Getting Started With Machine Learning

© 2018 Minitab, Inc.

Machine Learning

What is machine learning

A machine learning algorithm “teaches” a computer to recognize patterns using available data.

Data is usually split into a training set and a test set:

- Training (or learn) data is used to create the model.
- Test data is used to assess model performance.

Popular machine learning techniques include tools you are likely already familiar with such as regression and logistic regression. Tree-based techniques, such as CART and TreeNet, are also frequently used for machine learning.

When to use machine learning

Machine learning techniques can be supervised or unsupervised:

- Use supervised machine learning techniques when you have a response, or target, variable (Y) and multiple predictors (X’s). Supervised machine learning tools include Regression, Binary Logistic Regression, CART, Random Forests, TreeNet, and MARS.
- Use unsupervised machine learning techniques when no response exists and you want to find groupings in your data. Unsupervised machine learning tools include Principal Components Analysis and Cluster Analysis.

Why use machine learning

Use machine learning techniques to answer questions such as:

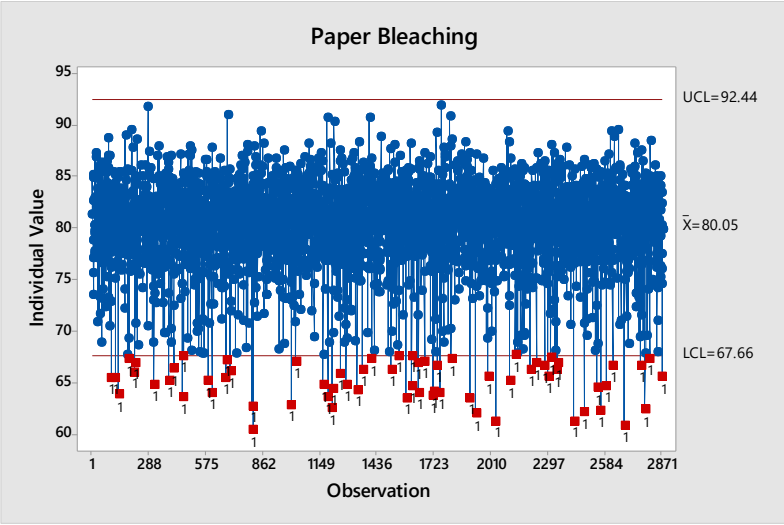
- What is the root cause of process defects?
- What factors contribute to excess variation in my process?
- What is the predicted cost of manufacturing a custom order?
- What conditions indicate machine degradation?

Minitab®
© 2018 Minitab, Inc.

Example: Pulp Bleaching Process

Problem

A paper manufacturer needs to use current process data to determine which factors are contributing to excessive variation in a pulp bleaching process. The process is very non-stable, creating an unacceptable defect rate.



Data collection

Process sensors collect data in real time. The results are saved in a CSV file for offline analysis.

Tools

- ► Graphs
- ► CART

Data set

PulpBrightness.csv

Variable	Description
Production Rate	Predictor
Discharge pH	Predictor
Caustic	Predictor
MgSO4	Predictor
H2O2	Predictor
O2	Predictor
Conductivity	Predictor
Unbleached	Predictor
Brightness	Continuous Response (Target)
Low	Categorical Response (Target) – Indicates when Brightness falls below the lower control limit

Minitab®
© 2018 Minitab, Inc.

Importing data

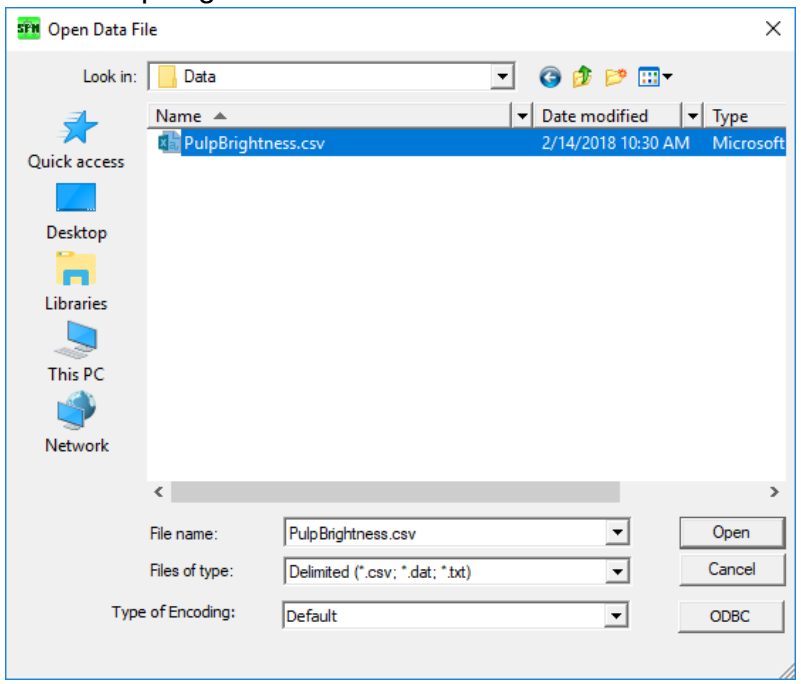
Importing Data

You can import data from many different file types or connect directly to your database via ODBC. To use ODBC, you may need to install ODBC drivers on your system.

The data for this example is in a CSV file.

Open Data File

1. Choose **File > Open > Data File**.
2. From **Files of type**, choose **Delimited (*.csv; *.dat; *.txt)**.
3. Select PulpBrightness.csv as shown below.



4. Click **Open**.

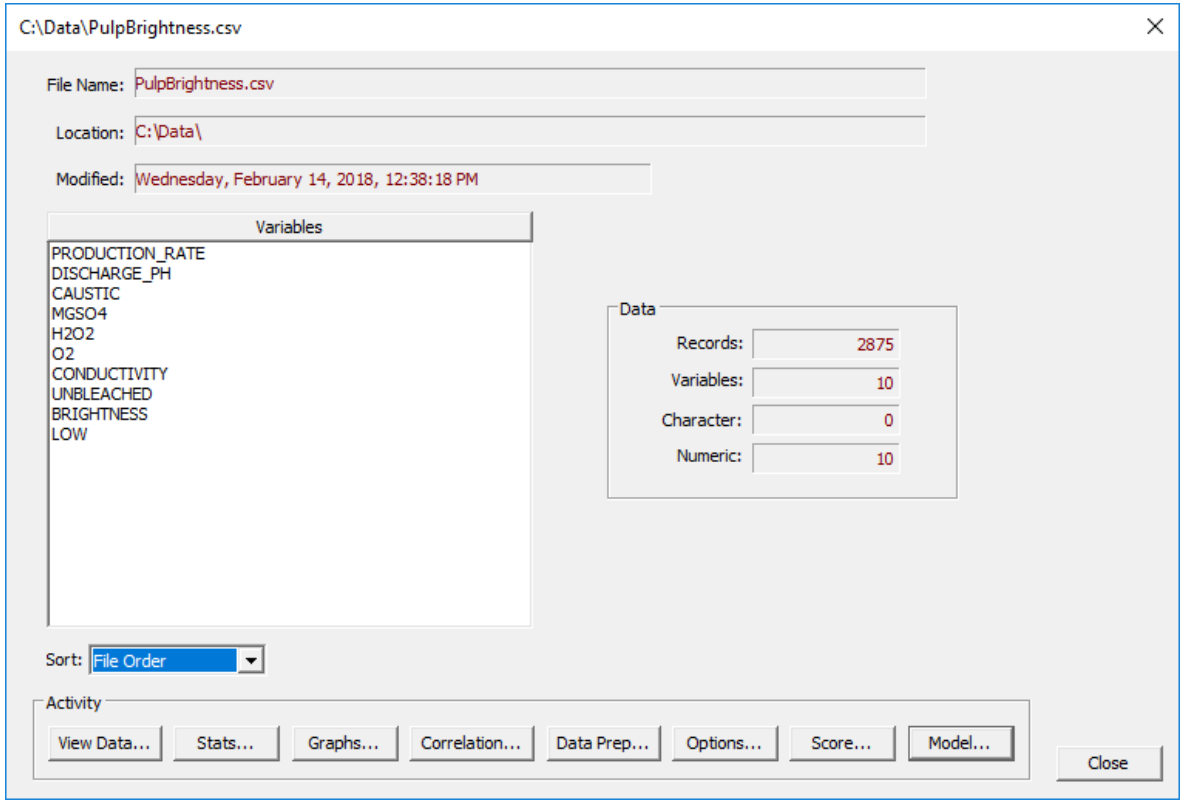
© 2018 Minitab, Inc.

Importing data

By default, you will see an alphabetical list of variables in the data source. To put these variables in the order they appeared in the original file, choose **File Order** from the **Sort** drop-down.

Note that this data set contains 2875 observations and 10 variables.

You can view your data, summary statistics, and graphs. For this example, we will begin by viewing some simple graphs.



© 2018 Minitab, Inc.

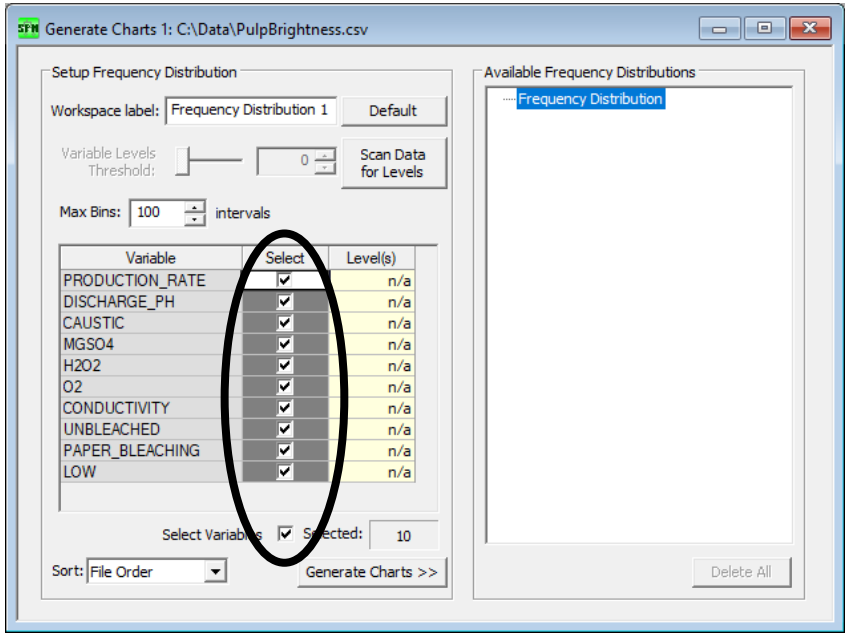
Visualizing data

Graphs are an important preliminary step in data analysis. They enable you to examine the data and to identify patterns, relationships, and potential problems.

Because a relatively small number of variables exist, select all of them to view in a graph.

Graphs

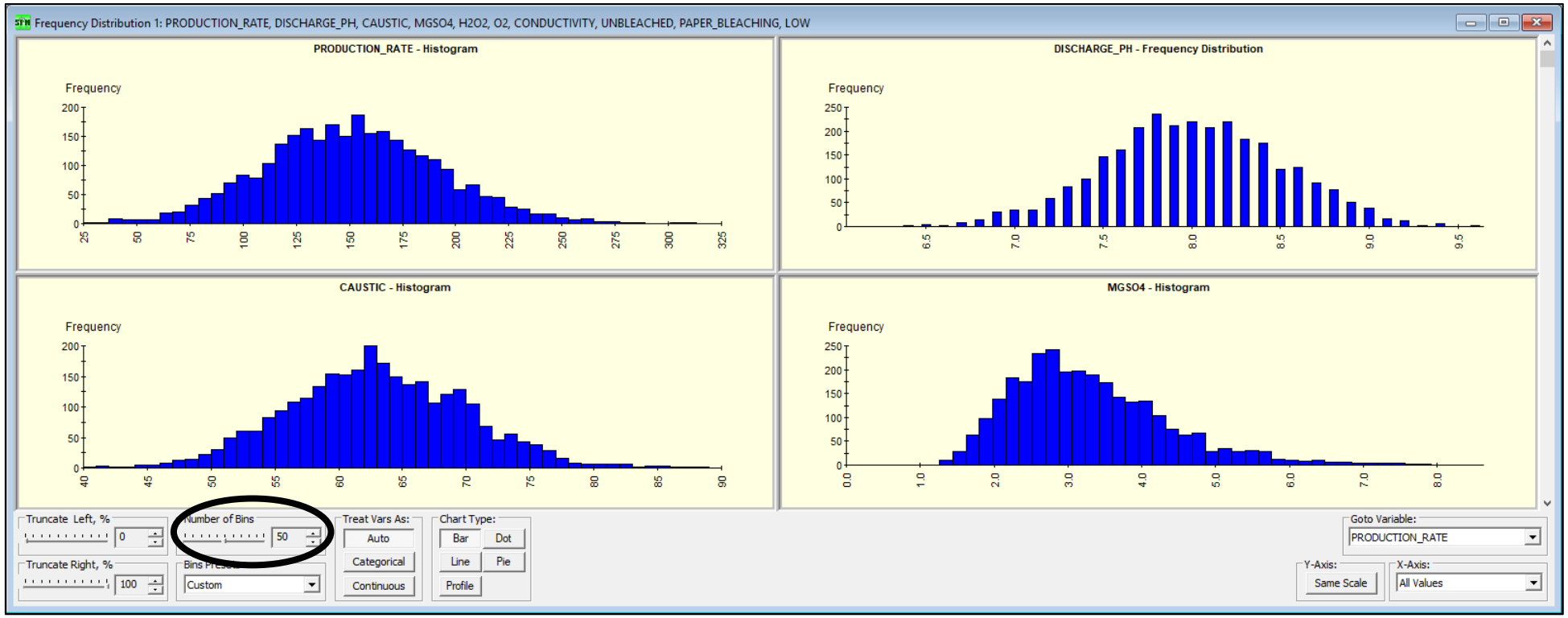
1. Click the **Graphs** button.
2. From **Sort**, choose **File Order**.
3. Click the **Select** heading to select all the variables, then check **Select Variables** as shown below.



4. Click **Generate Charts**.

© 2018 Minitab, Inc.

Visualizing data



The predictors in this data set all happen to be continuous. Histograms are very useful for looking at the shape, center, and spread of continuous variables.

Use the controls at the bottom of the Graph window to interact with these graphs. For example, by increasing the **Number of Bins** using the slider at the bottom of the window, some additional patterns emerge.

Notice that:

- ► Production rate, discharge pH, and caustic appear to be roughly bell-shaped, while MGSO4 is right-skewed.
- ► Because more rounding occurs with the pH measurement, the discharge pH appears slightly more discrete than the other predictor variables.

Minitab
© 2018 Minitab, Inc.

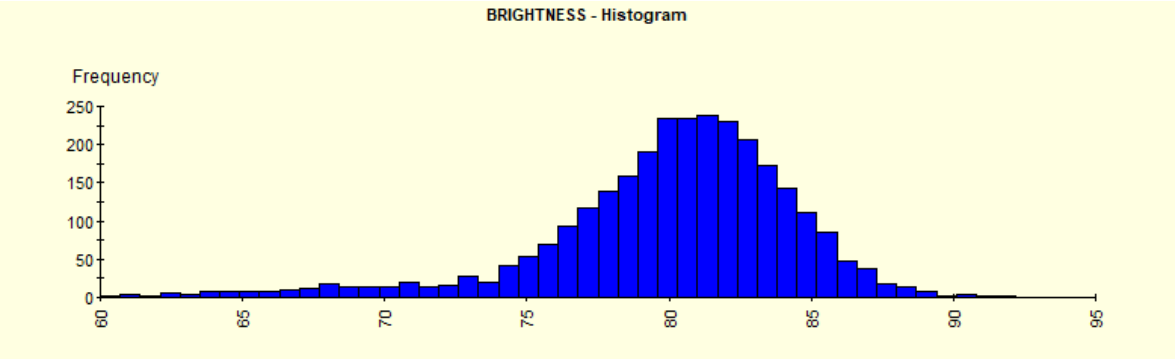
Visualizing data

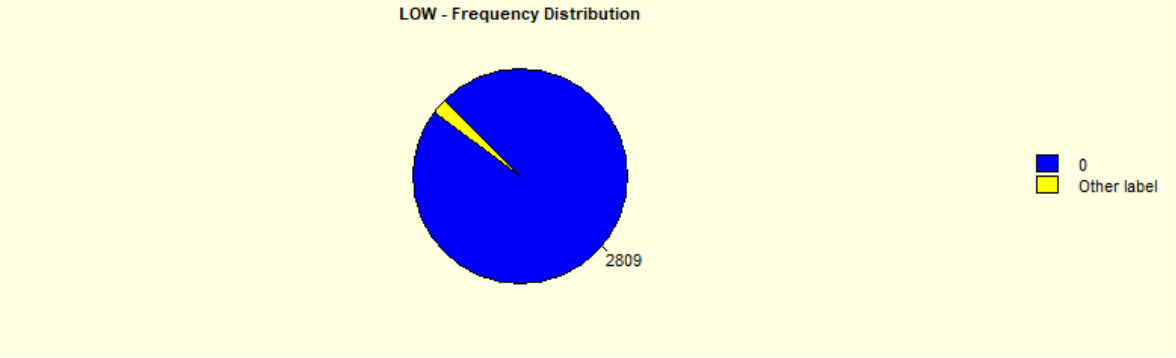
Scroll down in the window or choose BRIGHTNESS from the **Goto Variable** drop-down to visualize the two response (target) variables in this data set.

Brightness is a continuous target. Notice the long left tail. These are the points where the process data fall below the lower control limit.

Low is an indicator variable that identifies when the Brightness measure falls below the lower control limit. Because this variable is categorical, it is useful to view it in a Pie Chart. To switch from Histograms to Pie Charts, click **Pie** under **Chart Type**. Notice that only a small proportion of the cases fall below the lower specification limit.

Next, we will investigate the correlation among the continuous variables.





Minitab
© 2018 Minitab, Inc.

Correlation

What is correlation

Correlation measures the degree of linear association between two continuous variables. The correlation between two variables will always be between -1 and +1.

- A positive correlation indicates that both variables tend to increase or decrease together. The stronger the positive relationship, the closer to +1 the correlation will be.
- A negative correlation indicates that as one variable increases, the other tends to decrease. The stronger the negative relationship, the closer to -1 the correlation will be.

When to use correlation

Use correlation when you want to determine whether two continuous variables are linearly related. The correlation will not tell you whether the variables are related in a nonlinear fashion.

Why use correlation

Correlation answers questions such as:

- Which variables are highly correlated with the response?
- Which variables are highly correlated with each other?

Minitab®
© 2018 Minitab, Inc.

Assessing correlations

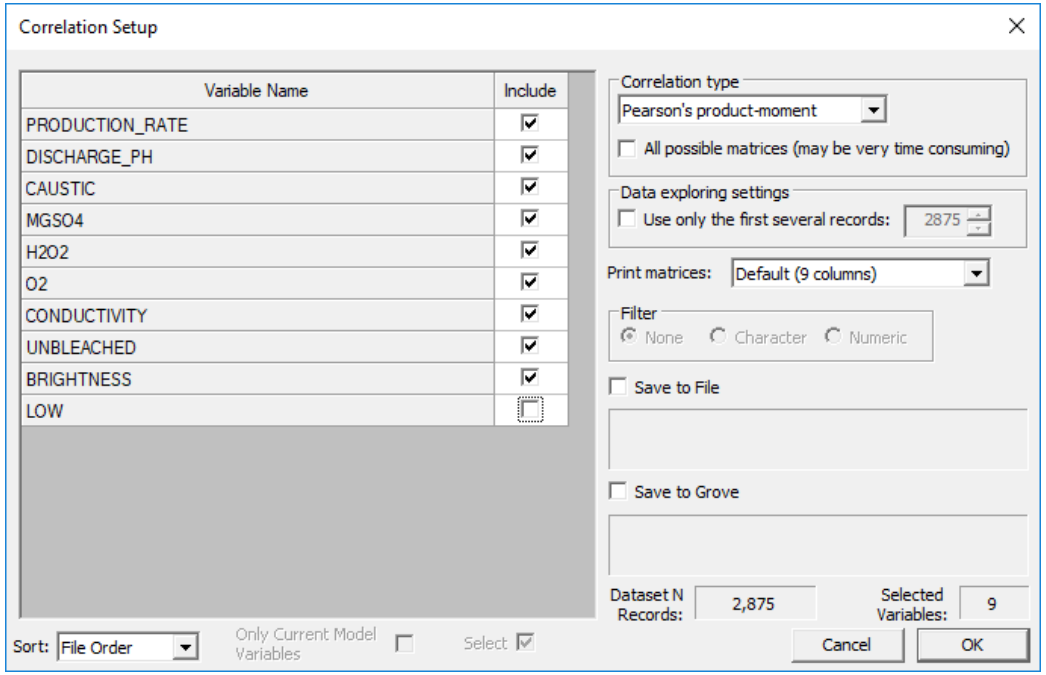
It is useful to look at a matrix of correlation statistics to get an overview of the linear relationships between the continuous variables before moving to formal modeling techniques. Note that correlation only measures linear relationships. Relationships may also follow non-linear patterns.

For larger data sets, you may want to limit the number of variables to look by carefully selecting the variables to include, or limit the number of rows by checking **Use only the first several record** and choosing a smaller number of observations

Because LOW is a categorical variable, leave it unchecked.

Correlation

1. Choose **Explore > Correlation**.
2. From **Sort**, choose **File Order**.
3. Check all of the variables except LOW as shown below.



4. Click **Generate Charts**.

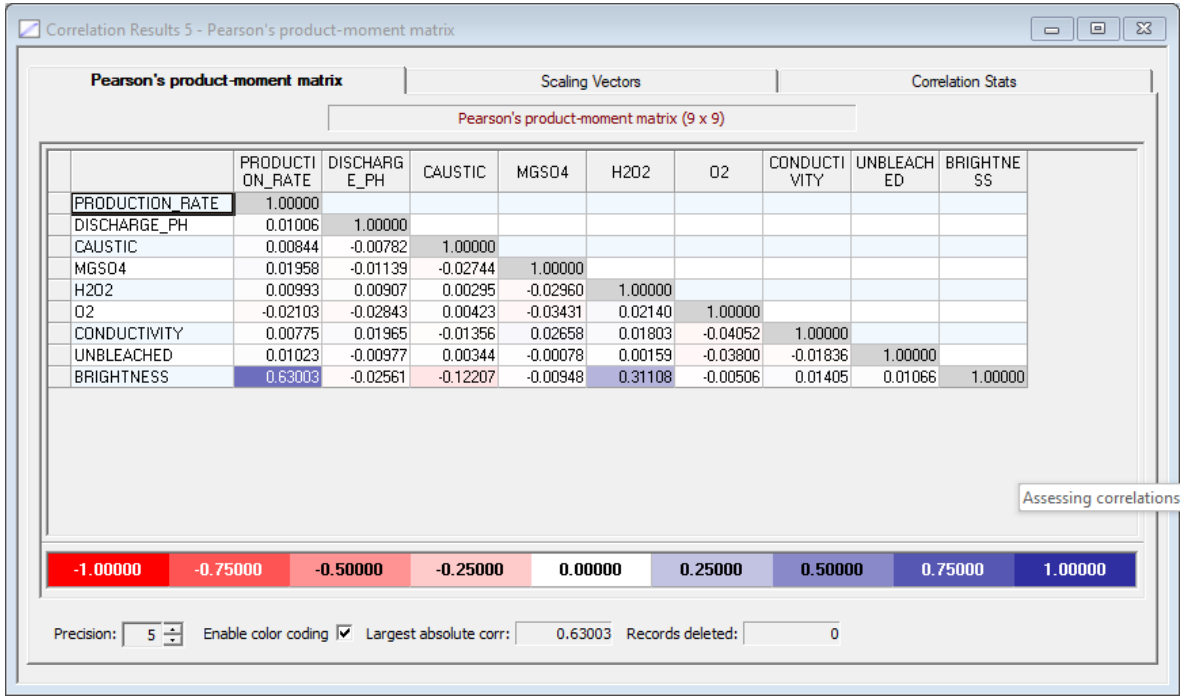
© 2018 Minitab, Inc.

Assessing correlations

The correlation matrix shows that the strongest linear relationship with the brightness target variable is the production rate (r = 0.6003). H2O2 also appears to be linearly related with brightness (r= 0.31108).

While some correlation exists among the predictor variables, none of those correlations appear to be very strong. When predictors are highly correlated, the importance of one predictor in a model can depend on which other predictors are in the model.

Next we will use a CART model to further assess which of these variables contribute to excess variation in brightness.



	PRODUCTION_RATE	DISCHARGE_PH	CAUSTIC	MGSO4	H2O2	O2	CONDUCTIVITY	UNBLEACHED	BRIGHTNESS
PRODUCTION_RATE	1.00000								
DISCHARGE_PH	0.01006	1.00000							
CAUSTIC	0.00844	-0.00782	1.00000						
MGSO4	0.01958	-0.01139	-0.02744	1.00000					
H2O2	0.00993	0.00907	0.00295	-0.02960	1.00000				
O2	-0.02103	-0.02843	0.00423	-0.03431	0.02140	1.00000			
CONDUCTIVITY	0.00775	0.01965	-0.01356	0.02658	0.01803	-0.04052	1.00000		
UNBLEACHED	0.01023	-0.00977	0.00344	-0.00078	0.00159	-0.03800	-0.01836	1.00000	
BRIGHTNESS	0.63003	-0.02561	-0.12207	-0.00948	0.31108	-0.00506	0.01405	0.01066	1.00000

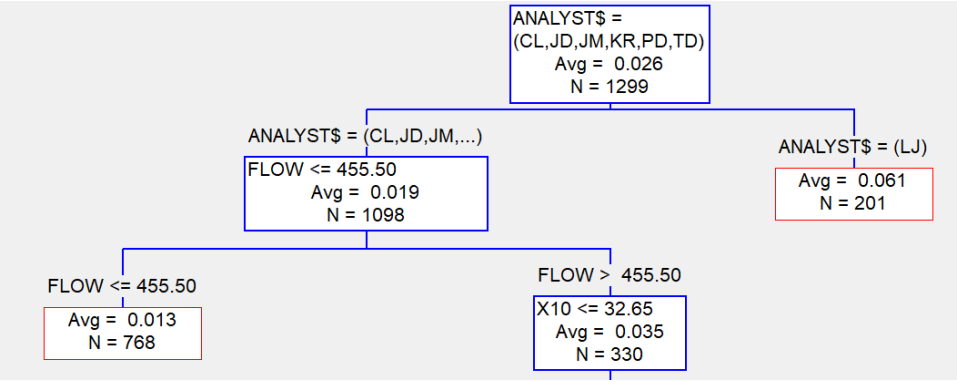
Minitab
© 2018 Minitab, Inc.

CART

What is CART

CART is an acronym for **C**lassification **a**nd **R**egression **T**rees. CART is a decision tree algorithm that works by creating a set of yes/no rules that splits the response or target variable into partitions based on the predictor settings.

The splits on factor settings are performed sequentially, one factor at a time, in order of their influence on the target variable. The first split is on the single factor that has the strongest effect on separating the target variable into two groups. Each subsequent split further refines the target variable into more precise groupings. This partitioning of the data optimizes the accuracy in prediction of the target variable given new settings of the predictors. When the partitioning is complete, the resulting prediction is the target variable mean (continuous data) or event probability (categorical data) in the final grouping or node for the combination of predictor settings.



When to use CART

Use CART to:

- Identify the most important predictors of a response or target
- Discover combinations of predictor settings that are most likely to lead to a specific outcome
- Visualize your findings
- Create business rules that are easy to understand, use and apply to your process in real time

Why use CART

CART answers questions such as:

- What is the root cause of the defects in my process?
- What rules should I put in place to monitor my process that identify when to perform corrective action?

Minitab®
© 2018 Minitab, Inc.

Fitting a CART model

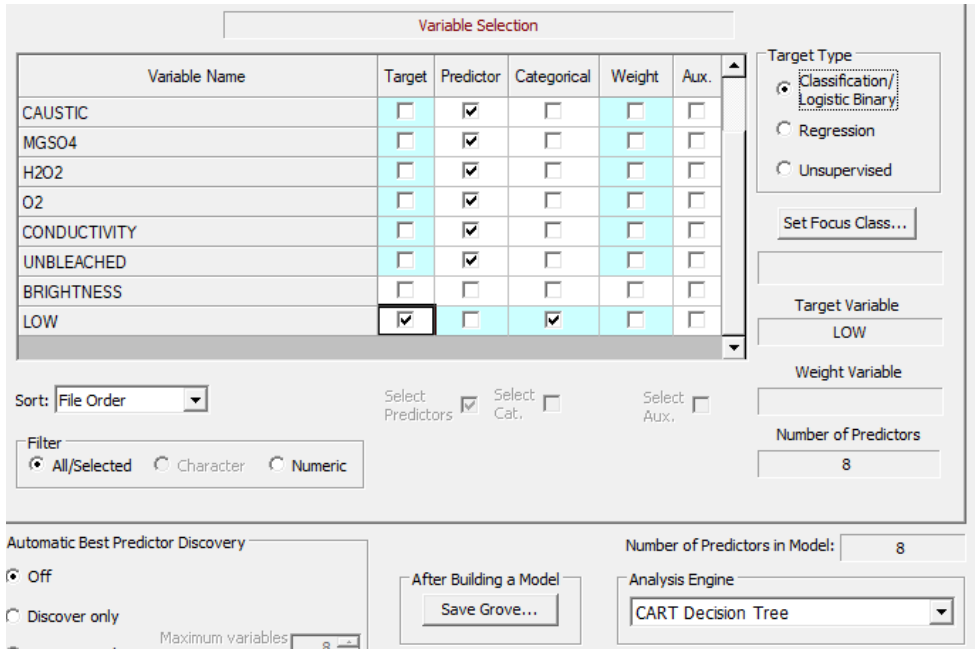
Recall that the LOW variable represents observations falling below the lower control limit on a control chart. These observations represent when the brightness measure fell below its expected range. Identifying the root cause associated with these out-of-control points allows us to determine the source of this excessive variation in the process.

LOW is a categorical target variable, so the **Target Type** is **Classification/Logistic Binary**. When you select LOW as your target, SPM will automatically check the **Categorical** box for this variable.

Because BRIGHTNESS is a response, not a predictor, you will leave this unchecked for this first model.

CART Model Setup

1. Choose **Model > Construct Model**.
2. From **Sort**, choose **File Order**.
3. From **Target Type**, choose **Classification/Logistic Binary**.
4. Check **Low** in the **Target** column.
5. Check the 8 predictors in the **Predictor** column.
6. From **Analysis Engine**, choose **Cart Decision Tree**.



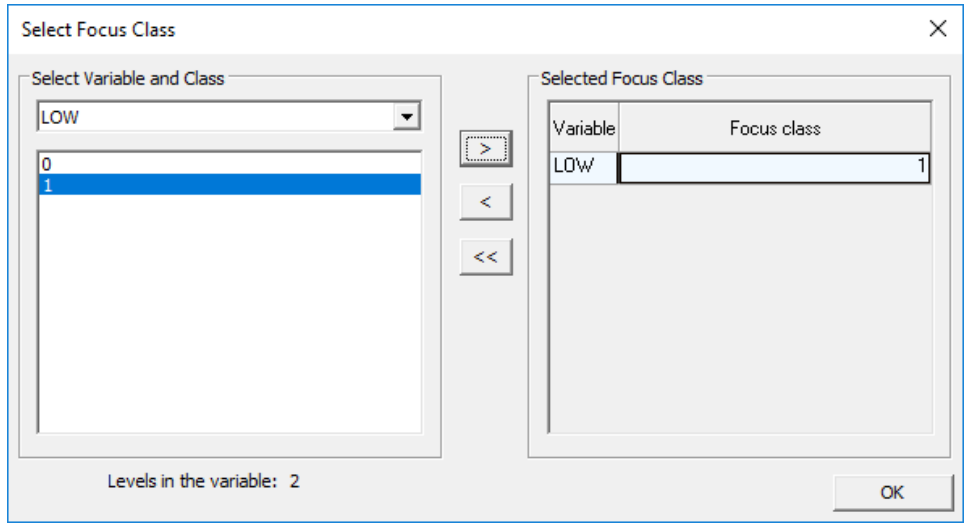
© 2018 Minitab, Inc.

Fitting a CART model

The target variable, Low, is an indicator variable where 1 indicates the observation was outside of statistical control. Because we are looking for the root cause of a point being outside of statistical control, 1 is the focus class for this target variable.

CART Model Setup (continued)

1. Click **Set Focus Class**.
2. Complete the dialog as shown below.



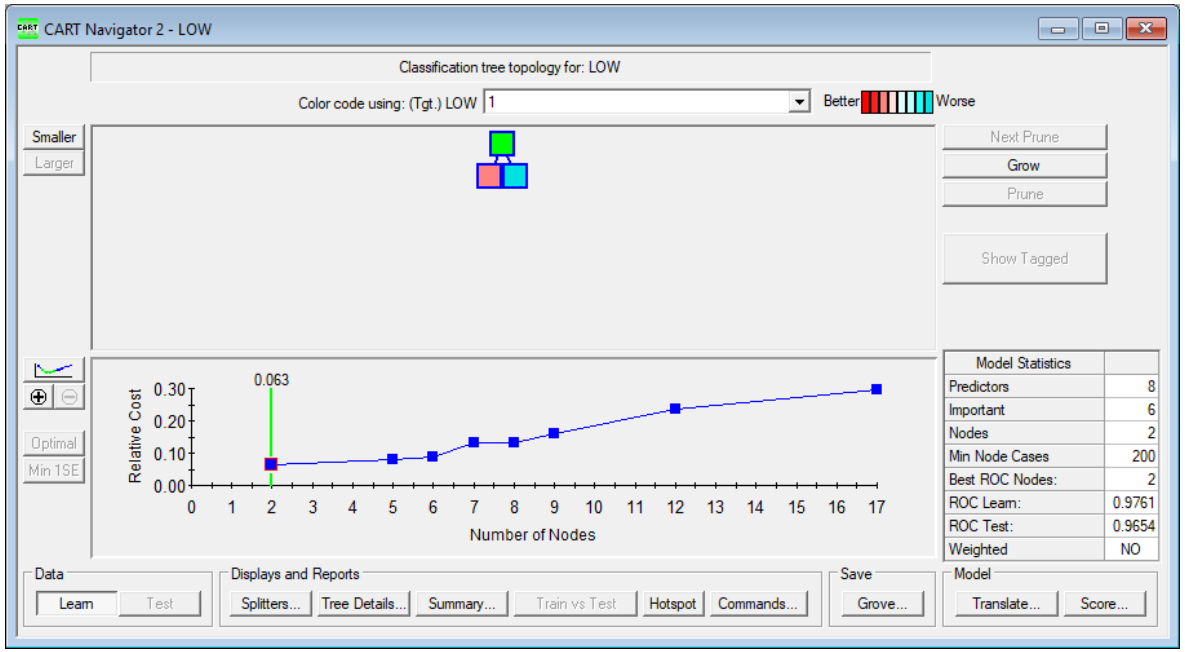
3. Click **OK**.
4. Click **Start**.

© 2018 Minitab, Inc.

Fitting a CART model

Recall that machine learning techniques involve splitting the data into learn and test sets where the test set is used to evaluate the model's predictive ability. By default, the CART model uses 10-fold cross validation. With 10-fold cross validation, the observations are randomly split into 10 groups or folds. To estimate the error in the prediction, 10 iterations are run where at each iteration, 9 of the 10 folds are used to create a CART decision tree and the remaining fold is held out as the test set to assess the model accuracy.

The "optimal" model is the model that minimizes the relative cost, or error rate of the tree with regard to predicting the value of the target for the holdout data. Here, the 2-node tree has the lowest relative cost.



© 2018 Minitab, Inc.

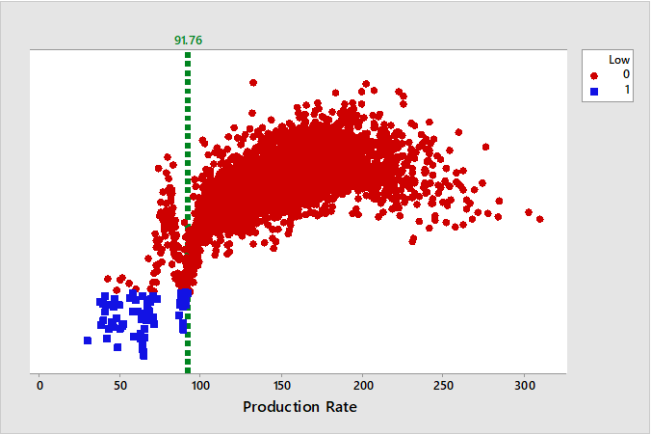
Viewing the tree

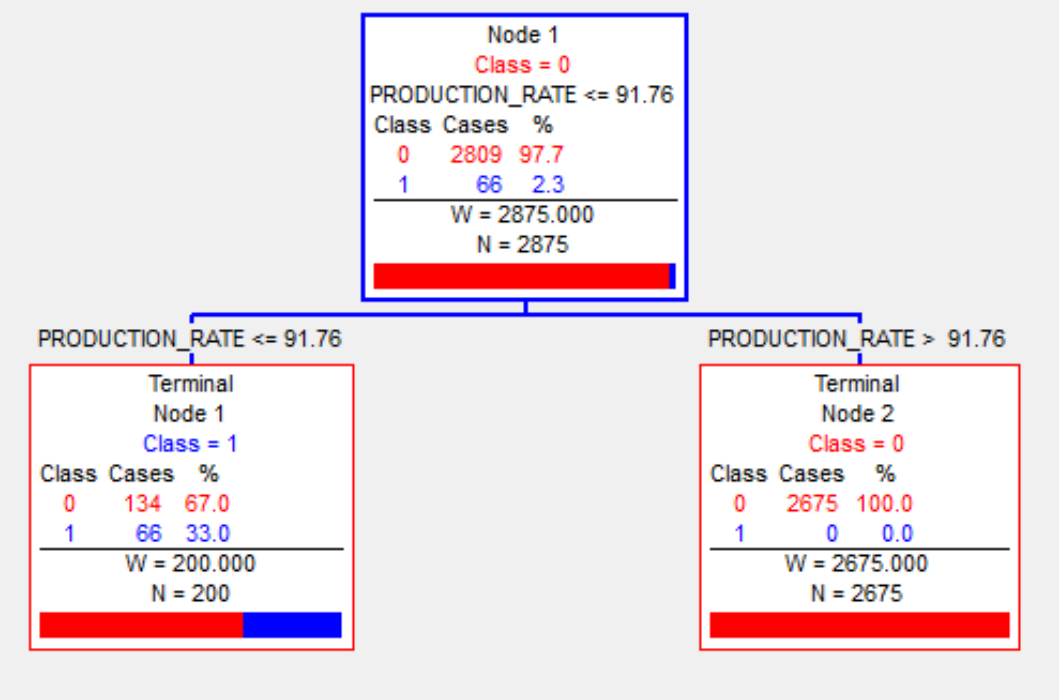
Under **Displays and Reports**, click **Tree details**.

The simple 2-node tree splits the data by Production Rate. All of the observations that are unusually low (outside of statistical control) had a production rate less than or equal to 91.76. So, our first rule is:

If production rate <= 91.76, then the estimated probability of the process being out of control is relatively high (33%). If production rate > 91.76, then the process is likely in statistical control.

The Minitab graph below explains why this rule works. The CART model finds the vertical line corresponding to production rate that best separates the Low = 0 from the Low = 1 group.



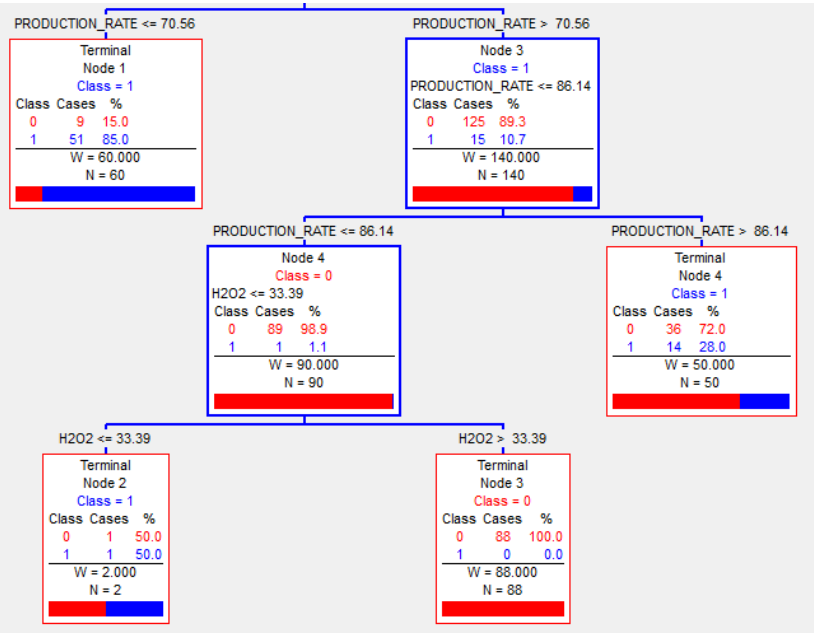


© 2018 Minitab, Inc.

Growing the CART model

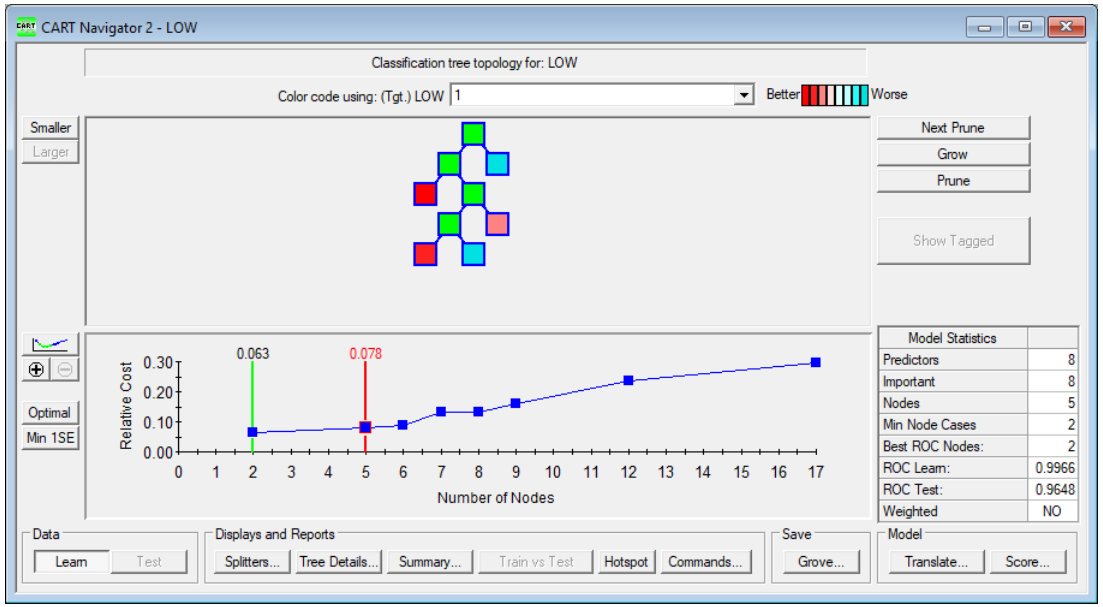
For these data, the model with the smallest error in prediction is the simple, 2-node model that splits on production rate. However, you can grow a larger tree for further investigation.

If you break down production rate even further, there is an 85% chance that the process will be out of control if the production rate is less than 70.56. Also, notice that the next new variable that CART splits is H2O2. Here, H2O2 only matters when the production rate falls between 70.56 and 86.14.



CART Model Grow

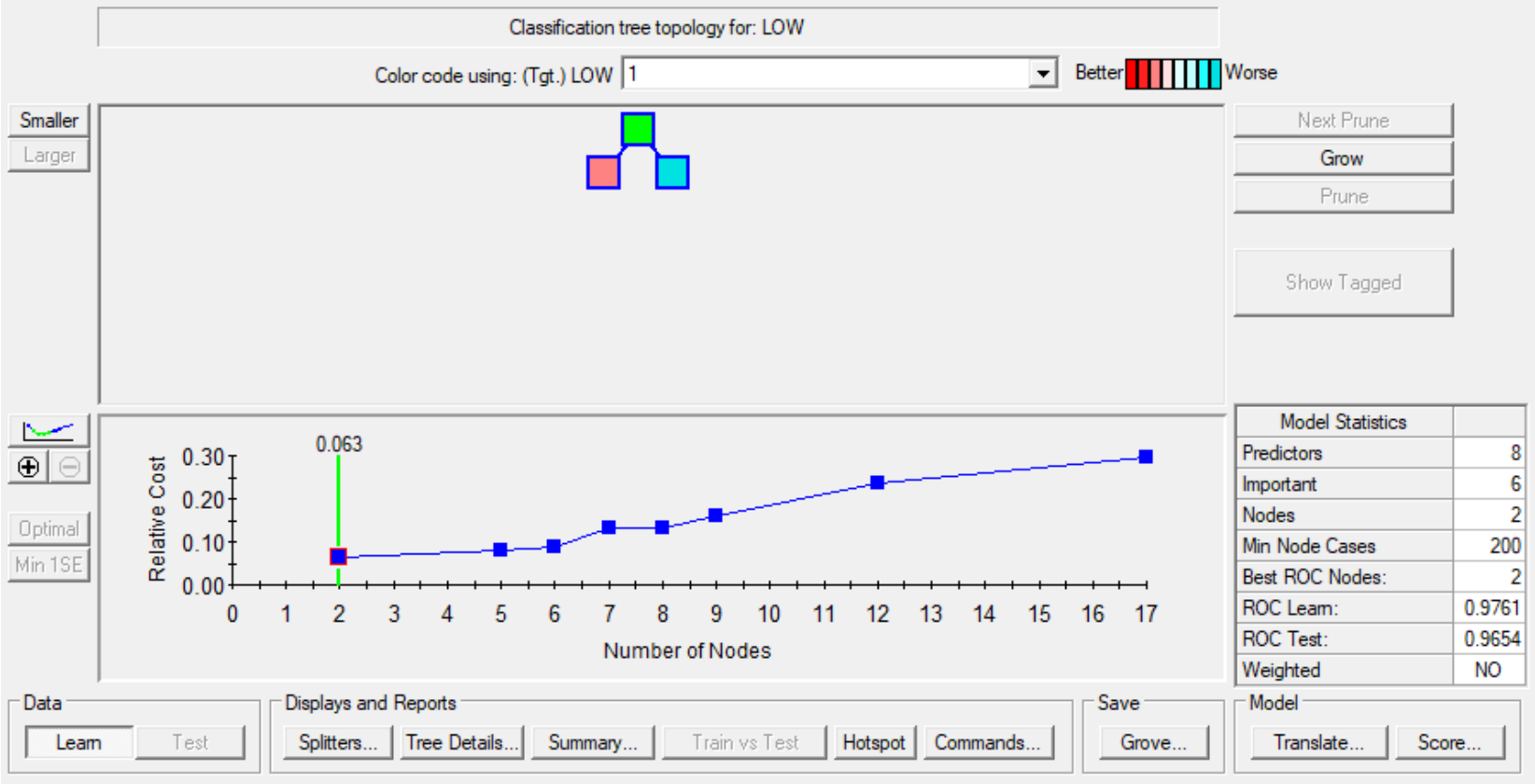
1. Return to the **CART Navigator** window. If you can't find it, choose **Window > Cart Navigator – LOW**.
2. Click **Grow**.
3. Click **Tree Details** if the tree isn't currently visible. You may want to arrange your **CART Navigator** and **Tree Details** windows so that both are visible together on your screen.



Minitab
© 2018 Minitab, Inc.

Pruning the CART model

Click **Prune** to return to the 2-node tree. This is the Tree we will use to more thoroughly assess the model's predictive ability.



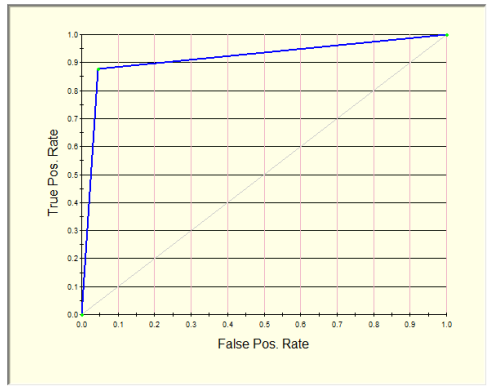
Minitab
© 2018 Minitab, Inc.

Assessing the model

The 10-fold cross-validation procedure allows for evaluation of model accuracy based on hold-out data not used to create the model. The **Summary** tab displays popular measures of model accuracy.

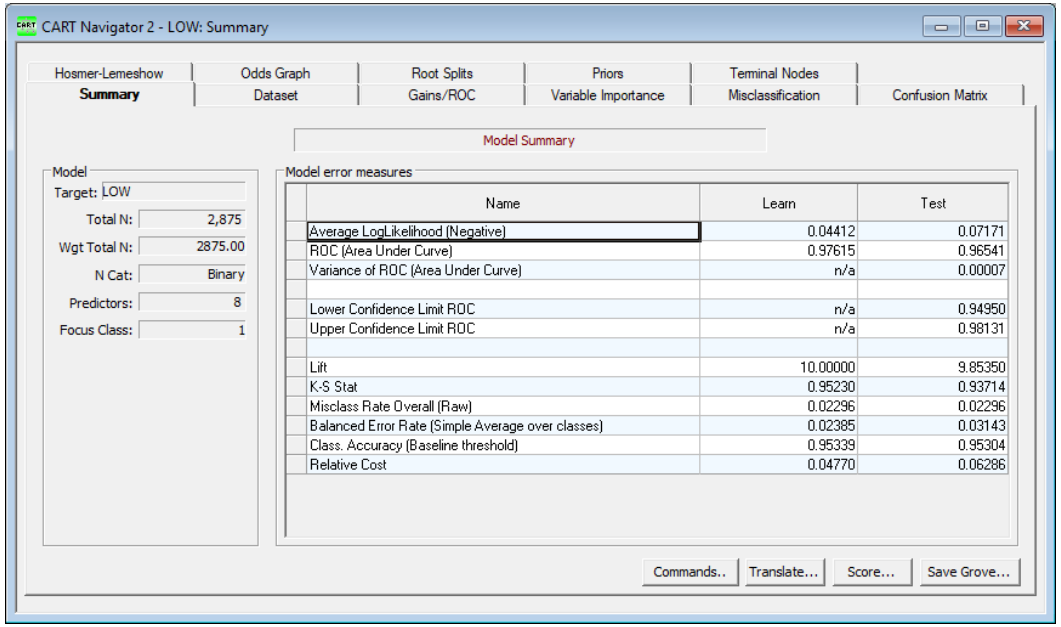
Two popular measures of model accuracy are the area under the Receiver Operating Characteristics (ROC) curve and the lift. These values compare how the classification rules from the model compare to simply classifying observations at random. You can visualize these values in the ROC graph of the true positive rate vs. the false positive rate. (This graph appears on the **Gains/ROC** tab.) If observations were assigned to groups (LOW = 0 or LOW = 1) at random, the graph of the these values would be a straight line at a 45 degree angle.

- ► The ROC graph displays the trade-off between benefits (true positives) and costs (false positives)
- ► Lift is the “lift” of the curve above the random (straight-line) model



Cart Tree Summary

1. Under **Displays and Reports**, click **Summary**.
2. Notice the **Model error measures** as shown below.



Name	Learn	Test
Average LogLikelihood (Negative)	0.04412	0.07171
ROC (Area Under Curve)	0.97615	0.96541
Variance of ROC (Area Under Curve)	n/a	0.00007
Lower Confidence Limit ROC	n/a	0.94950
Upper Confidence Limit ROC	n/a	0.98131
Lift	10.00000	9.85350
K-S Stat	0.95230	0.93714
Misclass Rate Overall (Raw)	0.02296	0.02296
Balanced Error Rate (Simple Average over classes)	0.02385	0.03143
Class. Accuracy (Baseline threshold)	0.95339	0.95304
Relative Cost	0.04770	0.06286

Minitab
© 2018 Minitab, Inc.

Assessing the model

Cart Tree Misclassification

1. Click the **Misclassification** tab.
2. Notice the **Model error measures** as shown below.

Misclassification

Learn Sample

Class	N Cases	N Mis-Classed	Pct. Error	Cost
0	2,809	134	4.77%	0.04770
1	66	0	0.00%	0.00000

Test Sample

Class	N Cases	N Mis-Classed	Pct. Error	Cost
0	2,809	134	4.77%	0.04770
1	66	1	1.52%	0.01515

- ► The data set contained 2875 observations. LOW = 0 (data point in statistical control) for 2809 observations. Low = 1 (data point outside of statistical control for 66 observations).
- ► 134 observations were misclassified in the learn sample. All of these had true values of Low = 0 and were classified as Low = 1.
- ► 135 observations were misclassified in the test sample created via hold-out data from the 10-fold cross-validation procedure. 134 of these actually had values of Low = 0 and were misclassified as Low = 1. Only 1 observation was outside the control limits (Low = 1) and misclassified as Low = 0.

For the pulp brightness data, we are interested in determining the root cause of the process being out of control, so we are most interested in “catching” the cases where Low = 1. The CART Model only missed one out-of-control observation in the holdout data.

Minitab®
© 2018 Minitab, Inc.

Assessing the model

A confusion matrix is another way to view the misclassification.

- 2675 observations were correctly classified as Low = 0
- 1 observation was incorrectly classified as Low = 0
- 65 observations were correctly classified as Low = 1
- 134 observations were incorrectly classified as Low = 1

The CART model classified 98.48% of the focus class (Low = 1) correctly.

Cart Tree Confusion Matrix

1. Click the **Confusion Matrix** tab.
2. Notice the **Predicted Classes** shown below.

Confusion Matrix

Actual Class	Total Class	Percent Correct	Predicted Classes 0 N = 2676	Predicted Classes 1 N = 199
0	2,809	95.23%	**2,675**	134
1	66	98.48%	1	**65**
Total:	2,875			
Average:		96.86%		
Overall % Correct:		95.30%		
Specificity		95.23%		
Sensitivity/Recall		98.48%		
Precision		32.66%		
F1 statistic		49.06%		

Minitab®
© 2018 Minitab, Inc.

Assessing variable importance

CART Tree Variable Importance

1. Click the **Variable Importance** tab.
2. Notice the relative importance of the predictors.

Variable Importance

Variable	Score
PRODUCTION_RATE	100.00
UNBLEACHED	3.56
O2	2.70
CAUSTIC	1.77
H2O2	1.46
CONDUCTIVITY	1.06

The **Variable Importance** score displays the relative importance of each variable in the current decision tree model. For these data, production rate was a much stronger predictor of an out-of-control point than any of the remaining potential predictors.

Now we will compare the results from this simple CART model to a more sophisticated TreeNet Gradient Boosting model using Brightness as the target variable.

Minitab®
© 2018 Minitab, Inc.

TreeNet Gradient Boosting

What is TreeNet Gradient Boosting

TreeNet is a gradient boosting algorithm that sequentially fits smaller CART trees, then uses the error from the previous tree to create a new prediction and update the model. The approach of fitting a model to the errors from a previous model is known as boosting. The amount of information carried from one model to the next is controlled by the learn rate. A smaller learn rate means the model learns more slowly because there is less weight given to the information from the previous model.

When to use TreeNet Gradient Boosting

When fitting a single CART tree with many nodes, there is a tendency for the tree to adapt very closely to the learn data (low bias), but not predict well for new data (high variance). The purpose of gradient boosting is to improve predictive accuracy by fitting a sequence of small trees with only a few nodes that have high bias, but low variability. Because each iteration “learns” from the mistakes (or errors) from the previous iteration, the bias eventually dies out making the final model exceptionally accurate in its predictions.

The disadvantage to using TreeNet over CART is the final model is a “black box” and not easily interpretable. However, with a TreeNet model you can still assess variable importance and the nature of the relationships between specific predictor variables and the target variables.

Why use TreeNet Gradient Boosting

TreeNet and CART models solve the same problems, but:

- CART models are easier to interpret and visualize
- TreeNet models are typically superior with regard to accuracy in prediction

It is common to use both techniques together to more fully understand your process.

TreeNet Gradient Boosting answers questions such as:

- What process variables have the most influence on my response or target variable?
- What is the nature of the relationship between key predictors and the target variable?
- What interactions exist between key predictor variables?
- What settings of the predictors appear to optimize the target variable?

Minitab®
© 2018 Minitab, Inc.

Fitting a TreeNet model

If **Model** does not appear in your menu, try closing the window you are in. The menu items change depending on which window is currently open in the application

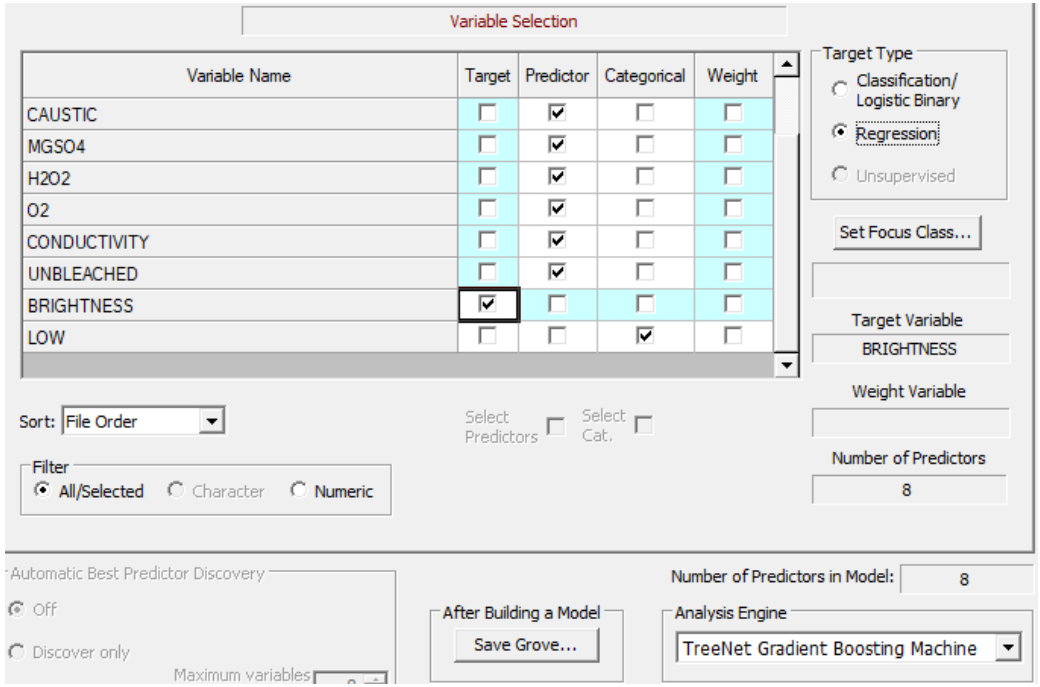
You can use TreeNet to fit a categorical target such as the variable LOW that we used in the previous example, but the goal now is to investigate the process more closely using the specific value of Brightness.

In general, both CART and TreeNet:

- ► Handle both categorical and continuous response or target variables
- ► Handle both categorical and continuous predictor variables

TreeNet

1. Close the **CART Summary** window.
2. Choose **Model > Construct Model**.
3. Under **Target Type**, choose **Regression**.
4. Under **Target** uncheck **LOW**, then check **Brightness**.
5. Verify that all 8 predictors are checked.
6. From **Analysis Engine**, choose **TreeNet Gradient Boosting Machine**.



Minitab
© 2018 Minitab, Inc.

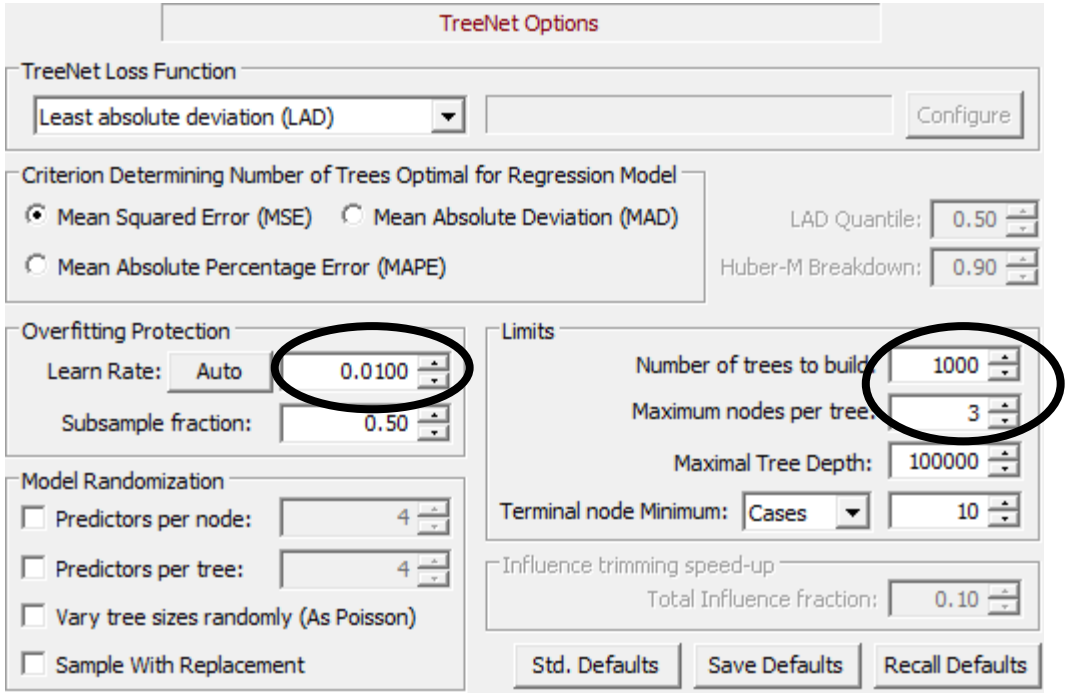
Fitting a TreeNet model

TreeNet models have several options that you may want to tweak. These options are called tuning parameters. You can also use the features in the Automate tab to let SPM find the optimal tuning parameters for you.

Because this data set is relatively small with only 8 predictors, a higher learn rate and smaller tree size works slightly better than the default settings. It is also reasonable to increase the number of trees to build because of the relatively small data set size. With very large data sets, the computation time becomes a consideration when modifying some of the tuning parameters.

TreeNet

1. Click the **TreeNet** tab.
2. In **Learn Rate**, enter 0.01.
3. In **Number of trees to build**, enter **500**.
4. In **Maximum nodes per tree**, enter 3.



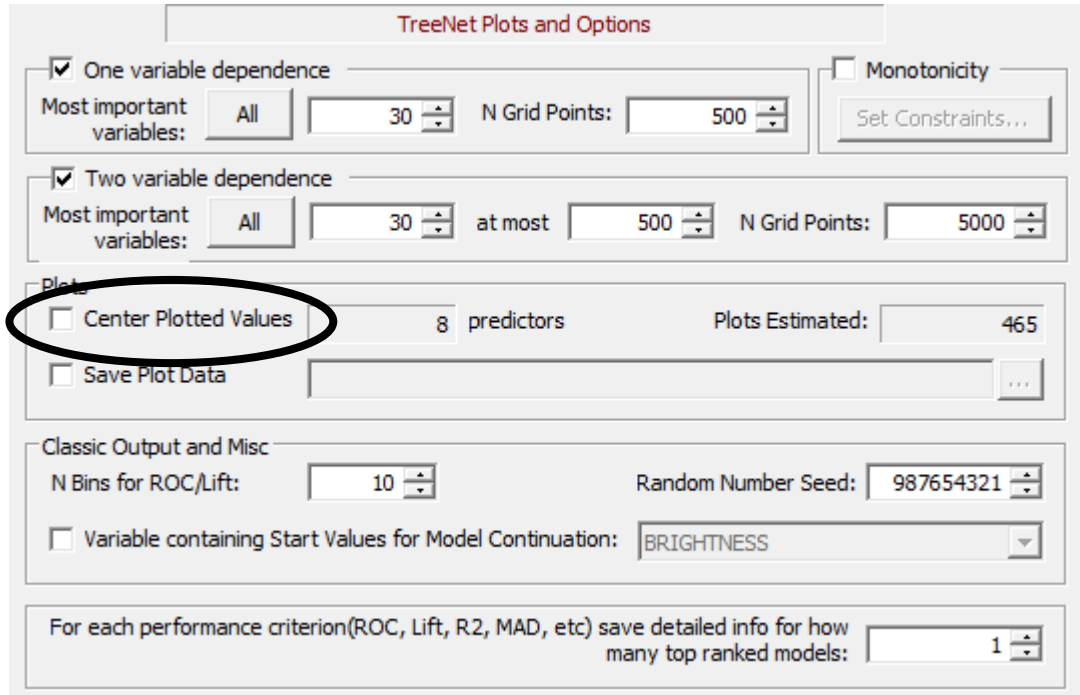
Minitab
© 2018 Minitab, Inc.

Fitting a TreeNet model

Centering the plotted values will let you compare the strength of the effects of different predictors. When the values are centered, they will no longer be in units of the target variable. For the pulp bleaching data, we will not center the plotted values so that we can see the relationships between Brightness and specific predictors in the units of the data.

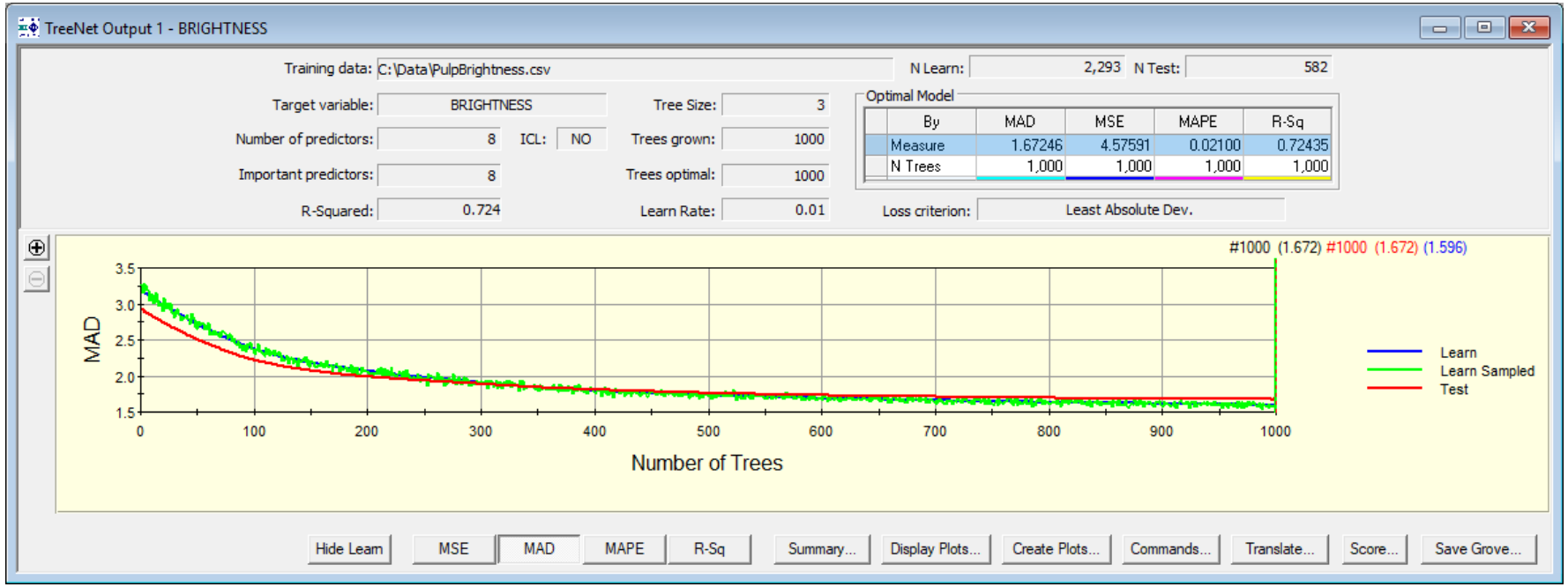
TreeNet

1. Click the **Plots&Options** tab.
2. Under **Plots**, uncheck **Center Plotted Values**.
3. Click **Start**.



Minitab
© 2018 Minitab, Inc.

Assessing the TreeNet model



By default, TreeNet separates the observations into a learn and test set at random with 20% of the observations held out in the test set that is not used to create the model. The prediction error compares the actual observed value to the prediction obtained when the observation is not used to create the model. These differences, or errors, can be measured as a mean square error (MSE), a mean absolute deviation (MAD), or a mean absolute percent error (MAPE). The R-Sq value also measures how well the model predicts for new observations by determining how well the model fits for the hold-out data. The best model in terms of accuracy in prediction maximizes R-Sq and minimizes MSE, MAD, and MAPE.

Note that because of the random components to the TreeNet algorithm, your results will not match these exactly, but they should be close.

Minitab
© 2018 Minitab, Inc.

TreeNet Model Variable Importance

TreeNet Model Variable Importance

1. Click **Summary.**
2. Choose the **Variable Importance** tab.

Variable Importance

Variable	Score
PRODUCTION_RATE	100.00
H2O2	63.42
CAUSTIC	36.06
O2	8.58
DISCHARGE_PH	7.03
UNBLEACHED	6.90
MGSO4	6.70
CONDUCTIVITY	6.50

Production Rate remains the most important predictor with respect to the variation in Brightness. Additionally, H2O2 and Caustic appear to have strong influence on the Brightness.

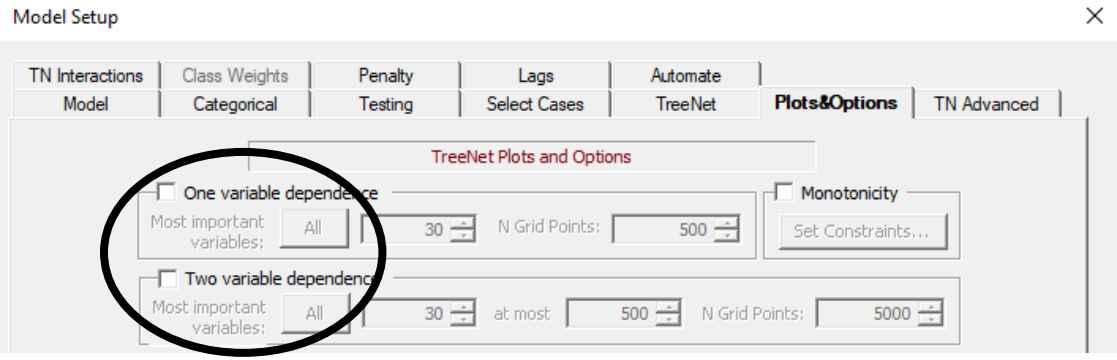
We can use the Partial Dependency Plots to investigate these relationships further.

Minitab®
© 2018 Minitab, Inc.

TreeNet Partial Dependency Plots

The partial dependency plots allow you to visualize the nature of the relationship between each predictor variable and the response when the other predictors are held constant.

By default, TreeNet executes the calculations for the one-variable and two-variable dependency plots for the most important variables when you fit the model. However, for larger data, you may want to turn this feature off during the model fitting stage in the **Plots&Options** tab of the **Model Setup** dialog to reduce computation time. Instead, you can create these plots after model fitting is complete using the **Create Plots** button in the TreeNet output window.

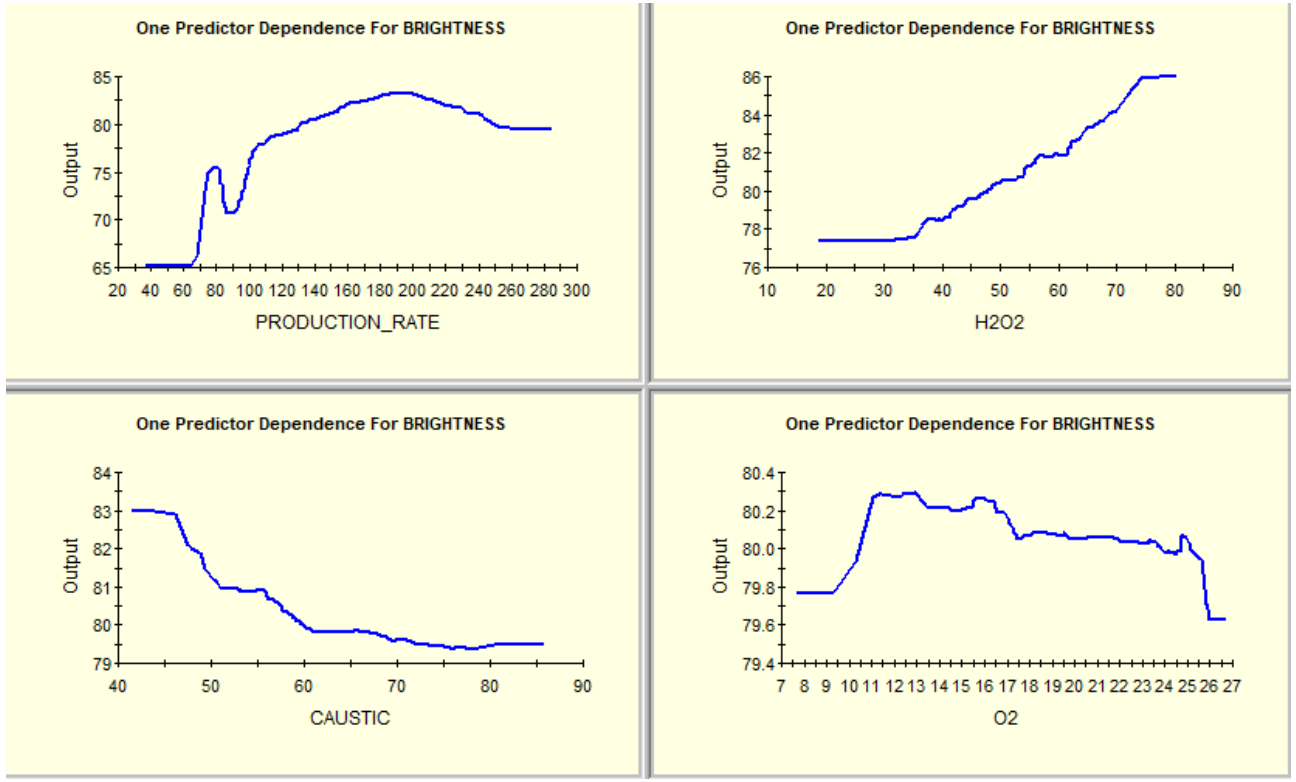


TreeNet Model Display Plots

1. Close the Summary window and return to the TreeNet window. (Choose **Window > TreeNet Output -- Brightness.**)
2. Click **Display Plots**.
3. **Under One Variable Dependence for** BRIGHTNESS, double-click **Trees = 1000 (Optimal)**.

Minitab
© 2018 Minitab, Inc.

TreeNet Partial Dependency Plots

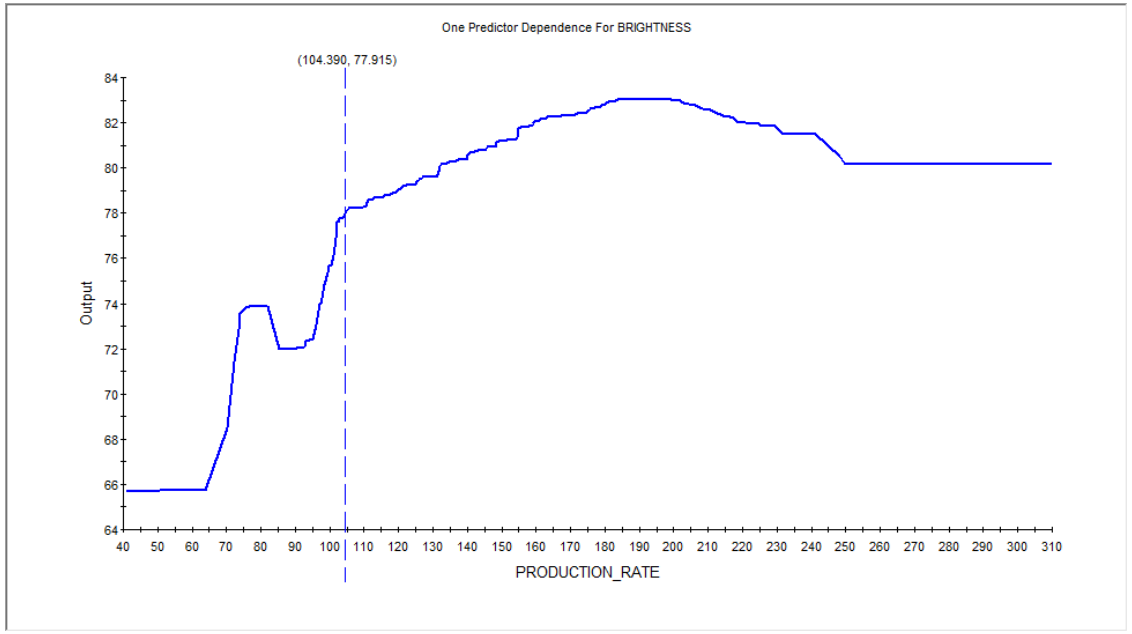


The partial dependency plots allow you to investigate the affect of one predictor on the target variable while holding the other predictors constant. These are displayed in order of importance. Notice that the Y-scale is different for each variable. As you can see by the Y-scale, production rate contributes much more to the variation in Brightness than any other variable. On the other hand, the variation in Brightness due to O2 is relatively minimal. This is consistent with what we found in the Variable Importance graph.

Minitab®
© 2018 Minitab, Inc.

TreeNet Partial Dependency Plots

Double-click on the partial dependency plot for PRODUCTION_RATE to interact with this graph.



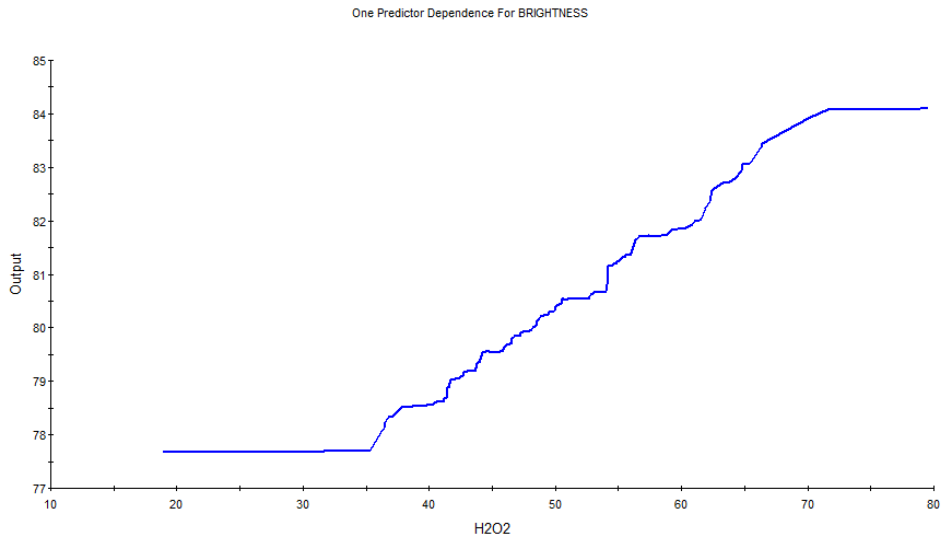
Notice that:

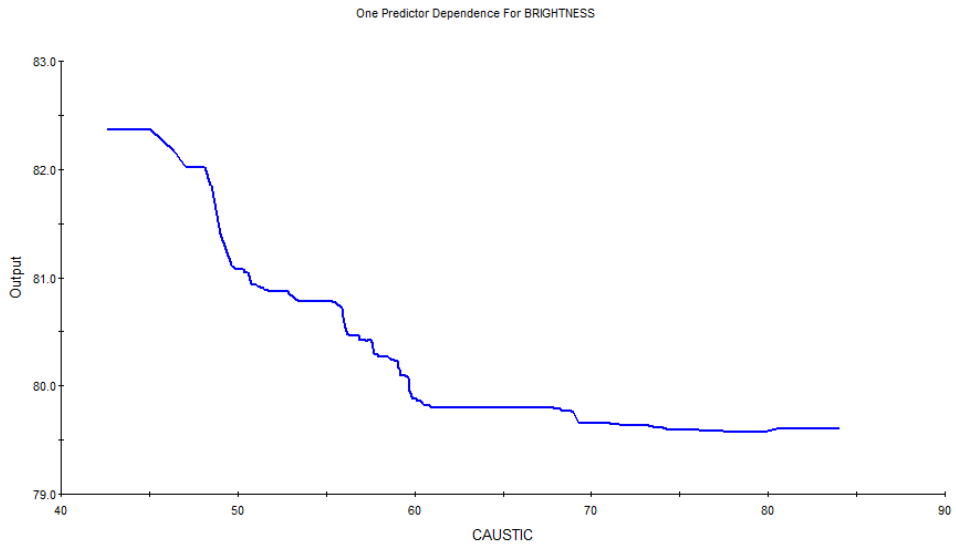
- The relationship is very nonlinear, so a traditional regression model would likely not fit well.
- Maintaining a production rate roughly above 105 would likely reduce the variation in Brightness.
- A production rate of around 130 would result in a target value = 80 for brightness, assuming all other variables could be held constant and no interactions between variables exist.

Minitab®
© 2018 Minitab, Inc.

TreeNet Partial Dependency Plots

Double-click on the partial dependency plots for H2O2 and CAUSTIC





Notice that:

- The relationship between brightness and H2O2 is roughly linear within a certain range.
- Brightness tends to decrease until caustic reaches a value around 60, then tends to level off. Maintaining caustic above 60 would likely reduce the variation in brightness.
- To achieve a target Brightness of 80, H202 should be set around 48 and caustic should be kept above 60, assuming other variables can be held constant and no interactions exist between these variables.

Minitab®
© 2018 Minitab, Inc.

Final considerations

Summary and conclusions

The CART decision tree model identified production rate as being the predictor that has the most influence on a brightness falling below its expected range on the control chart (LOW = 1). Monitoring this variable more closely in real time to insure that it remains above 91.76 should greatly reduce the variation in the process.

The TreeNet model used process data to find the variables that have the most impact on Brightness. Specifically, production rate was by the far the largest contributor to the variation in brightness, but H2O2 and caustic seem to have influence as well.

A follow-up designed experiment (DOE) should now be done using the factors Production Rate, H2O2, and Caustic to find their optimal settings and investigate potential interactions between those variables. The partial dependency plots from TreeNet provide good information about where to center this experiment as well as where to set the levels for these variables to quickly find the optimal settings for hitting a target value for brightness.

Additional considerations

A CART decision tree is a good way to start investigating process data. While a more sophisticated TreeNet model is likely to result in better predictive accuracy, the CART tree visualization and its corresponding simple rules are easy to interpret and explain.

To further increase the predictive accuracy of your model, try using some of the automates on the **Automate** tab. While many of the default settings are good choices, you can automate the selection of various tuning parameters to quickly improve the accuracy of your predictions.

Minitab®
© 2018 Minitab, Inc.