Paper 151-2011

Data mining using JMP®

Murphy Choy, School of Information Systems, SMU, Singapore

ABSTRACT

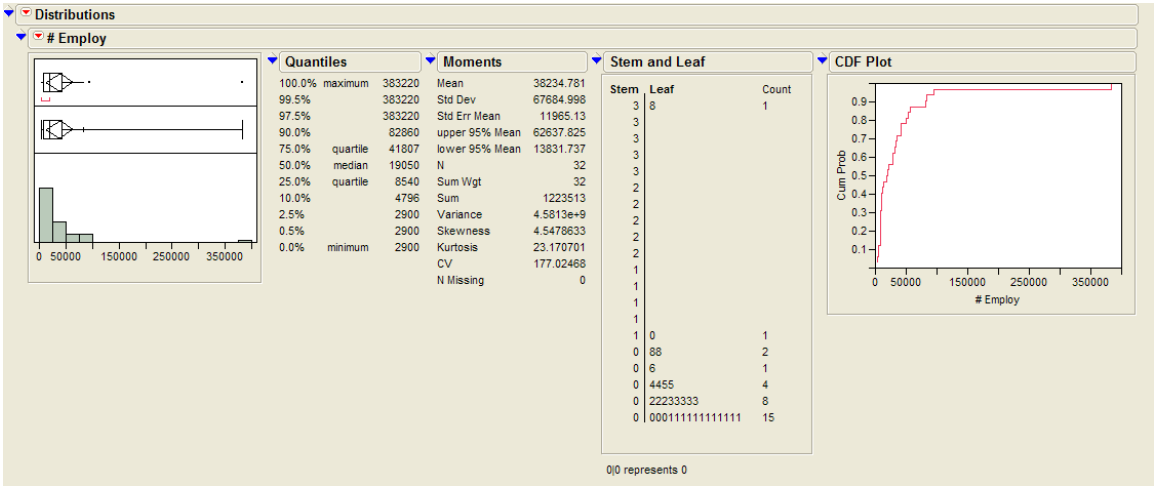
Data mining is increasingly adopted by companies who seek to maximize their overall operational efficiency and profitability. However, most companies are unwilling to invest huge capital to procure data mining software. With the worsening economy and drying up of credits and capital, companies are hard pressed to come up with the capital for such facilities which can help them streamline their process and improve the bottom line. JMP® is a simple and cheap alternative created by SAS® for six sigma and quality control. In addition, it has many features which can be adapted for data mining. In this paper, we will explore some of the options that JMP has to offer to companies who want to do serious data mining at an affordable price.

INTRODUCTION

Data mining has become an increasingly important tool in shaping business strategy and customer relationship management. With such increasing demand for analytical tools, a variety of high powered data mining tools have been developed to handle the requirements for individual sectors. However, there is a lack of solutions for small and medium enterprises to start business analytics endeavors. Very often, these enterprises rely on open source software or even base level programming to do their business analytics. However, with the availability of JMP, even SME will be able to do their business analytics easily.

EXPLORATORY DATA ANALYSIS

Exploratory data analysis is done in a graphical manner in JMP. The amount of information is very rich and the graphics are excellent. There is also a list of choices for the type of analysis that a modeler might wish to have. Using the exploratory data analysis section, we will be able to have a sense of the distribution of the data values.



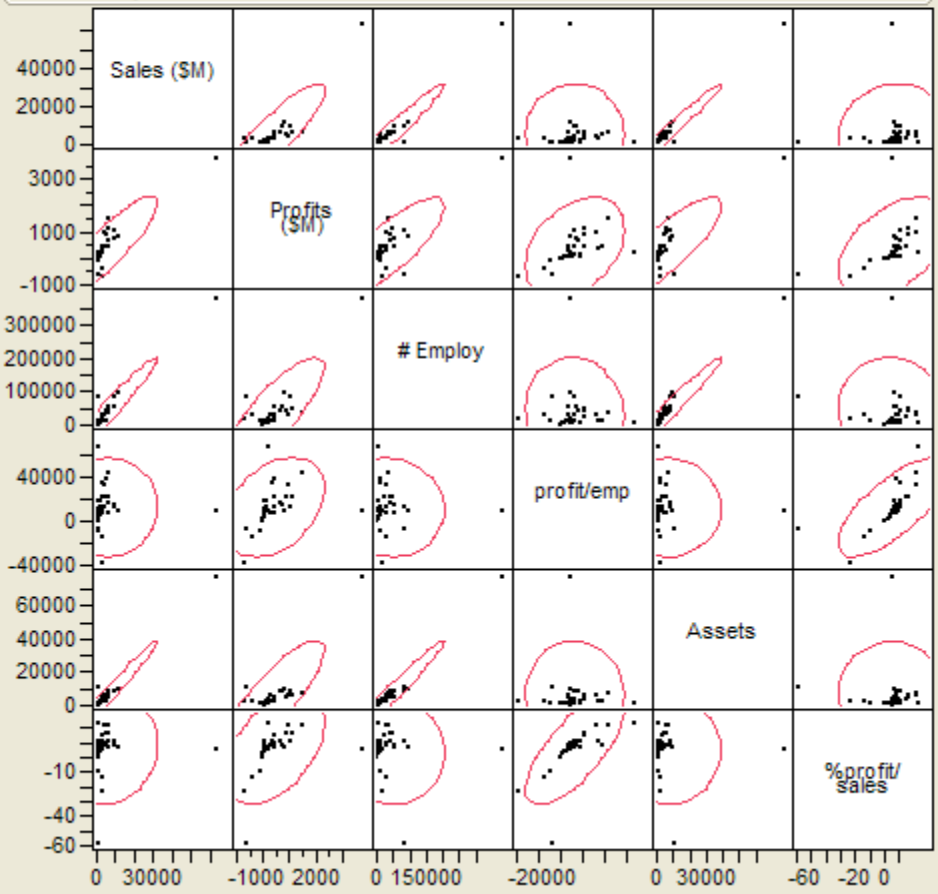
There are also options for analyzing the correlations between the different variables and understanding how some of the variables relate to one another. Under the multivariate options, there is a particular option called multivariate which is basically a correlation analysis with scatter plots.

Multivariate

Correlations

	Sales ($M)	Profits ($M)	# Employ	profit/emp	Assets	%profit/sales
Sales ($M)	1.0000	0.8745	0.9734	0.0152	0.9846	0.0790
Profits ($M)	0.8745	1.0000	0.8192	0.3704	0.8301	0.4604
# Employ	0.9734	0.8192	1.0000	-0.0626	0.9813	-0.0737
profit/emp	0.0152	0.3704	-0.0626	1.0000	-0.0151	0.7130
Assets	0.9846	0.8301	0.9813	-0.0151	1.0000	-0.0251
%profit/sales	0.0790	0.4604	-0.0737	0.7130	-0.0251	1.0000

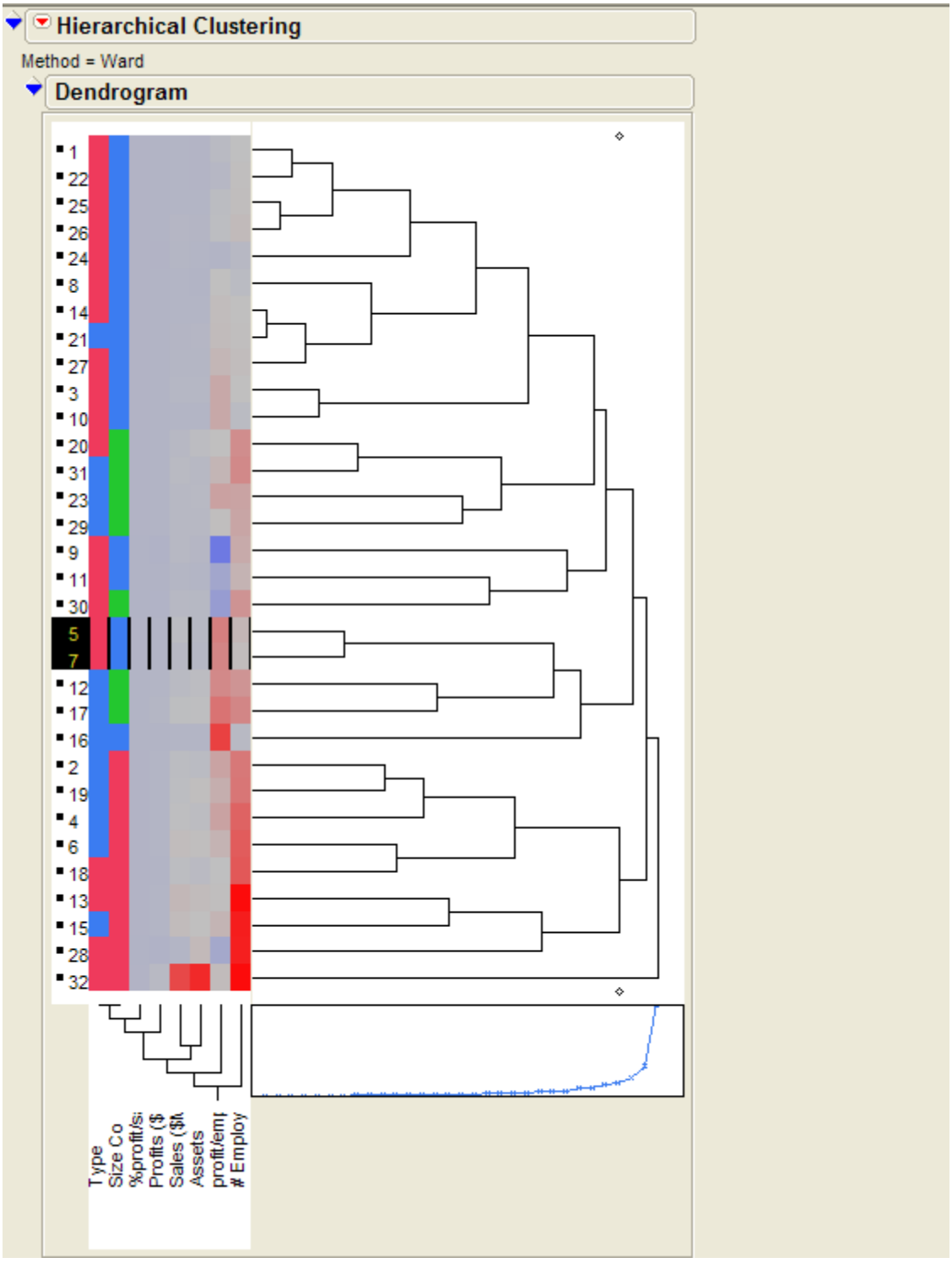
Scatterplot Matrix



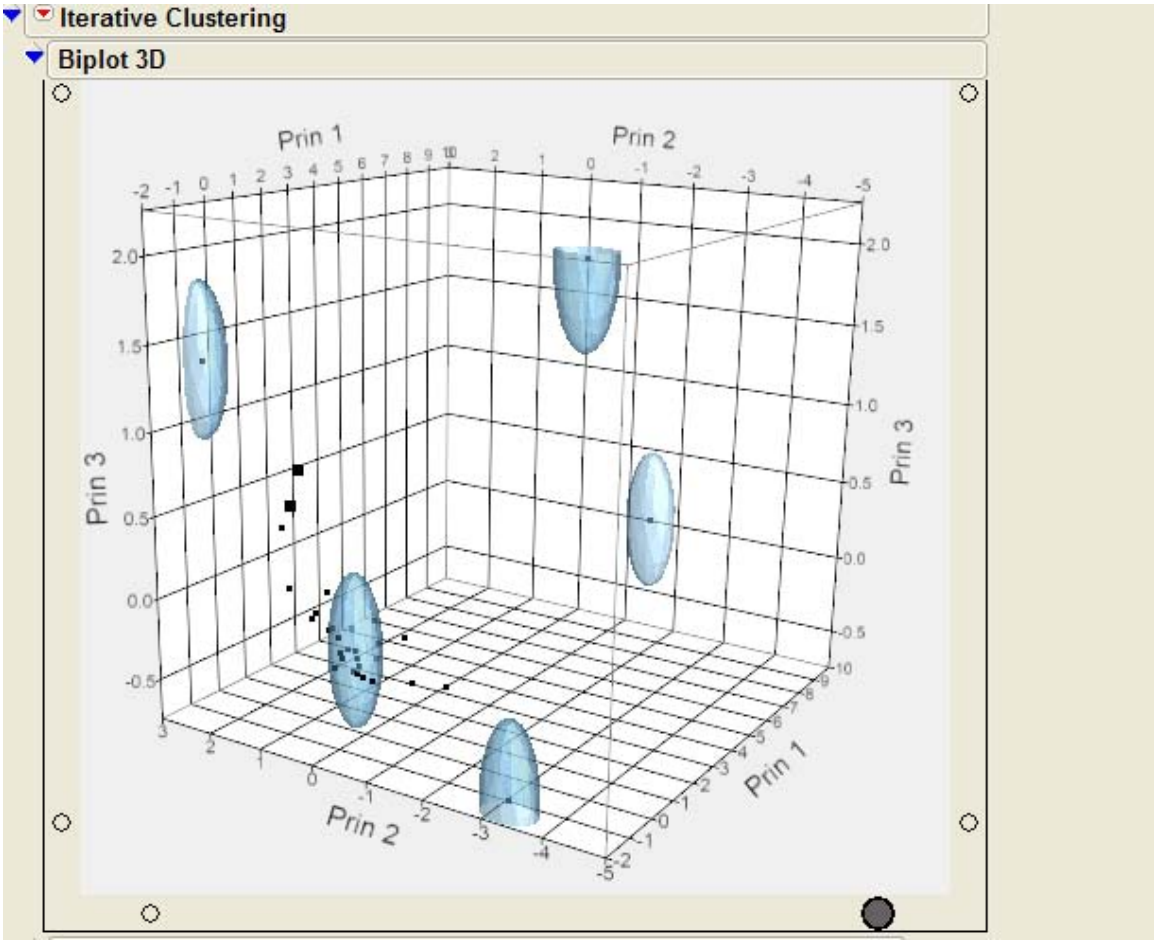
Multivariate analyses are useful in helping the modeler to understand the data. At the same time, it can help to uncover some interesting facts which would otherwise not be known.

CLUSTER ANALYSIS

Cluster analysis is also widely used as a tool to explore the data, especially in terms of grouping certain classes of observations together based on the selected variables. There are two types of clustering approach, the first one being the hierarchical approach and the second one being K-means. Both approaches are implemented in JMP.



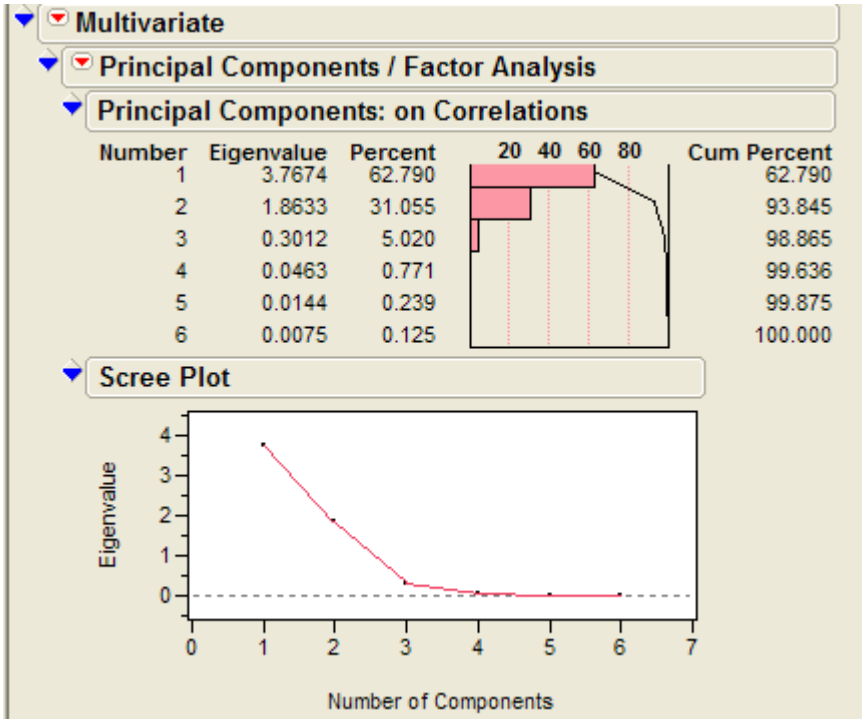
Hierarchical clustering is usually done in a direct attempt to understand the dimensions in the data and find out how many natural occurring clusters exist in the data. This is usually done first as most people will have very vague ideas about what they have in the data. At the same time, it helps to have a sense of how the observations are different from one another. JMP can also implement a distance graph which can be used to determine the number of clusters in the population using the L-intersection technique.



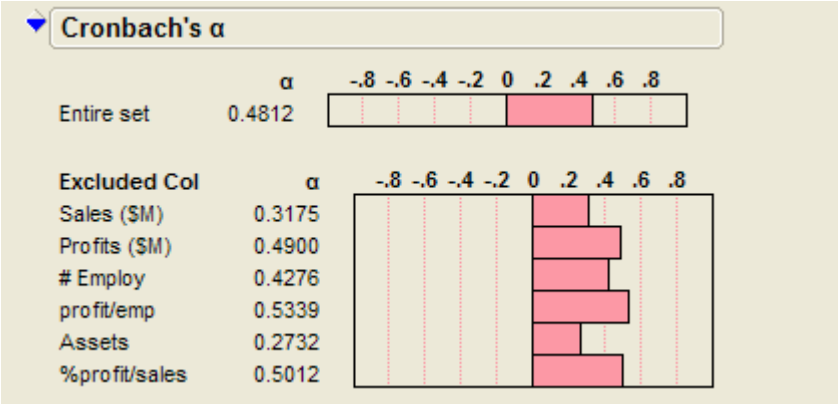
After having a good idea about the number of clusters in the data, it will be good to then separate the data into the desired natural number of clusters. Hierarchical clustering, however, is not always necessary as the aim of the separation might be to make the groups more distinct, in which case the number of clusters depends on the user's requirement. As for K – Means clustering, it has the additional advantage of detecting outliers as these observations have a tendency to have values that are far beyond the normal ranges.

PRINCIPAL COMPONENT ANALYSIS

Occasionally, the modeler may desire a simplification of the data structure to make it easier for people to understand and able to explain sufficient amount of the data variation. This leads to principal component analysis which is a critical aspect of variable reduction and exploration. Principal component analysis is also an important aspect of data understanding as it seeks to reduce the variables into common ones related to one another.

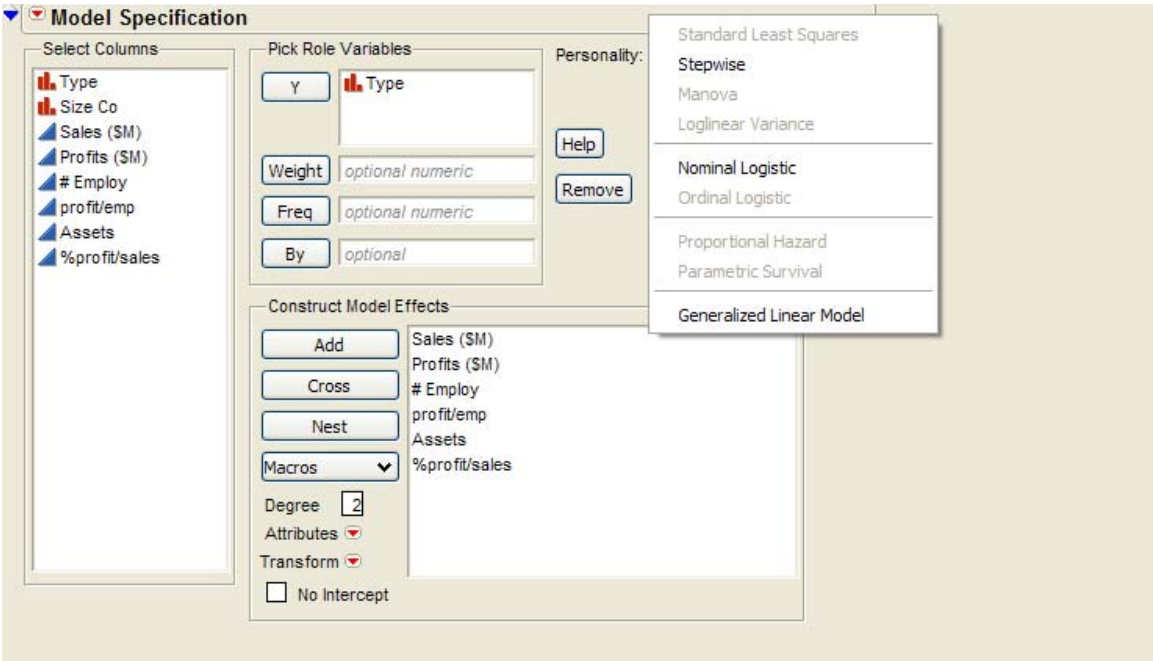


JMP produces excellent outputs with scree plot and cumulative percentages calculated. The graphical display helps in highlighting important factors and other stuff. JMP also calculates the Cronbach's alpha which is an indicator of the reliability of the data and particularly important to psychology and pharmaceutical studies.

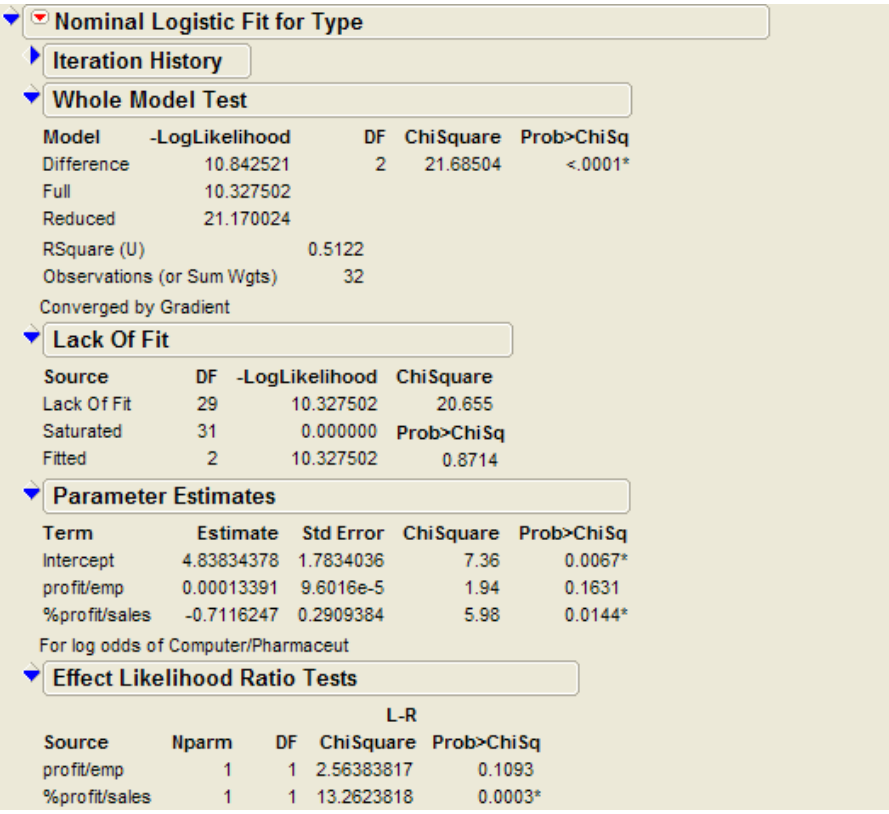


REGRESSION MODELS: LOGISTICS AND LEAST SQUARE

Regression models are the main work horses of data mining. Most of the models built for direct marketing, credit risk scorecards and other propensity scoring models are regression models. There are several types of regression models which can be built within JMP itself. The classical least square regression model and the logistic regression model are just two of the many models. At the same time, JMP also provides very good interactive features which allows users to select variables on their own.



JMP also provides many useful and critical statistics for the evaluation of the efficiency and effectiveness of the regression models.



Nominal Logistic Fit for Type

Iteration History

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	10.842521	2	21.68504	<.0001*
Full	10.327502			
Reduced	21.170024			

RSquare (U)	0.5122
Observations (or Sum Wgts)	32

Converged by Gradient

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	29	10.327502	20.655
Saturated	31	0.000000	Prob>ChiSq
Fitted	2	10.327502	0.8714

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	4.83834378	1.7834036	7.36	0.0067*
profit/emp	0.00013391	9.6016e-5	1.94	0.1631
%profit/sales	-0.7116247	0.2909384	5.98	0.0144*

For log odds of Computer/Pharmaceut

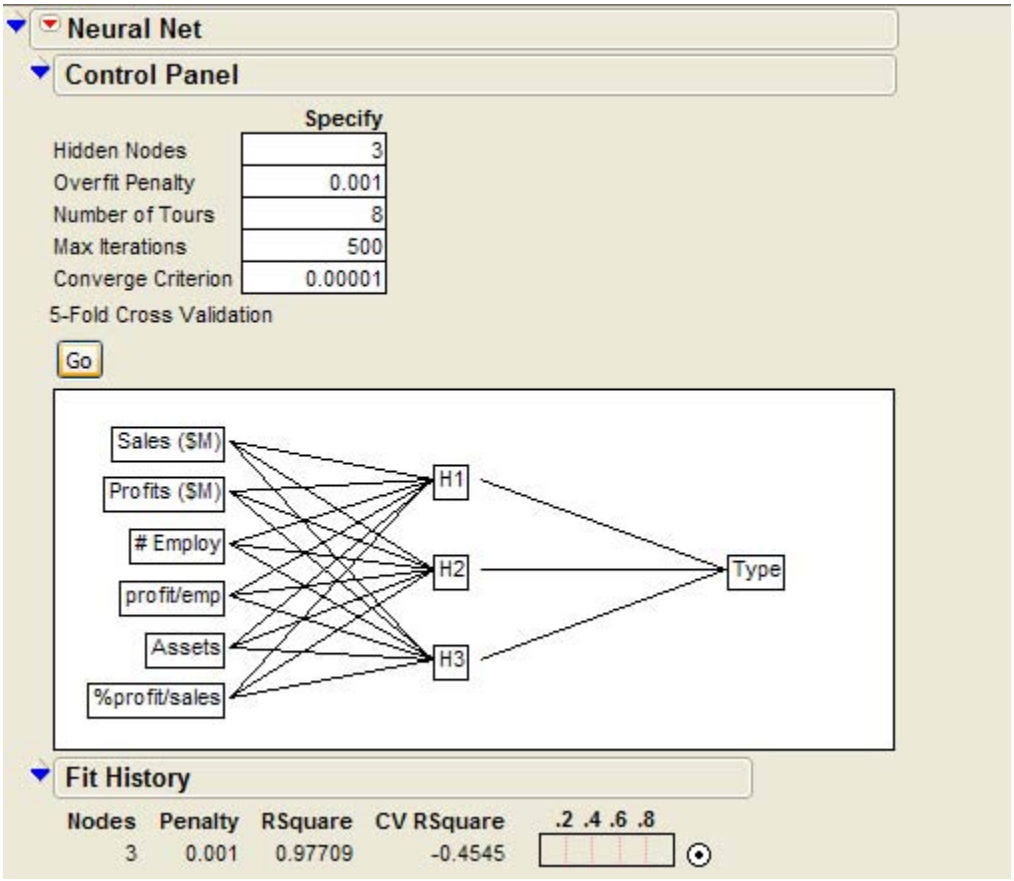
Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
profit/emp	1	1	2.56383817	0.1093
%profit/sales	1	1	13.2623818	0.0003*

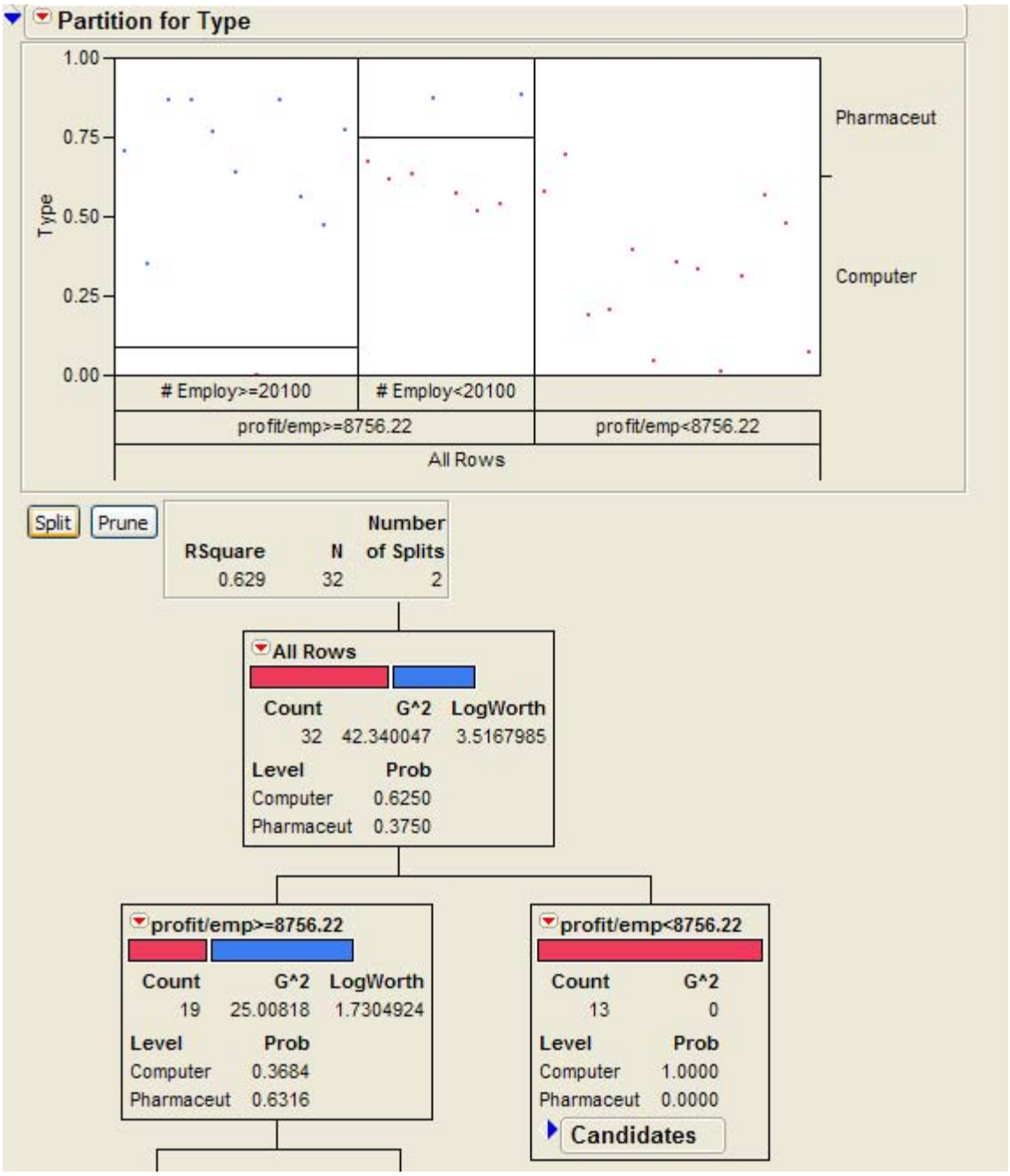
NEURAL NETWORKS AND DECISION TREE

Neural nets are usually black box models which are difficult to comprehend despite their power in handling non linear models. In JMP, the process is simplified by the diagram that represents the neural network built

on the data. These simplifications make it easier for more experienced users to rapidly adjust the neural network to predict the model better and at the same time allows them to modify things easily.



Decision trees are models that partition the data in smaller section which then allows them to predict targets or to conduct some form of segmentation based on certain targets. Decision tree are particularly popular among data miners as they are simple to comprehend and the results are highly adjustable which can then modeled around business sense and also statistical importance. JMP provides an excellent interface with the ability to grow the tree interactively with the choice of variables and the value cuts. At the same time, the model parameters can be set easily using the interface. The elaborate tree diagram with ROC curves and Lift curves also help to assess the power of the model.



CONCLUSION

JMP is a very powerful software package that allows users to create excellent models at a relatively cheap price. This is important to companies who wish to acquire business analytics power but do not wish to spend too much of their budget on it initially.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Murphy Choy
Enterprise: School of Information Systems, Singapore Management University
Address: 80 Stamford Road

City, State ZIP: Singapore 178902
Work Phone: +65-92384058
E-mail: goladin@gmail.com/murphychoy@smu.edu.sg