영남대학교 Yeungnam University

데이터마이닝 리포트

HOUSING 데이터 분석

강의명 : 데이터마이닝 입문

교수님 : 이성원 교수님

학번 : 21410785

이름 : 황 희

◆ 목 차 ◆

1. 서 론

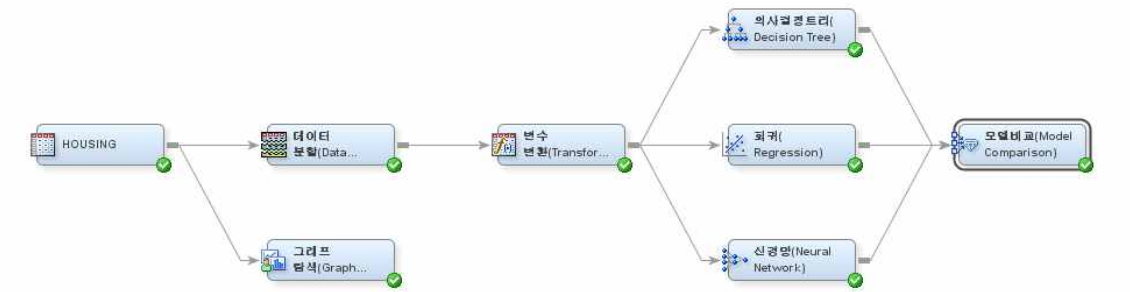
가. 연구 목적

SAS Enterprise Miner를 이용하여 HOUSING 데이터를 분석하고자 한다. 사용 할 변수는 다음과 같다.

변수	내용	측도
CRIM	범죄율	Interval
ZN	주택용 부지의 비율	Interval
INDUS	중대형 버스의 비율	Interval
CHAS	강에 인접해 있는지의 여부	Binary
NOX	산화질소의 농두	Interval
RM	방의 평균개수	Interval
AGE	1940년 이전의 주택비율	Interval
DIS	근무중심까지의 가중거리	Interval
RAD	주요 도로까지의 접근성	Interval
TAX	세율	Interval
PTRATION	초중등학교 교사의 비율	Interval
B	흑인 비율	Interval
LSTAT	중하류층의 비율	Interval
MEDV	평균 주택가격 ($1000)	Interval

이 중, 목표변수는 MEDV이고 RM의 변수변환 전, 후를 비교하여 MEDV에 RM이 어떤 영향을 끼치는지 확인해 볼것이다. 다시말해 방의 평균개수가 주택의 평균가격에 얼마나 영향을 미치는지 확인해 본다.

나. 다이어그램



분석은

-> 의사결정트리

HOUSING데이터 -> 데이터 분할 -> 변수변환 -> 회귀 -> 모델비교

-> 그래프탐색 -> 신경망

다음과 같은 순으로 진행된다.

2. 본 론

가. 변수변환 전 결과 분석

◆ 기본적인 변수 편집

이름	역할	레벨	리포트	순서	제거	하한	상한
AGE	Input	Interval	아니요		아니요	.	.
B	Input	Interval	아니요		아니요	.	.
CHAS	Input	BINARY	아니요		아니요	.	.
CRIM	Input	Interval	아니요		아니요	.	.
DIS	Input	Interval	아니요		아니요	.	.
INDUS	Input	Interval	아니요		아니요	.	.
LSTAT	Input	Interval	아니요		아니요	.	.
MEDV	Target	Interval	아니요		아니요	.	.
NOX	Input	Interval	아니요		아니요	.	.
PTRATIO	Input	Interval	아니요		아니요	.	.
RAD	Input	Interval	아니요		아니요	.	.
RM	Input	Interval	아니요		아니요	.	.
TAX	Input	Interval	아니요		아니요	.	.
ZN	Input	Interval	아니요		아니요	.	.

- 타겟 데이터는 MEDV(평균 주택가격), CHAS(강 주변위치 여부)는 BINARY 변수로 지정.

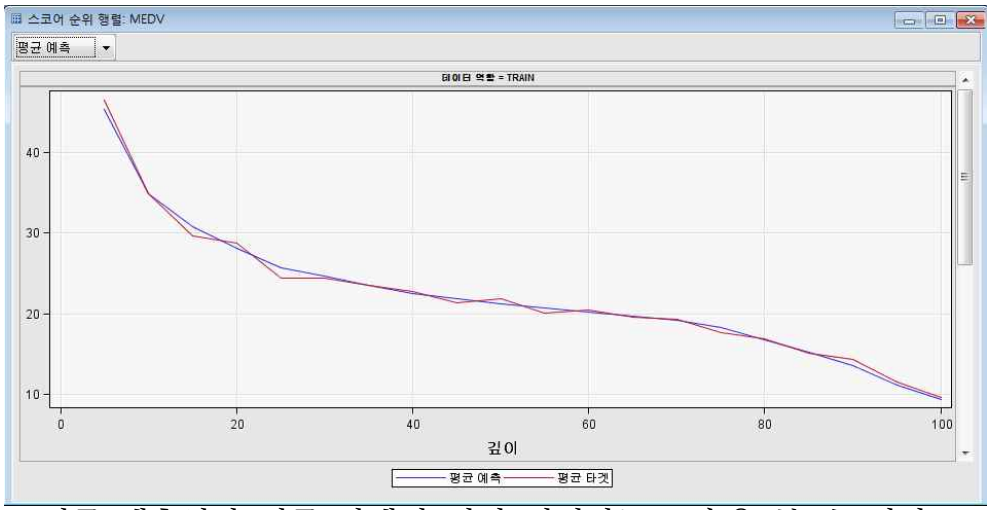
◆ 데이터 분할노드의 옵션

일반	
노드 ID	Part
가져온 데이터	
내보낸 데이터	
노트	
분석	
변수	
출력 유형	데이터
분할 방법(Partitioning M	기본
난수초기값	12345
⊟데이터셋 할당	
분석용(Training)	70.0
평가용(Validation)	30.0
검증용(Test)	0.0
리포트	
Interval 타겟	예
Class 타겟	예

- 분석용 데이터 70%, 평가용 데이터 30%로 설정

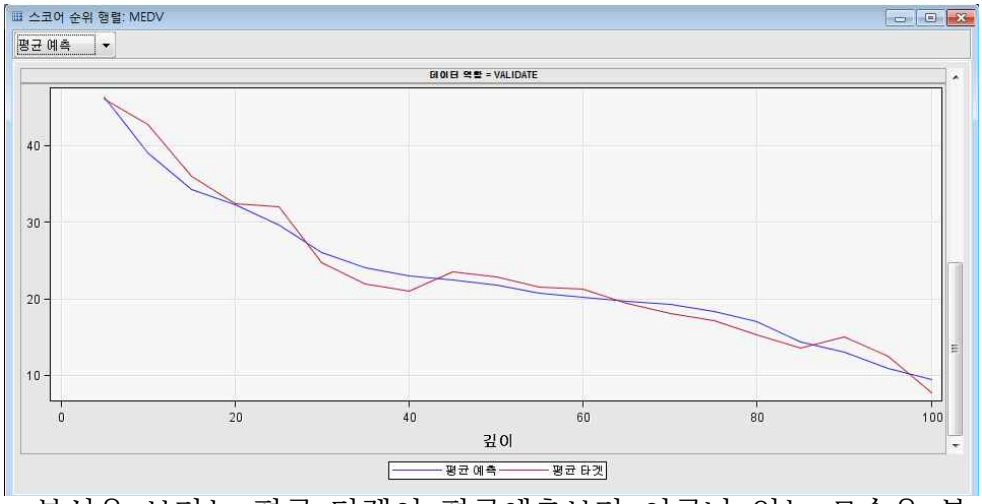
1. 신경망 분석

스코어 순위 행렬 (분석용)



- 평균 예측선과 평균 타겟이 거의 일치하는 모습을 볼 수 있다.

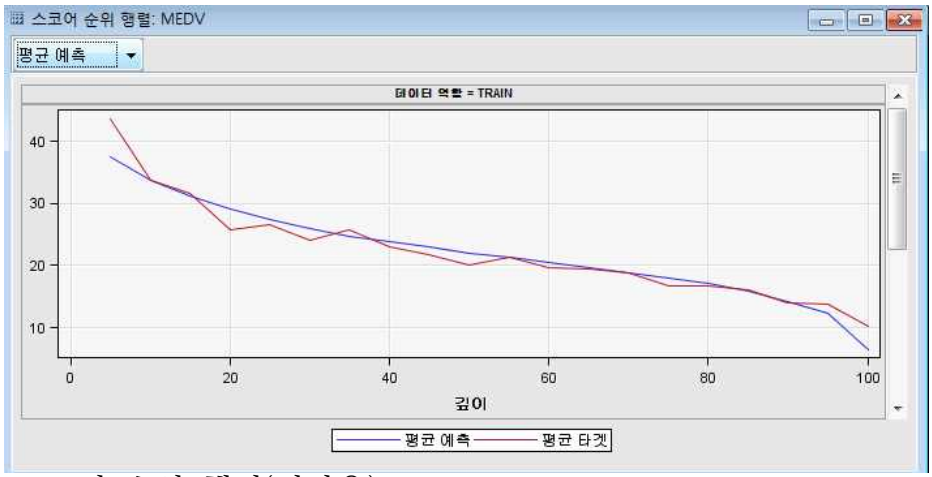
스코어 순위 행렬(평가용)



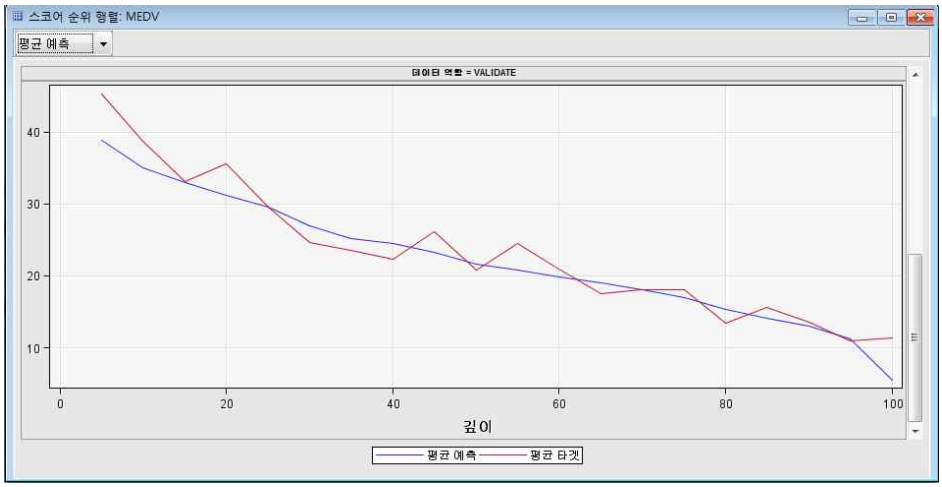
- 분석용 보다는 평균 타겟이 평균예측보다 어긋나 있는 모습을 볼 수 있다.

2. 회귀 분석

스코어 순위 행렬(분석용)



스코어 순위 행렬(평가용)



분산분석표

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
AGE	1	18.9142	0.97	0.3248
B	1	162.3673	8.35	0.0041
CHAS	1	139.5762	7.18	0.0077
CRIM	1	238.3808	12.26	0.0005
DIS	1	892.7026	45.90	<.0001
INDUS	1	8.7782	0.45	0.5022
LSTAT	1	839.2257	43.15	<.0001
NOX	1	264.4321	13.60	0.0003
PTRATIO	1	682.4788	35.09	<.0001
RAD	1	311.8008	16.03	<.0001
RM	1	1227.2535	63.10	<.0001
TAX	1	191.0545	9.82	0.0019
ZN	1	210.9258	10.84	0.0011

- 유의수준 α를 0.05로 생각할 때, AGE(1940년 이전의 주택비율)와 INDUS(중대형 버스의 비율)은 P-값이 모두 0.05를 초과하므로 MEDV와 상관이 없다고 생각한다.

적합통계량

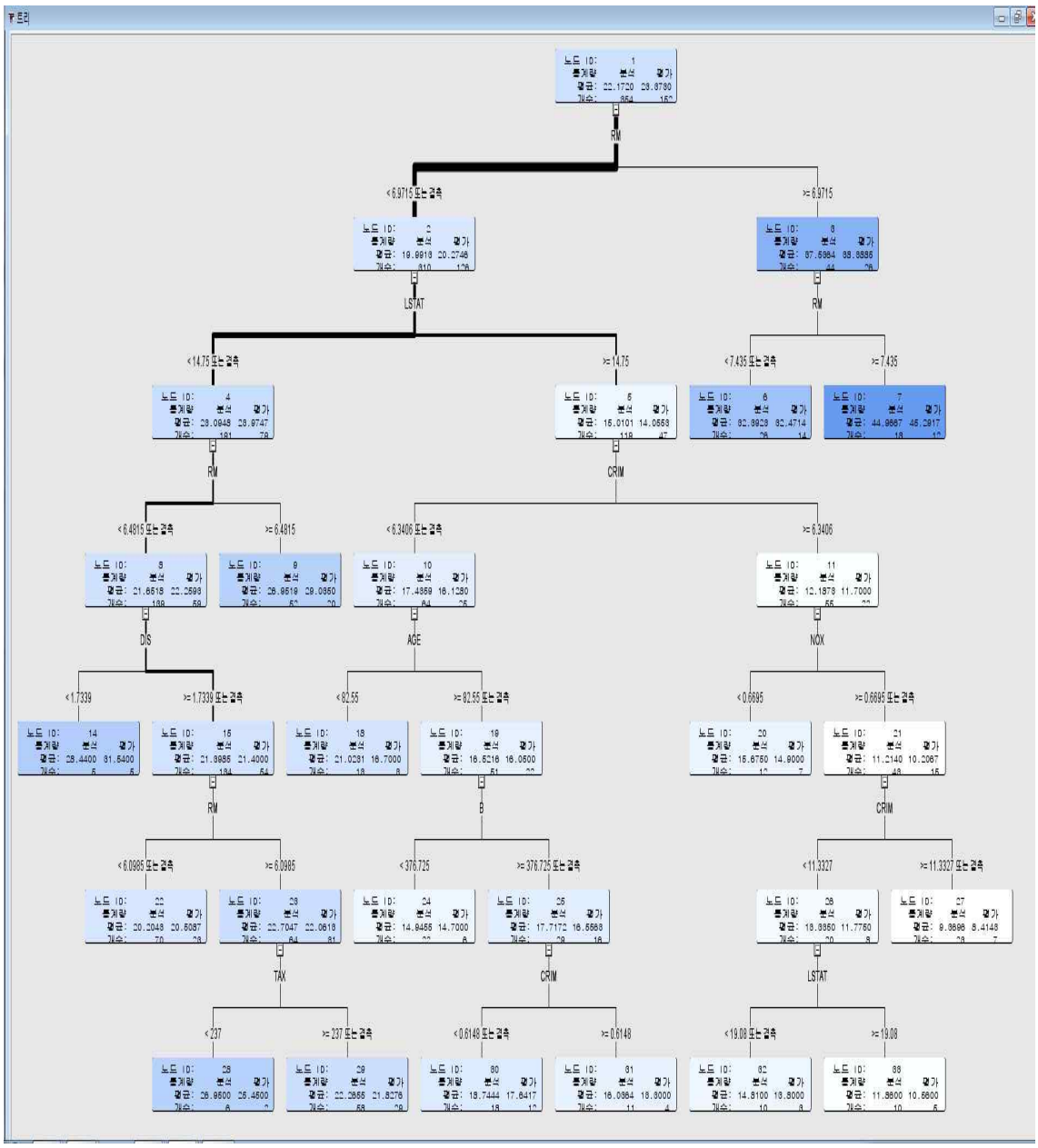
적합통계량

타겟=MEDV 타겟 레이블=' '

적합통계량	통계량 레이블	분석	평가
AIC	Akaike's Information Criterion	1064.34	.
ASE	Average Squared Error	18.68	31.09
AVERR	Average Error Function	18.68	31.09
DFE	Degrees of Freedom for Error	340.00	.
DFM	Model Degrees of Freedom	14.00	.
DFT	Total Degrees of Freedom	354.00	.
DIV	Divisor for ASE	354.00	152.00
ERR	Error Function	6613.07	4725.58
FPE	Final Prediction Error	20.22	.
MAX	Maximum Absolute Error	25.77	28.81
MSE	Mean Square Error	19.45	31.09
NOBS	Sum of Frequencies	354.00	152.00
_NW_	Number of Estimate Weights	14.00	.
RASE	Root Average Sum of Squares	4.32	5.58
RFPE	Root Final Prediction Error	4.50	.
RMSE	Root Mean Squared Error	4.41	5.58
SBC	Schwarz's Bayesian Criterion	1118.51	.
SSE	Sum of Squared Errors	6613.07	4725.58
SUMW	Sum of Case Weights Times Freq	354.00	152.00

- MSE = 19.45 , 31.09

3. 의사결정트리



변수 중요도

변수 중요도

변수 이름	레이블	분리 규칙 개수	중요도	평가 중요도	분석 중요도에 따른 평가 비율
RM		4	1.0000	1.0000	1.0000
LSTAT		2	0.5713	0.5612	0.9822
CRIM		3	0.2641	0.1794	0.6792
DIS		1	0.1270	0.2222	1.7496
AGE		1	0.1190	0.0000	0.0000
NOX		1	0.1123	0.1005	0.8955
TAX		1	0.0897	0.0551	0.6140
B		1	0.0805	0.0000	0.0000

의사결정트리에서 변수 RM(평균 방의 갯수)가 가지를 나누는데 가장 큰 역할을 함을 알 수 있다.

의사결정트리 해석

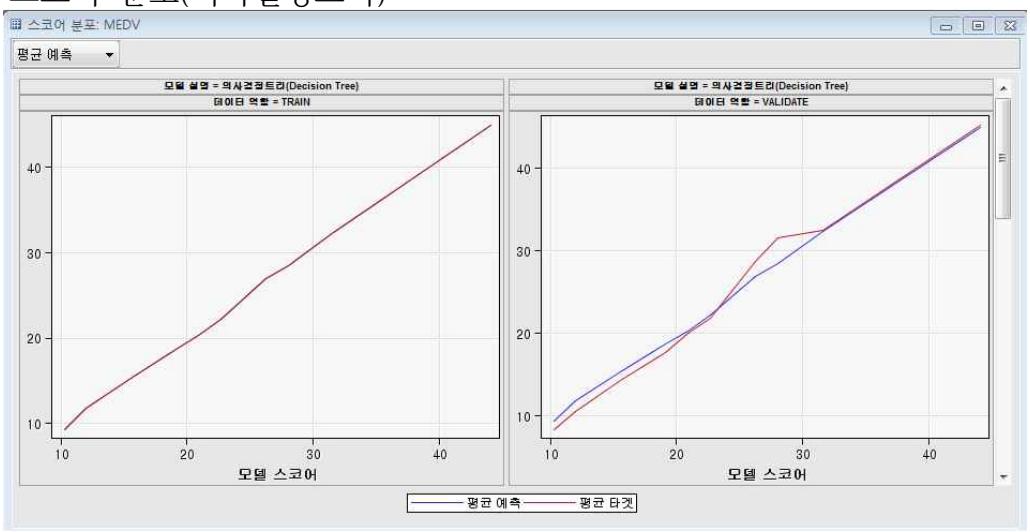
① RM(평균 방의 개수)이 >=7.435 일때, MEDV(주택 평균가격)가 44.99($1000)로 가장 평균이 높은 것을 알 수 있다.

② 또한 RM(평균 방의 개수)가 <6.95 또는 결측이고, LSTAT(중하류층의 비율)이 >=14.75 이며, CRIM(범죄율)이 >=6.34 이고, NOX(산화질소의 농도)가 >=0.6695 이상인 주택 중 CRIM(범죄율)이 >=11.33 일때 MEDV(주택평균가격)가 9.35($1000)으로 가장 낮은 것을 알 수 있다.

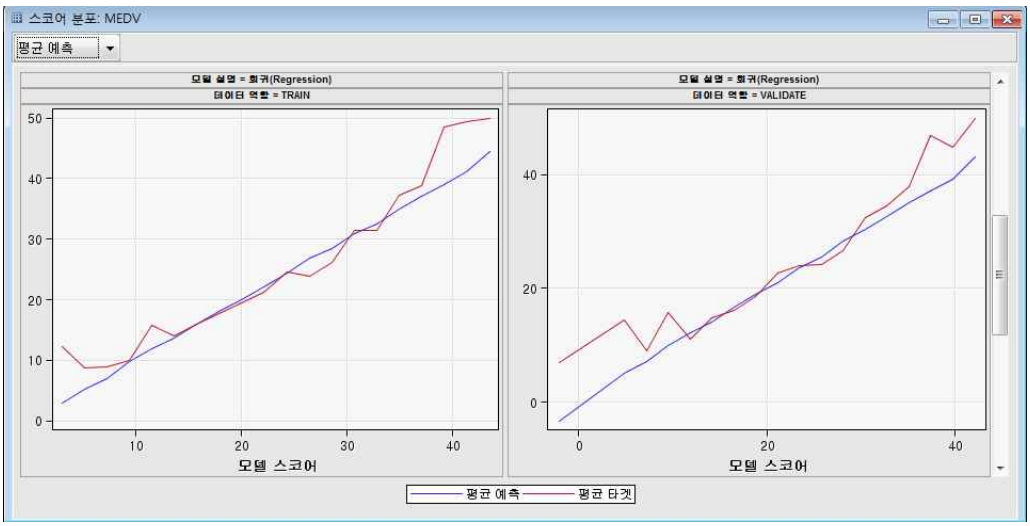
- 따라서 MDEV(주택평균가격)가 RM(평균 방의 개수)에 따라 크게 변하며, LSTAT(중하류층의 비율), CRIM(범죄율)순으로 크게 좌우됨을 알 수 있다.

4. 모델비교

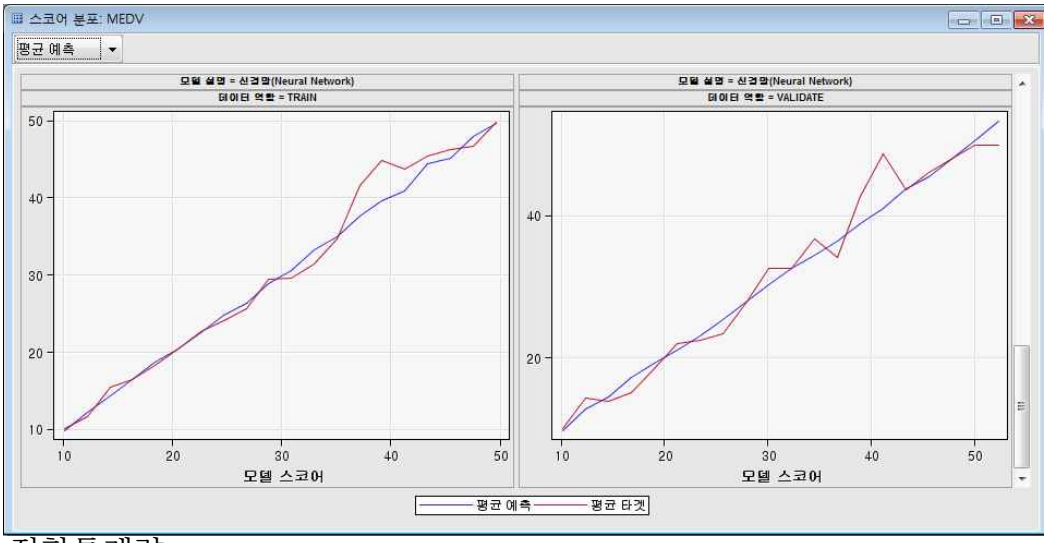
스코어 분포(의사결정트리)



스코어 분포(회귀)



스코어 분포(신경망)



적합통계량

적합통계량 테이블
타겟: MEDV

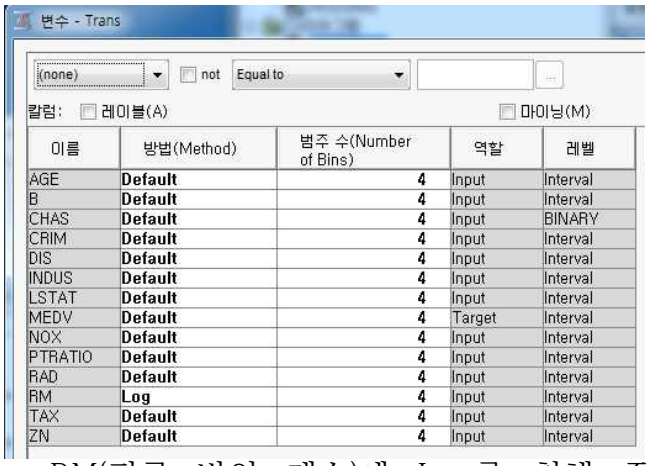
데이터 역할=Train

통계량	Neural	Tree	Reg
Train: Akaike's Information Criterion	794.77	.	1064.34
Train: Average Squared Error	7.28	12.55	18.68
Train: Average Error Function	7.28	.	18.68
선택 기준: Valid: Average Squared Error	14.14	27.25	31.09
Train: Degrees of Freedom for Error	308.00	.	340.00
Train: Model Degrees of Freedom	46.00	.	14.00
Train: Total Degrees of Freedom	354.00	354.00	354.00
Train: Divisor for ASE	354.00	354.00	354.00
Train: Error Function	2577.39	.	6613.07
Train: Final Prediction Error	9.46	.	20.22
Train: Maximum Absolute Error	13.21	23.07	25.77
Train: Misclassification Rate	.	.	.
Train: Mean Square Error	8.37	.	19.45
Train: Sum of Frequencies	354.00	354.00	354.00
Train: Number of Estimate Weights	46.00	.	14.00
Train: Root Average Sum of Squares	2.70	3.54	4.32
Train: Root Final Prediction Error	3.07	.	4.50
Train: Root Mean Squared Error	2.89	.	4.41
Train: Schwarz's Bayesian Criterion	972.76	.	1118.51
Train: Sum of Squared Errors	2577.39	4442.09	6613.07
Train: Sum of Case Weights Times Freq	354.00	.	354.00
Train: Number of Wrong Classifications	.	.	.

- 신경망, 의사결정트리, 회귀의 스코어 분포 ASE(선택 기준)는 각각 14.14 27.25 31.09 로 신경망의 ASE가 가장 적으므로 MEDV의 예측은 신경망의 스코어 분포를 가장 잘 따른다고 할 수 있다.

나. 변수변환 후 결과 분석

◆ 변수변환



이름	방법(Method)	범주 수(Number of Bins)	역할	레벨
AGE	**Default**	**4**	Input	Interval
B	**Default**	**4**	Input	Interval
CHAS	**Default**	**4**	Input	BINARY
CRIM	**Default**	**4**	Input	Interval
DIS	**Default**	**4**	Input	Interval
INDUS	**Default**	**4**	Input	Interval
LSTAT	**Default**	**4**	Input	Interval
MEDV	**Default**	**4**	Target	Interval
NOX	**Default**	**4**	Input	Interval
PTRATIO	**Default**	**4**	Input	Interval
RAD	**Default**	**4**	Input	Interval
RM	**Log**	**4**	Input	Interval
TAX	**Default**	**4**	Input	Interval
ZN	**Default**	**4**	Input	Interval

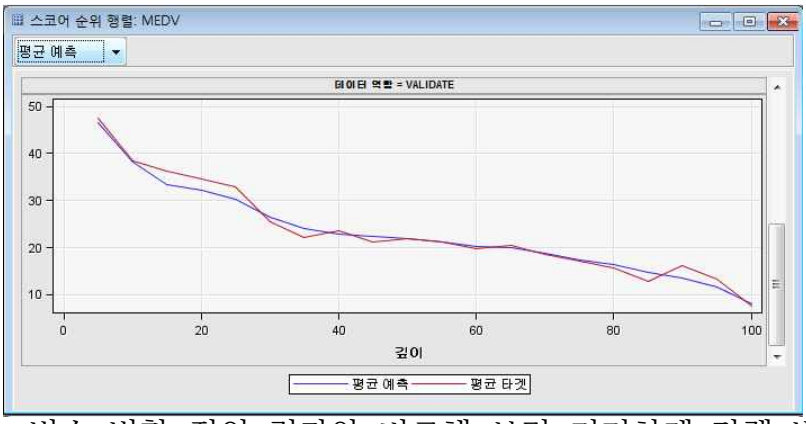
- RM(평균 방의 개수)에 Log를 취해 줌으로써 주택의 방의 개수변화가 MEDV(평균 주택가격)에 얼마나 영향을 주는지 알아보도록 하겠다.

1. 신경망 분석

스코어 순위 행렬(분석용)

- 변수 변환 전의 결과와 비교해 보았을 때 큰 차이를 보이진 않는다.

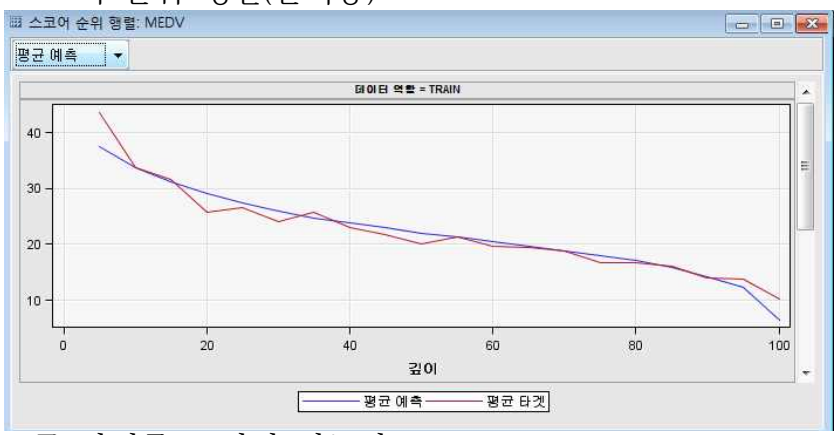
스코어 순위 행렬(평가용)



- 변수 변환 전의 결과와 비교해 보면 미미하게 타겟 변수를 잘 예측 한 것을 알수 있다.

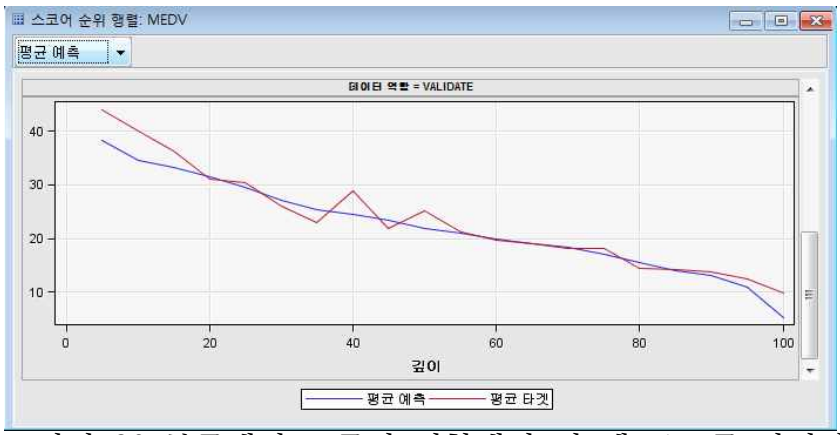
2. 회귀분석

스코어 순위 행렬(분석용)



- 큰 차이를 보이진 않는다.

스코어 순위 행렬(평가용)



- 깊이 20 부근에서 조금더 적합해진 것 빼고는 큰 차이가 없다.

분산분석표

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
AGE	1	6.6268	0.32	0.5701
B	1	147.5756	7.20	0.0077
CHAS	1	144.0595	7.03	0.0084
CRIM	1	232.4654	11.34	0.0008
DIS	1	999.6933	48.76	<.0001
INDUS	1	13.0110	0.63	0.4262
LOG_RM	1	869.5258	42.41	<.0001
LSTAT	1	1039.2179	50.69	<.0001
NOX	1	299.1146	14.59	0.0002
PTRATIO	1	775.5841	37.83	<.0001
RAD	1	364.5918	17.78	<.0001
TAX	1	208.3246	10.16	0.0016
ZN	1	261.1567	12.74	0.0004

- RM의 변화에 따라 다른 변수의 Sum of Squares도 변한다. 따라서 RM은 다른 설명변수와도 연관성이 있다고 볼 수 있다.

적합통계량

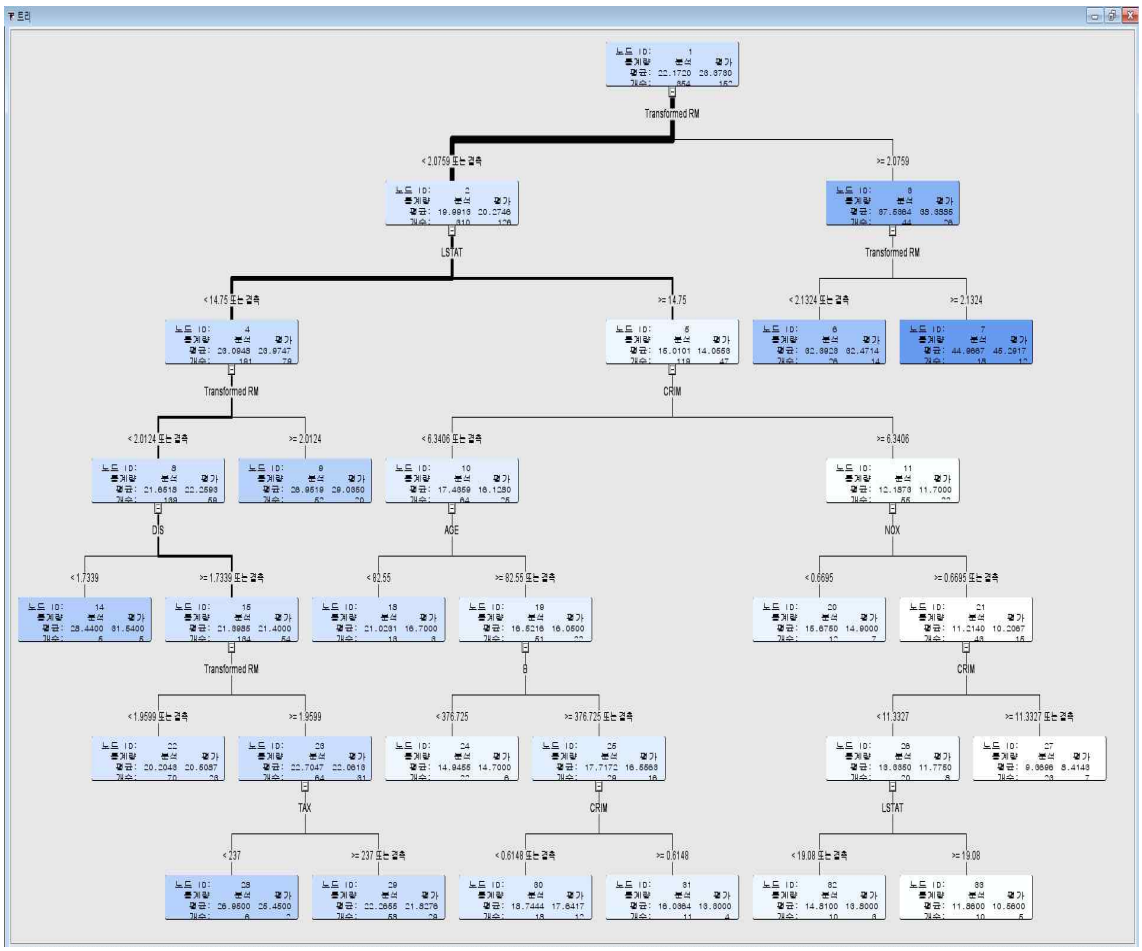
적합통계량

타겟=MEDV 타겟 레이블=' '

적합통계량	통계량 레이블	분석	평가
AIC	Akaike's Information Criterion	1082.99	.
ASE	Average Squared Error	19.69	32.25
AVERR	Average Error Function	19.69	32.25
DFE	Degrees of Freedom for Error	340.00	.
DFM	Model Degrees of Freedom	14.00	.
DFT	Total Degrees of Freedom	354.00	.
DIV	Divisor for ASE	354.00	152.00
ERR	Error Function	6970.79	4902.30
FPE	Final Prediction Error	21.31	.
MAX	Maximum Absolute Error	25.00	27.57
MSE	Mean Square Error	20.50	32.25
NOBS	Sum of Frequencies	354.00	152.00
_NW_	Number of Estimate Weights	14.00	.
RASE	Root Average Sum of Squares	4.44	5.68
RFPE	Root Final Prediction Error	4.62	.
RMSE	Root Mean Squared Error	4.53	5.68
SBC	Schwarz's Bayesian Criterion	1137.16	.
SSE	Sum of Squared Errors	6970.79	4902.30
SUMW	Sum of Case Weights Times Freq	354.00	152.00

- MSE = 20.50 , 32.25

3. 의사결정트리



변수 중요도

변수 중요도

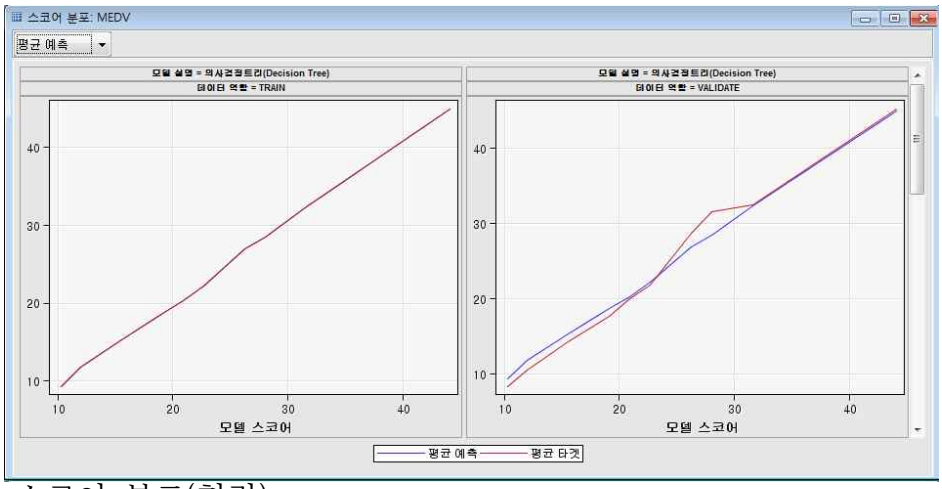
변수 이름	레이블	분리 규칙 개수	중요도	평가 중요도	분석 중요도에 따른 평가 비율
LOG_RM	Transformed RM	4	1.0000	1.0000	1.0000
LSTAT		2	0.5713	0.5612	0.9822
CRIM		3	0.2641	0.1794	0.6792
DIS		1	0.1270	0.2222	1.7496
AGE		1	0.1190	0.0000	0.0000
NOX		1	0.1123	0.1005	0.8955
TAX		1	0.0897	0.0551	0.6140
B		1	0.0805	0.0000	0.0000

- 변수 중요도는 RM의 변화에 관계가 없다.

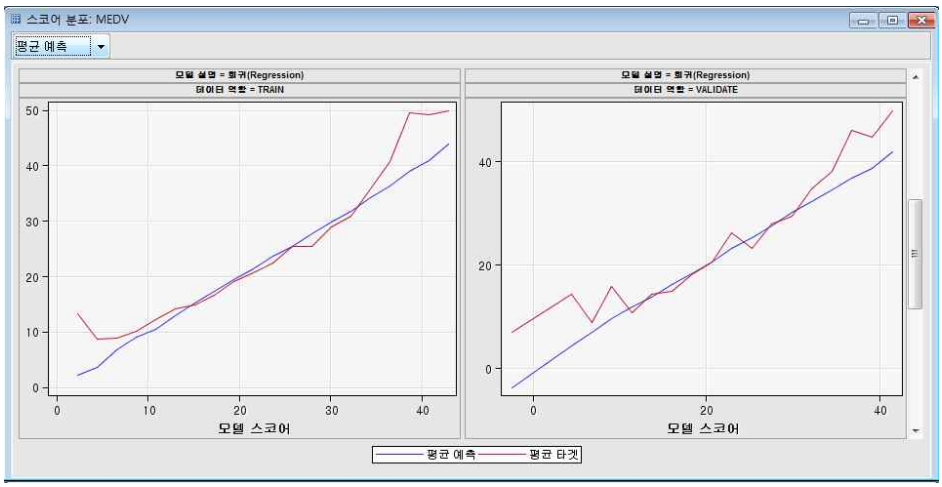
또한 의사결정트리에서 RM은 별 영향을 미치지 않는 것으로 보인다.

4. 모델비교

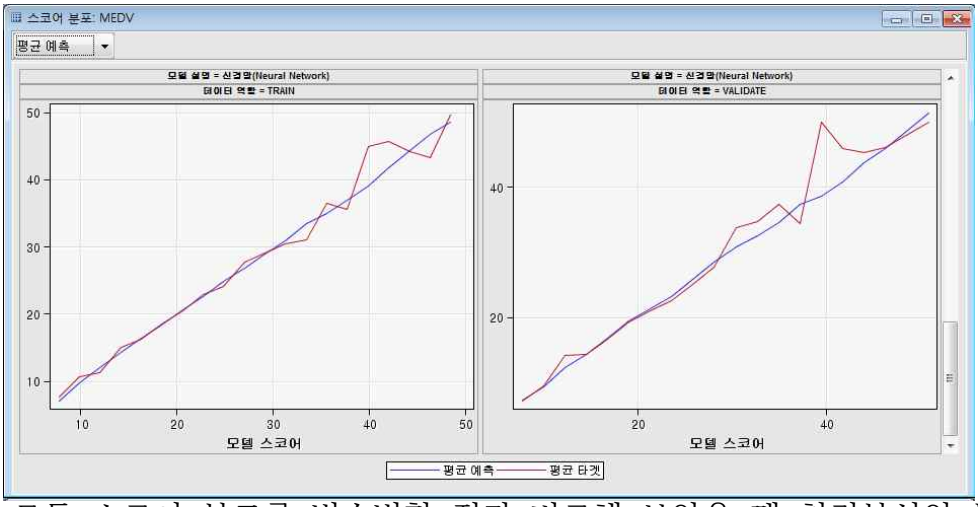
스코어 분포(의사결정트리)



스코어 분포(회귀)



스코어 분포(신경망)



- 모든 스코어 분포를 변수변환 전과 비교해 보았을 때 회귀분석의 분석용 데이터의 스코어 분포에서 타겟을 더 잘 적합한 예측을 했음이 보인다.

적합통계량

적합통계량 테이블
타겟: MEDV

데이터 역할=Train

통계량	Neural	Tree	Reg
Train: Akaike's Information Criterion	814.39	.	1082.99
Train: Average Squared Error	7.70	12.55	19.69
Train: Average Error Function	7.70	.	19.69
선택 기준: Valid: Average Squared Error	16.34	27.25	32.25
Train: Degrees of Freedom for Error	308.00	.	340.00
Train: Model Degrees of Freedom	46.00	.	14.00
Train: Total Degrees of Freedom	354.00	354.00	354.00
Train: Divisor for ASE	354.00	354.00	354.00
Train: Error Function	2724.24	.	6970.79
Train: Final Prediction Error	9.99	.	21.31
Train: Maximum Absolute Error	14.15	23.07	25.00
Train: Misclassification Rate	.	.	.
Train: Mean Squared Error	8.84	.	20.50
Train: Sum of Frequencies	354.00	354.00	354.00
Train: Number of Estimated Weights	46.00	.	14.00
Train: Root Average Squared Error	2.77	3.54	4.44
Train: Root Final Prediction Error	3.16	.	4.62
Train: Root Mean Squared Error	2.97	.	4.53
Train: Schwarz's Bayesian Criterion	992.38	.	1137.16
Train: Sum of Squared Errors	2724.24	4442.09	6970.79
Train: Sum of Case Weights Times Freq	354.00	.	354.00
Train: Number of Wrong Classifications	.	.	.

3. 결론

가. HOUSING 데이터에 대한 분석 결과

타겟변수 = MDEV(주택 평균 가격)
의사결정트리에서 가지를 나누는데 중요도 1순위 = RM(평균 방 개수)
다음으로 큰 영향을 미치는 주요 변수 = LSTAT(중하류층 의 비율), CRIME(범죄율)
MDEV와 관계없는 설명변수 = AGE(1940년대 이전의 주택비율), INDUS(중대형 버스의 비율)

따라서, 주택 평균가격은 방의 개수가 많을 수록, 중하류층의 비율이 적을수록, 범죄율이 적을수록 가장 높아진다고 볼 수 있다. 또한 1940년대 이전의 비율과 중대형 버스의 비율은 주택 평균가격의 변동과 관계가 없다.

또한, RM의 변수변환을 통해 방의 개수 변화는 MEDV를 예측하는데 큰 변화를 주지 못함을 알 수 있다.

감사합니다.