

Six Sigma, Data Mining, and the JMP[®] Partition Platform

World Conference on Quality and Improvement
May 16, 2005

Philip J. Ramsey, Ph.D., Mia L. Stephens, MS, Marie Gaudard, Ph.D.

North Haven Group, <http://www.northhavengroup.com>

pramsey@northhavengroup.com, 603-672-5651

mstephens@northhavengroup.com, 207-363-5739

mgaudard@northhavengroup.com, 352-560-0312

About NHG

North Haven Group (NHG) is a limited liability company registered in the state of New Hampshire, providing comprehensive consulting and training for industry and service organizations.

NHG provides consulting and support for Six Sigma quality programs to improve manufactured products and business processes.

With over 50 years of combined experience, the partners of NHG provide a unique combination of outstanding academic credentials and expertise in the application of statistical techniques and continuous improvement methods.

About NHG

NHG specialties include:

- Six Sigma implementation and deployment
- Six Sigma Champion, Blackbelt, and Greenbelt certification
- Lean Six Sigma training and consultation
- Design of Experiments (DOE)
- Failure Mode Effects Analysis (FMEA)
- Measurement Systems Analysis (MSA)
- Process Stabilization and Control
- Continuous Improvement and Problem Solving
- Data Mining

Talk Outline

- Introduction
- What is Data Mining?
- JMP[®] and Data Mining
- Traditional Statistics vs Data Mining
- Partitioning
- Example: The Press Band Data
- Partitioning and Six Sigma Training
- Case Studies
- Summary

What is Data Mining?

Data mining refers to a collection of techniques used primarily to analyze large observational data sets.

Unlike traditional “statistical” techniques, data mining methods do not utilize hypothesis testing for assessing model fit or decision-making.

Data mining methods include neural nets, classification and regression trees, clustering algorithms, as well as other methods.



John Wilder Tukey
1915 - 2000



*We are drowning in information
and starving for knowledge.*

- Rutherford D. Roger

What Is Data Mining?

Data Mining is the analysis of large observational datasets with the goal of finding unsuspected relationships.

By “large” datasets, we mean either a *large number of records*, or a *large number of variables measured on each record*.

The other key word in the definition is “observational”.

- Often, data sets used in data mining were collected for purposes other than those of the data mining study.
- Consequently, data mining datasets usually consist of convenience samples, rather than random, samples.

What Is Data Mining?

The initial applications of data mining were in customer research -
Who buys what? If customers buy A, are they likely to buy B
(market basket analysis)?

Note that biological research using ‘microarrays’ results in datasets
with large numbers of variables, where data mining techniques
can be extremely useful.

Such large data sets are characterized by complex observations.

They require extensive pre-processing.

Data sets with hundreds of thousands of variables have been
analyzed using data mining techniques.

What Is Data Mining?

The types of relationships that one seeks in a data mining study can be categorized into two main structures:

- Global models, or
- Local patterns.

A **global model** defines a structure that applies (globally) to all points in the data set.

Typical examples of models are:

- Predictive models
- Classification models

What Is Data Mining?

A **local pattern** is a relationship that applies in a restricted region of the variable values. For example,

- In a marketing study, one might learn that 90% of customers who buy a high-end yogurt product also buy high-end ice cream.
- In a study of accounts receivable data, one might learn that a certain group of customers do not fit the general pattern in terms of payment and returns.

(Detecting deviations from the general pattern is called *anomaly detection*, and is useful in studies of fraud.)

What Is Data Mining?

Types of algorithms typically associated with data mining include:

- Multiple linear and logistic regression,
- Classification and regression trees,
- Neural nets,
- Clustering algorithms, and
- Association rules.

However, other techniques are often used as well, such as:

- Extensive display and visualization tools,
- Variable reduction techniques, and
- Bayesian methods.

JMP® and Data Mining

JMP® provides the user with a number of these data mining tools for **global modeling**:

- Multiple linear and logistic regression
- Classification and regression trees (the Partition platform)
- Neural nets

For **local modeling**, JMP provides a cluster analysis platform.

In terms of **visualization**, almost all JMP analyses are supported by extensive display and visualization tools.

JMP[®] and Data Mining

JMP[®]'s Neural Net platform fits a neural net with one hidden layer to a continuous or nominal response.

JMP[®]'s Partition platform is a classification and regression tree-fitting methodology.

Other tree-fitting methodologies, found in high-end and very expensive data mining packages, are CART[™], CHAID[™], C5.0.

Since convenience data sets are often messy and unruly, JMP's display capabilities support the user in data-cleaning.

Traditional Statistics Vs Data Mining

In classical statistical modeling, the complete set of data is used to develop the model.

Significant model predictors are chosen based on hypothesis testing.

The quality of model predictions is assessed using prediction intervals.

Overfitting is prevented through the use of statistical tests and diagnostics based on the underlying model assumption.

Data mining techniques such as classification and regression tree analysis and neural nets **do not support hypothesis testing.**

Traditional Statistics Vs Data Mining

Models based on data mining techniques are usually validated on independent data.

Often, the complete data set is split into a *training* or *development data set* and an *evaluation data set*.

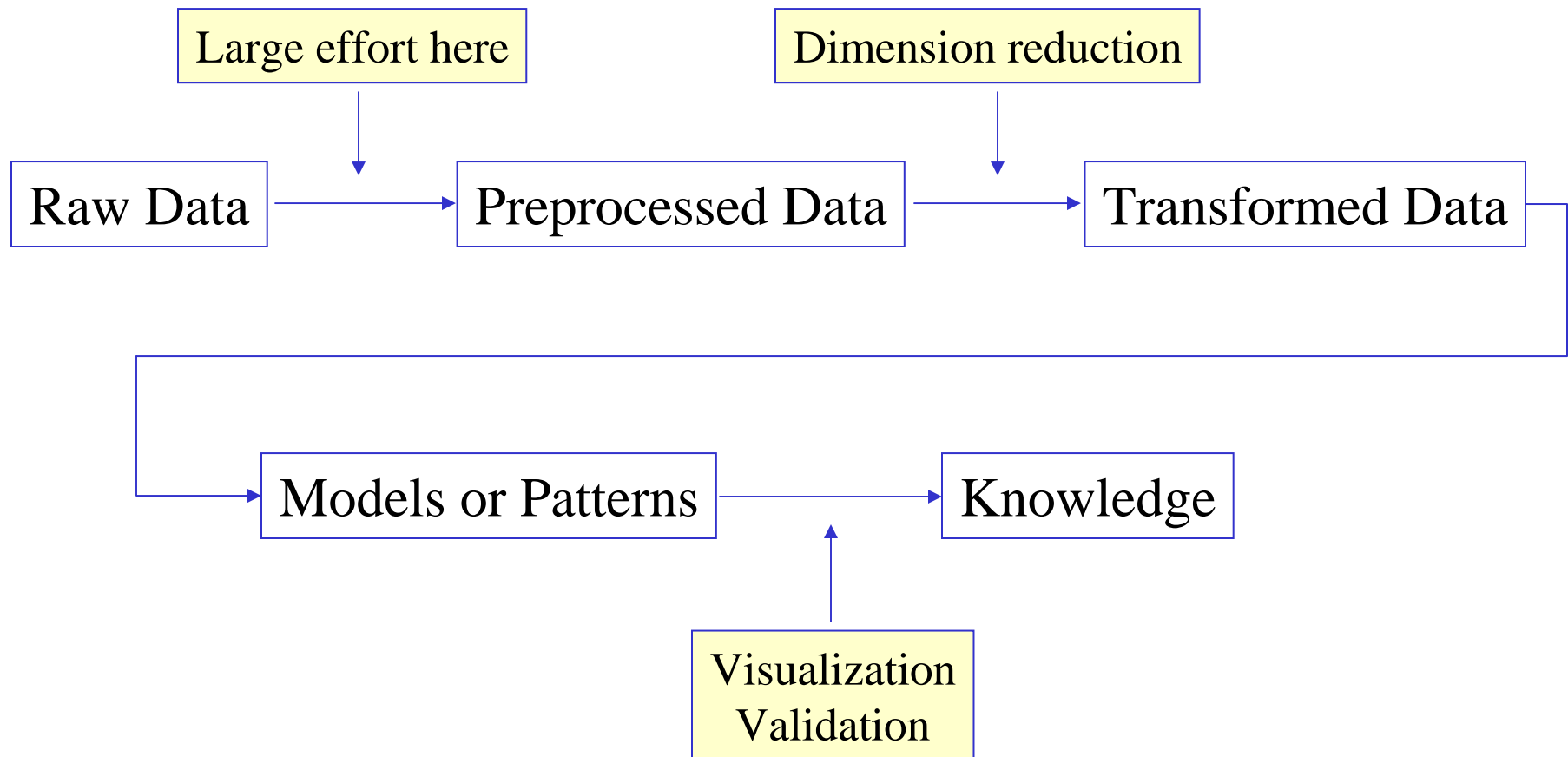
Models are developed using the development data, and they are evaluated on the evaluation data (often a 70/30 or 60/40 split).

With large datasets, it becomes very easy to model noise.

Since most data mining analyses are predictive, it is important not to model idiosyncrasies of the training data.

Use of an independent evaluation data set guards against overfitting.

Traditional Statistics Vs Data Mining



Partitioning

JMP's Partition platform is a version of Classification and Regression Tree Analysis.

Both response and factors can be either continuous or nominal.

Continuous factors are **split** into **two partitions according to cutting values**.

Nominal factors are **split** into **two groups of levels**.

If the **response is continuous**, the splits are constructed so as to maximize the separation of the two groups, as measured by the sums of squares due to the differences between means.

In this case, the fitted values are the means within groups.

Partitioning

If the **response is categorical**, the splits are determined by maximizing the likelihood ratio chi-square statistic, reported in the JMP output as “ G^2 ”, or a related value called the LogWorth.

In this case, the fitted values are the estimated probabilities within groups.

JMP’s Partition platform is extremely useful for both exploring relationships and for modeling.

JMP’s Partition platform provide only a minimal criterion to determine when to stop building a tree (a **stopping rule**). We have actually found this advantageous.

Example: The Press Band Data

We will illustrate JMP's partition platform using a dataset from the rotogravure printing business.

In this printing process:

- An engraved copper cylinder is rotated in a bath of ink,
- Excess ink is removed,
- Paper is pressed against the inked image,
- Once the job is complete, the engraved image is removed from the cylinder, and the cylinder is reused.

Example: The Press Band Data

A defect called banding can sometimes occur, ruining the product.

Banding consists of grooves that appear in the cylinder at some point during the print run.

Once detected, the run is halted, and the cylinder is removed and repaired.

This process can take from one-half to six hours.

Understanding the conditions that lead to banding is critical and could save a printer enormous amounts of money.

Example: The Press Band Data

We will use a dataset that contains observational data on banding.

This dataset can be found at <http://ftp.ics.uci.edu/pub/machine-learning-databases/> and is called `cylinder-bands`.

The dataset consists of 540 records and 39 variables.

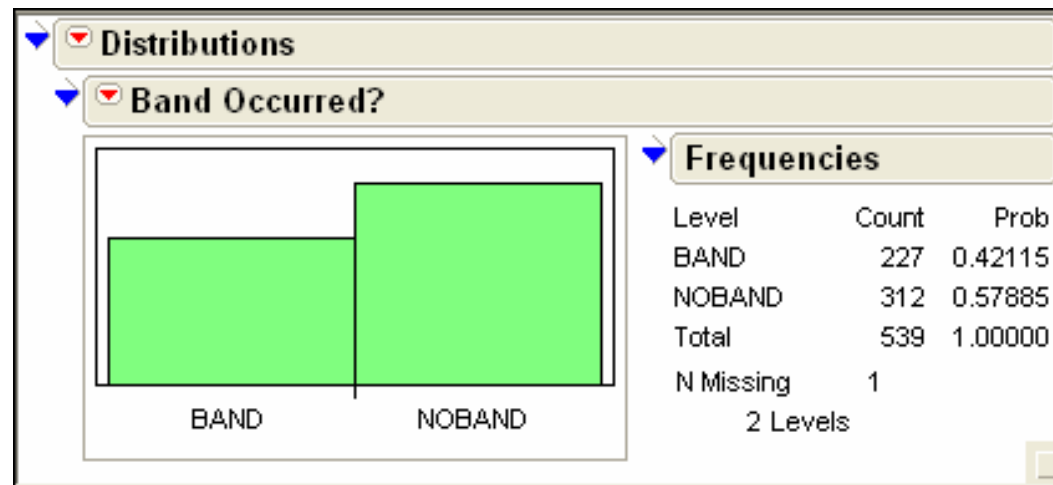
Part of the dataset is shown on the following slide.

The target variable is “Band Occurred?”, whose values are “BAND” and “NOBAND”.

| PressBandingClean2 | | 37.0 Cols | | | | | | |
|--------------------|---|------------|----------|------------|--------------|--------------|----------------|--|
| Partition | | Date | Date M/Y | Job Number | Cylinder No. | Customer | grain screened | |
| 1 | • | 03/30/1990 | 03/1990 | 23040 | X750 | GUIDEPOSTS | YES | |
| 2 | • | 04/09/1990 | 04/1990 | 34683 | G467 | ECKERD | NO | |
| 3 | • | 04/09/1990 | 04/1990 | 25416 | X203 | TVGUIDE | YES | |
| 4 | • | 04/14/1990 | 04/1990 | 34545 | O21 | TARGET | NO | |
| 5 | • | 04/17/1990 | 04/1990 | 36858 | T313 | EXXON | YES | |
| 6 | • | 04/18/1990 | 04/1990 | 36053 | J68 | WARDS | NO | |
| 7 | • | 04/18/1990 | 04/1990 | 36053 | J42 | WARDS | NO | |
| 8 | • | 04/18/1990 | 04/1990 | 36858 | F329 | EXXON | YES | |
| 9 | • | 04/25/1990 | 04/1990 | 34664 | G496 | BURDINES | YES | |
| 10 | • | 04/26/1990 | 04/1990 | 34545 | O6 | TARGET | NO | |
| 11 | • | 04/26/1990 | 04/1990 | 34545 | O14 | TARGET | NO | |
| 12 | • | 05/05/1990 | 05/1990 | 47103 | T244 | MODMAT | YES | |
| 13 | • | 05/07/1990 | 05/1990 | 47103 | M93 | MODMAT | NO | |
| 14 | • | 05/07/1990 | 05/1990 | 47103 | M260 | MODMAT | YES | |
| 15 | • | 05/07/1990 | 05/1990 | 47103 | T383 | MODMAT | NO | |
| 16 | • | 05/07/1990 | 05/1990 | 47103 | T78 | MODMAT | YES | |
| 17 | • | 05/07/1990 | 05/1990 | 47103 | M4 | MODMAT | YES | |
| 18 | • | 05/07/1990 | 05/1990 | 36928 | M432 | HOMESHOPPING | NO | |
| 19 | • | 05/07/1990 | 05/1990 | 36928 | M257 | HOMESHOPPING | NO | |
| 20 | • | 05/09/1990 | 05/1990 | 47103 | F242 | MODMAT | YES | |
| 21 | • | 05/10/1990 | 05/1990 | 47103 | F672 | MODMAT | YES | |
| 22 | • | 05/11/1990 | 05/1990 | 47103 | M260 | MODMAT | YES | |
| 23 | • | 05/14/1990 | 05/1990 | 47103 | F679 | MODMAT | YES | |
| 24 | • | 05/17/1990 | 05/1990 | 36054 | X400 | WARDS | NO | |
| 25 | • | 05/17/1990 | 05/1990 | 34752 | X776 | TOYSRUS | NO | |
| 26 | • | 05/17/1990 | 05/1990 | 34752 | X713 | TOYSRUS | NO | |
| 27 | • | 05/18/1990 | 05/1990 | 34402 | I331 | AUSTADS | YES | |
| 28 | • | 05/24/1990 | 05/1990 | 36648 | F227 | JAMESWAY | NO | |
| 29 | • | 06/02/1990 | 06/1990 | 36859 | F590 | NATLWLDLIFE | YES | |
| 30 | • | 06/03/1990 | 06/1990 | 36859 | F670 | NATLWLDLIFE | YES | |
| 31 | • | 06/06/1990 | 06/1990 | 36859 | F331 | NATLWLDLIFE | YES | |
| 32 | • | 06/06/1990 | 06/1990 | 36859 | F571 | NATLWLDLIFE | YES | |

Example: The Press Band Data

The bar graph shown below, generated from the JMP Distribution Platform, indicates that banding occurred in 42% of jobs.



We would like to find which conditions lead to banding.

Example: The Press Band Data

Note that a Six Sigma team will often construct Pareto charts at this point, using data from the runs where banding occurred.

Note that these ignore information on when banding did NOT occur.

Also, the available “predictors” for banding consist of 11 nominal (categorical) variables and 18 continuous variables.

Stand-alone Pareto charts of the nominal variables could easily overlook complex relationships and interactions among these variables.

Example: The Press Band Data

Box plots might be used to explore the relationships between the continuous variables and “Band Occurred?”, but again, complex interactions would be ignored.

A logistic regression model could be constructed, but including all the predictors is not possible.

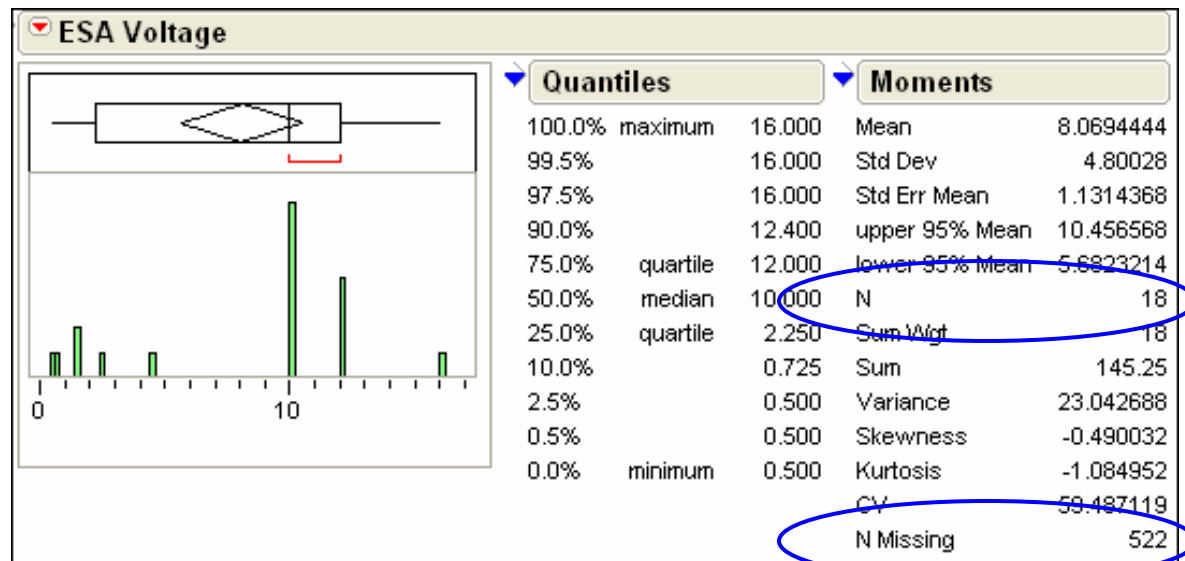
This is because of the many nominal variables, and the fact that numerous cells are not populated.

However, a classification tree, with “Band Occurred?” as target, can be easily constructed, and it can provide rich information on the conditions that lead to banding.

Example: The Press Band Data

We note, at this point, that observational data sets must always be examined for data integrity before they are analyzed.

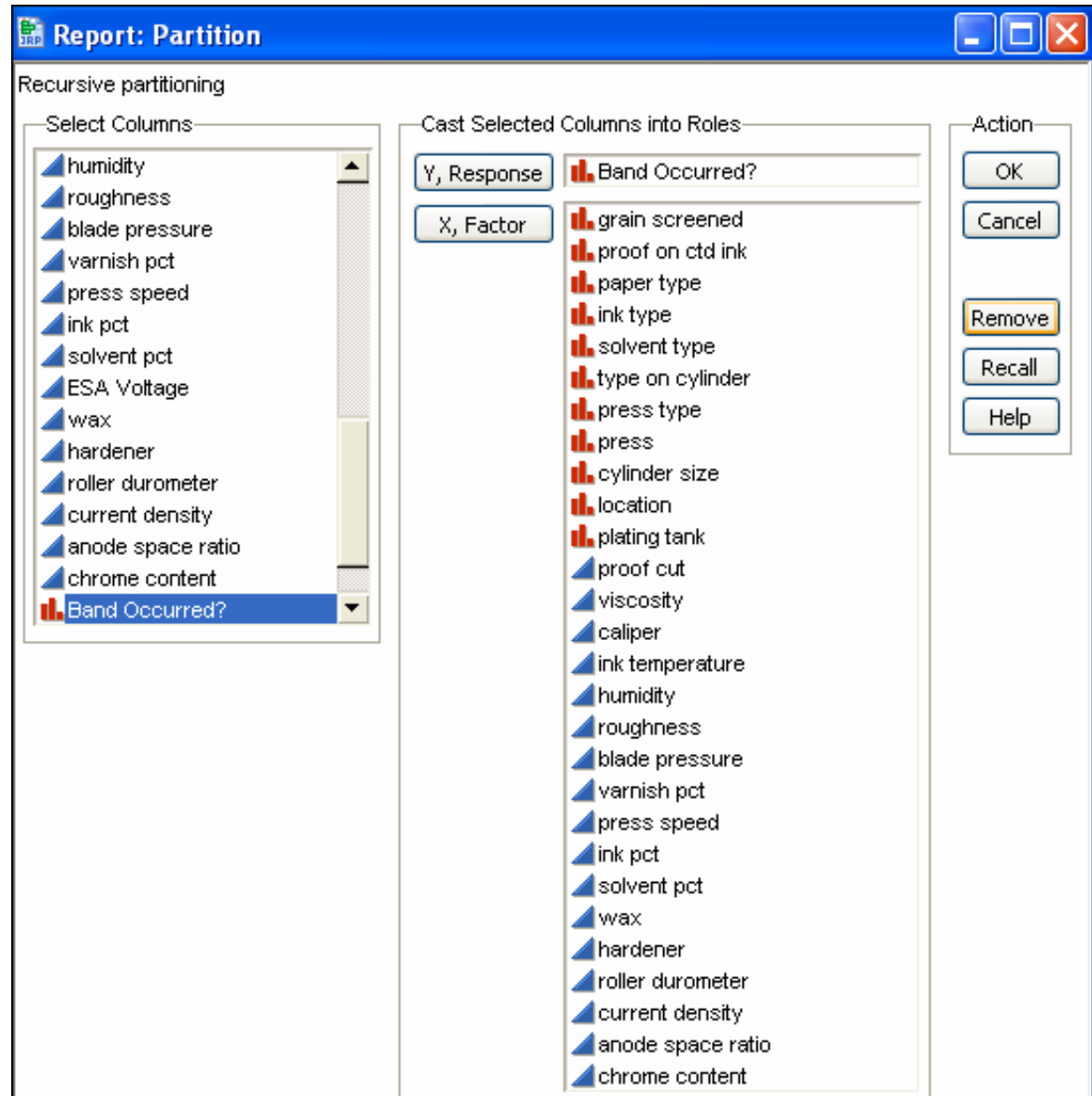
For example, the variable ESA Voltage is missing for all but 18 records:



It is unreasonable to include this variable in modeling efforts, and it will not be included in our classification model.

A classification model is requested using the Partition menu in JMP.

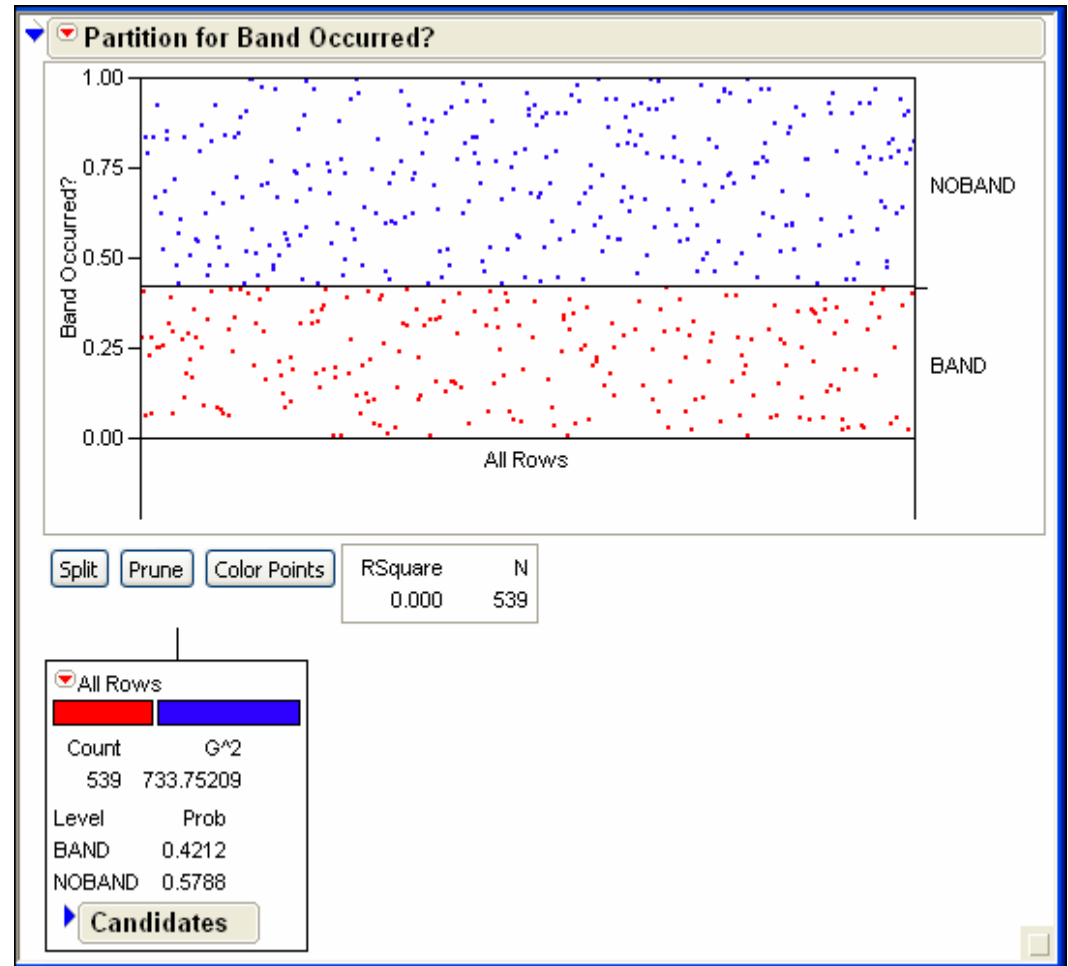
“Band Occurred?” is the response, and the 28 variables are input as candidate predictors.



The Partition report to the right opens.

Points corresponding to the runs are jittered in such a way that runs with banding are red and in the area of the graph beneath the horizontal divider at 42.12%.

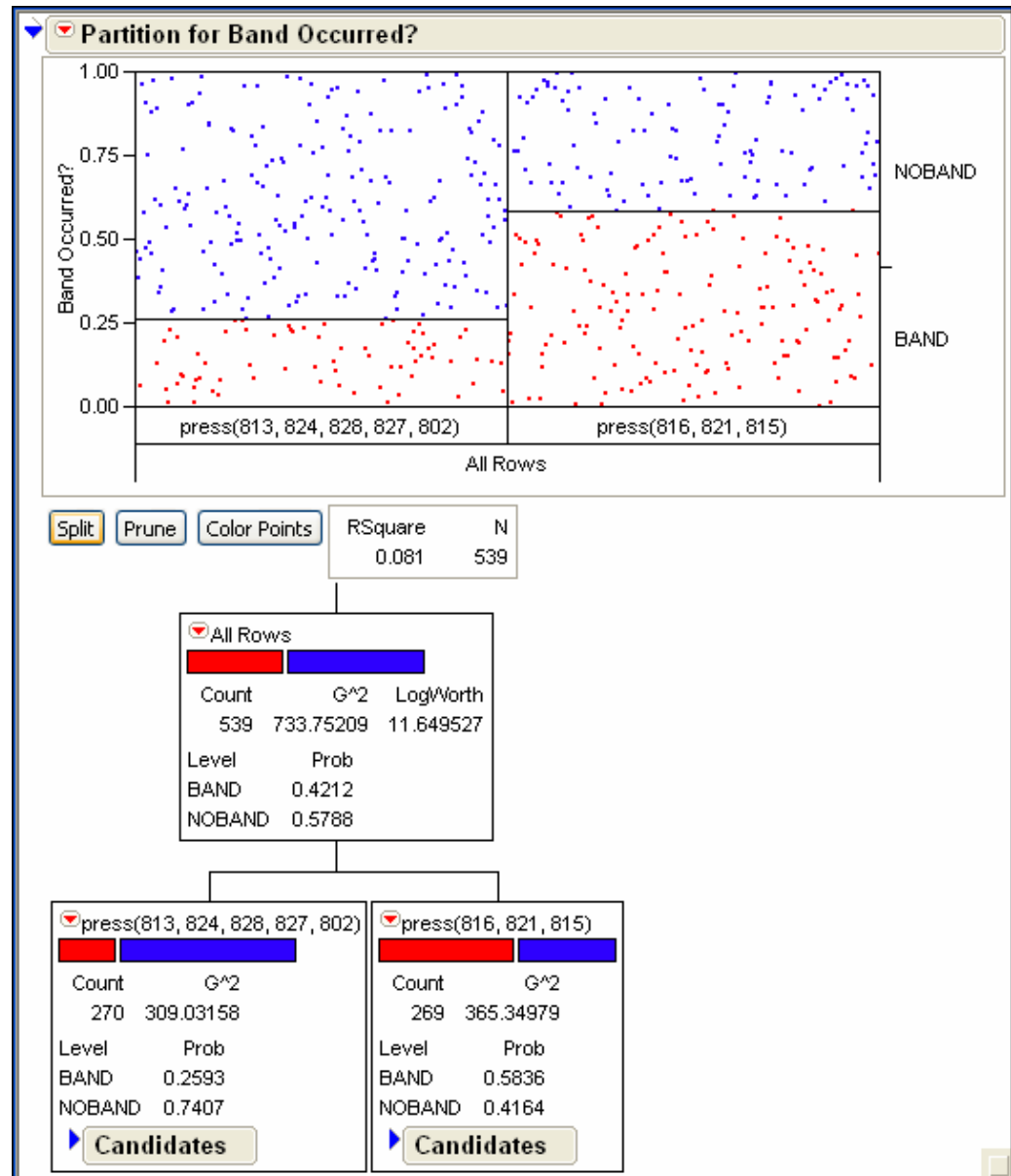
Blue points (no banding) are shown above the line.



We split once.

Note that the variable “press” is chosen as the splitting variable.

The split places four presses in a group where “NOBAND” is predicted, and the three other presses in a group where “BAND” is predicted.



Had we opened the Candidates list before splitting, we would have seen Candidate G² and LogWorth values.

Take the variable “grain screened” as an example. All possible splits of “grain screened” are obtained.

For each possible split, the likelihood ratio chi-square value for a test of independence of “Band Occurred?” versus “grain screened” (with the given split) is obtained.

The G² value is the largest of these likelihood ratio chi-square values.

| Term | Candidate G ² | LogWorth |
|-------------------|--------------------------|-------------|
| grain screened | 15.26317452 | 4.02908560 |
| proof on ctd ink | 1.84018649 | 0.75713796 |
| paper type | 41.19685301 | 9.09263749 |
| ink type | 23.36747644 | 5.47317019 |
| solvent type | 0.12447377 | 0.03691410 |
| type on cylinder | 6.74183269 | 2.02605134 |
| press type | 19.21404721 | 4.05141802 |
| press | 59.37071878 * | 11.64952731 |
| cylinder size | 0.65493009 | 0.17425329 |
| location | 7.48196255 | 1.17734960 |
| plating tank | 0.77105555 | 0.42034133 |
| proof cut | 10.64919339 | 1.90329222 |
| viscosity | 11.42286651 | 2.01782601 |
| caliper | 2.43769360 | 0.24417435 |
| ink temperature | 13.62015533 | 2.47926676 |
| humidity | 17.09997024 | 3.44683674 |
| roughness | 9.47150999 | 1.80321356 |
| blade pressure | 7.18955741 | 1.04488922 |
| varnish pct | 5.01276163 | 0.28628205 |
| press speed | 42.28087082 | 11.11336985 |
| ink pct | 17.17727140 | 3.38586103 |
| solvent pct | 10.72213806 | 1.45229444 |
| wax | 6.74466686 | 1.01416710 |
| hardener | 8.87487939 | 1.50836683 |
| roller durometer | 17.99548697 | 3.99976992 |
| current density | 16.98014698 | 3.89053048 |
| anode space ratio | 2.78235275 | 0.11752418 |
| chrome content | 8.38271198 | 2.19021045 |

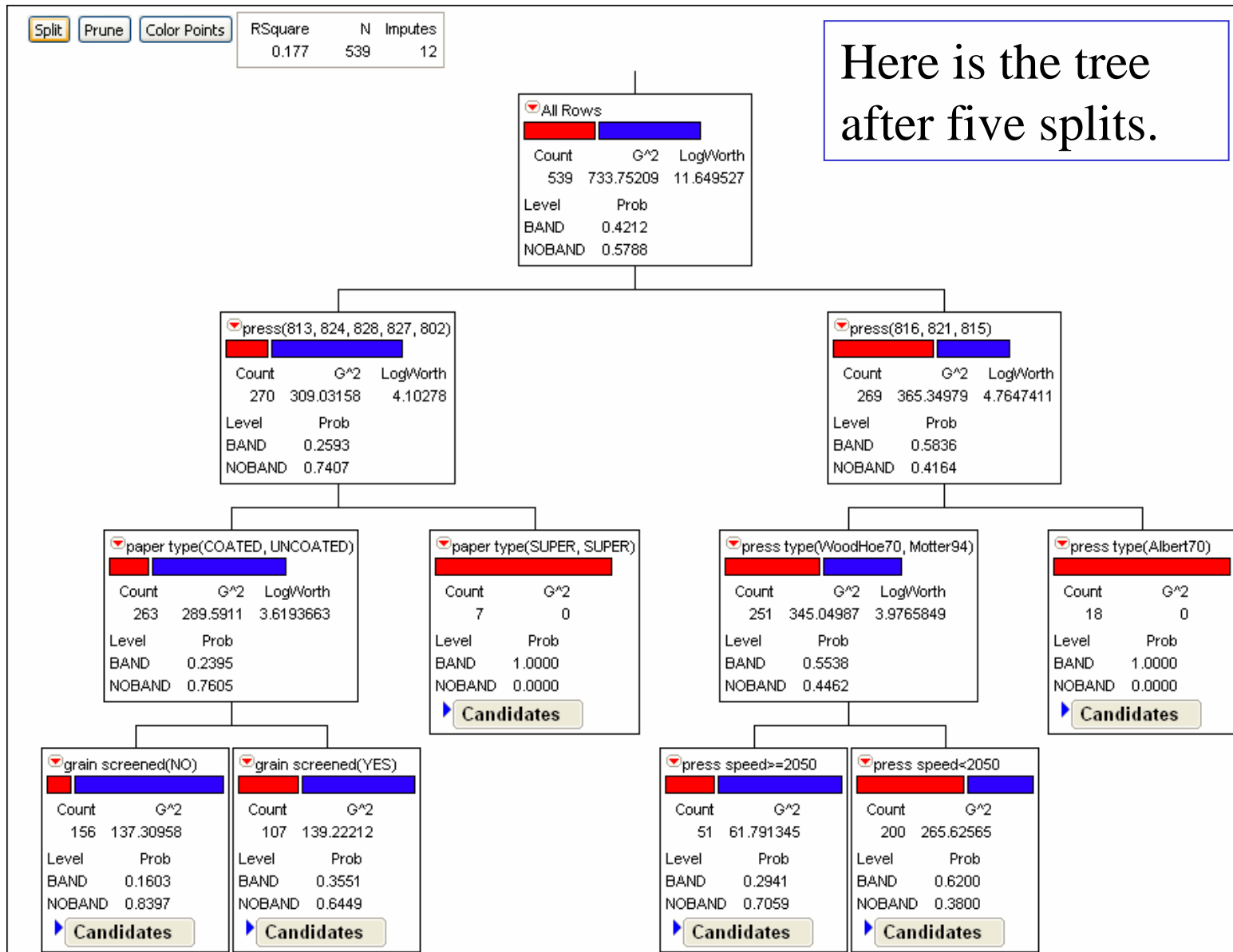
The LogWorth values are the logs of adjusted p-values for the chi-square test of independence. These are adjusted to account the number of ways that splits can occur.

For a particular variable, the LogWorth value corresponding to the split that gives the largest such value is the one shown.

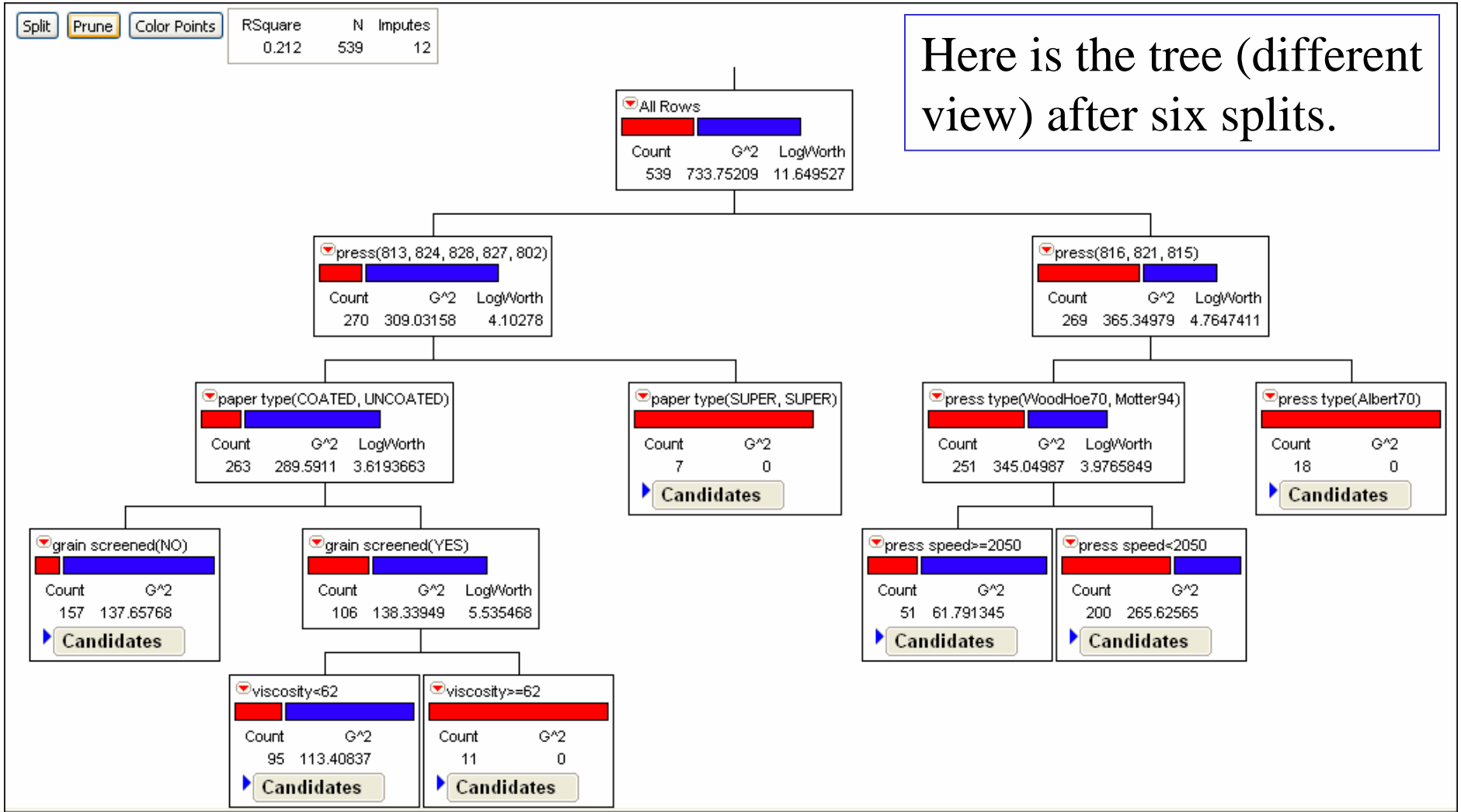
For either of these criteria, the larger the value, the more “significant” the split.

The variable “press” shows the largest values on both criteria.

| ▼ All Rows | | |
|-------------------|--------------------------|-------------|
| Count | G ² | |
| 539 | 733.75209 | |
| Level | Prob | |
| BAND | 0.4212 | |
| NOBAND | 0.5788 | |
| ▼ Candidates | | |
| Term | Candidate G ² | LogWorth |
| grain screened | 15.26317452 | 4.02908560 |
| proof on ctd ink | 1.84018649 | 0.75713796 |
| paper type | 41.19685301 | 9.09263749 |
| ink type | 23.36747644 | 5.47317019 |
| solvent type | 0.12447377 | 0.03691410 |
| type on cylinder | 6.74183269 | 2.02605134 |
| press type | 19.21404721 | 4.05141802 |
| press | 59.37071878 * | 11.64952731 |
| cylinder size | 0.65493009 | 0.17425329 |
| location | 7.48196255 | 1.17734960 |
| plating tank | 0.77105555 | 0.42034133 |
| proof cut | 10.64919339 | 1.90329222 |
| viscosity | 11.42286651 | 2.01782601 |
| caliper | 2.43769360 | 0.24417435 |
| ink temperature | 13.62015533 | 2.47926676 |
| humidity | 17.09997024 | 3.44683674 |
| roughness | 9.47150999 | 1.80321356 |
| blade pressure | 7.18955741 | 1.04488922 |
| varnish pct | 5.01276163 | 0.28628205 |
| press speed | 42.28087082 | 11.11336985 |
| ink pct | 17.17727140 | 3.38586103 |
| solvent pct | 10.72213806 | 1.45229444 |
| wax | 6.74466686 | 1.01416710 |
| hardener | 8.87487939 | 1.50836683 |
| roller durometer | 17.99548697 | 3.99976992 |
| current density | 16.98014698 | 3.89053048 |
| anode space ratio | 2.78235275 | 0.11752418 |
| chrome content | 8.38271198 | 2.19021045 |



Here is the tree (different view) after six splits.



Example: The Press Band Data

Note that splits have occurred both on nominal and continuous predictors.

Note also that 12 values have been imputed.

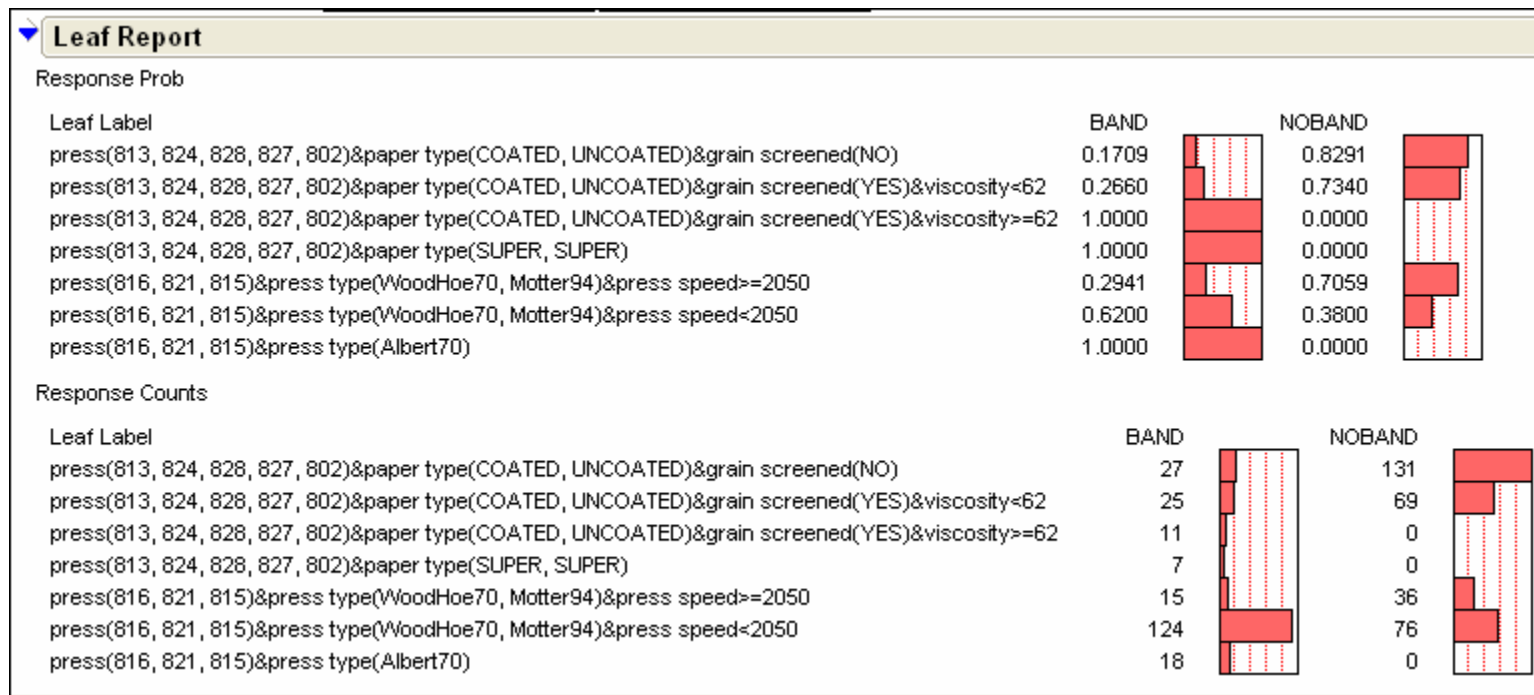
This means that certain variables used in the analysis had missing values, and so values, computed according to a selected rule, have been substituted.

The user controls splitting; at each split, JMP provides the best splitting variable and grouping of levels of that variable.

Example: The Press Band Data

As trees get large, they become visually intractable.

JMP provides a Leaf Report, which gives the rule set and a graphic display of the terminal nodes' discriminatory ability.



Example: The Press Band Data

Formulas for the predicted probabilities, leaf numbers, and leaf labels (rule set) can be saved to columns in the JMP data table.

| | Band Occurred? | Prob(Band Occurred?==BAND) | Prob(Band Occurred?==NOBAND) | Leaf Number Formula | Leaf Label Formula |
|----|----------------|----------------------------|------------------------------|---------------------|--|
| 1 | BAND | 0.62 | 0.38 | 6 | press(816, 821, 815)&press type(WoodHoe70, Motter94)&p |
| 2 | NOBAND | 0.62 | 0.38 | 6 | press(816, 821, 815)&press type(WoodHoe70, Motter94)&p |
| 3 | BAND | 0.62 | 0.38 | 6 | press(816, 821, 815)&press type(WoodHoe70, Motter94)&p |
| 4 | NOBAND | 0.17088608 | 0.82911392 | 1 | press(813, 824, 828, 827, 802)&paper type(COATED, UNCO |
| 5 | BAND | 0.26595745 | 0.73404255 | 2 | press(813, 824, 828, 827, 802)&paper type(COATED, UNCO |
| 6 | NOBAND | 0.29411765 | 0.70588235 | 5 | press(816, 821, 815)&press type(WoodHoe70, Motter94)&p |
| 7 | NOBAND | 0.29411765 | 0.70588235 | 5 | press(816, 821, 815)&press type(WoodHoe70, Motter94)&p |
| 8 | BAND | 0.26595745 | 0.73404255 | 2 | press(813, 824, 828, 827, 802)&paper type(COATED, UNCO |
| 9 | BAND | 0.62 | 0.38 | 6 | press(816, 821, 815)&press type(WoodHoe70, Motter94)&p |
| 10 | NOBAND | 0.17088608 | 0.82911392 | 1 | press(813, 824, 828, 827, 802)&paper type(COATED, UNCO |
| 11 | NOBAND | 0.17088608 | 0.82911392 | 1 | press(813, 824, 828, 827, 802)&paper type(COATED, UNCO |
| 12 | BAND | 0.26595745 | 0.73404255 | 2 | press(813, 824, 828, 827, 802)&paper type(COATED, UNCO |

Example: The Press Band Data

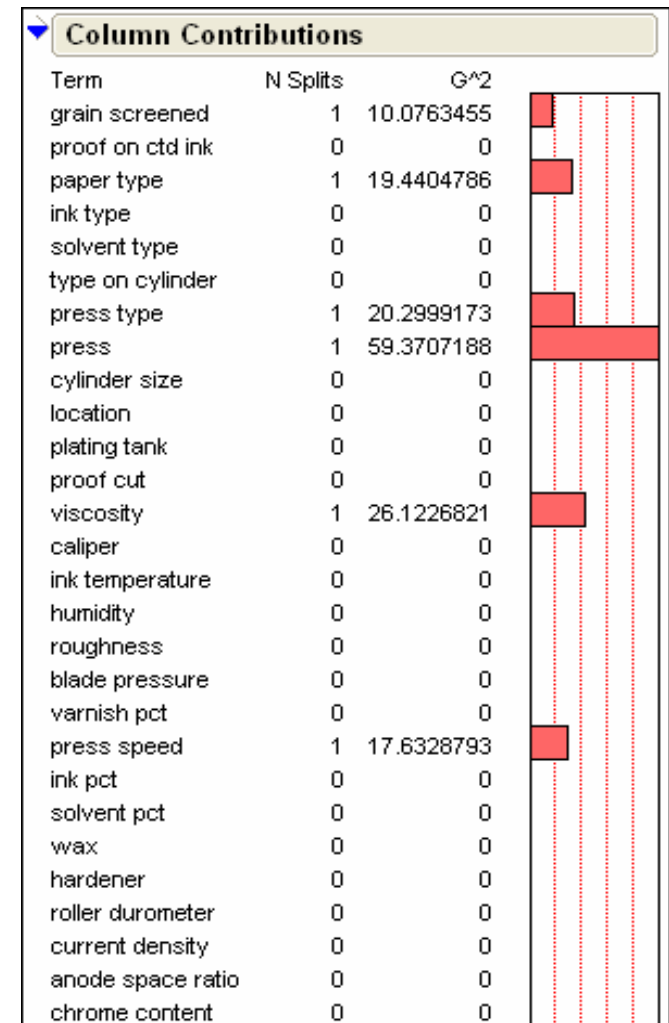
Here is the formula for the predicted probabilities.

Note that the formula simply follows the splits to terminal nodes, and then assigns to a job that falls in that terminal node, the proportion of banding that was observed in that node.



Example: The Press Band Data

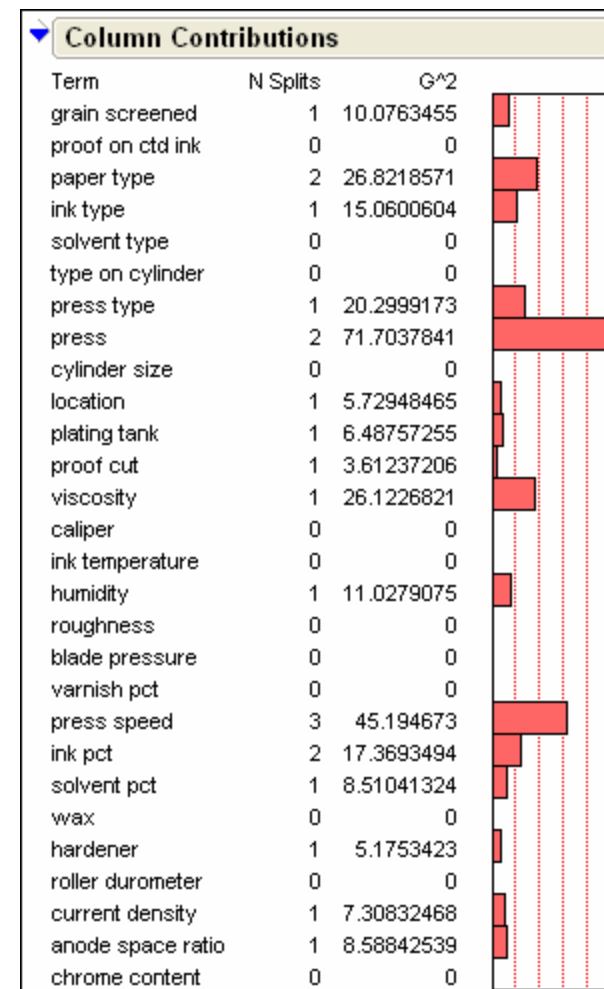
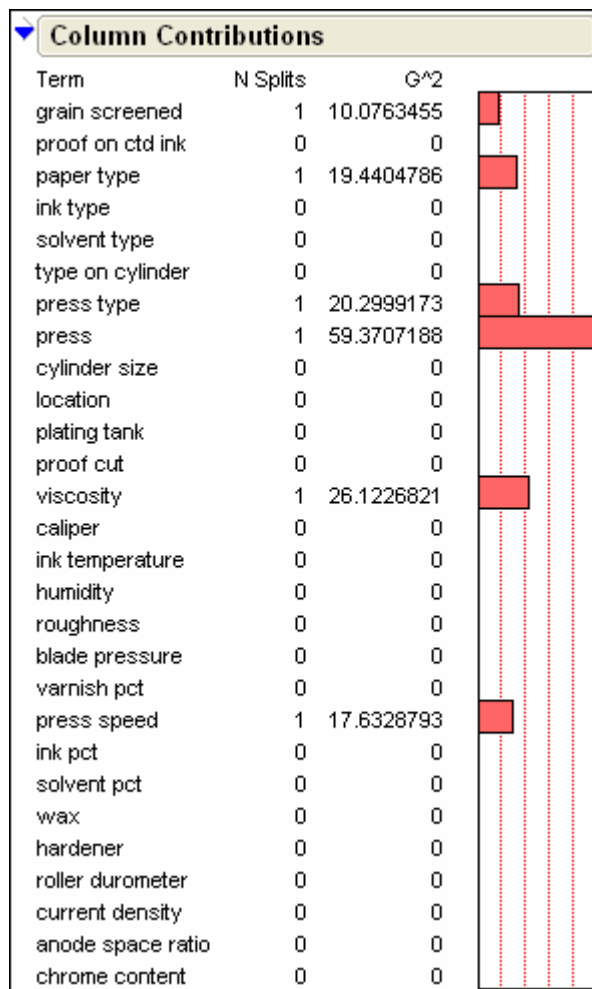
JMP also provides a Column Contributions plot, to help determine the influence of the variables on the response.



Example: The Press Band Data

The outputs to the right shows contributions after six splits (left) and 21 splits (right).

Note that, at some point, we begin to split on variables that seem to contribute little in terms of discrimination.



Example: The Press Band Data

We return to our analysis based on six splits.

The ability of the model to correctly classify jobs as affected or not affected by banding can be assessed using a Lift Chart and/or a Receiver Operating Characteristic (ROC) curve.

To understand the lift chart, think of the predicted probabilities of BAND being sorted in descending order.

Each value of the probability of BAND is thought of as a **cut point for the decision to classify** a record as BAND.

Example: The Press Band Data

For example, one of the predicted probabilities for BAND is 0.62.

Suppose we classify any run whose predicted value is 0.62 or greater as having BAND, and any run with probability less than 0.62 as having NOBAND.

Is this classification scheme better than chance alone?

Can we answer this question for all possible cut points?

Note that, if the model were no better than random, then about 5% of the BAND runs would appear in the top 5% of probabilities, 10% of the BAND runs would appear in the top 10% of probabilities, etc.

Example: The Press Band Data

In other words, if the model were randomly ranking runs, in the top X% of probability rankings, we would expect to see X% of the BAND runs.

Thus, the ratio

$(\text{proportion of BAND runs in the top X\%})/X\%$

would be 1.0.

If the ranking were NOT random, that is, if we have a strong model for predicting BAND, then we would expect to see higher proportions of the BAND runs in the more highly ranked groupings of runs.

Example: The Press Band Data

In the ranking by predicted probabilities, **the value of the lift chart at an X% portion of the population is the proportion of BAND runs that appear in the top X% of the population, divided by X%.**

If this ratio for, say, the top 20% of the population were 3, then the proportion of BAND runs identified by the top 20% of runs, as ranked by the model probabilities, would be three times what one would expect by chance alone.

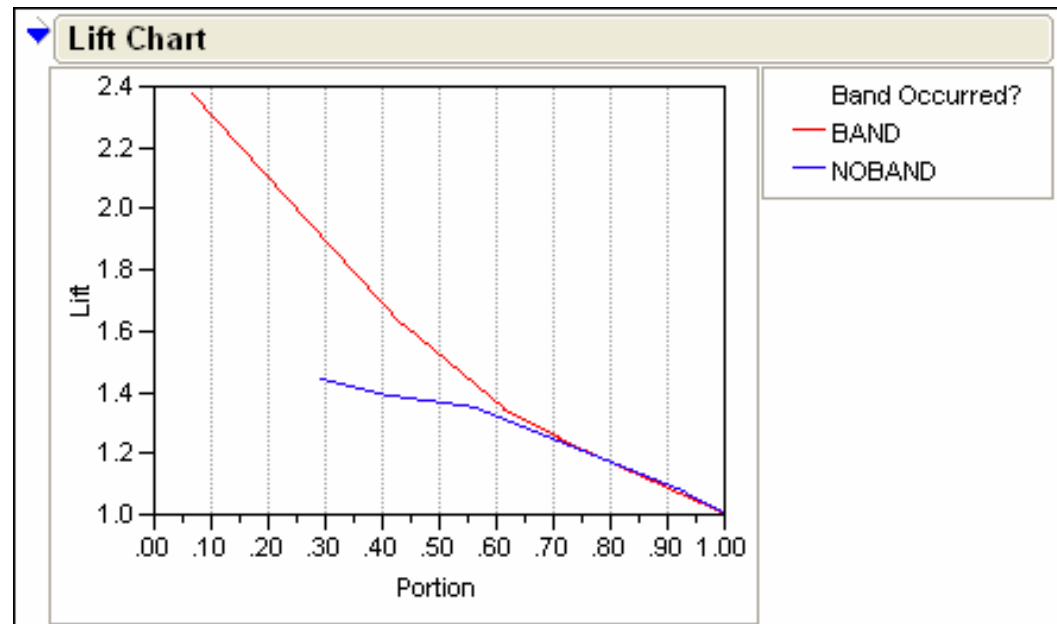
This is the **lift** provided by the model at X%.

Example: The Press Band Data

The value of the lift curve for BAND at a population portion of 0.10 is about 2.3.

The 0.10 portion refers to the top decile of runs, ranked by predicted probabilities.

The value at 0.10 indicates that, in the top decile of predicted probabilities, 2.3 times as many jobs are correctly identified as BAND as would be identified by chance alone.



Example: The Press Band Data

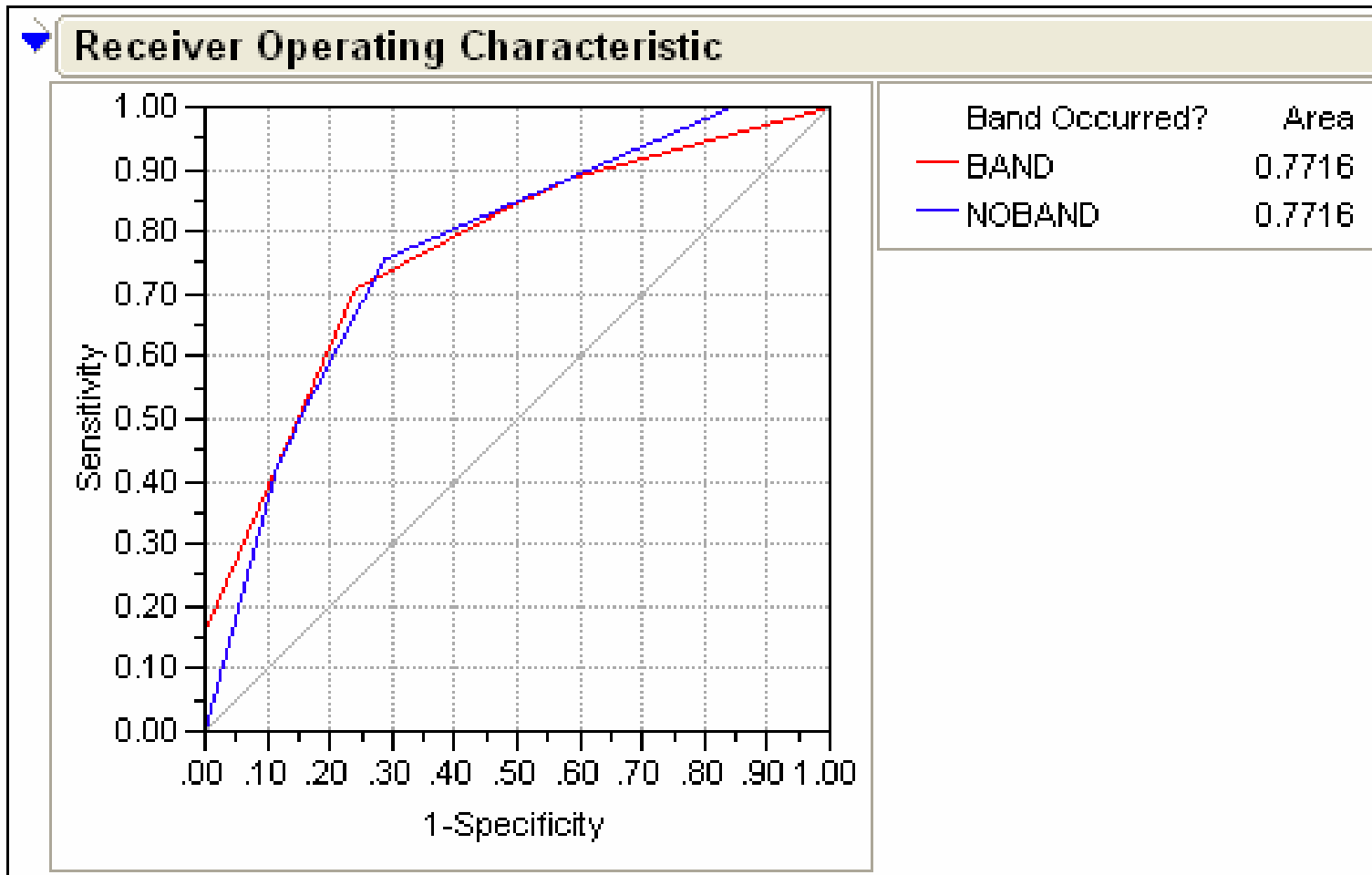
The **Receiver Operating Characteristic (ROC)** curve is also based on the idea of treating the probabilities as cut points for a classification scheme.

For a given cut point, the ROC curve plots the proportion of correct classifications (**hit rate, or true positive ratio**) on the Y axis and the proportion of incorrect classifications (**false alarm rate, or false positive ratio**) on the X axis.

An ideal model has a hit rate of 1 and a false alarm rate of 0.

The closer the curve is to the left and upper corners of the graph, the better the model.

Example: The Press Band Data



Example: The Press Band Data

When does one stop splitting?

JMP allows the user to input a minimum node size.

Splitting on a node ends when that size is reached.

However, splitting until terminal nodes reach the minimum size is not wise, as this will result in noise, rather than structure, being modeled.

If the goal of the analysis is predictive modeling, it is strongly recommended that an evaluation set be removed from the data before modeling is begun.

Example: The Press Band Data

Model development should be done on the training set.

Here, one can select a number of candidate models based on criteria such as minimum change in LogWorth or R^2 , column contributions, or lift charts.

Then these models can be evaluated on the evaluation set, and a best model chosen.

The evaluation set helps guard against overfitting!

Note that the question of when to stop splitting is less of an issue with exploratory use of the Partition platform.

Example: The Press Band Data

We note that JMP has many features that facilitate use of the Partition platform. We list just a few here.

Suppose that you have split six times and have produced a lift chart and a leaf report.

When you split once more, the lift chart and leaf report update automatically – there is no need to regenerate these.

As we have seen, formulas can be saved to columns. These can then be applied to new records, or copied and pasted into a new data table that contains new records.

JMP's row state data type allows one to easily track development and evaluation samples.

Partitioning and Six Sigma Training

We introduce the Partition platform in both our Green Belt and Black Belt training.

It is useful in both manufacturing and transactional projects, but is especially useful in transactional areas.

We introduce partitioning after covering multiple linear and logistic regression.

At this point, participants have familiarity with modeling and can appreciate the novelty, flexibility, and uses of this procedure.

Partitioning and Six Sigma Training

Multiple linear regression works well when the predictors and the response are linearly related.

However, relationships are not always linear.

In fact, they can be much more complex.

Multiple linear regression can be adversely affected by outliers and unruly distributions, both for the predictors and response.

Multiple linear regression does not deal well with nominal predictors that have many levels (for example, Part Number, Distribution Center, Sales Region).

Partitioning and Six Sigma Training

Large observational datasets often consist of unruly data.

Even small data sets often exhibit complex interactions and nonlinear behavior.

In these situations, partition modeling can provide better models than multiple linear or logistic regression.

Partitioning and Six Sigma Training

But, not only do regression and classification trees give an intuitive way to model data.

They are especially useful for **exploring data**.

In particular, when a predictor has numerous levels, Partition modeling can be used to group those levels into broader categories.

These broader categories may be of interest on their own, or they may be used as predictors in a traditional regression analysis.

Partitioning and Six Sigma Training

We provide the following guidance to Six Sigma project teams:

A Six Sigma project team should consider using tree-based methods when:

- There is a large observational data set to explore; or
- The team's data contains one or more multi-level nominal variables; or
- The data is unruly (many outliers, missing data); or
- The data may contain complex interactions.

Partitioning and Six Sigma Training

To summarize, we introduce Partition methods to Green Belts and Black Belts for two reasons:

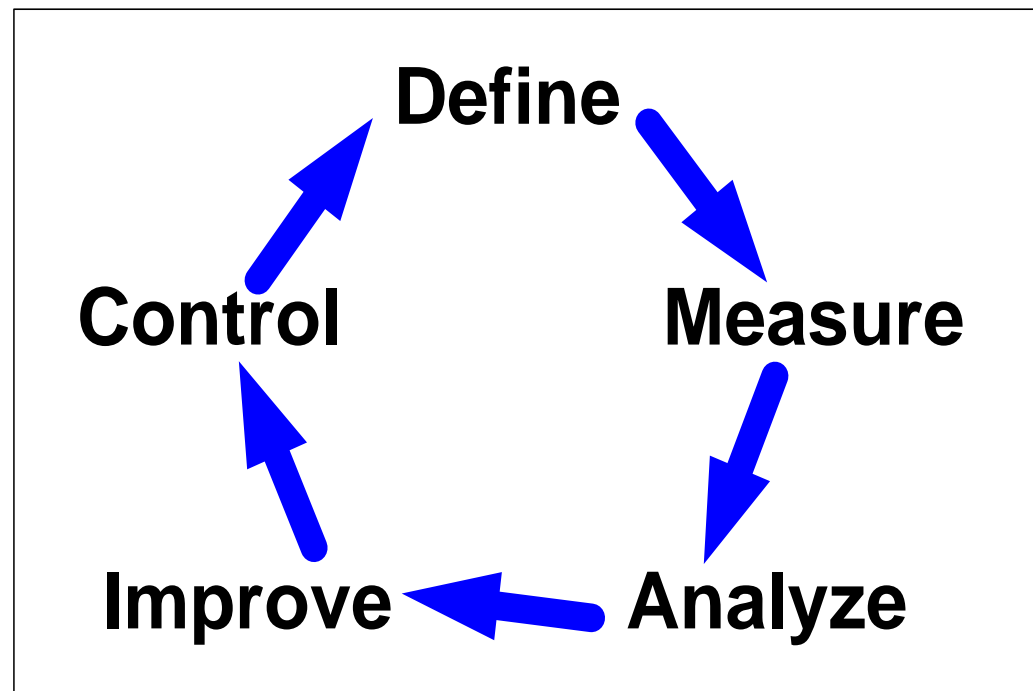
- They allow the building of better models.
- They assist in data exploration.

As a side benefit, the methods are intuitive and easily understood by Six Sigma project team members.

Partitioning and Six Sigma Training

The Partition platform can be used in the Define phase of the DMAIC cycle to help a Black Belt better define a project.

It can also be used in the Analyze and Improve phases to establish causal relationships and identify potential solutions.



Case Study 1 - Tooling Task Time

A Six Sigma team was charged with improving the accuracy of time allocation for various tooling tasks (Task Time).

There were 25 such tasks.

An observational data set, consisting of about 2800 records, was obtained for analysis.

There were 16 variables that might be useful for modeling.

Some of these were nominal variables with over 100 levels.

The JMP Partition platform was used for exploratory examination of the data.

Case Study 1 - Tooling Task Time

Partition analysis indicated that the two major determiners of Task Time were the **type of product** for which the job was conducted and the **skill level used** for the job.

Also of interest were a job difficulty measure, the plant for which the job was completed, the location where the job was completed, and the team's area of specialization.

Using the Partition platform, it was also possible to group types of tasks that took similar amounts of time.

Case Study 1 - Tooling Task Time

This information could have been used in deriving regression tree or traditional regression models for Task Time within these task groupings (rather than deriving 25 separate models).

Note that this analysis is integral to the Analyze phase of the DMAIC cycle.

The exploratory information revealed may also point the way to further data collection.

Case Study 2 - Defect Reduction

A Six Sigma team is studying the occurrence of a product defect.

Although the defect is rare, its occurrence costs exceed \$10,000 per incident.

Since a large number of processing factors and raw material factors were suspected of causing the defect, the team decided to run an experiment to isolate the exact cause.

To reduce the number of candidate factors for the experiment to a manageable number of meaningful factors, the team used a Partition analysis on an observational database for the product that listed process and quality information.

Case Study 2 - Defect Reduction

The response of interest was whether or not the defect occurred.

This was recorded as PASS or FAIL.

The database contained over 6000 records.

Nine process and raw material factors (five continuous, four nominal), believed to include the root causes of the defect, were used as inputs to the Partition analysis.

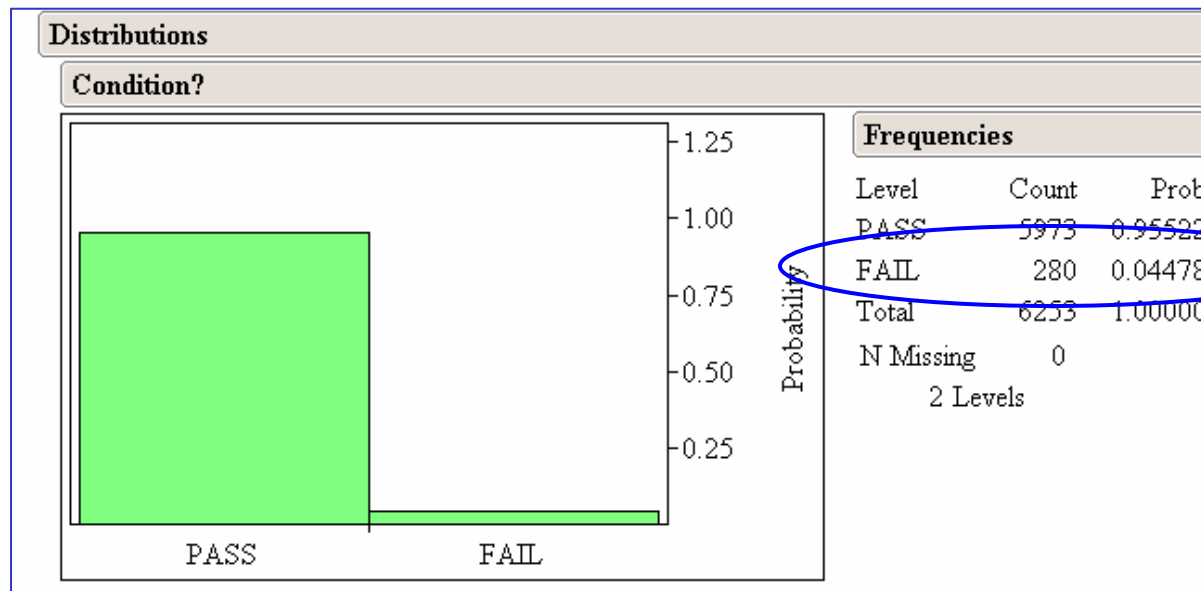
From the Partition analysis, three factors were selected for the experiment.

The experiment led to root cause identification and elimination of the defect.

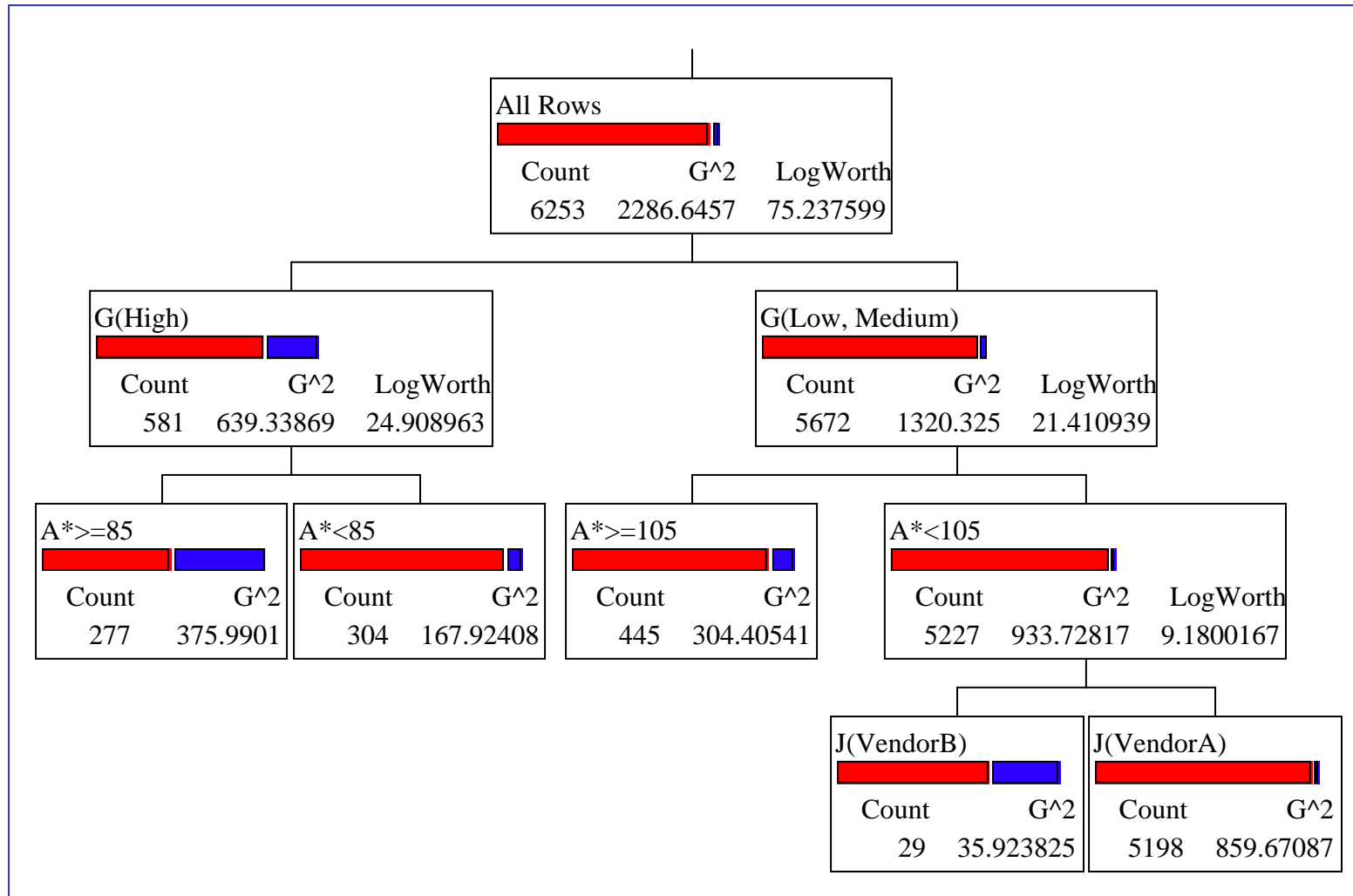
Case Study 2 - Defect Reduction

The defect occurrence rate was determined in the Distribution platform of JMP®.

The estimated occurrence rate is $< 5\%$, but had large cost implications.



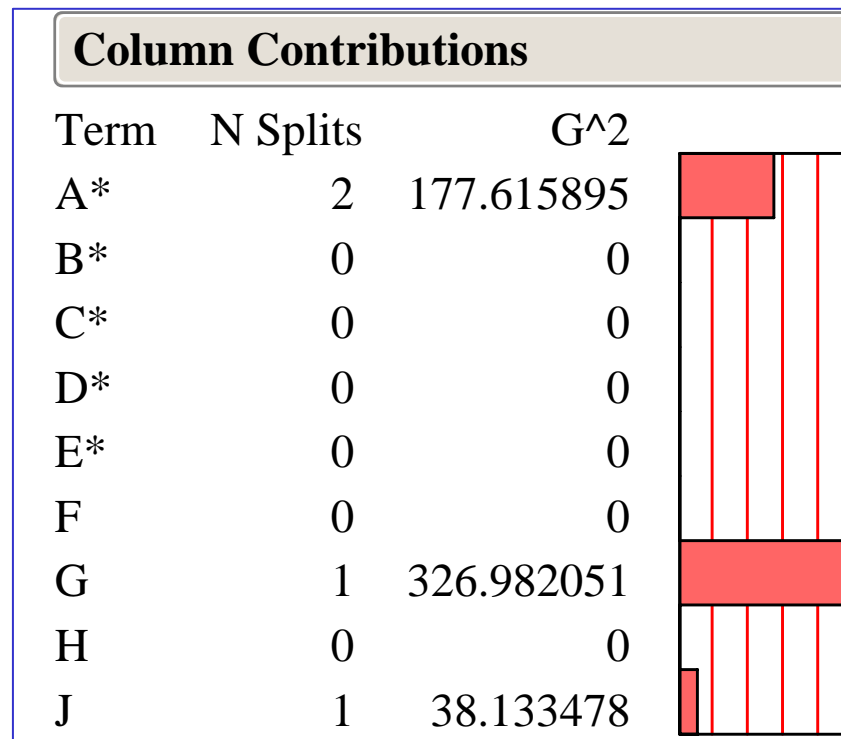
Case Study 2 - Defect Reduction



Case Study 2 - Defect Reduction

From the Column Contributions report the team decided to perform a 2^3 factorial experiment with factors A*, G, and J.

As mentioned earlier, the experiment led to identification of the root cause.



Summary

Data mining is a modern technique to analyze large data sets, or even small data sets, with unruly variable distributions.

In Six Sigma, data mining is a neglected tool, which can be of great value in achieving project success in many phases of the DMAIC cycle.

JMP[®] provides data mining analysis in the Partition platform, and provides useful visualizations to coincide with the analysis.

Two Six Sigma case studies have been shown where data mining was invaluable in achieving project success.