly Agre
18	Strongly Agr	Agree	Agree	Agree	Strongly Agre

FIGURE 10. Selecting Interactive BOOSTED TREES algorithm from the menu.



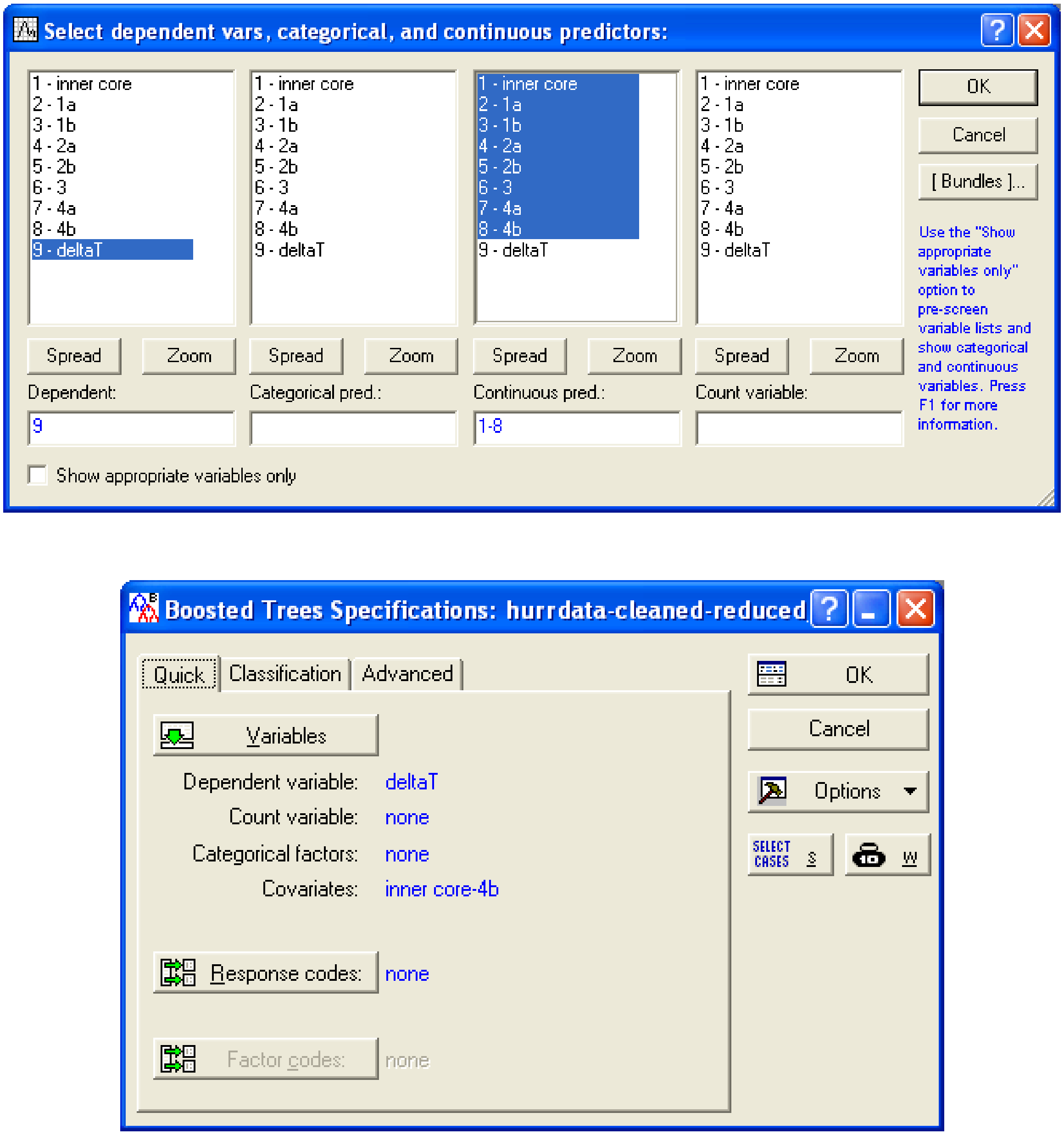
Type of analysis: Classification Analysis

FIGURE 11. Selecting CLASSIFICATION ANALYSIS as our "type of analysis" from the BOOSTED TREES initial dialog.



Dependent variable: none
Count variable: none
Categorical factors: none
Covariates: none
Response codes: none
Factor codes: none

FIGURE 12. The INTERACTIVE BOOSTED TREES initial dialog.



Use the "Show appropriate variables only" option to pre-screen variable lists and show categorical and continuous variables. Press F1 for more information.

Dependent: 9

Categorical pred.:

Continuous pred.: 1-8

Count variable:

☐ Show appropriate variables only

Dependent variable: deltaT

Count variable: none

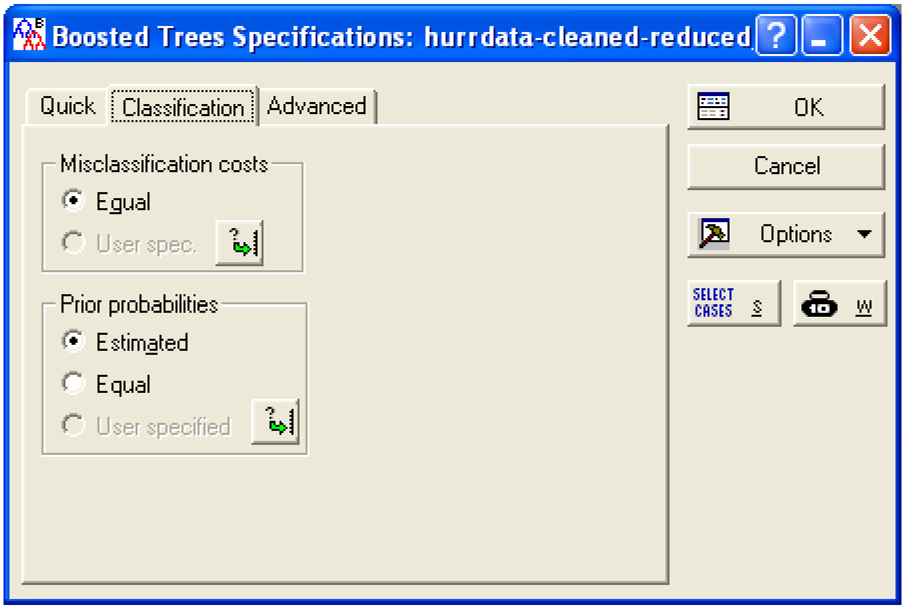
Categorical factors: none

Covariates: inner core-4b

Response codes: none

Factor codes: none

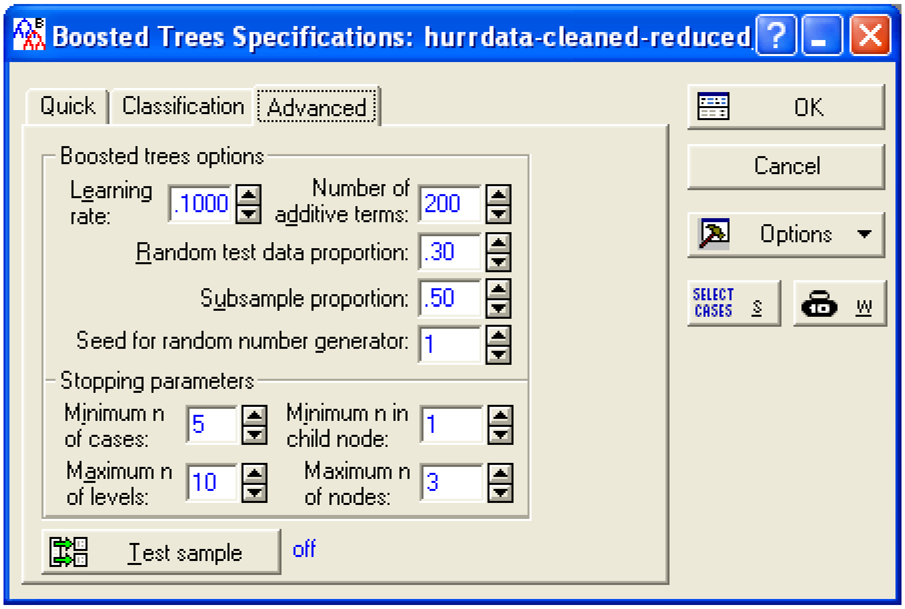
FIGURE 13. Variables selected are “showing as entered” on the BT DIALOG.



Misclassification costs
● Equal
○ User spec.

Prior probabilities
● Estimated
○ Equal
○ User specified

FIGURE 14. Classification Tab is left at it’s default values.



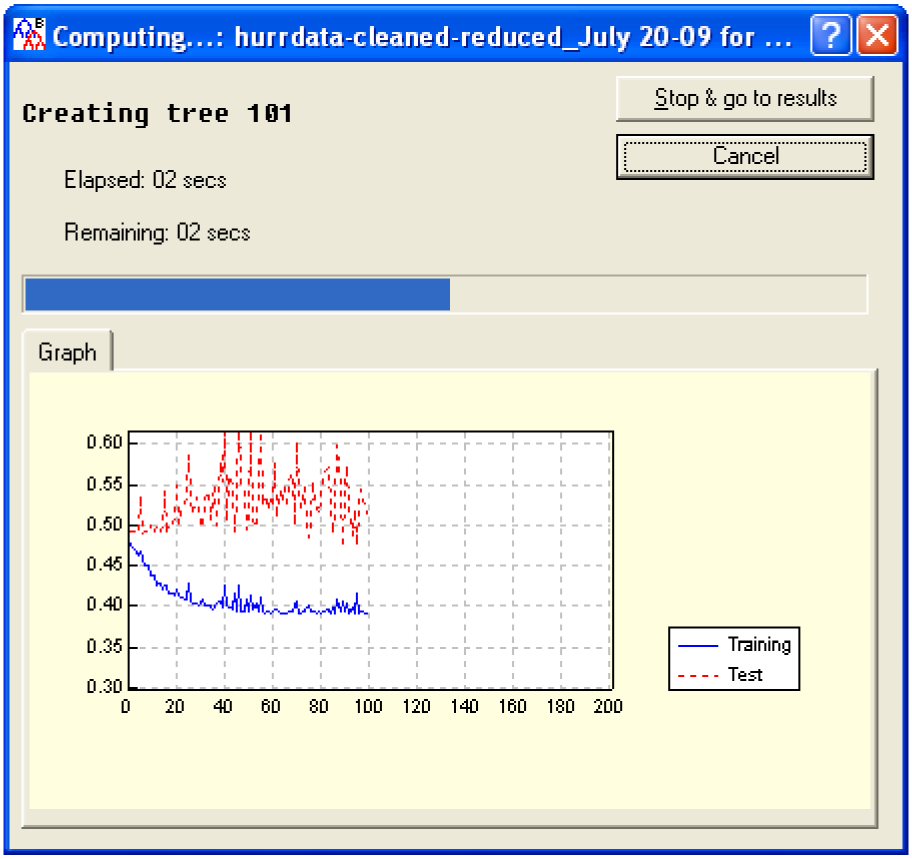
Boosted trees options
Learning rate: .1000
Number of additive terms: 200
Random test data proportion: .30
Subsample proportion: .50
Seed for random number generator: 1

Stopping parameters
Minimum n of cases: 5
Minimum n in child node: 1
Maximum n of levels: 10
Maximum n of nodes: 3

Test sample off

FIGURE 15. The Advanced Tab is also left at it’s default values for this analysis.

Hit **<u>OK</u>** to run, which will produce a series of windows as shown in the figures that follow.



Creating tree 101

Elapsed: 02 secs

Remaining: 02 secs

FIGURE 16. Active Dialog that is visible for a few seconds [or minutes, depending on data size] of the BOOSTED TREES growing process where both the TRAINING data [blue color] and TEST data [red color] "models" are being graphed as per number of trees being grown; when it hits a point where things "level off" for both this will be selected as the "End Point" for an optimal Boosted Trees model.

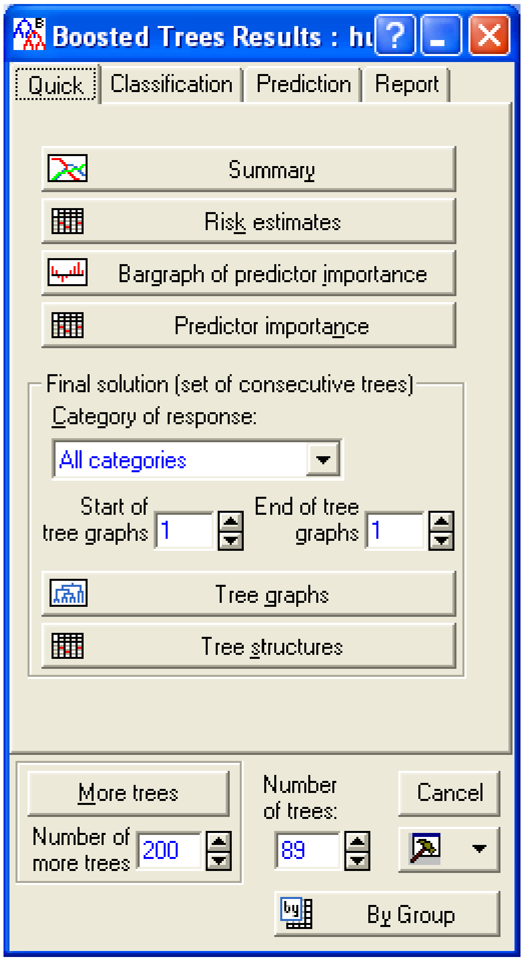
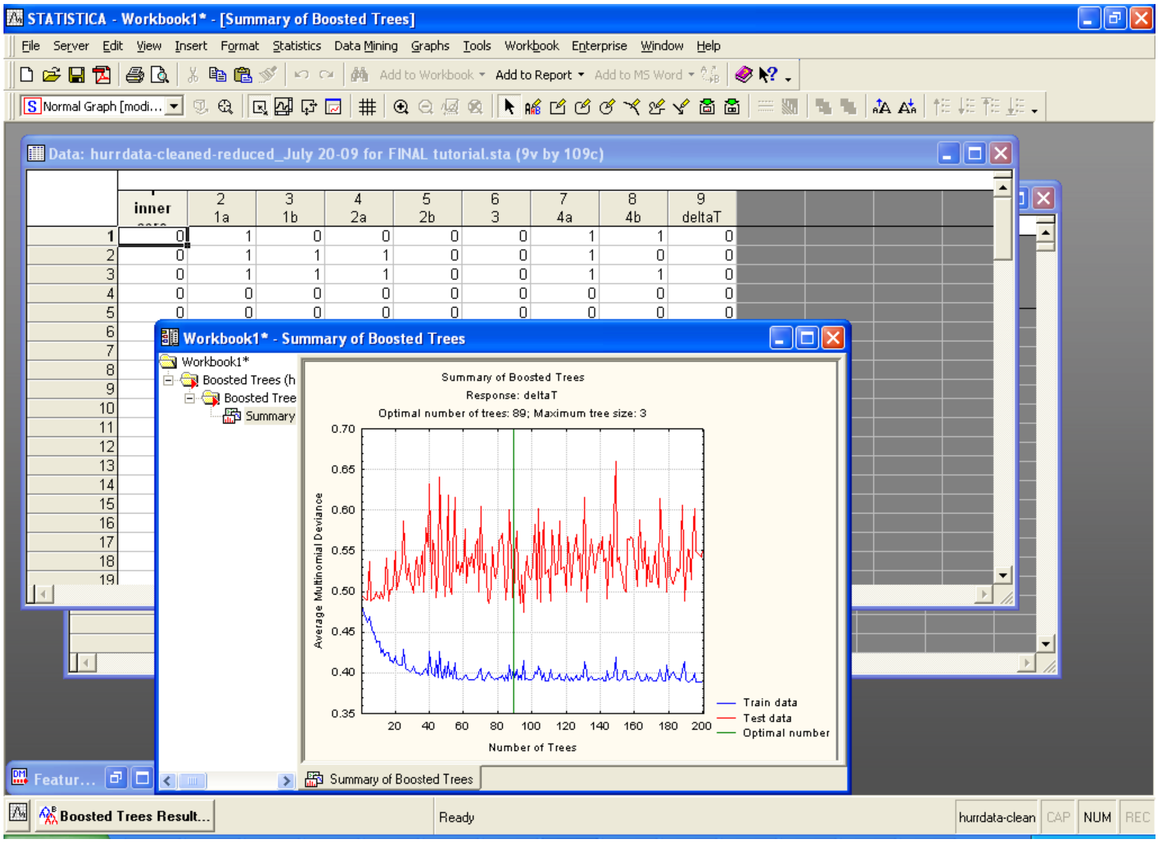


FIGURE 17. When the Boosted Trees computations concludes, the above RESULTS DIALOG will pop on the computer screen.


STATISTICA - Workbook1* - [Summary of Boosted Trees]
File Server Edit View Insert Format Statistics Data Mining Graphs Tools Workbook Enterprise Window Help
Add to Workbook Add to Report Add to MS Word
Normal Graph [modi...
Data: hurrdata-cleaned-reduced_July 20-09 for FINAL tutorial.sta (9v by 109c)
Workbook1* - Summary of Boosted Trees
Workbook1*
Boosted Trees (h
Boosted Tree
Summary
Summary of Boosted Trees
Response: deltaT
Optimal number of trees: 89; Maximum tree size: 3
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
Average Multinomial Deviance
20
40
60
80
100
120
140
160
180
200
Number of Trees
Train data
Test data
Optimal number
Summary of Boosted Trees
Featur...
Boosted Trees Result...
Ready
hurrdata-clean
CAP
NUM
REC

	inner	2 1a	3 1b	4 2a	5 2b	6 3	7 4a	8 4b	9 deltaT
1	0	1	0	0	0	0	1	1	0
2	0	1	1	1	0	0	1	0	0
3	0	1	1	1	0	0	1	1	0
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0

FIGURE 18. Clicking on the SUMMARY button on the RESULTS DIALOG will give a workbook with the result in the left panel of the RESULTS DIALOG, as seen above, will show the final Boosted Trees modeling with the Green Vertical Line being the point of optimum growth for this dataset.

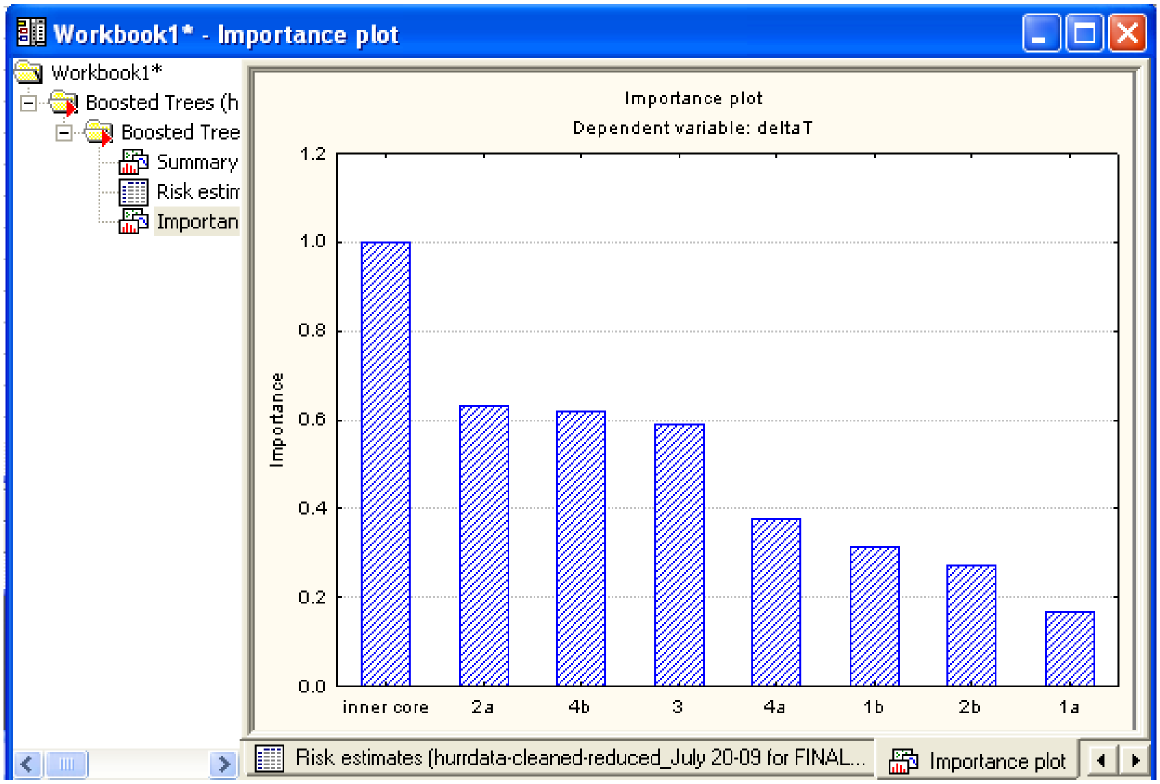
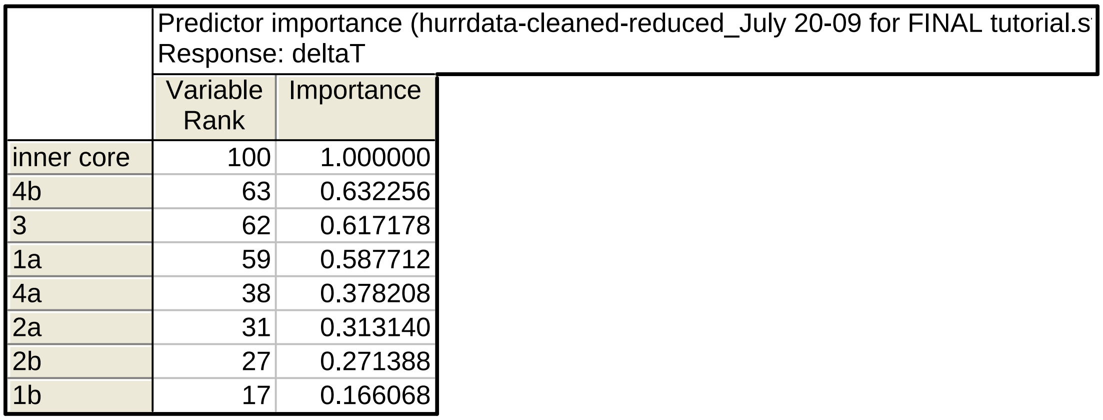


FIGURE 19. Clicking on the PREDICTOR IMPORTANCE button on the RESULTS DIALOG will give the Importance Plot as seen above; this indicates that the "Inner Core" is the most important variable in predicting a hurricane, with the '2a', '4b', and '3' variables being the next in importance, almost equal as they have about the same heights; then the remaining variables trail off in importance. From this we might try using just the top 4 predictor variables in further modeling, as this might give us the highest Accuracy Score in Hurricane Prediction.



Predictor importance (hurrdata-cleaned-reduced_July 20-09 for FINAL tutorial.s
Response: deltaT

	Variable Rank	Importance
inner core	100	1.000000
4b	63	0.632256
3	62	0.617178
1a	59	0.587712
4a	38	0.378208
2a	31	0.313140
2b	27	0.271388
1b	17	0.166068

FIGURE 20. Another way of looking at the Predictor Importance is in the table where the predictor variables are "ranked" on basis of 100 being highest.

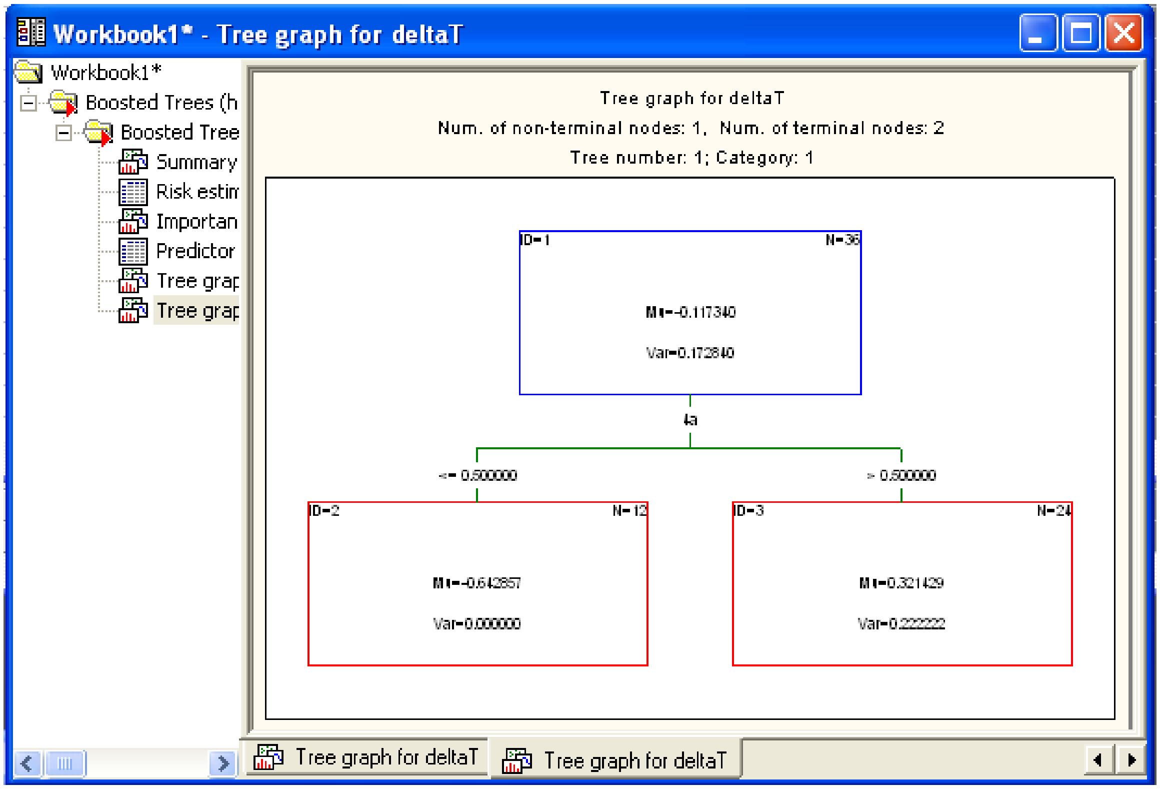


FIGURE 21. Clicking on the TREE GRAPH on the RESULTS DIALOG gives the pictorial representation of the "boosted trees" model.

Now, if we want to deploy this model to new data that has been subsequently collected, we need to follow the steps below:

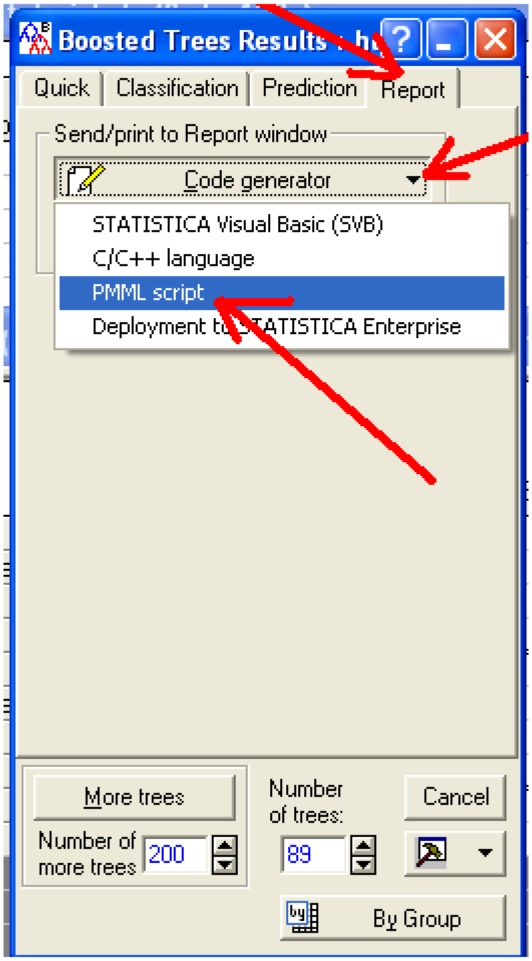
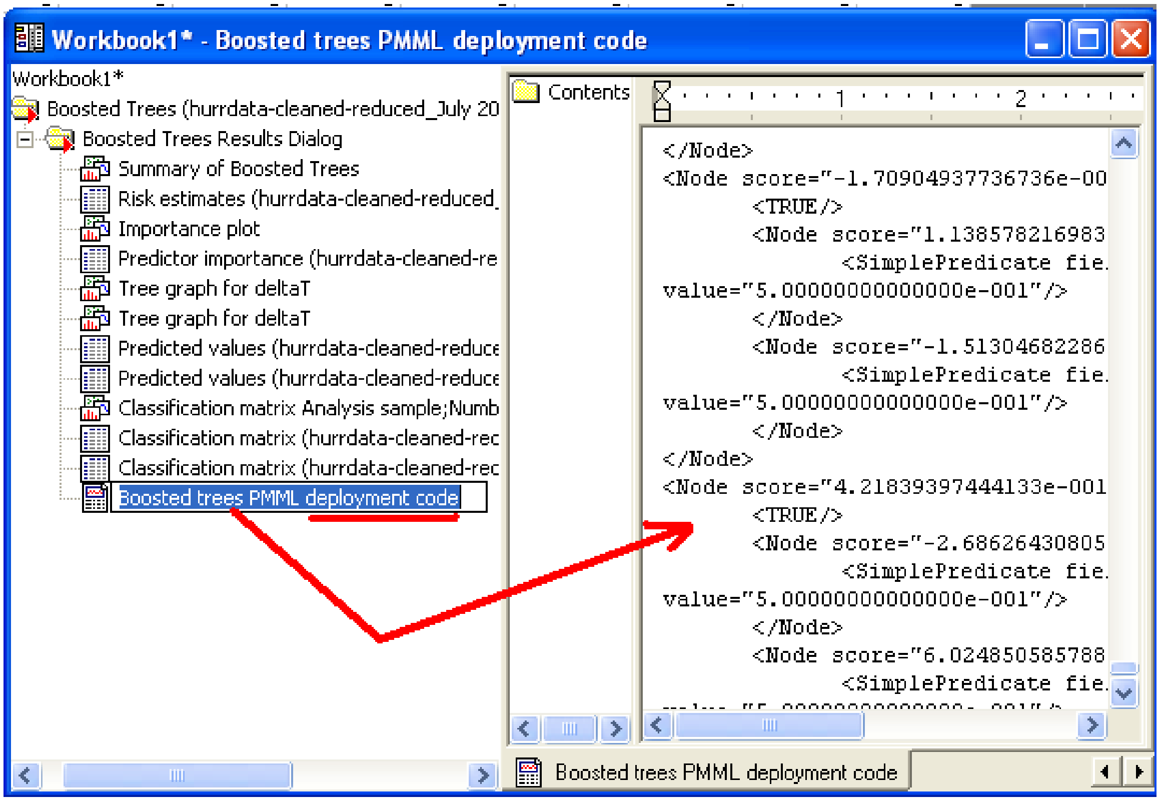


FIGURE 22. Using the REPORT TAB on the Boosted Trees RESULTS DIALOG, select PMML script for the method of creating the model for use in future deployment to new data.

Using the **Report** tab, click on the **Code Generator** button above, then, release on **PMML script**, to produce it in a workbook window.


Workbook1* - Boosted trees PMML deployment code
Workbook1*
Boosted Trees (hurrdata-cleaned-reduced_July 20
Boosted Trees Results Dialog
Summary of Boosted Trees
Risk estimates (hurrdata-cleaned-reduced_
Importance plot
Predictor importance (hurrdata-cleaned-re
Tree graph for deltaT
Tree graph for deltaT
Predicted values (hurrdata-cleaned-reduce
Predicted values (hurrdata-cleaned-reduce
Classification matrix Analysis sample;Numb
Classification matrix (hurrdata-cleaned-rec
Classification matrix (hurrdata-cleaned-rec
Boosted trees PMML deployment code
Contents
Boosted trees PMML deployment code

```
</Node>
<Node score="-1.70904937736736e-00
        <TRUE/>
        <Node score="1.138578216983
                <SimplePredicate fie
value="5.00000000000000e-001"/>
        </Node>
        <Node score="-1.51304682286
                <SimplePredicate fie
value="5.00000000000000e-001"/>
        </Node>
</Node>
<Node score="4.21839397444133e-001
        <TRUE/>
        <Node score="-2.68626430805
                <SimplePredicate fie
value="5.00000000000000e-001"/>
        </Node>
        <Node score="6.024850585788
                <SimplePredicate fie
```

FIGURE 23. Workbook results from selecting the PMML deployment; here the name BOOSTED TREES PMML DEPLOYMENT CODE is in the left panel, and the actual code is in the right panel.

Then, selecting the **Boosted trees PMML deployment code** in the left panel above produces the following window where we choose **Save Item(s) As…,** and save the file as indicated, remembering the directory and file name.

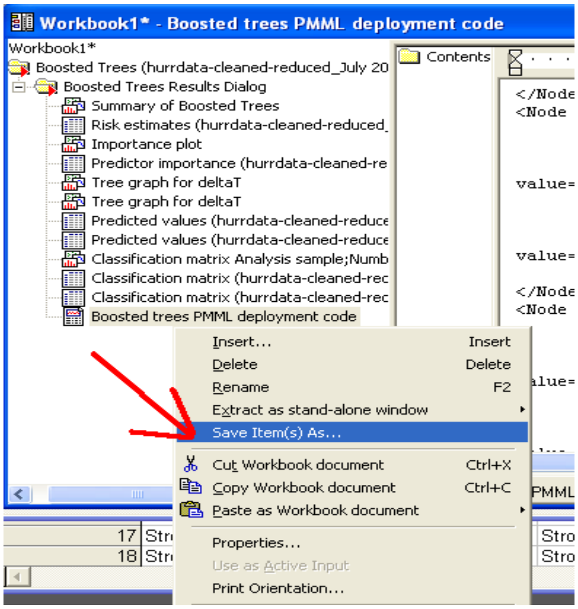
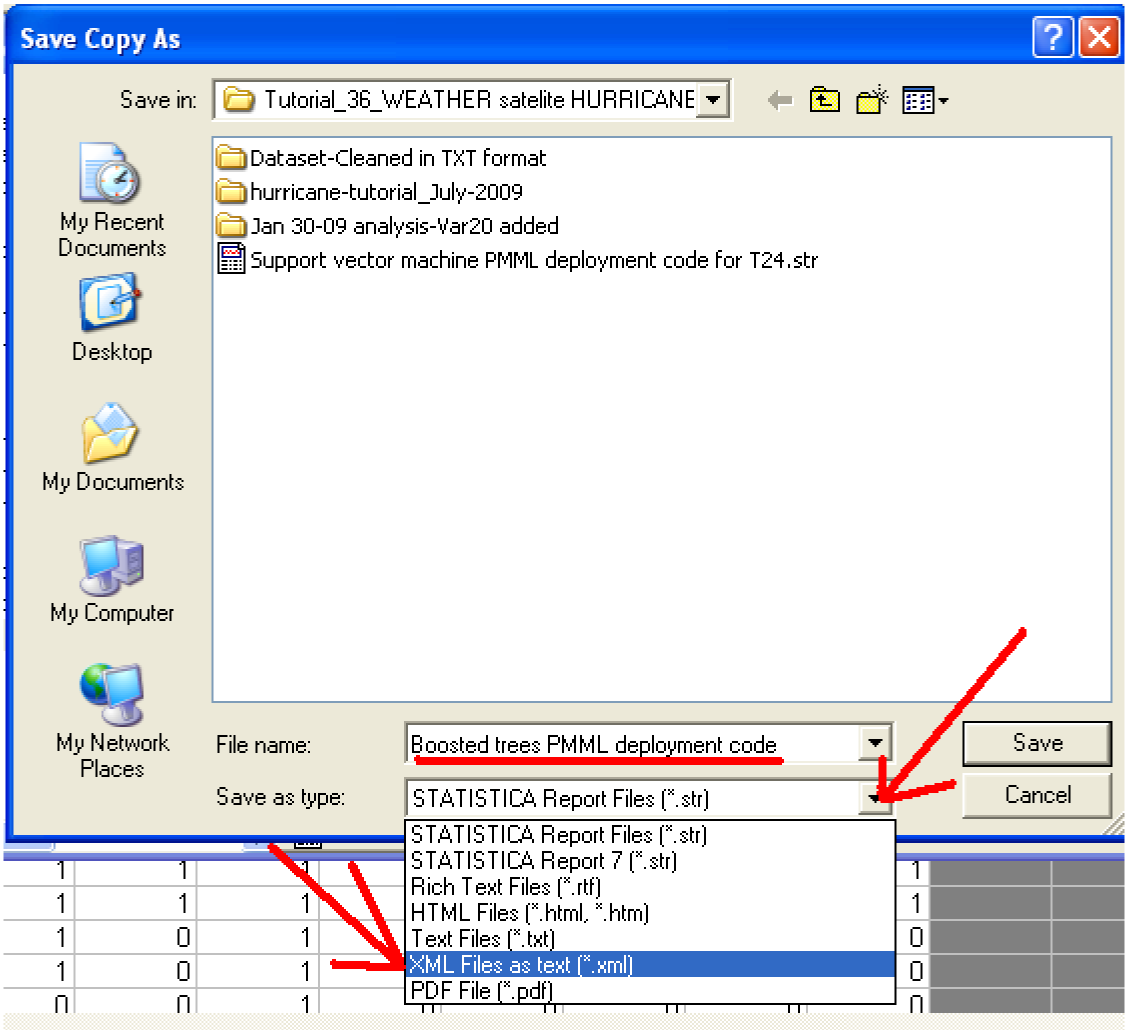

Workbook1* - Boosted trees PMML deployment code
Workbook1*
Boosted Trees (hurrdata-cleaned-reduced_July 20
Boosted Trees Results Dialog
Summary of Boosted Trees
Risk estimates (hurrdata-cleaned-reduced_
Importance plot
Predictor importance (hurrdata-cleaned-re
Tree graph for deltaT
Tree graph for deltaT
Predicted values (hurrdata-cleaned-reduce
Predicted values (hurrdata-cleaned-reduce
Classification matrix Analysis sample;Numb
Classification matrix (hurrdata-cleaned-rec
Classification matrix (hurrdata-cleaned-rec
Boosted trees PMML deployment code
Contents
</Node
<Node
value=
value=
</Node
<Node
Insert… Insert
Delete Delete
Rename F2
Extract as stand-alone window
Save Item(s) As…
Cut Workbook document Ctrl+X
Copy Workbook document Ctrl+C
Paste as Workbook document
Properties…
Use as Active Input
Print Orientation…
PMML
17 Str
18 Str
Stro
Stro

FIGURE 24. By right-clicking on the BOOSTED TREES PMML DEPLOYMENT CODE file name a “flying menu” appears; on this flying menu select SAVE ITEM AS, and release, to bring up a folder / file dialog, as shown below [you may have to browse for the right one….].


Save Copy As
My Recent Documents
Desktop
My Documents
My Computer
My Network Places
Dataset-Cleaned in TXT format
hurricane-tutorial_July-2009
Jan 30-09 analysis-Var20 added
Support vector machine PMML deployment code for T24.str
Save
Cancel
STATISTICA Report Files (*.str)
STATISTICA Report 7 (*.str)
Rich Text Files (*.rtf)
HTML Files (*.html, *.htm)
Text Files (*.txt)
XML Files as text (*.xml)
PDF File (*.pdf)

Save in: Tutorial_36_WEATHER satelite HURRICANE

File name: Boosted trees PMML deployment code

Save as type: STATISTICA Report Files (*.str)

FIGURE 25. Select the SML FILES AS TEXT (“.xml”) to save the deployment code into this type of xml file.

Now we open the **Rapid Deployment Module**

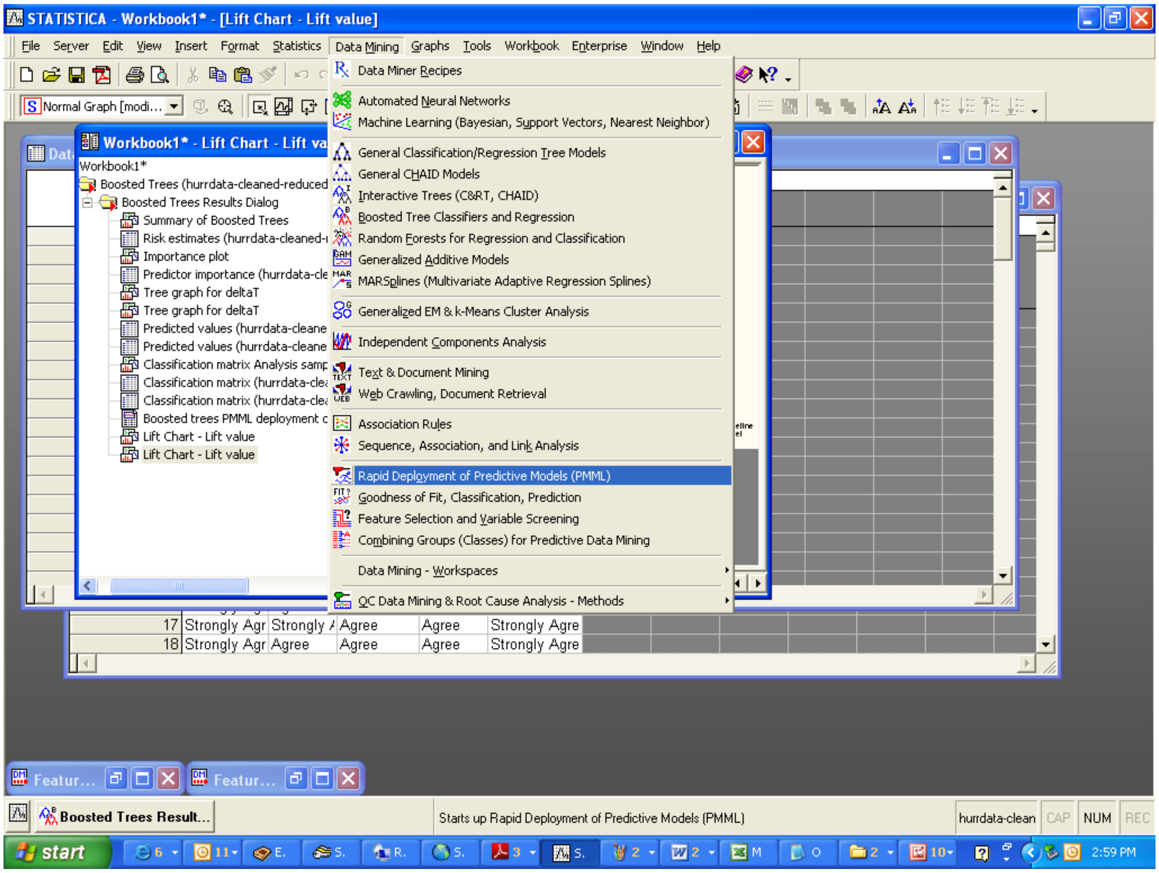
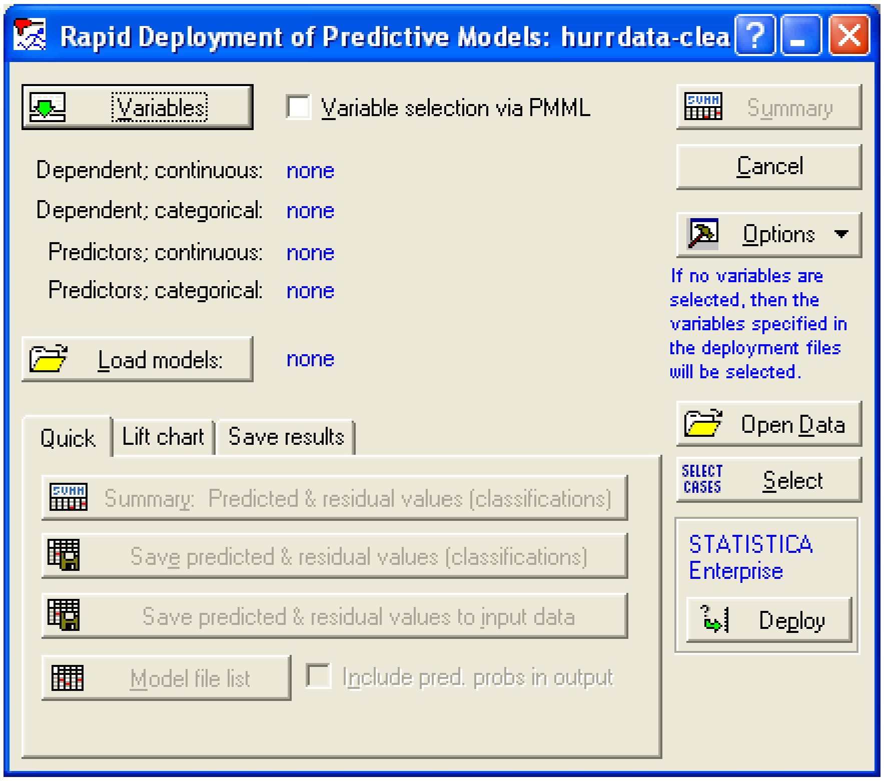

STATISTICA - Workbook1* - [Lift Chart - Lift value]
File Server Edit View Insert Format Statistics Data Mining Graphs Tools Workbook Enterprise Window Help
Data Miner Recipes
Automated Neural Networks
Machine Learning (Bayesian, Support Vectors, Nearest Neighbor)
General Classification/Regression Tree Models
General CHAID Models
Interactive Trees (C&RT, CHAID)
Boosted Tree Classifiers and Regression
Random Forests for Regression and Classification
Generalized Additive Models
MARSplines (Multivariate Adaptive Regression Splines)
Generalized EM & k-Means Cluster Analysis
Independent Components Analysis
Text & Document Mining
Web Crawling, Document Retrieval
Association Rules
Sequence, Association, and Link Analysis
Rapid Deployment of Predictive Models (PMML)
Goodness of Fit, Classification, Prediction
Feature Selection and Variable Screening
Combining Groups (Classes) for Predictive Data Mining
Data Mining - Workspaces
QC Data Mining & Root Cause Analysis - Methods
Workbook1* - Lift Chart - Lift va
Workbook1*
Boosted Trees (hurrdata-cleaned-reduced
Boosted Trees Results Dialog
Summary of Boosted Trees
Risk estimates (hurrdata-cleaned-
Importance plot
Predictor importance (hurrdata-cle
Tree graph for deltaT
Tree graph for deltaT
Predicted values (hurrdata-cleane
Predicted values (hurrdata-cleane
Classification matrix Analysis samp
Classification matrix (hurrdata-cle
Classification matrix (hurrdata-cle
Boosted trees PMML deployment c
Lift Chart - Lift value
Lift Chart - Lift value
17 Strongly Agr Strongly A Agree Agree Strongly Agre
18 Strongly Agr Agree Agree Agree Strongly Agre
Featur...
Featur...
Boosted Trees Result...
Starts up Rapid Deployment of Predictive Models (PMML)
hurrdata-clean CAP NUM REC
start
2:59 PM

FIGURE 26. Now select the RAPID DEPLOYMENT of PREDICTIVE MODELS (PMML) selection, as show above.



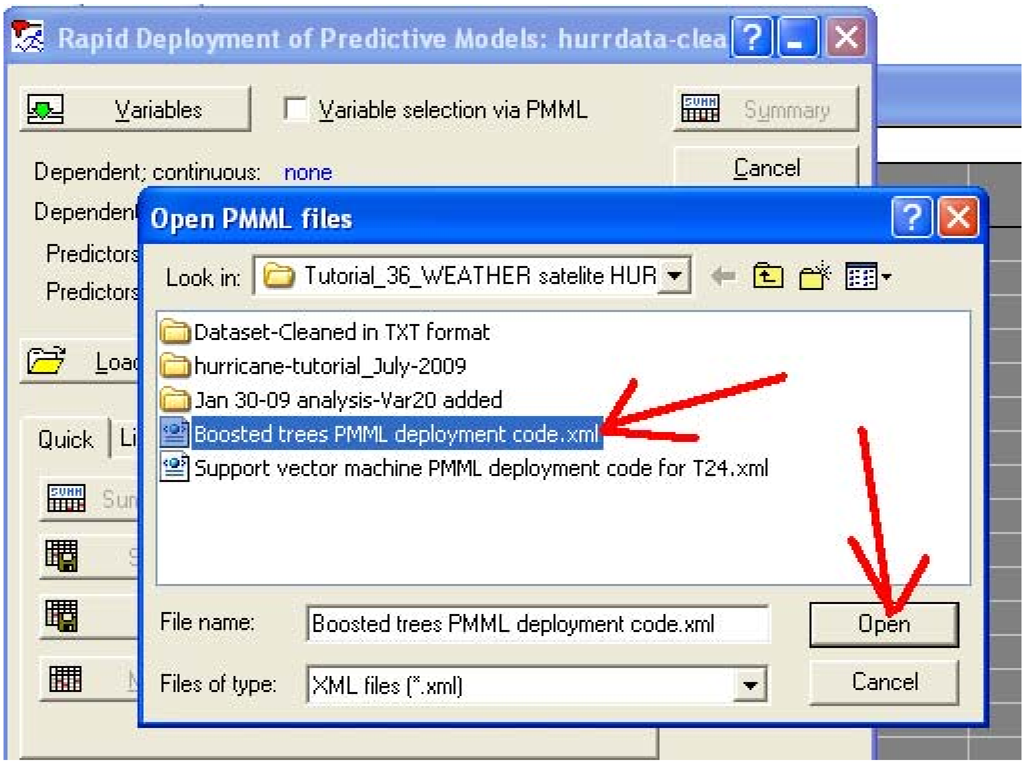
Rapid Deployment of Predictive Models: hurrdata-clea
Variables
☐ Variable selection via PMML
Summary
Cancel
Options
Load models:
Quick
Lift chart
Save results
Summary: Predicted & residual values (classifications)
Save predicted & residual values (classifications)
Save predicted & residual values to input data
Model file list
☐ Include pred. probs in output
Open Data
SELECT CASES
Select
STATISTICA Enterprise
Deploy

Dependent; continuous: none
Dependent; categorical: none
Predictors; continuous: none
Predictors; categorical: none
Load models: none

If no variables are selected, then the variables specified in the deployment files will be selected.

FIGURE 27. The RAPID DEPLOYMENT OF PREDICTIVE MODELS dialog.

Then click on the **Load Models** button, and choose the items as indicated below.



☐ Variable selection via PMML

Dependent; continuous: none

Open PMML files

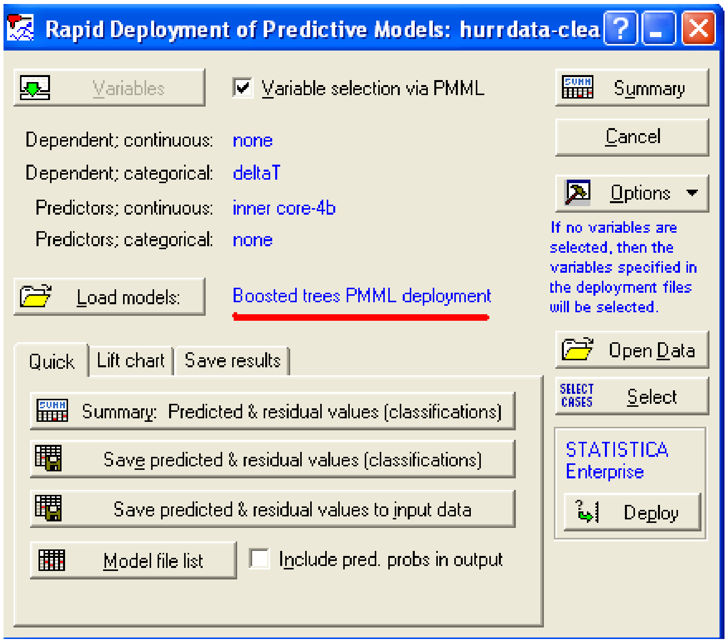
Look in: Tutorial_36_WEATHER satelite HUR

Dataset-Cleaned in TXT format
hurricane-tutorial_July-2009
Jan 30-09 analysis-Var20 added
Boosted trees PMML deployment code.xml
Support vector machine PMML deployment code for T24.xml

File name: Boosted trees PMML deployment code.xml

Files of type: XML files (*.xml)

FIGURE 28. Select the BOOSTED TREES PMML DEPLOYMENT CODE.XML that you have saved away from your BT Modeling, and click OPEN to put into the RAPID DEPLOYMENT process.


Rapid Deployment of Predictive Models: hurrdata-clea
Variables
Summary
Cancel
Options
Load models:
Open Data
Select
STATISTICA Enterprise
Deploy
Quick
Lift chart
Save results
Summary: Predicted & residual values (classifications)
Save predicted & residual values (classifications)
Save predicted & residual values to input data
Model file list

☑ Variable selection via PMML

Dependent; continuous: none
Dependent; categorical: deltaT
Predictors; continuous: inner core-4b
Predictors; categorical: none

Load models: Boosted trees PMML deployment

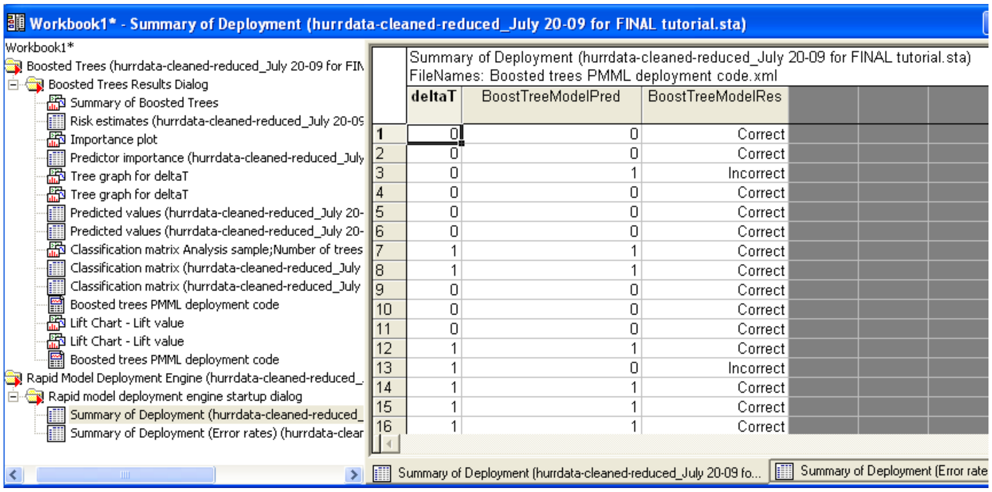
If no variables are selected, then the variables specified in the deployment files will be selected.

☐ Include pred. probs in output

FIGURE 29. The RAPID DEPLOYMENT DIALOG after the model has been loaded.

Note that the variables have already been selected (this is because our dataset was sitting open, behind the Rapid Deployment Screen, and selected as the "active spreadsheet input". We could have brought up a 2nd dataset, with the same variable names and same basic format, and it would have selected those variables from this new dataset.

Now hit the **Summary** button to get the results.


Workbook1* - Summary of Deployment (hurrdata-cleaned-reduced_July 20-09 for FINAL tutorial.sta)
Workbook1*
Boosted Trees (hurrdata-cleaned-reduced_July 20-09 for FIN
Boosted Trees Results Dialog
Summary of Boosted Trees
Risk estimates (hurrdata-cleaned-reduced_July 20-09
Importance plot
Predictor importance (hurrdata-cleaned-reduced_July
Tree graph for deltaT
Tree graph for deltaT
Predicted values (hurrdata-cleaned-reduced_July 20-
Predicted values (hurrdata-cleaned-reduced_July 20-
Classification matrix Analysis sample;Number of trees
Classification matrix (hurrdata-cleaned-reduced_July
Classification matrix (hurrdata-cleaned-reduced_July
Boosted trees PMML deployment code
Lift Chart - Lift value
Lift Chart - Lift value
Boosted trees PMML deployment code
Rapid Model Deployment Engine (hurrdata-cleaned-reduced_
Rapid model deployment engine startup dialog
Summary of Deployment (hurrdata-cleaned-reduced_
Summary of Deployment (Error rates) (hurrdata-clear
Summary of Deployment (hurrdata-cleaned-reduced_July 20-09 fo...
Summary of Deployment (Error rate

Summary of Deployment (hurrdata-cleaned-reduced_July 20-09 for FINAL tutorial.sta)
FileNames: Boosted trees PMML deployment code.xml

	deltaT	BoostTreeModelPred	BoostTreeModelRes
1	0	0	Correct
2	0	0	Correct
3	0	1	Incorrect
4	0	0	Correct
5	0	0	Correct
6	0	0	Correct
7	1	1	Correct
8	1	1	Correct
9	0	0	Correct
10	0	0	Correct
11	0	0	Correct
12	1	1	Correct
13	1	0	Incorrect
14	1	1	Correct
15	1	1	Correct
16	1	1	Correct

FIGURE 30. The new data has been "scored" by the Boosted Tree model as to whether it is a DELTA-T "yes" or "no" to a hurricane, and then giving in the 3rd column whether this prediction was "correct" or "incorrect". [e.g. if 85% of the cases are CORRECT then we'd say that we have an 85% Accuracy Rate in predicting "Delta-T"……….].

V. CONCLUSIONS

Of course, summarizing the data in the manner that we are suggesting is where the physical insight into the problem comes into play. What this tutorial aims to show is that once the data has been prepared properly, (i) the process of building a model is systematic and straightforward with the help of the software, and (ii) the statistical (data mining) approach that we are pursuing can be fruitful.

Bibliography:

1) "Tropical Cyclone Intensity Analysis and Forecasting from Satellite Imagery"
Vernon F Dvorak, Monthly Weather Review 103, 420 (1975)

2) "The Dvorak Technique Explained", NOAA,
http://www.ssd.noaa.gov/PS/TROP/dvorak.html

3) "Development of an Objective Scheme to Estimate Tropical Cyclone Intensity from Digital Geostationary Satellite Infrared Imagery"
Christopher S. Velden, Timothy L. Olander and Raymond M. Zehr,
http://cimss.ssec.wisc.edu/tropic/research/products/dvorak/odt.html