Monthly User Guide from JMP Korea

제 8호 (2018년 3월)

Partition [Decision Tree, 의사 결정 나무]

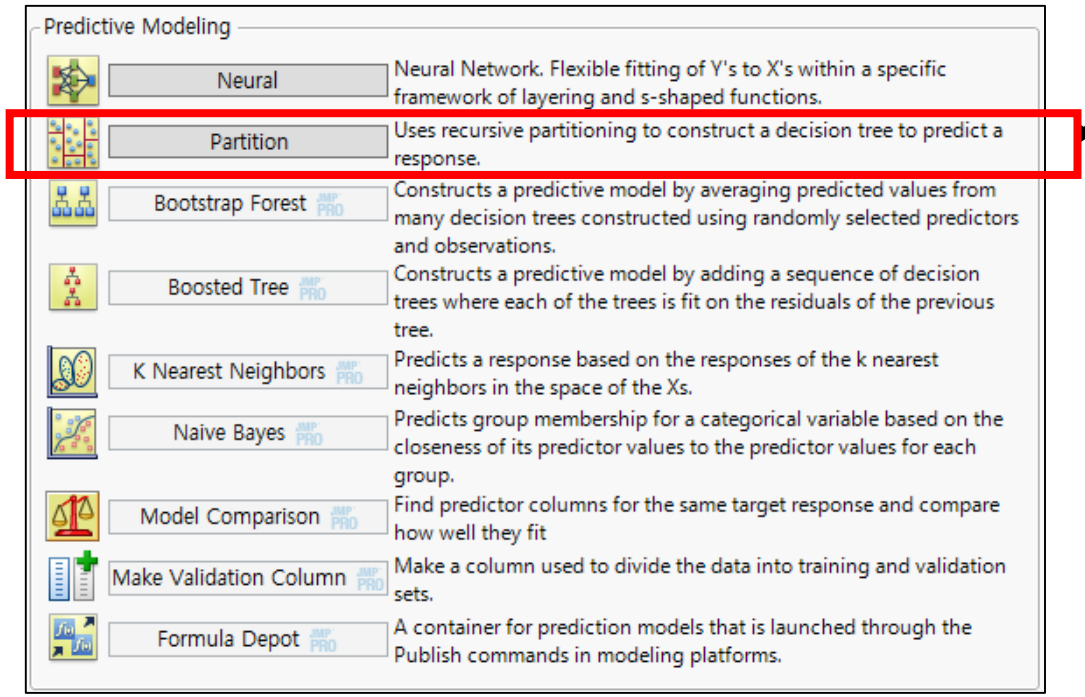
* 본 Guide 는 매월 네 번째 수요일에 발행됩니다 (다음 호는 2018년 4/28(수)일에 발행 예정입니다)
** Monthly User Guide 지난 호는 다음 Site 를 참조하세요. https://www.jmp.com/ko_kr/newsletters.html
*** **JMP 14 Version 이 출시되었습니다만, 본 Guide 는 당분간 13 Version 기준으로 작성될 예정입니다**
**** 본 Guide 의 내용과 관련한 문의는 Ikju.Shin@jmp.com 으로 문의 바랍니다

Predictive Modeling in JMP

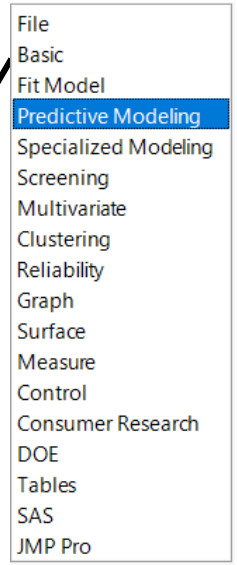
1. Predictive Modeling 은 Predictive Modeling, Exploratory Modeling 및 Data Mining 기법 등 매우 다양한 이름으로 불립니다
   (경우에 따라서는 'Big data 분석 방법' 이라는 다소 주관적인 다른 이름으로 불리기도 합니다).
   규칙 발견, 분류 및 모델링 등을 목적으로 보통은 자동화된 방법을 이용하여 대용량의 데이터를 분석하는 통칭하는 개념이라 할 수 있겠습니다.

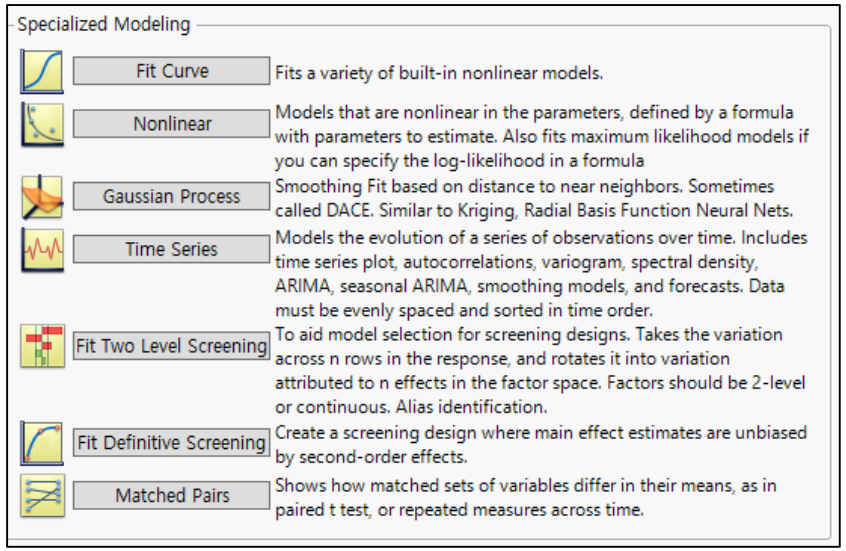
2. JMP Pro 에는 매우 다양한 Predictive Modeling 기법이 포함되어 있고, JMP 에는 Partition 과 Neural Network 이 포함되어 있는 데 여기서는 Partition 에 대해서 소개하도록 하겠습니다 (이론 및 분석 알고리즘에 대한 설명은 제외하고, JMP 활용 측면만 설명)

JMP 에서 Predictive Modeling 기법



View / JMP Starter





Partition [Decision Tree]

1. 출력 변수(Y) 가 존재할 경우 많이 사용
   1) 출력 변수 Y 를 설명하기 위한 인자의 모든 조합을 찾음
      (all possible splits of predictors to best predict the response)
   2) 이러한 조합(분기) 은 의사 결정 나무의 형태로, 최적의 모델이 만들어질 때까지 반복적으로 이루어짐
      (These splits (or partitions) of the data are done recursively to form a tree of decision rules.
      The splits continue until the desired fit is reached)

2. 이러한 분기(Split, Partition) 방법은
   1) 분기의 통계적 알고리즘에 따라 CART*, CHAID** 라고 불리우기도 하고
   2) 분기의 최종 모습을 본 따 Decision Tree 또는 통칭하여 Partition Model 로도 불림.

3. Y 및 X 의 모든 Data 유형(continuous, categorical-ordinal & nominal) 에 활용될 수 있으며,
   일반적으로 반응치 Y 가 categorical data 일 때, 보다 유용하다고 알려져 있다.

4. Decision Tree 의 장단점
   1) 장점 : Data에 대해 사전 정보가 충분하지 않을 때 많이 사용, 복잡한 문제에 대해 쉽게 해석이 가능
   2) 단점 : Overfitting 이 발생할 가능성이 높음.
      첫 번째 분기 기준으로 계속 분기하므로 첫 번째 분기가 잘못되어 있으면 계속 잘못됨

5. Decision Tree 분석 결과의 활용
   1) 각 분류 범주(Split)별로 Cost, Risk, 실현 가능성 등을 종합적으로 고려하여 평가하고
   2) 이를 근거로 분류(Classification) 깊이를 선택하거나 예측(Prediction) 해야 함

* CART(Classification & Regression Tree), ** CHAID(Chi-Squared Automatic Interaction Detection)

Partition [Decision Tree]

1. Sample file : diabetes.jmp
 1) 당뇨병과 관련하여 수집된 data
 2) 본 사례에서는 반응치(response) 로 Y(ordinal) 사용 : 혈당 수치를 상/중/하로 구분
 3) Age ~ Glucose 까지 10개의 factor
 4) 분석 목적 : 혈당에 영향을 미치는 주요 인자 선별

2. JMP Menu
 1) analyze / predictive modeling / partition
  Y : Y(ordinal)
  X : Age ~ Glucose 까지 10개의 factor
 2) 오른쪽 결과는
  -'Split' 를 한 번 click
  -'Color points' click
  (click 하게 되면 상단 그림이 color 화되고 해당 icon 은 사라짐)
  -하단 'candidates' 를 펼친 결과임

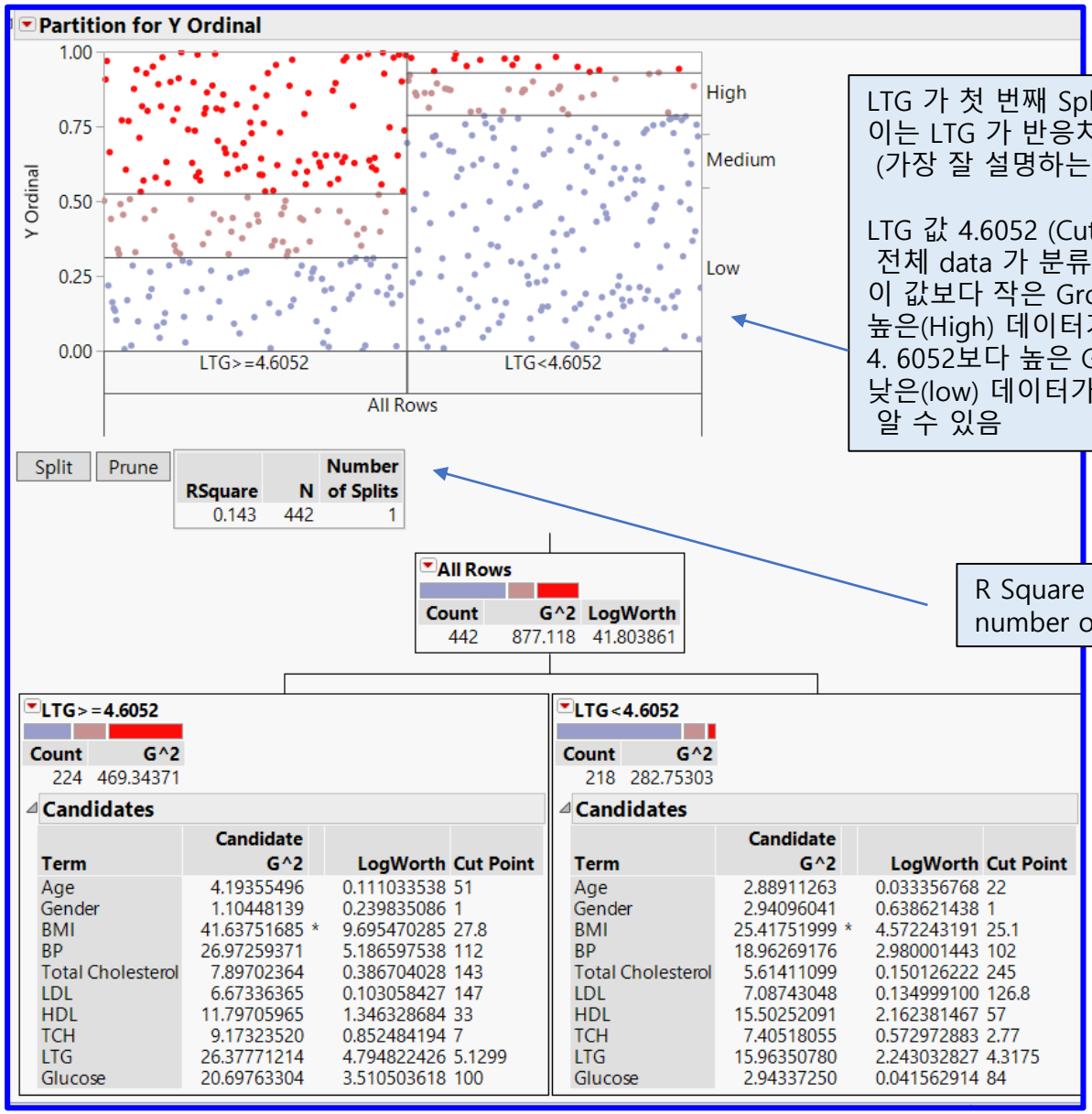
Split : 분기
Prune : 가지치기 직전 Split 제거 'Split' 의 역순

Count : 분석 대상으로 선택된 data 의 개수

G^2 : 해당 변수를 가지고 Split 한 경우의 G^값(값이 클 수록 영향을 많이 주는 변수)
-Y 가 연속형일 경우 제곱합(SS)이 출력됨
- * : 다음 Split 의 기준이 되는 변수 (그 다음으로 중요한 변수라는 뜻)

LogWorth : - log(p-value), 이 값이 높을 수록 출력변수를 잘 설명하는 변수임.

Cut Point : 해당 변수의 Split 기준 값. 분기점(split point) 를 나타냄



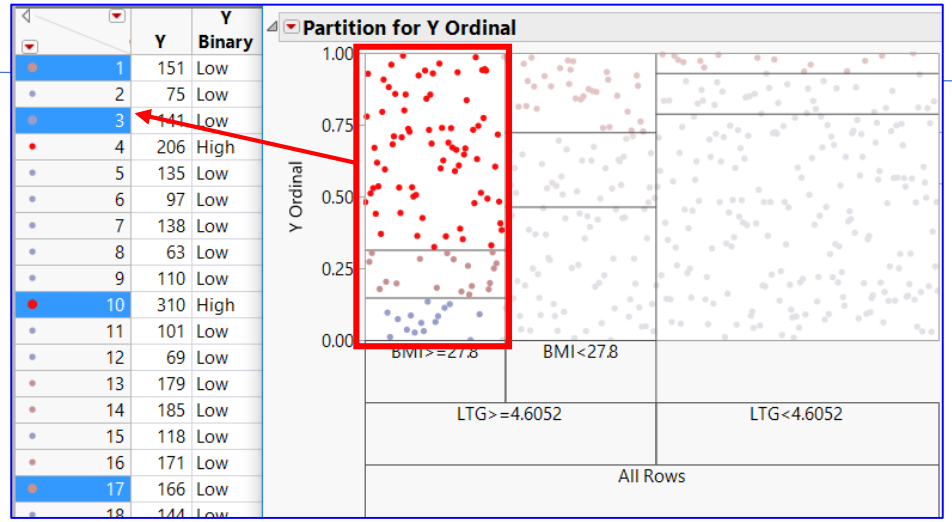
LTG 가 첫 번째 Split 인자로 선정됨.
이는 LTG 가 반응치를 가장 잘 분류하는 (가장 잘 설명하는) 인자라는 뜻.

LTG 값 4.6052 (Cut Point) 를 기준으로 전체 data 가 분류되고
이 값보다 작은 Group 의 경우 혈당이 높은(High) 데이터가 약 50% 정도 되지만
4. 6052보다 높은 Group 의 경우 혈당이 낮은(low) 데이터가 약 80% 정도 됨을 알 수 있음

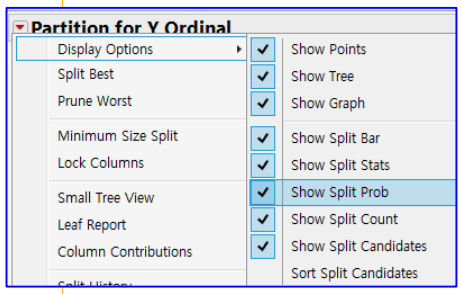
R Square : 현재의 R^2 값
number of Splits : 분기 횟수

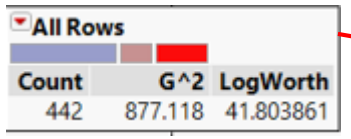
Partition [Decision Tree]

3. 그래프에서 특정 영역을 선택하면 data table 에서 이에 해당되는 data 가 별도 표시됨을 알 수 있다 (Interaction !!!)

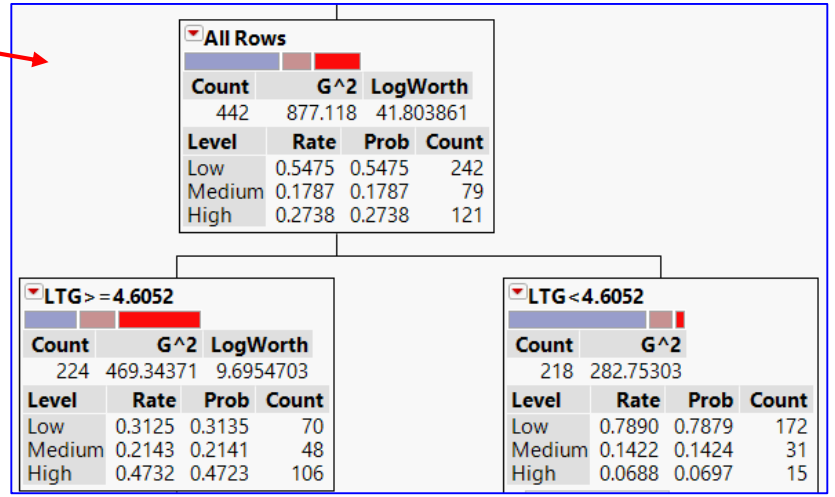


4. Y 가 Categorical data 일 경우,
show split prob 및 show split count 를 선택하면 Split 된 Group 별로 Data 의 Probability 및 Count 를 확인할 수 있다

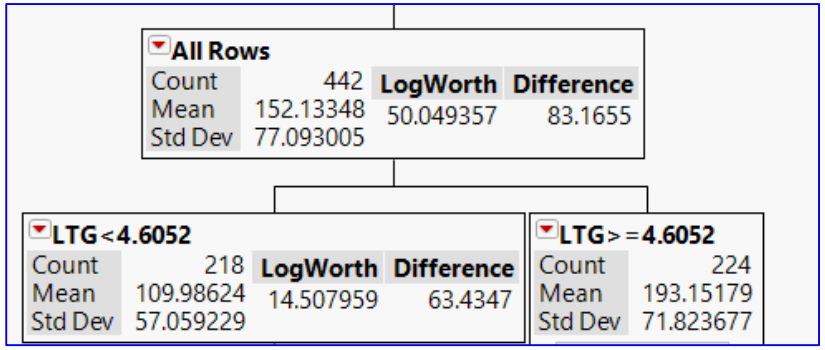




show split prob, show split count 를 선택하면 Split Group 별 정보가 오른쪽처럼 바뀌게 된다



5. 만약, Y 가 Continuous data 라면
Split 된 Group 별로
Data 의 평균과 표준편차 값이 표시 된다

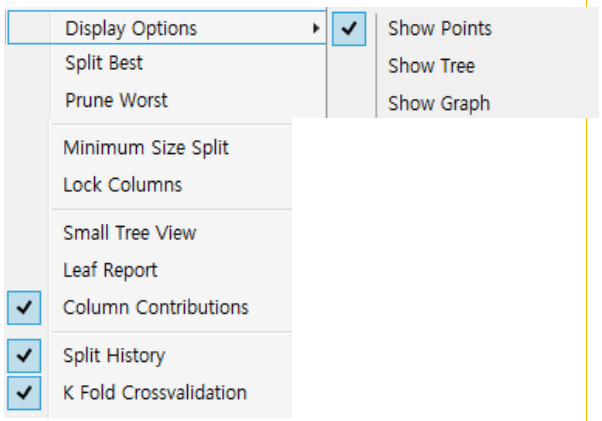


Partition [Decision Tree]

6. 얼마만큼 Split 해야 하는가 ?
(반응치에 중요한 영향을 주는 변수는 무엇인가)
1) Partition for Y Ordinal 의 왼쪽 빨간색 역삼각형 click 후 'Column contributions' 실행
: 반응치에 영향을 많이 주는 순서대로 X 인자에 대한 정보가 표시된다

2) 이론적으로는 (Data 개수 – 1) 만큼 split 하여 완벽한(!) 모델은 만들 수 있으나, overfitting 의 문제로 이러한 모델을 만드는 것은 현실성이 없다

3) K Fold cross validation 개념을 활용하여 적절한 split 개수를 확인할 수 있다
-display options 에서 show tree 와 show graph 를 선택 해제
-Column contributions, split history, K Fold cross validation 을 선택
-그런 다음 split icon 을 계속 누르다 보면 K Fold R-sq 값이 작아지는 지점이 발생하는데, 여기서부터 overfitting 이 발생한다고 볼 수 있다.



Partition for Y Ordinal

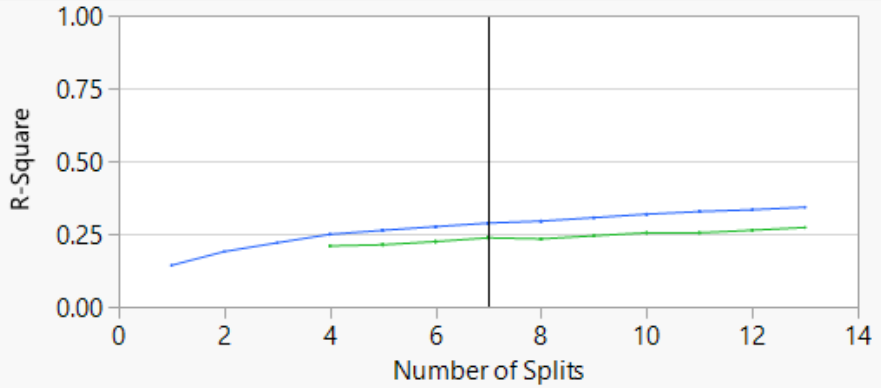
Split | Prune | Go | Color Points

RSquare	N	Number of Splits
0.288	442	7

Column Contributions

Term	Number of Splits	G^2	Portion
LTG	3	149.164044	0.5871
BMI	2	67.0550368	0.2639
BP	1	26.6327744	0.1048
Glucose	1	11.2326879	0.0442
Age	0	0	0.0000
Gender	0	0	0.0000
Total Cholesterol	0	0	0.0000
LDL	0	0	0.0000
HDL	0	0	0.0000
TCH	0	0	0.0000

Split History



K-Fold in Green

Crossvalidation

k-fold		-2LogLike	RSquare
5	Folded	669.107348	0.2372
	Overall	623.03346	0.2876

예제에서는 Split 이 8회가 되면
K Fold R-sq 값이 작아짐을 알 수 있다

RSquare	N	Number of Splits
0.294	442	8

Crossvalidation

k-fold		-2LogLike	RSquare
5	Folded	672.24628	0.2336
	Overall	616.251031	0.2943

<K-Fold Cross validation 에 대해서는 아래 내용 참조>

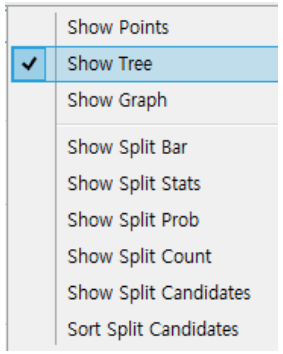
Randomly divides the original data into K subsets. In turn, each of the K sets is used to validate the model fit on the rest of the data, fitting a total of K models. The final model is selected based on the cross validation RSquare, where a stopping rule is imposed to avoid overfitting the model. This method is useful for small data sets, because it makes efficient use of limited amounts of data.

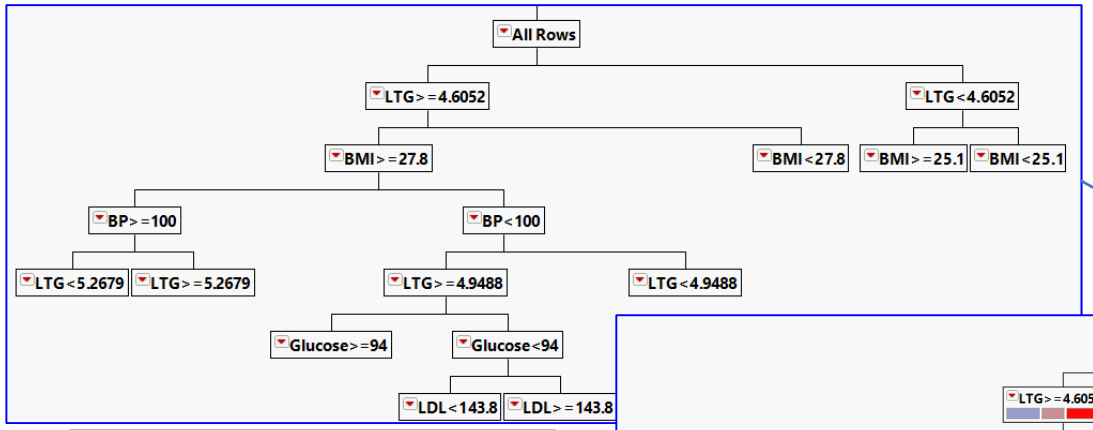
Number of folds.
2LogLike or SSE
Gives twice the negative log-likelihood (-2LogLikelihood) values when the response is categorical. Gives sum of squared errors (SSE) when the response is continuous. The first row gives results averaged over the folds. The second row gives results for the single model fit to all observations.
RSquare
The first row gives the RSquare value averaged over the folds. The second row gives the RSquare value for the single model fit to all observations

Partition [Decision Tree]

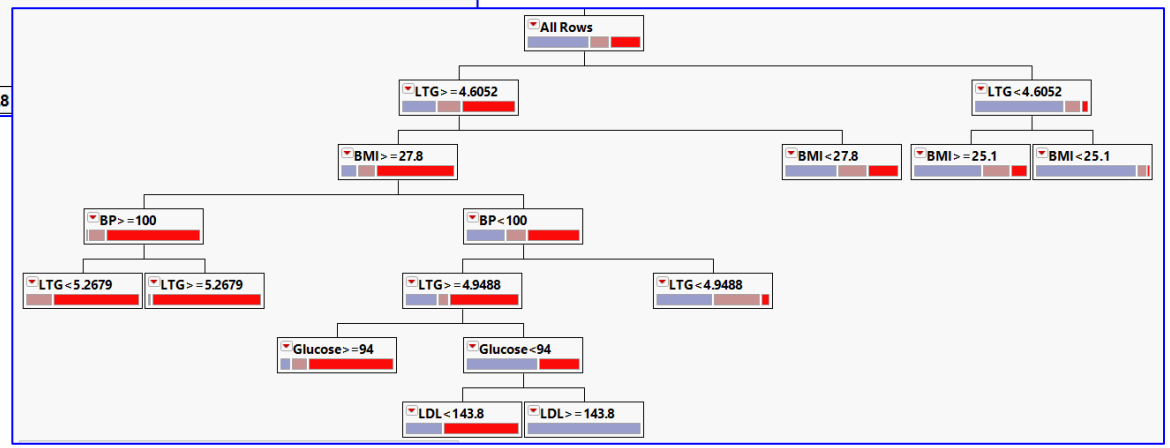
7. Split 회수가 많거나, 인자 수가 많으면 Graph 가 복잡해 지는 데 아래와 같은 기능을 이용하여 Graph 를 간략화할 수 있다

1) Display option 에서 'show tree' 만 선택

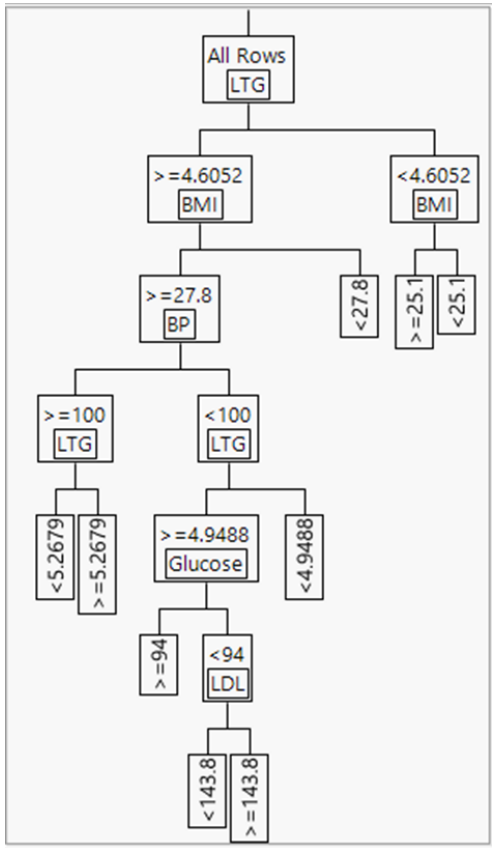




Show split bar 를 추가적으로 선택



2) Partition for Y Ordinal 의 왼쪽 빨간색 역삼각형 click 후 'small tree view' 를 선택한 후의 결과



Partition [Decision Tree]

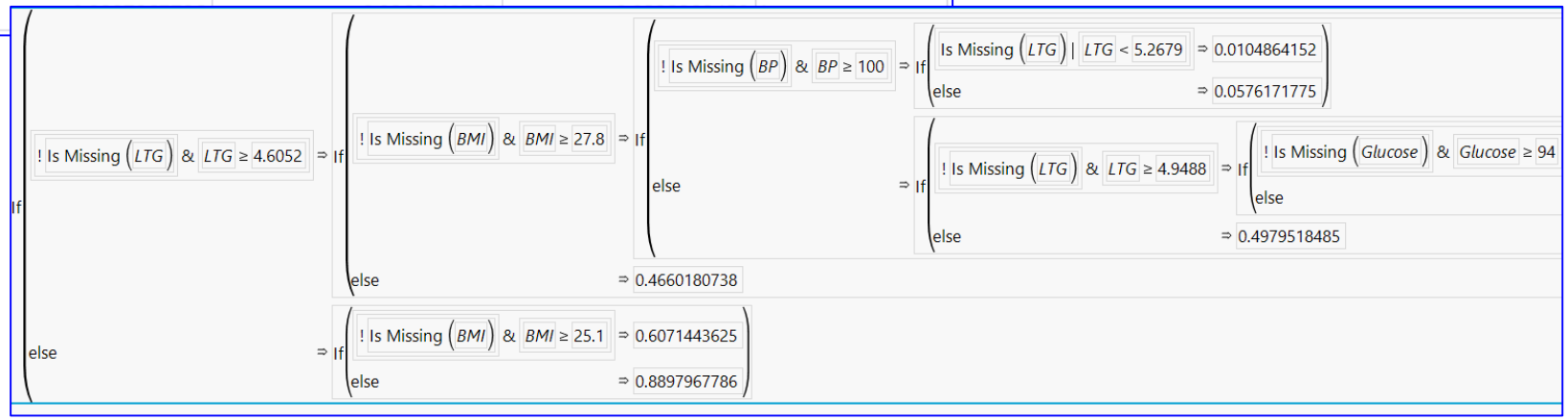
8. 분석 결과의 저장

1) Partition for Y Ordinal 의 왼쪽 빨간색 역삼각형 click 후
save column / save prediction formula 클릭
2) data table 에 반응치에 대해 추정한 확률 값이 생성됨
3) Column header 에서 마우스 오른쪽 클릭 후 formula 를 선택하면 수식을 확인할 수 있음

4) 만약, 반응치가 continuous data 라면 아래와 같이 Y 에 대한 추정 값으로 표현됨

Y Predictor
231.34090909
96.30994152
178.21212121
162.68103448
96.30994152
96.30994152
96.30994152
96.30994152
159.74468085
178.21212121
96.30994152
159.74468085
96.30994152
162.68103448
96.30994152
162.68103448
231.34090909

Prob(Y Ordinal==Low)	Prob(Y Ordinal==Medium)	Prob(Y Ordinal==High)	Most Likely Y Ordinal
0.0104864152	0.2423352132	0.7471783717	High
0.8897967786	0.0869649757	0.0232382458	Low
0.4979518485	0.3985573976	0.1034907539	Low
0.4660180738	0.257968141	0.2760137852	Low
0.8897967786	0.0869649757	0.0232382458	Low
0.8897967786	0.0869649757	0.0232382458	Low
0.8897967786	0.0869649757	0.0232382458	Low
0.6071443625	0.2396887074	0.1531669301	Low
0.6071443625	0.2396887074	0.1531669301	Low
0.3531724529	0.0224569776	0.6243705695	High
0.8897967786	0.0869649757	0.0232382458	Low
0.6071443625	0.2396887074	0.1531669301	Low
0.8897967786	0.0869649757	0.0232382458	Low
0.4660180738	0.257968141	0.2760137852	Low
0.8897967786	0.0869649757	0.0232382458	Low
0.4660180738	0.257968141	0.2760137852	Low
0.0104864152	0.2423352132	0.7471783717	High
0.4660180738	0.257968141	0.2760137852	Low
0.6071443625	0.2396887074	0.1531669301	Low



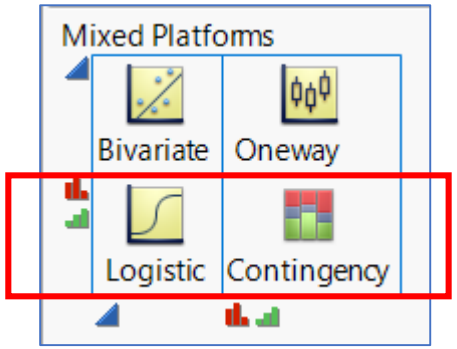
참고 : 다른 방법들(1) – fit Y by X

본 예제에서의 분석 목적은 반응치 Y(ordinal) 을 가장 잘 설명하는 인자를 찾아내는 것이다.

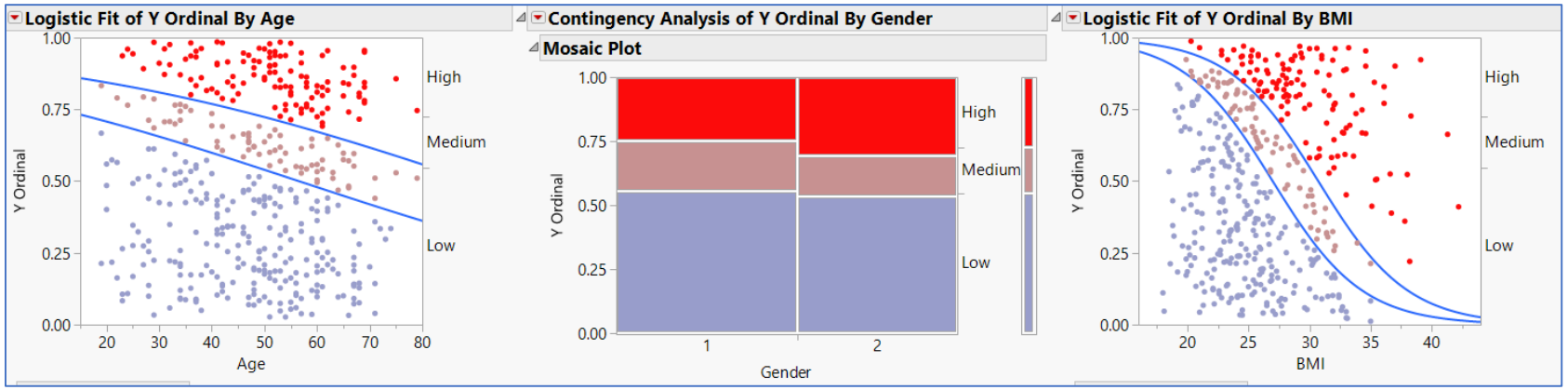
이러한 목적에 부합하는
통계적 분석 방법은 매우 다양하며,
JMP 또한 다양한 방법을 지원하고 있는 데,
가장 대표적인 것 두 가지에 대해 간략히 소개하면

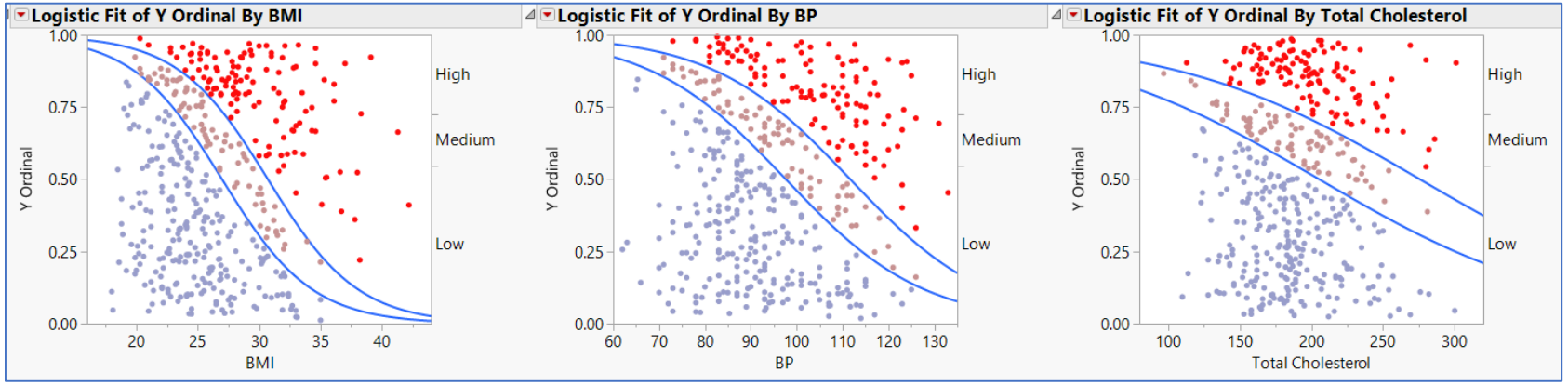
1. Analyze / Fit Y by X

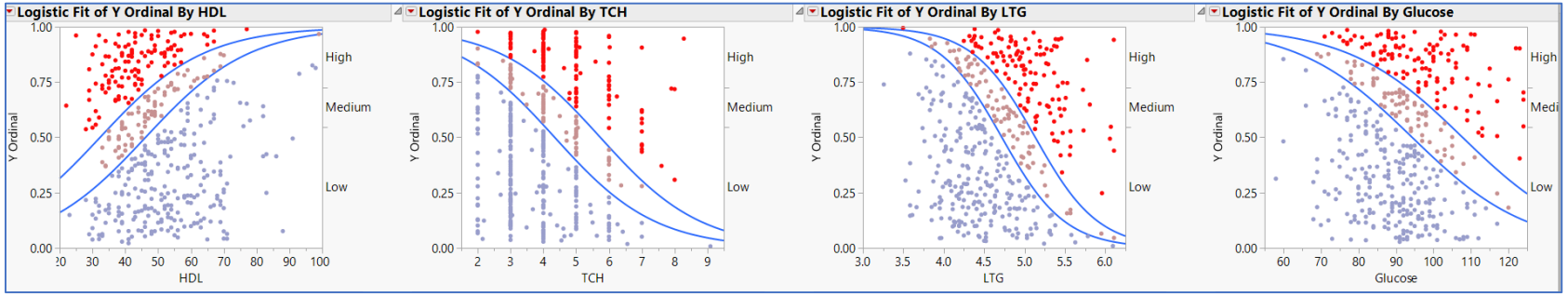
1) 인자별로 반응치 Y 에 대한 관련성을 분석한다
2) 분석 대상 Data 의 Modeling type(연속형, 서열형, 명목형) 을 JMP 가 자동으로 인식하여 이에 적합한 통계 방법을 자동 선택하여 분석해 준다
3) 인자별 분석 결과에 대해 그래프만 정리하면 오른쪽과 같다



본 예제 data 의 modeling type 을 JMP 가 인식하여 붉은 색 네모안의 분석 방법을 자동으로 선택한다.







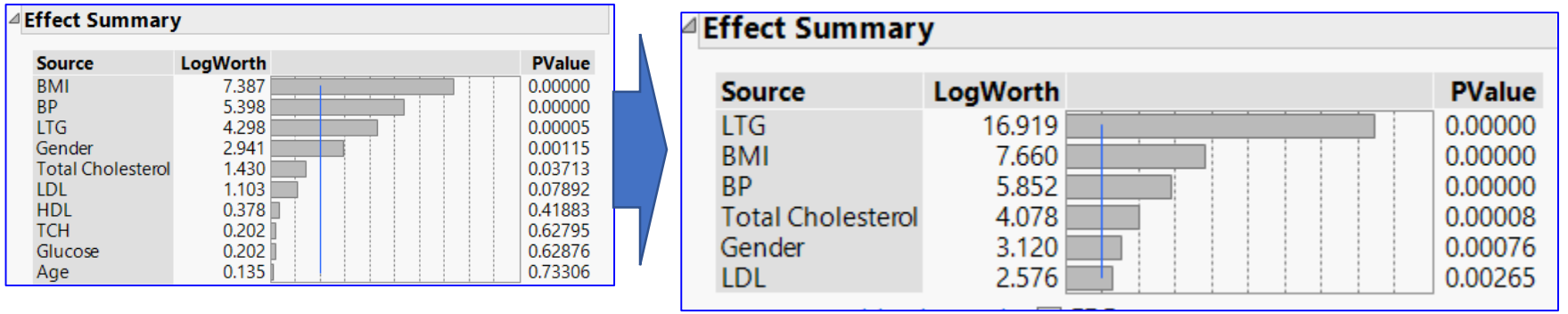
참고 : 다른 방법들(2) – Fit Model

2. Analyze / Fit Model
   1) fit Y by X 와 달리, Fit Model 에서는 선택된 모든 인자를 종합적으로 고려하여, Y 에 대한 영향도를 분석한다

   2) 선형 관계 뿐만 아니라 곡률 효과 및 인자들 간의 상호작용 등을 분석할 수 있다

   3) Prediction Profiler, Simulation 기능 등을 이용하여 보다 정교한 분석을 할 수 있다

<인자 선별>



<Prediction Profiler>