OTHER PREDICTIVE MODELING CONSIDERATIONS

TOPICS

- Dimension Reduction
- Recoding
- Disproportionate Sampling
- Confusion Matrix and Cutoff Values

Common Problems with Data

Missing data (or missing values) occur when a value for a variable is not available for an observation. Understanding why values are missing is important and can lead to different strategies for dealing with the missingness. Data may be **missing at random** within a variable, or **missing completely at random** across variables. In these cases, analyses may not be seriously impacted by the missingness unless a large number of observations have missing values. However, values that are missing **not at random** are cause for concern. This occurs, for example, when a question on a survey is purposefully skipped or when sensitive information is omitted, potentially leading to biased analyses or incorrect conclusions.

Draft Chapter 3, Working with Data, Building Better Models with JMP Pro (SAS), 2015.

Dirty or ***messy data*** are data that are inaccurate, have errors or typos, or are not complete. Data may be incorrectly coded, have inconsistent capitalization, abbreviations and spacing. Records may be duplicated, or variables may be redundant or highly correlated. For categorical variables, there may be an overwhelming number of categories, some of which have few values. Continuous data may be highly skewed or multi-modal, and can have extreme observations or clumps of observations.

Draft Chapter 3, Working with Data, Building Better Models with JMP Pro (SAS), 2015.

Incomplete data can relate to missing variables or not having enough data to perform an analysis. If critical variables are missing, the predictive model will most likely not perform well. If there aren't enough observations, it may not be possible to estimate important model parameters and the model predictions may not be very precise.

Draft Chapter 3, Working with Data, Building Better Models with JMP Pro (SAS), 2015.

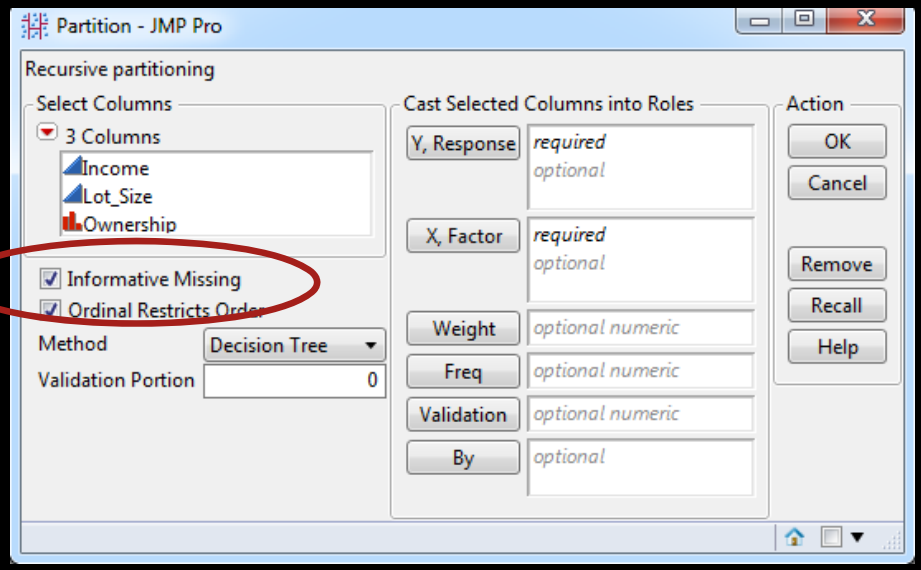
A flip side of this is ***extremely large data sets*** in terms of **many variables** or **many observations**. While this isn't a shortcoming of the data or a data quality issue per se, it can pose a challenge for modeling. Having too many variables can be problematic when the variables are correlated with one another, are redundant, or don't provide any useful information about the response. However, variable reduction methods can be applied, and "modern" modeling techniques are effective in dealing with a large number of observations and predictors.

Draft Chapter 3, Working with Data, Building Better Models with JMP Pro (SAS), 2015.

Incorrectly formatted data are data in the wrong form or format for analysis. This can apply to the data table as a whole, or to the formatting of variables in the data table. For example, data might be stored in separate columns while an analysis requires data stacked in one column. Or individual variables in JMP might have the incorrect modeling type. Since the modeling type drives the analysis in JMP, having an incorrect modeling type may lead to the wrong analysis. Another issue relates to dates and times: Columns containing date or time measurements need to be formatted as date or time variables (in **Column Info**) in order to perform calculations, such as elapsed time.

Draft Chapter 3, Working with Data, Building Better Models with JMP Pro (SAS), 2015.

When a continuous variable has missing values, imputation is often used to replace the missing values with substituted values. Imputation in JMP Pro is available from the **Explore Missing Values** option under **Cols > Modeling Utilities**, and from many modeling platforms.



Informative Missing is another approach for handling missing values. How the missing values are handled depends on whether the variables are continuous or categorical. For continuous variables, the model will include the original variable with the mean value imputed for missing values and a second variable indicating whether the value is missing or not. For categorical variables, an additional level will be included that indicates that the value for the variable is missing. You could create these columns yourself using recoding and the formula editor, but JMP platforms with the option for informative missing do this automatically.

Draft Chapter 3, Working with Data, Building Better Models with JMP Pro (SAS), 2015.

CELL CLASSIFICATION DATA

UCI Machine Learning Repository

Breast Cancer Wisconsin (Diagnostic) Data Set

http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

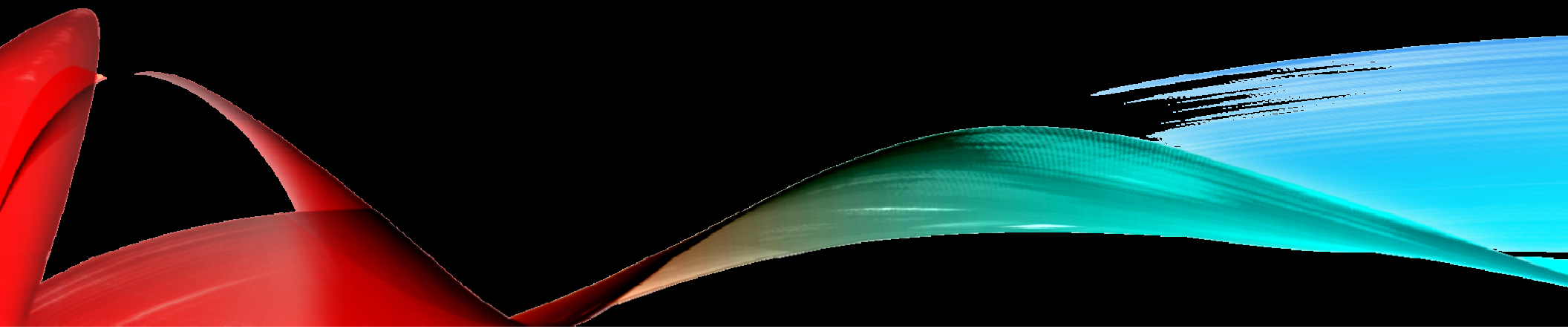
JMP Whitepaper

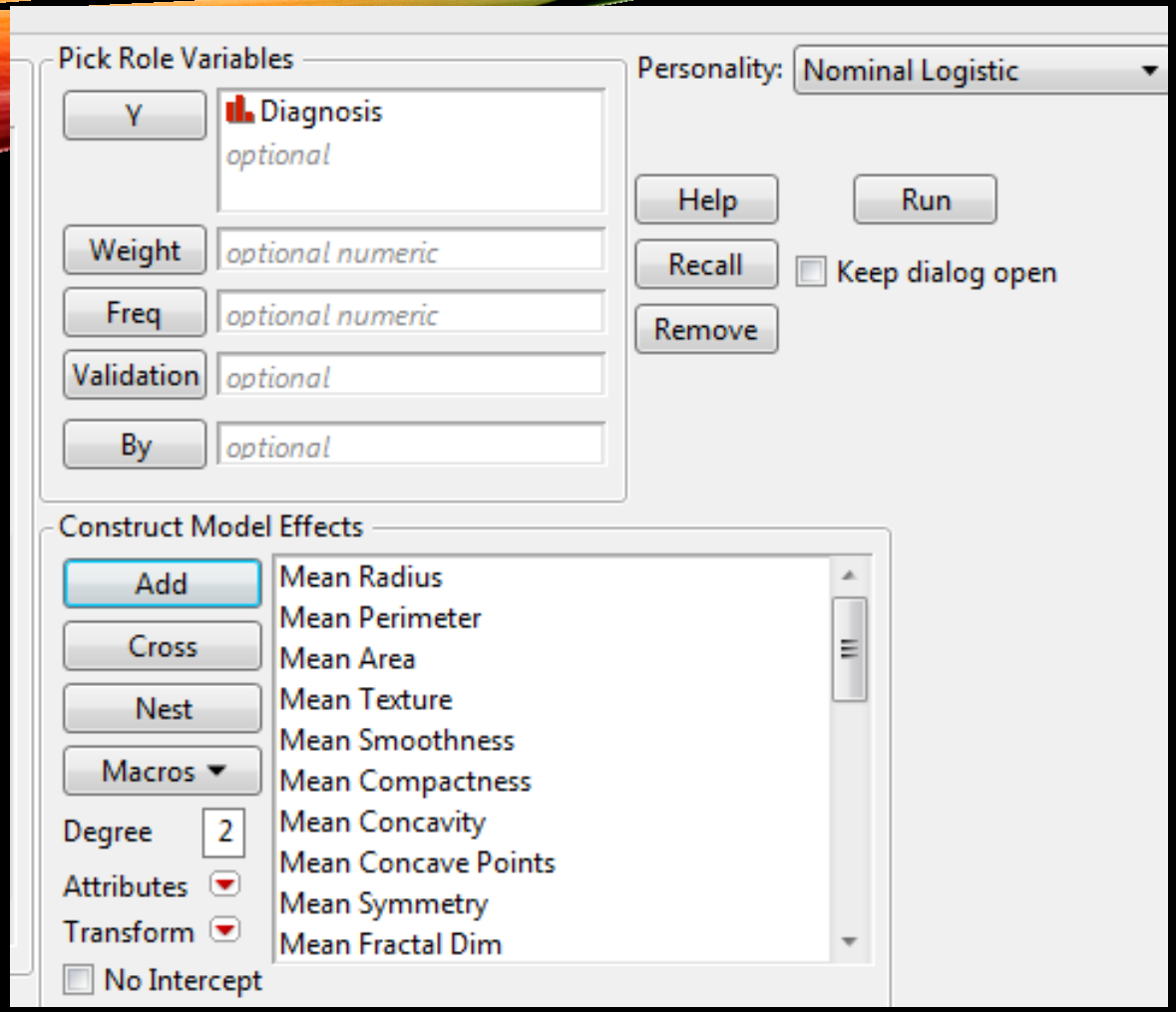
http://www.jmp.com/en_gb/whitepapers/classification-breast-cancer-cells-using-jmp.html

We use a publicly available data set, obtained from a study conducted at the University of Wisconsin [Mangasarian, Street, Wolberg]. The data set contains a total of 569 records, of which 357 tumors are benign and 212 are malignant. Each tumor contained multiple cells, and image analysis was used to determine the following ten features for each cell:

- Radius: Mean of distances from the center to points on the perimeter
- Texture: Standard deviation of gray-scale values
- Perimeter: Perimeter of the cell nucleus
- Area: Area of the cell nucleus
- Smoothness: Local variation in radius lengths
- Compactness: Perimeter2/area
- Concavity: Severity of concave portions of the contour
- Concave Points: Number of concave portions of the contour
- Symmetry: Symmetry of the cell nucleus
- Fractal dimension: Regularity of the boundary (the contour) of the nucleus

DIMENSION REDUCTION



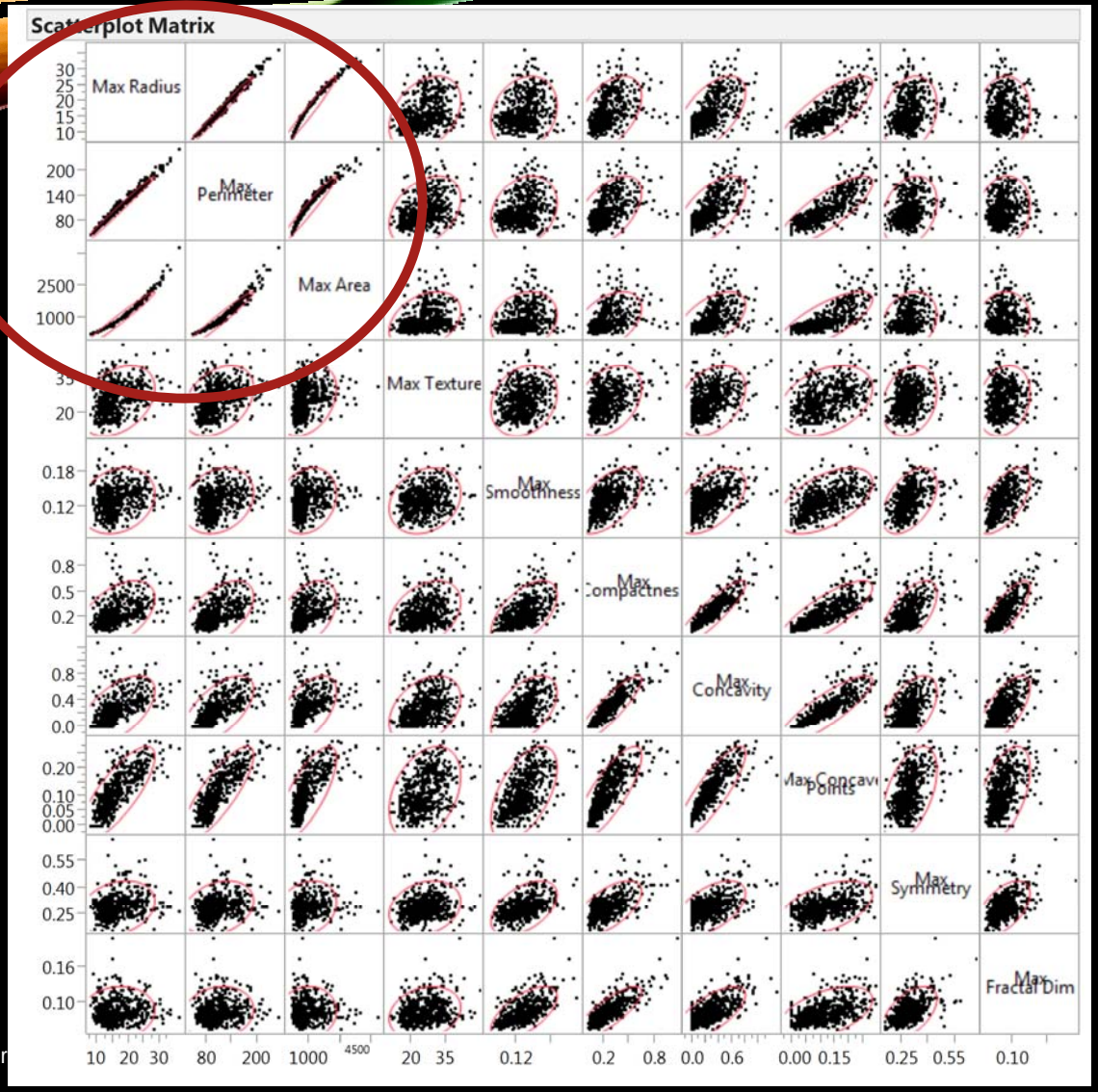

Pick Role Variables
Y
Diagnosis
optional
Weight
optional numeric
Freq
optional numeric
Validation
optional
By
optional
Personality:
Nominal Logistic
Help
Run
Recall
Keep dialog open
Remove
Construct Model Effects
Add
Cross
Nest
Macros
Degree
2
Attributes
Transform
No Intercept
Mean Radius
Mean Perimeter
Mean Area
Mean Texture
Mean Smoothness
Mean Compactness
Mean Concavity
Mean Concave Points
Mean Symmetry
Mean Fractal Dim

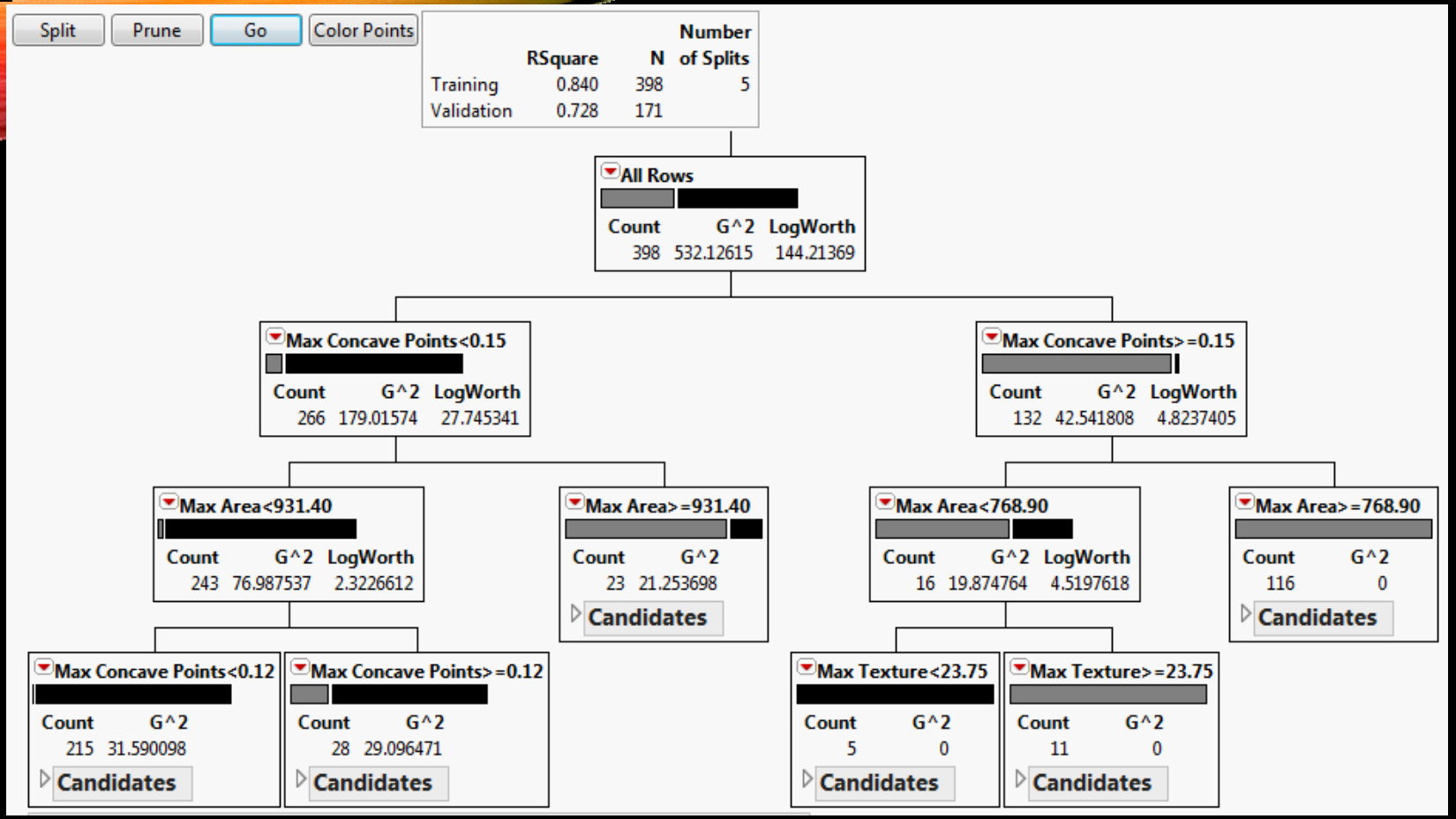
Parameter Estimates

Term		Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	Unstable	367.513222	759085.48	0.00	0.9996
Mean Radius		-3983.5312	625536.37	0.00	0.9949
Mean Perimeter		117.305899	69175.608	0.00	0.9986
Mean Area		29.7442813	4189.4968	0.00	0.9943
Mean Texture		92.831892			
Mean Smoothness		29132.3103			
Mean Compactness		-35588.378			
Mean Concavity		19006.8952			
Mean Concave Point		19276.1123			
Mean Symmetry		13777.522			
Mean Fractal Dim	Zeroed	44501.846			
Max Radius		1012.50982			
Max Perimeter		41.9162024			
Max Area		-6.5824333			
Max Texture	Unstable	31.096902			
Max Smoothness	Unstable	-2809.0486			
Max Compactness		-4776.3758			
Max Concavity		4100.55576			
Max Concave Points		4861.09829			
Max Symmetry		13804.8535			
Max Fractal Dim		33290.137	4728281.9	0.00	0.9944
SE Radius		2421.42572	1811741	0.00	0.9989
SE Perimeter		-795.32649	151007.59	0.00	0.9958
SE Area		67.0074574	18988.111	0.00	0.9972
SE Texture	Unstable	-102.96445	278151.29	0.00	0.9997
SE Smoothness	Zeroed	-65205.555	0	.	.
SE Compactness		59964.9745	2884479.1	0.00	0.9834
SE Concavity	Zeroed	-50577.293	0	.	.
SE Concave Points	Zeroed	197971.811	0	.	.
SE Symmetry	Zeroed	-60814.82	0	.	.
SE Fractal Dim	Zeroed	-466079.06	0	.	.

For log odds of M/B

An estimate can be unstable if highly correlated with other variables or there are more parameters in the model than can be predicted with the data and zeroed if the variable is a linear combination of other explanatory variables


Scatterplot Matrix
Max Radius
Max Perimeter
Max Area
Max Texture
Max Smoothness
Max Compactnes
Max Concavity
Max Concave Points
Max Symmetry
Max Fractal Dim


Split
Prune
Go
Color Points
RSquare N Number of Splits
Training 0.840 398 5
Validation 0.728 171
All Rows
Count G^2 LogWorth
398 532.12615 144.21369
Max Concave Points<0.15
Count G^2 LogWorth
266 179.01574 27.745341
Max Concave Points>=0.15
Count G^2 LogWorth
132 42.541808 4.8237405
Max Area<931.40
Count G^2 LogWorth
243 76.987537 2.3226612
Max Area>=931.40
Count G^2
23 21.253698
Candidates
Max Area<768.90
Count G^2 LogWorth
16 19.874764 4.5197618
Max Area>=768.90
Count G^2
116 0
Candidates
Max Concave Points<0.12
Count G^2
215 31.590098
Candidates
Max Concave Points>=0.12
Count G^2
28 29.096471
Candidates
Max Texture<23.75
Count G^2
5 0
Candidates
Max Texture>=23.75
Count G^2
11 0
Candidates

Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	0.8404	0.7280	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.9153	0.8391	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.1067	0.1731	∑ -Log(ρ[j])/n
RMSE	0.1666	0.2032	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.0604	0.0643	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.0327	0.0468	∑ (ρ[j]≠ρMax)/n
N	398	171	n

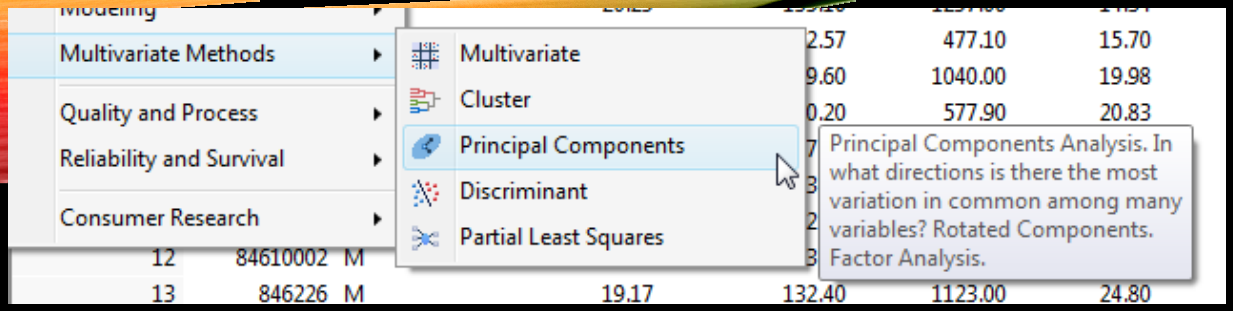
Confusion Matrix

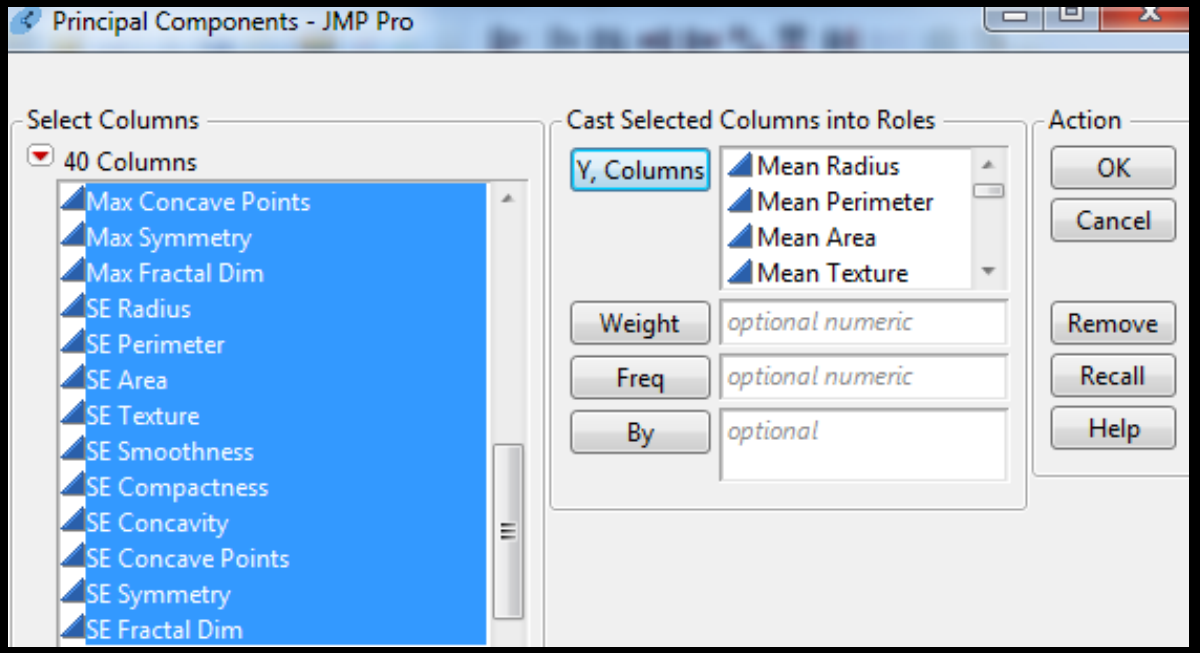
Actual / Predicted	Training M	Training B
M	146	9
B	4	239

Actual / Predicted	Validation M	Validation B
M	52	5
B	3	111

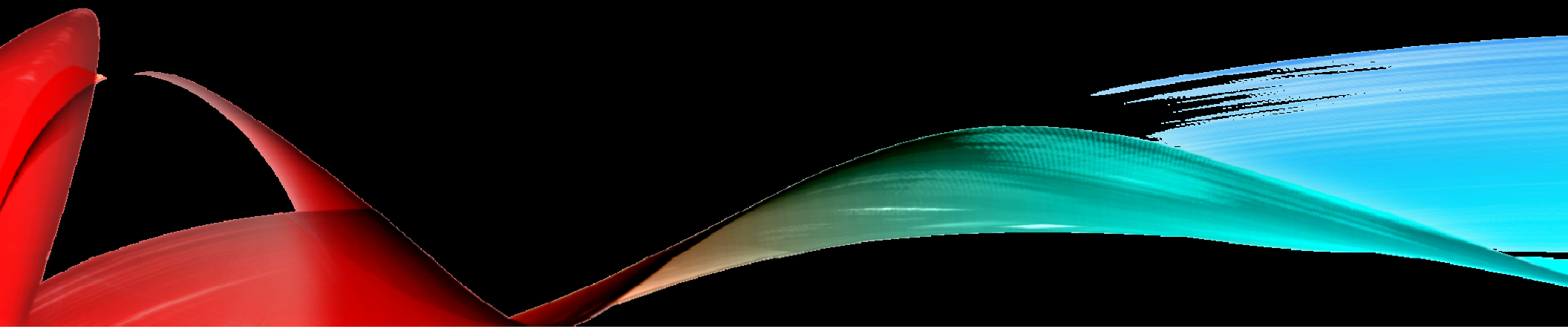
Column Contributions

Term	Number of Splits	G^2	Portion
Max Concave Points	2	326.869565	0.7261
Max Area	2	103.441553	0.2298
Max Texture	1	19.874764	0.0441
Max Concavity	0	0	0.0000
Max Smoothness	0	0	0.0000
Mean Texture	0	0	0.0000


Multivariate Methods
Quality and Process
Reliability and Survival
Consumer Research
Multivariate
Cluster
Principal Components
Discriminant
Partial Least Squares
Principal Components Analysis. In what directions is there the most variation in common among many variables? Rotated Components. Factor Analysis.
2.57 477.10 15.70
9.60 1040.00 19.98
0.20 577.90 20.83
12 84610002 M
13 846226 M 19.17 132.40 1123.00 24.80


Principal Components - JMP Pro
Select Columns
40 Columns
Max Concave Points
Max Symmetry
Max Fractal Dim
SE Radius
SE Perimeter
SE Area
SE Texture
SE Smoothness
SE Compactness
SE Concavity
SE Concave Points
SE Symmetry
SE Fractal Dim
Cast Selected Columns into Roles
Y, Columns
Mean Radius
Mean Perimeter
Mean Area
Mean Texture
Weight
optional numeric
Freq
optional numeric
By
optional
Action
OK
Cancel
Remove
Recall
Help

PRINCIPAL COMPONENT ANALYSIS



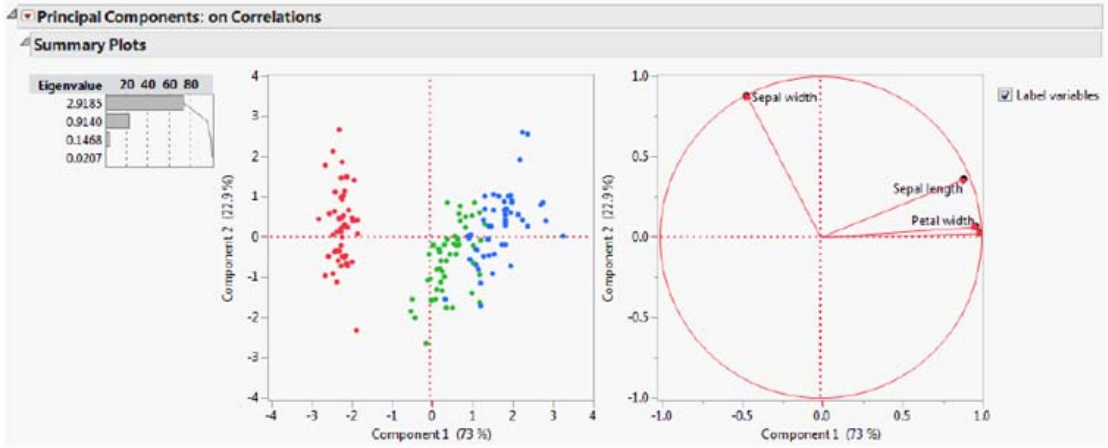
About Principal Component Analysis

The purpose of principal component analysis is to derive a small number of independent linear combinations (principal components) of a set of measured variables that capture as much of the variability in the original variables as possible. Principal component analysis is an exploratory data analysis tool and is also used for making predictive models.

The Principal Components platform also supports factor analysis. JMP offers several types of orthogonal and oblique factor analysis-style rotations to help interpret the extracted components.

For factor analysis, see the Factor Analysis chapter in the *Consumer Research* book.

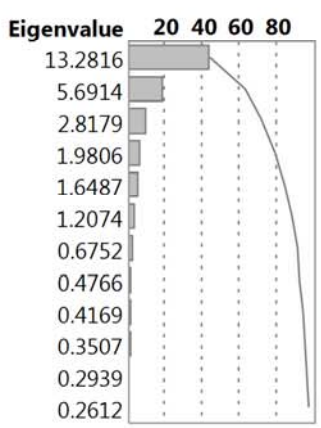
Figure 5.1 Example of Principal Components

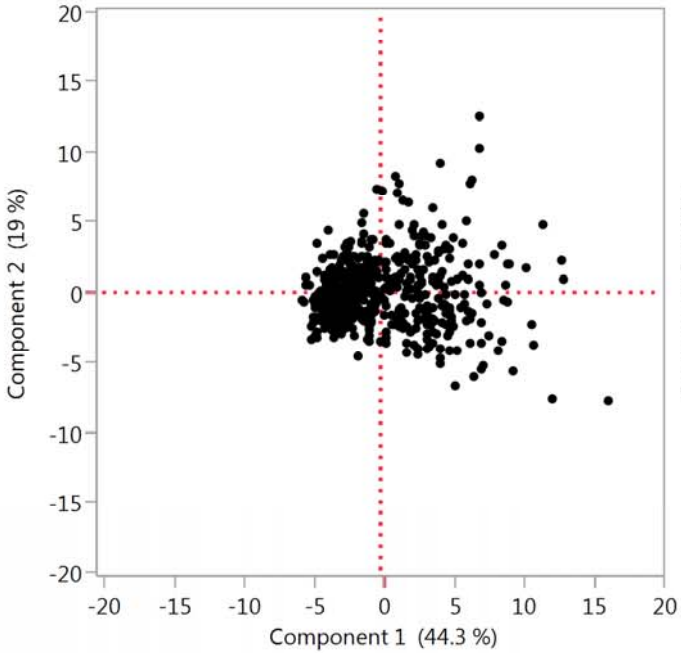


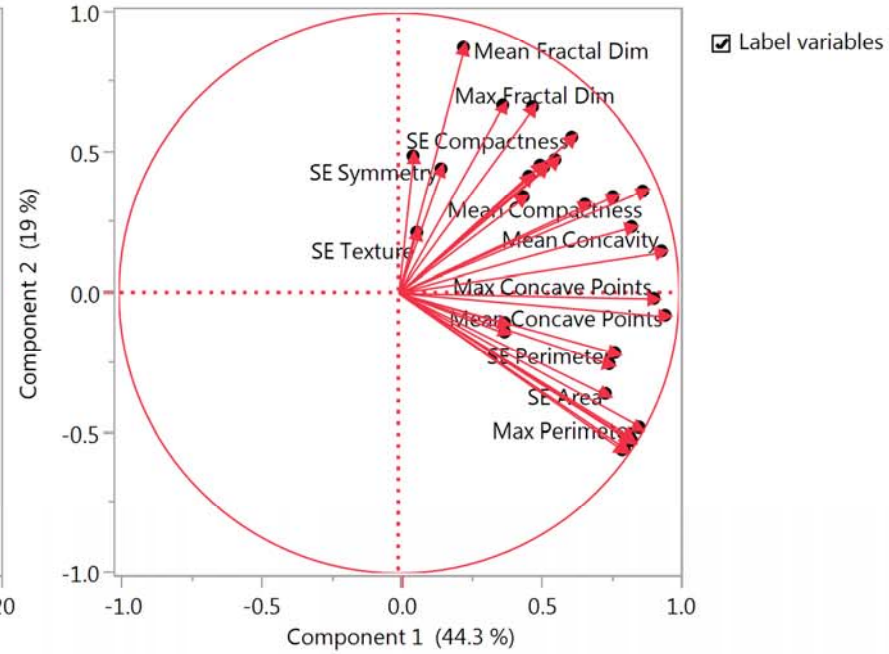
JMP Book: Multivariate Methods

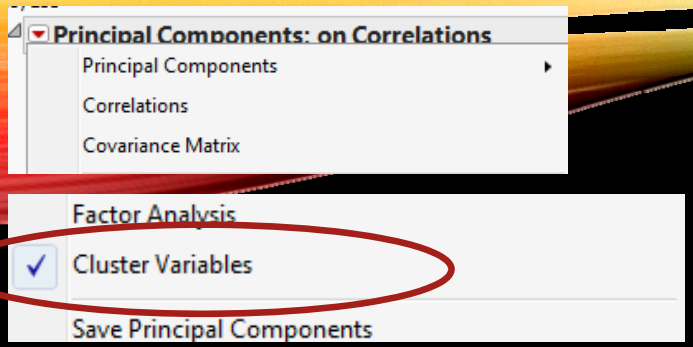
Principal Components: on Correlations

Summary Plots





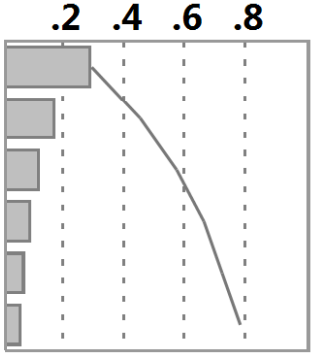




Variable Clustering

Cluster Summary

Cluster	Number of Members	Most Representative Variable	Cluster Proportion of Variation Explained	Total Proportion of Variation Explained
1	10	Mean Area	0.856	0.285
2	6	Max Concavity	0.817	0.163
3	5	SE Fractal Dim	0.726	0.121
6	4	Mean Smoothness	0.673	0.09
4	2	Mean Texture	0.956	0.064
5	3	SE Symmetry	0.605	0.06



Proportion of variation explained by clustering: 0.784

Mean Area
Max Concavity
SE Fractal Dim
Mean Smoothness
Mean Texture
SE Symmetry
Max Concave Points
Max Area
Max Texture

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	231.63379	5	463.2676	<.0001*
Full	34.42929			
Reduced	266.06307			

RSquare (U)	0.8706
AICc	81.0734
BIC	104.777
Observations (or Sum Wgts)	398

Measure	Training	Validation	Definition
Entropy RSquare	0.8706	0.9015	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.9327	0.9481	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.0865	0.0627	$\sum$ -Log(ρ[j])/n
RMSE	0.1490	0.1432	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.0475	0.0369	$\sum \lvert y[j]-\rho[j]\rvert/n$
Misclassification Rate	0.0276	0.0292	$\sum$ (ρ[j]≠ρMax)/n
N	398	171	n

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-36.819068	6.6336586	30.81	<.0001*
Mean Area	-0.0158764	0.0082493	3.70	0.0543
Max Concavity	6.68973963	1.8689154	12.81	0.0003*
Mean Texture	0.42072542	0.0975747	18.59	<.0001*
Mean Smoothness	141.383606	34.469812	16.82	<.0001*
Max Area	0.02799025	0.0072431	14.93	0.0001*

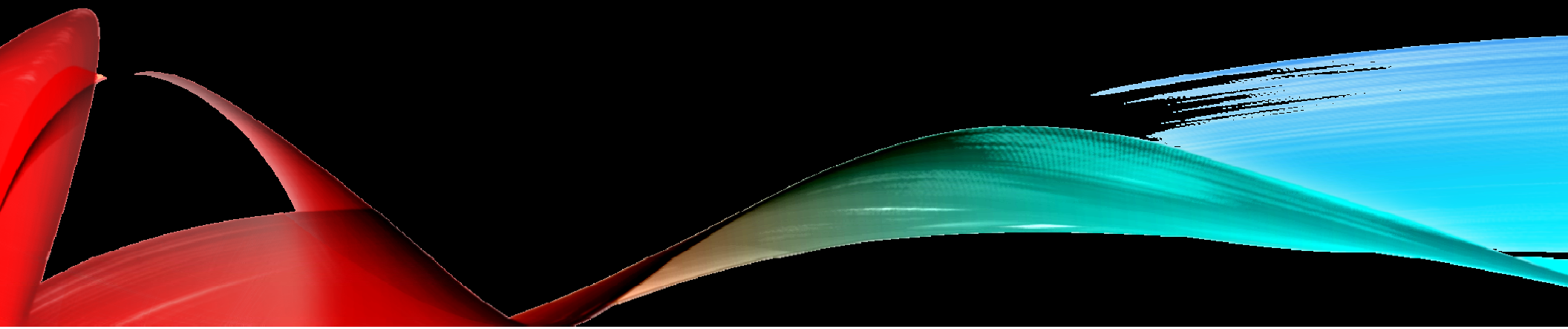
For log odds of M/B

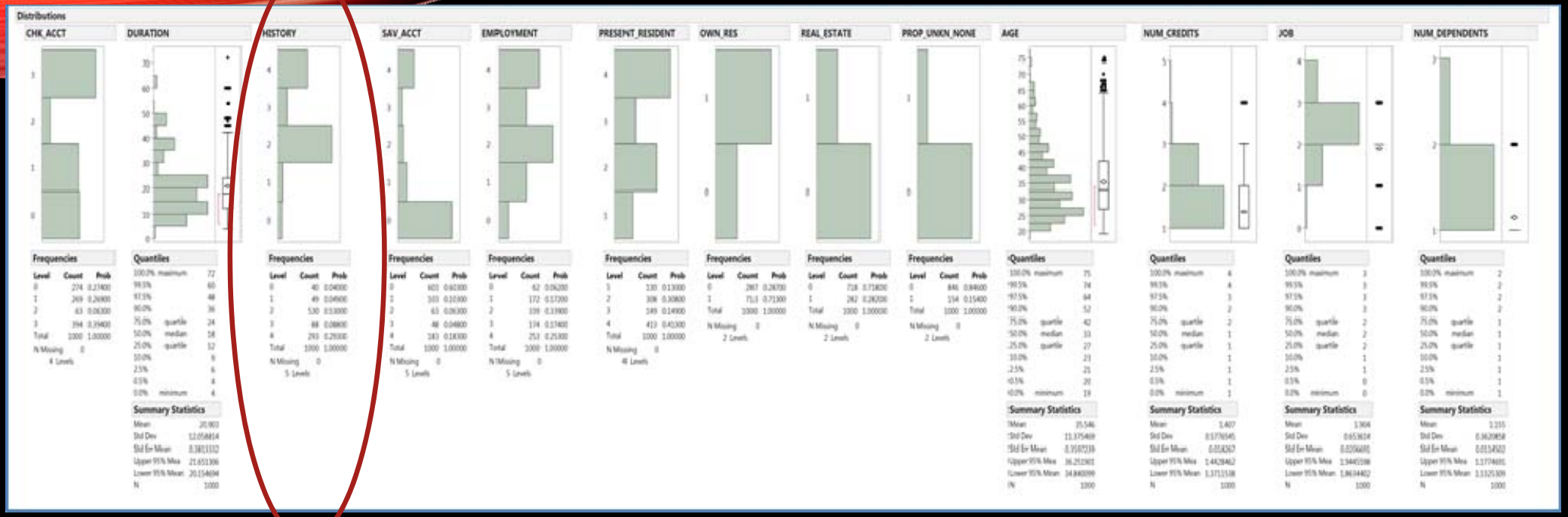
Confusion Matrix

Actual		Predicted	Actual		Predicted
Training	**M**	**B**	**Validation**	**M**	**B**
M	147	8	M	54	3
B	3	240	B	2	112

=3/57=5.3%

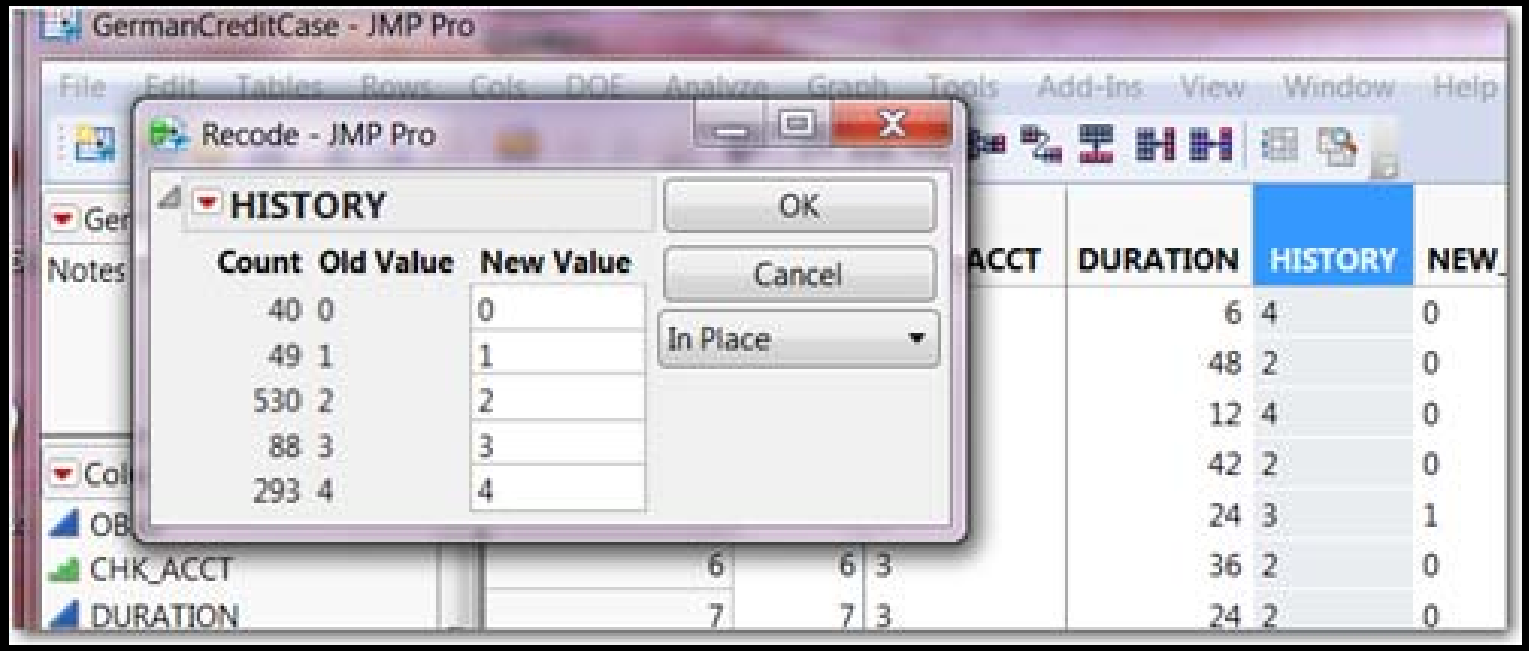
REDUCING CATEGORIES / COMBINING CATEGORIES



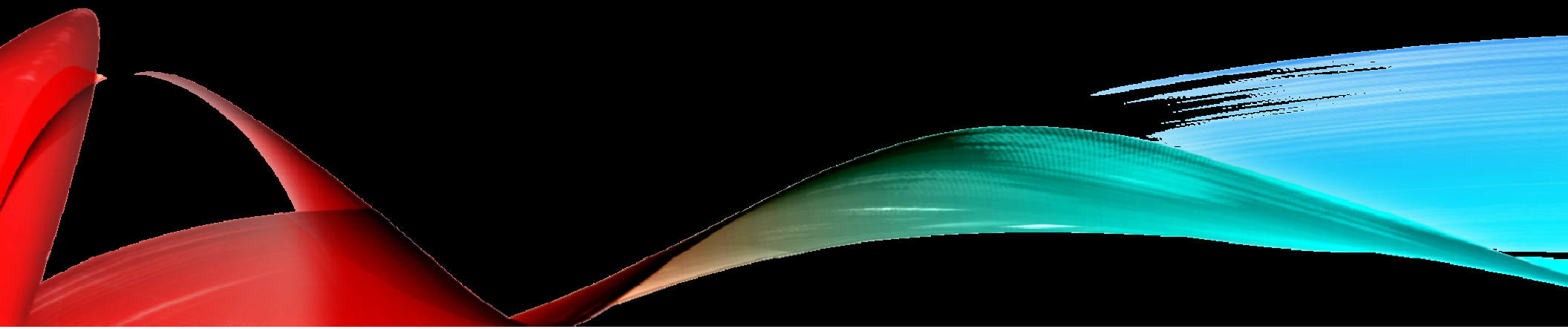


German Credit Case

Column > Recode



DISPROPORTIONATE SAMPLING



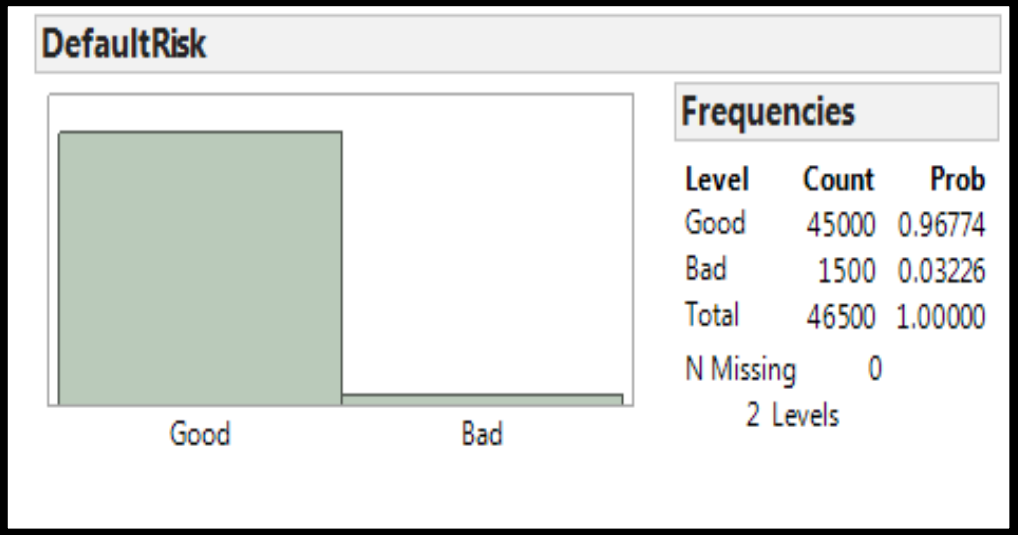
In certain situations there are a disproportionate number of “yes” responses to “no” responses when the response we want to model is a “yes”. (For example, see this blog post)

http://blogs.sas.com/content/jmp/2011/07/13/sampling-data-to-build-validate-and-test-predictive-models-when-target-event-occurs-infrequently/

When building models to predict a binary outcome variable, such as respond or not respond, the proportion in the desired category (respond) may be low. For example, in direct mail campaigns the response rate is often 1% or lower. In such cases, a predictive model is likely to learn how to predict the frequent outcome very well and the non-frequent outcome poorly.

There are various ways of ensuring better prediction of the non-frequent class, such as assigning higher weights to the infrequent events; stratified sampling to include a higher ratio of target events in the model building, selection and testing samples than would occur at random; or assigning different profit/costs associated with the different correct and incorrect predictions.

Suppose we want to predict customers who might default on a loan (based on a 2012 JMP Building Better Models presentation). The response variable shown is **DefaultRisk** with Good meaning the customer is a good risk and Bad meaning the customer is likely to default. As you can see there are 45,000 Good and 1500 Bad.



Bootstrap Forest for GB

Specifications

Target Column:	GB	Training rows:	27900
Validation Column:	Validation	Validation rows:	9300
		Test rows:	9300
Number of trees in the forest:	11	Number of terms:	23
Number of terms sampled per split:	5	Bootstrap samples:	27900
		Minimum Splits Per Tree:	10
		Minimum Size Split:	46

Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.2322	0.1621	0.1325	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.2582	0.1821	0.1494	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.1094	0.1194	0.1236	$\sum$ -Log(ρ[j])/n
RMSE	0.1672	0.1706	0.1717	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.0581	0.0593	0.0589	$\sum \lvert y[j]-\rho[j]\rvert/n$
Misclassification Rate	0.0323	0.0323	0.0323	$\sum$ (ρ[j]≠ρMax)/n
N	27900	9300	9300	n

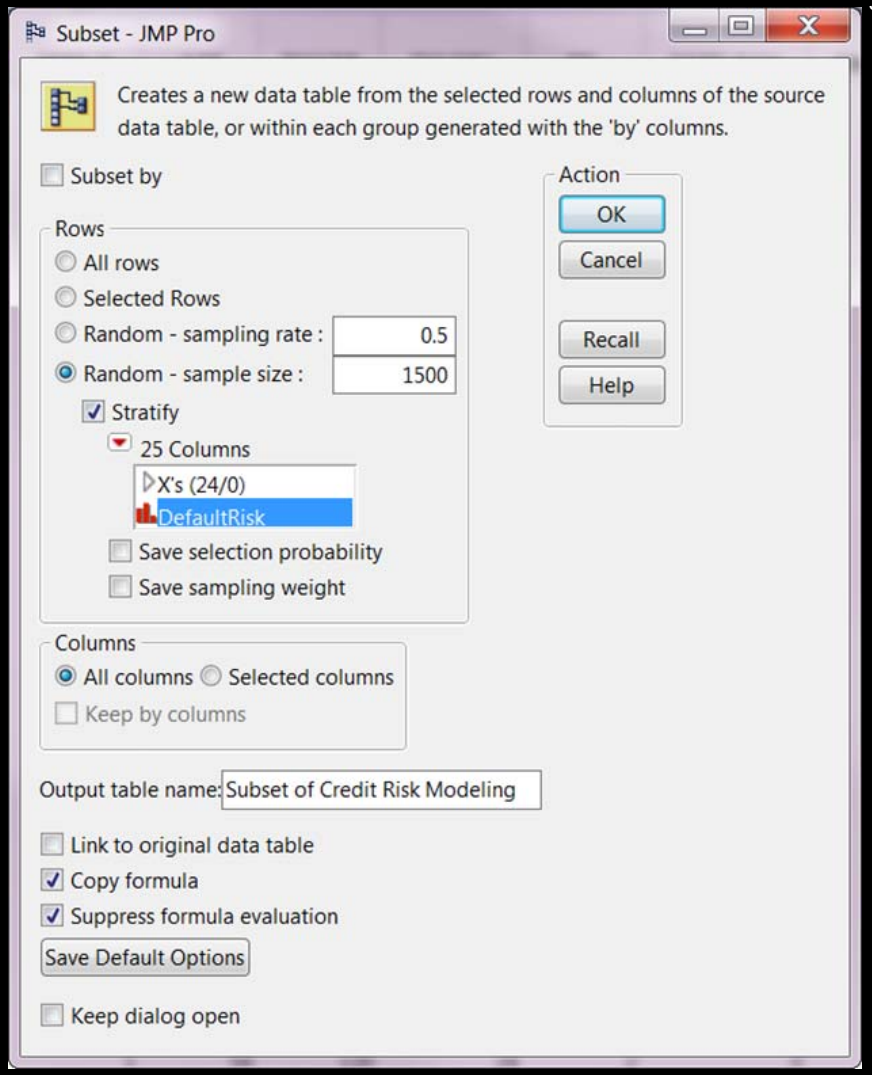
Confusion Matrix

Actual	Predicted		Actual	Predicted		Actual	Predicted	
Training	**0**	**1**	**Validation**	**0**	**1**	**Test**	**0**	**1**
0	27000	0	0	9000	0	0	9000	0
1	900	0	1	300	0	1	300	0

Step 1

Tables > Subset

- Select Random – sample size (enter number of "positives" – in this case 1500)
- Check Stratify
- Select the target "column" (response column – in this case DefaultRisk)
- Keep all other default values
- OK

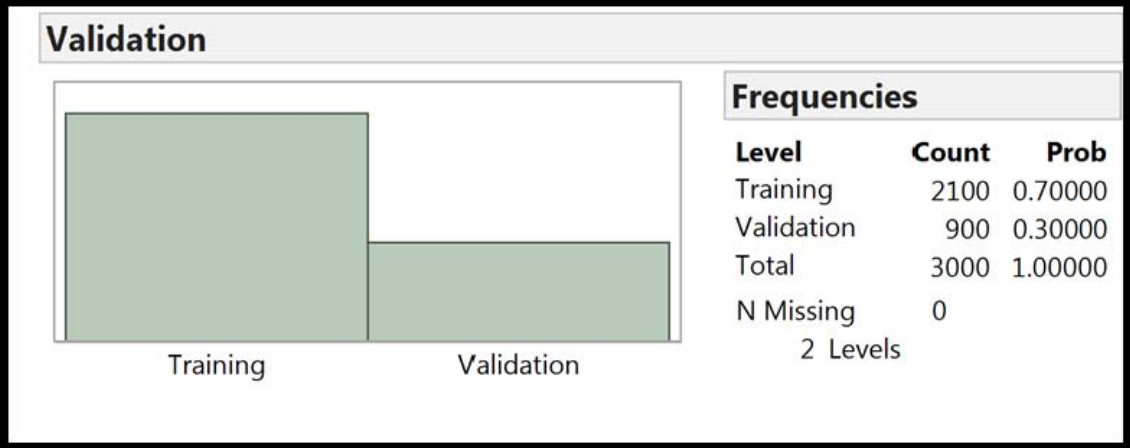


Step 2

Verify the new data table (subset of original data table) has twice the number of responses as the “random sample size” – in this case, 3000.

Step 2 - continued

Create validation column in the new subset tables with desired mix for training and validation (or training, validation and test).



Step 3 -

Develop model or models, and assess model performance

.

Bootstrap Forest for GB

Specifications

Target Column:	GB	Training rows:	2100
Validation Column:	validation2	Validation rows:	900
		Test rows:	0
Number of trees in the forest:	50	Number of terms:	23
Number of terms sampled per split:	5	Bootstrap samples:	2100
		Minimum Splits Per Tree:	10
		Minimum Size Split:	5

Overall Statistics

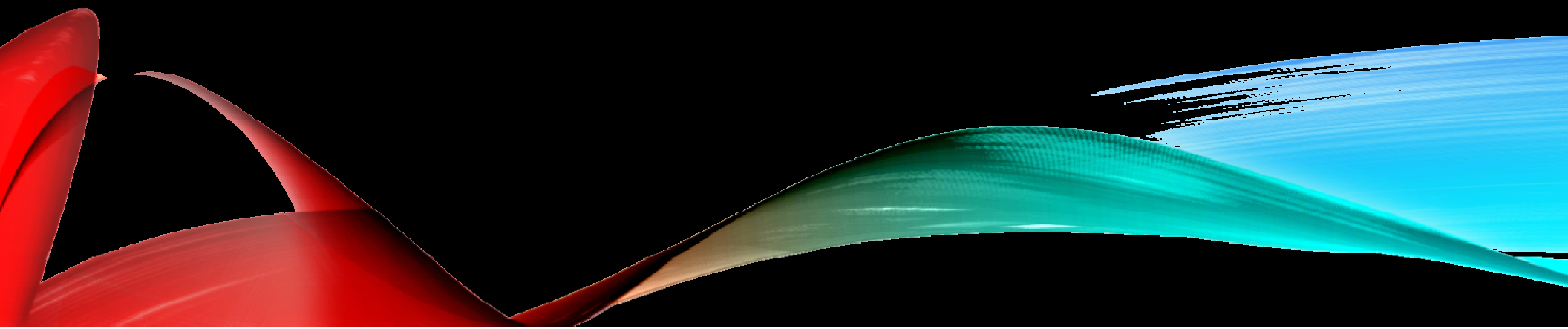
Measure	Training	Validation	Definition
Entropy RSquare	0.4040	0.1786	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.5717	0.2923	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.4130	0.5688	$\sum$ -Log(ρ[j])/n
RMSE	0.3505	0.4386	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.3231	0.4063	$\sum \lvert y[j]-\rho[j]\rvert/n$
Misclassification Rate	0.1048	0.2778	$\sum$ (ρ[j]≠ρMax)/n
N	2100	900	n

Confusion Matrix

Actual	Predicted	
Training	**0**	**1**
0	919	114
1	106	961

Actual	Predicted	
Validation	**0**	**1**
0	335	132
1	118	315

CUT OFF VALUES



Riding Mowers

Income	Lot_Size	Ownership	Lin[non-owner]	Prob[non-owner]	Most Likely
60	18.4	owner	1.55318	0.825373	non-owner
85.5	16.8	owner	0.268333	0.566684	non-owner
64.8	21.6	owner	-2.063034	0.112742	owner
61.5	20.8	owner	-0.926177	0.283701	owner
87	23.6	owner	-6.451653	0.001575	owner
110.1	19.2	owner	-4.771858	0.008394	owner
108	17.6	owner	-2.997008	0.047561	owner
82.8	22.4	owner	-4.829512	0.007927	owner
69	20	owner	-0.986593	0.271585	owner
93	20.8	owner	-4.418223	0.011912	owner
51	22	owner	-0.918697	0.285223	owner
81	20	owner	-2.316896	0.089733	owner
75	19.6	non-owner	-1.266233	0.219903	owner
52.8	20.8	non-owner	0.038292	0.509572	non-owner
64.8	17.2	non-owner	2.177594	0.898219	non-owner
43.2	20.4	non-owner	1.488046	0.815785	non-owner
84	17.6	non-owner	-0.336402	0.416684	owner
49.2	17.6	non-owner	3.521476	0.971293	non-owner
59.4	16	non-owner	3.932765	0.980787	non-owner
66	18.4	non-owner	0.888029	0.708483	non-owner
47.4	16.4	non-owner	4.877557	0.992442	non-owner
33	18.8	non-owner	4.16085	0.984645	non-owner
51	14	non-owner	6.791536	0.998878	non-owner
63	14.8	non-owner	4.690209	0.990899	non-owner

	non-owner	owner		
non-owner	10	2		16.67%
owner	2	10		16.67%

	A	B	C	D	E	F	G
1							
2							
3							
4							
5					G8	=IF(E8>$B$6,"non-owner","owner")	
6	Cutoff	0.5					
7	Income	Lot_Size	Ownership	Lin[non-owner]	Prob[non-owner]	Most Likely	**Classify**
8	60	18.4	owner	1.55318	0.825373	non-owner	non-owner
9	85.5	16.8	owner	0.268333	0.566684	non-owner	non-owner
10	64.8	21.6	owner	-2.063034	0.112742	owner	owner

Can copy/paste into EXCEL or
File > Save as > EXCEL

	A	B	C	D	E	F	G	H	I	J	K	L
6	Cutoff	0.5							non-owner		owner	
7	Income	Lot_Size	Ownership	Lin[non-owner]	Prob[non-owner]	Most Likely	**Classify**		TP	FP	TN	FN
8	60	18.4	owner	1.55318	0.825373	non-owner	non-owner		0	1	0	0
9	85.5	16.8	owner	0.268333	0.566684	non-owner	non-owner		0	1	0	0
10	64.8	21.6	owner	-2.063034	0.112742	owner	owner		0	0	1	0
11	61.5	20.8	owner	-0.926177	0.283701	owner	owner		0	0	1	0

TP =IF(AND($G8=$I$6,$C8=$I$6),1,0)

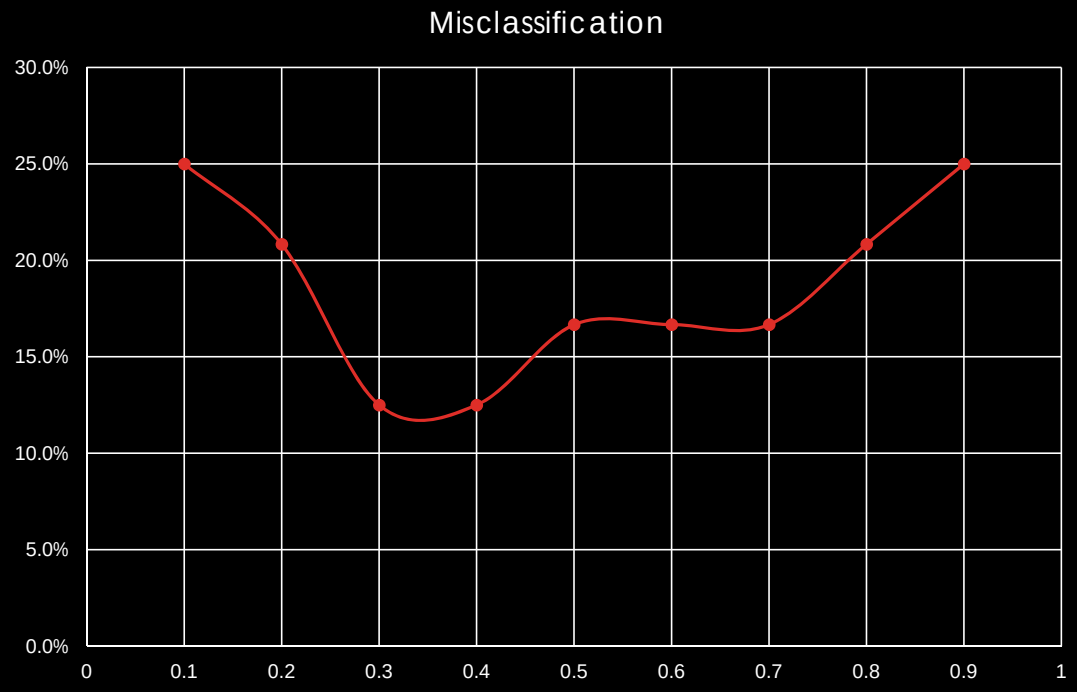
FP =IF(AND($G8=$I$6,$C8<>$I$6),1,0)

TN =IF(AND($G8=$K$6,$C8=$K$6),1,0)

FN =IF(AND($G8=$K$6,$C8<>$K$6),1,0)

	non-owner	owner		
non-owner	10	2		16.67%
owner	2	10		16.67%
			16.7%	

Misclassification	
	16.67%
0.1	25.0%
0.2	20.8%
0.3	12.5%
0.4	12.5%
0.5	16.7%
0.6	16.7%
0.7	16.7%
0.8	20.8%
0.9	25.0%



	non-owner	owner		
non-owner		20		
owner	20	8		160

Cost		
	$	160
0.1	$	168
0.2	$	156
0.3	$	140
0.4	$	140
0.5	$	160
0.6	$	168
0.7	$	168
0.8	$	188
0.9	$	216