

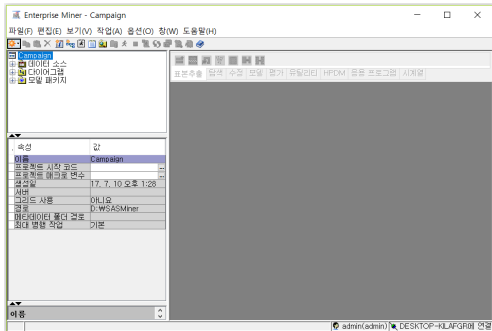
SAMPSIO.HMEQ data

Name	Model Role	Measurement	Description
BAD	Target	Binary	Bad = 1, client defaulted on the loan. Bad = 0, client paid off the loan.
CLAGE	Input	Interval	Age of the oldest credit line(in months)
CLNO	Input	Interval	Number of credit lines
DEBTINC	Input	Interval	Debt-to-income ratio
DELINQ	Input	Interval	Number of delinquent credit lines
DEROG	Input	Interval	Number of major derogatory reports
JOB	Input	Nominal	Occupational categories
LOAN	Input	Interval	Amount requested for the loan
MORTDUE	Input	Interval	Amount due on the existing mortgage
NINQ	Input	Interval	Number of recent credit inquiries
REASON	Input	Binary	DebtCon: debt consolidation, HomImp: home improvement.
VALUE	Input	Interval	Value of the current property
YOJ	Input	Interval	Years at the applicant's current job

- The SAMPSIO.HMEQ data set contains 5,960 observations.
- The goal is to find the clients who will default on a loan.

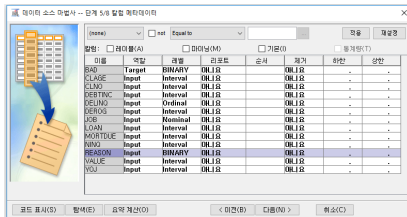
프로젝트 생성

- Run “SAS Enterprise Miner Workstation”
- Select “새로운 프로젝트”
 1. Input 프로젝트 이름, say “Loan”
 2. 디렉토리 지정, say, “d:\SASMiner”



라이브러리, 데이터 소스, 다이어그램 생성

- Select “파일” -> “새로 만들기” -> “라이브러리” (필요시)
 1. 라이브러리 이름 지정, say “MyLib”
 2. 디렉토리 지정, say, “d:\SASMiner”
- Select “파일” -> “새로 만들기” -> “다이어그램”
 1. 다이어그램 이름 지정, say, “SCORE”
- Select “파일” -> “새로 만들기” -> “데이터 소스”
 1. 찾아보기 -> Sampsiso -> HMEQ
 2. 데이터 소스 마법사 단계 5/8 칼럼 메타데이터

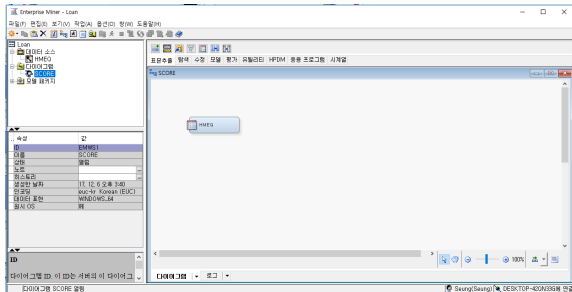


3. 통계량 계산: “요약 계산” 클릭 후, 통계량 확인

다수의 interval 변수에서 왜도가 크다.

4. 데이터 소스 마법사 단계 7/8

- 역할을 “Train”으로 변경하고 데이터 소스 마법사 종료
- 생성된 메타데이터를 다이어그램으로 끌어 온다.

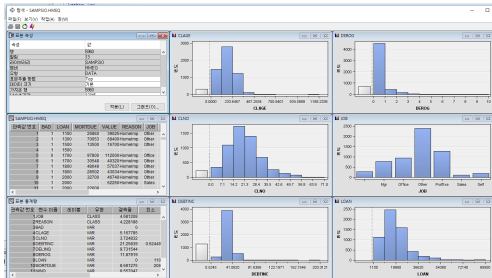


변수의 역할 및 속성

- “HMEQ” icon에 마우스를 클릭하고, 왼쪽 속성창에서 변수(컬럼 밑) 오른쪽 eclipse 클릭(변수 창 생성)
- 역할: target, input, rejected etc.
- 레벨: interval, ordinal, nominal, binary etc.
 1. 자동적으로 변수들의 레벨이 주어지지만 정확하지 않을 수 있다.
 2. 레이블(레이블 박스 체크)과 탐색을 참조하여 레벨 수정
- Example:
 1. BAD: input은 Target으로, Level은 Binary로 수정
 2. DELINQ: Level을 Ordinal로 수정
 3. REASON: Level을 Binary로 수정

변수들의 분포 탐색

- 변수 창에서 변수들의 이름에 마우스 왼쪽 버튼을 클릭하고 drag
- 윈도우 오른쪽 아래의 “탐색” 버튼 클릭



- Observations:
 - Bad = 0: 4771 cases(약 80%), Bad = 1: 1189 cases(약 20%)
 - interval 유형의 변수에서 다수의 결측치
 - interval 유형의 변수는 왜도가 크고, 특히 DELOG, NINQ, DELINQ는 한쪽으로 치우침이 매우 크다.

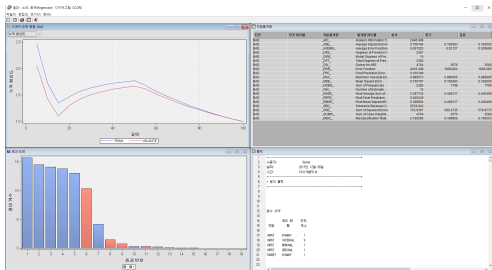
데이터셋 분할

- 분석용 데이터(training data): 모델을 추정에 사용
- 평가용 데이터(validation data): 모델의 성능 개선에 사용
- 검증용 데이터(test data): 모델의 정확도 등을 평가하는데 사용
 1. 표본추출 메뉴에서 “데이터 분할” 노드 icon을 다이어그램으로 끌어 오고 “HMEQ” icon과 연결
 2. 속성창에서 각 데이터 셋의 비율을 지정, say 40%, 30%, 30%



- Regression

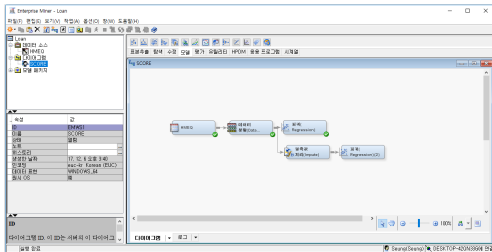
1. “모델”에서 “회귀분석” icon을 다이어그램으로 끌어 오고 “데이터 분할“ icon과 연결
2. 속성 창에서 모델 선택(say, 단계별 회귀) 설정
3. “회귀분석” 실행 후, 결과 확인

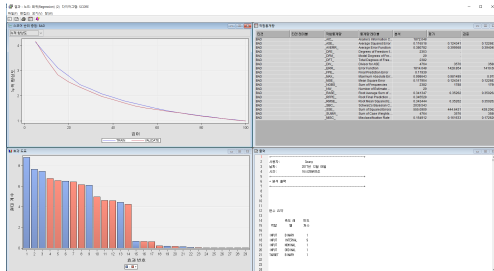


- Very bad Lift chart!!

결측치 처리

- Decision Tree와는 달리 로지스틱 회귀, Neural Network 등의 분류모형에서는 결측치에 대한 사전처리가 요구됨
 1. “수정”에서 “결측값 처리” icon을 다이어그램으로 끌어 오고 “데이터 분할” icon과 연결
 2. 속성 창에서 대체방법 설정
 3. “모델”에서 “회귀” icon을 다이어그램으로 끌어 오고 “결측값 처리” icon과 연결
 4. 속성 창에서 “모델선택”을 단계별 회귀로 설정

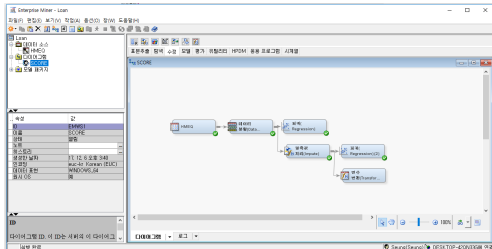




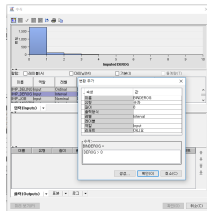
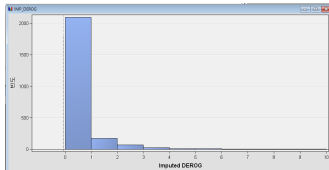
- 결측치 처리 후, 향상된 Lift chart
- 분류 능력의 향상을 위해 결측치의 처리는 매우 중요

변수변환

- 변수변환은 로지스틱 회귀, Neural Network 등의 분류모형의 분류 능력의 향상을 위한 또 하나의 중요 요소
 - “수정”에서 “변수변환”을 불러 “결측값 처리” icon과 연결
 - 대부분의 interval유형의 변수들은 대칭화를 위해 log변환이 요구
 - “변수변환” 속성 창에서 “interval 입력”을 “Log”로 변경
 - “변수변환” 실행 후, 결과 확인
 - DEROG(3.08)는 변수변환 후에도 치우침이 매우 큰 것으로 나타나, 다른 변환(이산화)가 요구된다.
 - DEBTINC는 대칭형이었으므로 변수변환 후, 왜도가 커졌다.
 - 한편, ordinal 변수인 DELINQ도 대부분의 데이터가 0에 모여 있어, 나머지 값을 하나의 범주로 묶을 필요가 있다.



- DEROG:



1. 속성창의 “수식” 오른쪽 eclipse 크릭
2. 변수 IMP_DEROG” 선택 후, 수식창 왼쪽 상단의 “생성” 버튼 크릭
3. “변수변환 추가” 창에서 이름을 “BINDEROG”으로 바꾸고, 수식란에 $DEROG > 0$ 입력
4. “확인” 버튼 크릭

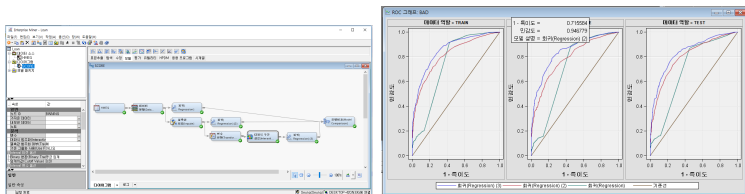
- DELINQ: Repeat this process for DELINQ

- DEROG(3.08)는 변수변환 후에도 치우침이 매우 큰 것으로 나타나, 다른 변환(이산화)가 요구된다.
- DEBTINC는 대칭형이었으므로 변수변환 후, 왜도가 커졌다.
- 한편, ordinal 변수인 DELINQ도 대부분의 데이터가 0에 모여 있어, 나머지 값을 하나의 범주로 묶을 필요가 있다.

Interactive Binning

- Interactive Binning은 수동으로 범주를 변경할 때 사용
 1. “수정”에서 “대화식 구간”을 끌어 “변수변환”과 연결하고
 2. “대화식 구간”을 실행 후, 속성창의 “대화식 범주화” 오른쪽 eclipse 클릭
 3. 변수 `IMP_DEROG` 선택 후, 수식창 왼쪽 상단의 “생성” 버튼 클릭
 4. 대화식 범주화 창에서 “계산된 역할” 란의 “rejected”를 “input”으로 수정
 5. 선택한 변수를 `LOG_IMP_NINQ`로 지정

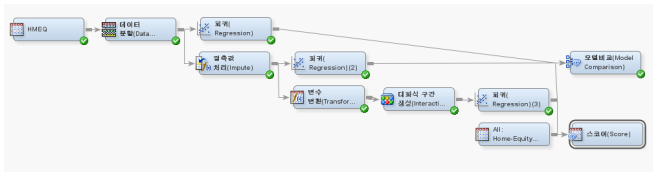
- 여러 개의 분류모형에서 최적 모형 선택
 - “평가”에서 “모델 비교” icon을 다이어그램으로 끌어 오고 이를 각 분류모형 icon과 연결
 - 모델 비교 실행



- Regression3 모형이 최적 모형 선택

Generating and Using Scoring Code

- Sampsio.DMAHMEQ는 Sampsio.HMEQ와 같은 변수들로 구성된 데이터로써 선택된 모형으로 해당 고객의 **default** 여부를 예정하고자 함.
 1. 새로운 데이터 소스 생성 (Sampsio.DMAHMEQ): 데이터 소스 마법사 단계 7/8에서 역할을 “Score”로 지정
 2. “평가” 메뉴에서 “Score” 노드를 끌어 선택된 최종모형과 연결
 3. 새로운 데이터 소스를 “Score” 노드와 연결
 4. 속성 창의 “스코어 데이터 유형”을 “데이터”로 변경 후, Score 노드 실행



5. 결과 창의 “최적화된 SAS 코드를 “파일” -> “저장”
6. SAS에서 저장된 코드를 불러 오고, 코드의 첫 부분과 마지막 부분에 각각 다음의 코드를 삽입하고 실행한다.

```
data predict ;  
  set sampio.dmahmeq ;
```

```
proc print data = predict(obs=10) ;  
  var bad p_BAD1 p_BAD0 ;  
run;
```



Obs	bad	P_BAD1	P_BAD0
1	1	0.76151	0.23849
2	1	0.58915	0.41085
3	1	0.68561	0.31439
4	1	0.99312	0.00688
5	0	0.58447	0.41553
6	1	0.15151	0.84849
7	1	0.76151	0.23849
8	1	0.13291	0.86709
9	1	0.54028	0.45972
10	1	0.73268	0.26732