DS 11주차 수업 팀과제

Linear Regression

팀 AOA

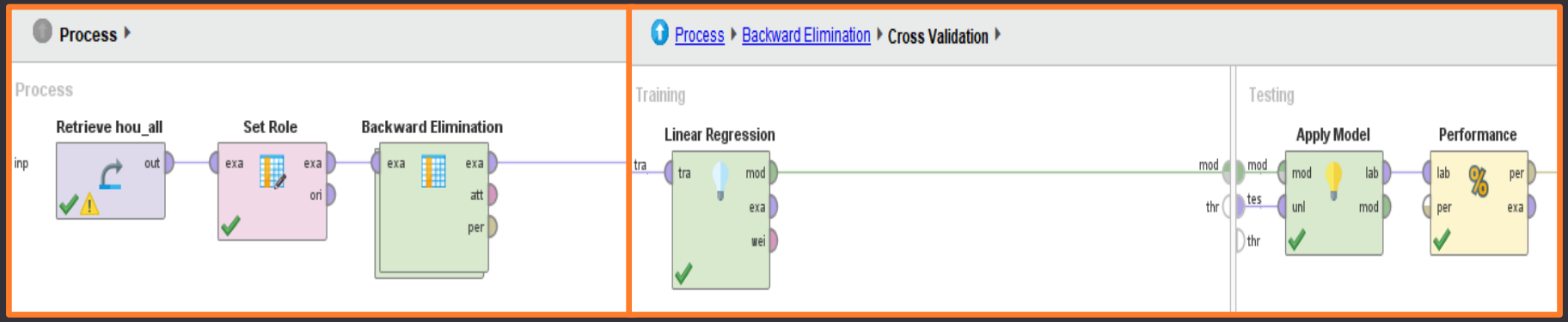
경영학과 2014109050 박한솔
경영학과 2015126062 신지수

01
Boston Housing Dataset

회귀 분석을 위해 **목표 변수가 연속형 수치형**인 데이터셋 Boston Housing을 사용한다.

모델 설계에서 방의 개수, 위치 등의 속성을 사용해 주택 가격의 중앙값(MEDV)을 예측하는 것이 목적이다.

02
후방 제거법



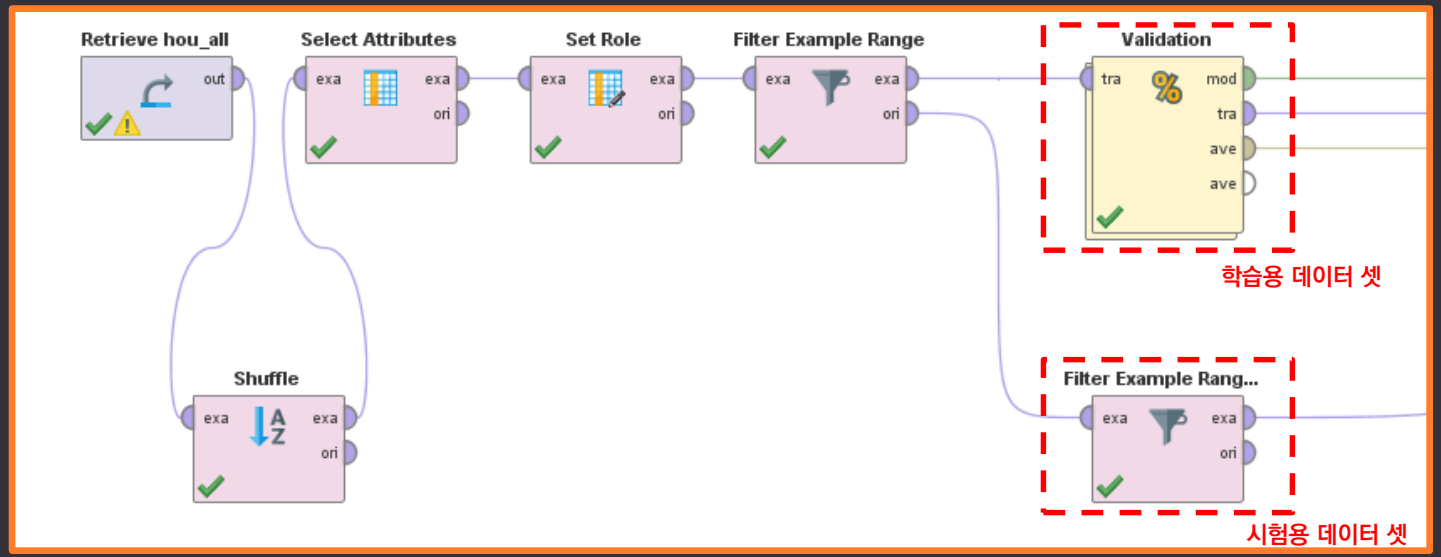
〈Backward Elimination Operator〉

최적의 변수 집합을 선택하기 위해 후방제거법을 사용한다.

특징 선택 시 Backward Elimination-Cross Validation 중첩된 로직으로 설계한다.

모델의 결과로 영향을 적게 미치는 속성 Age가 삭제되었다.

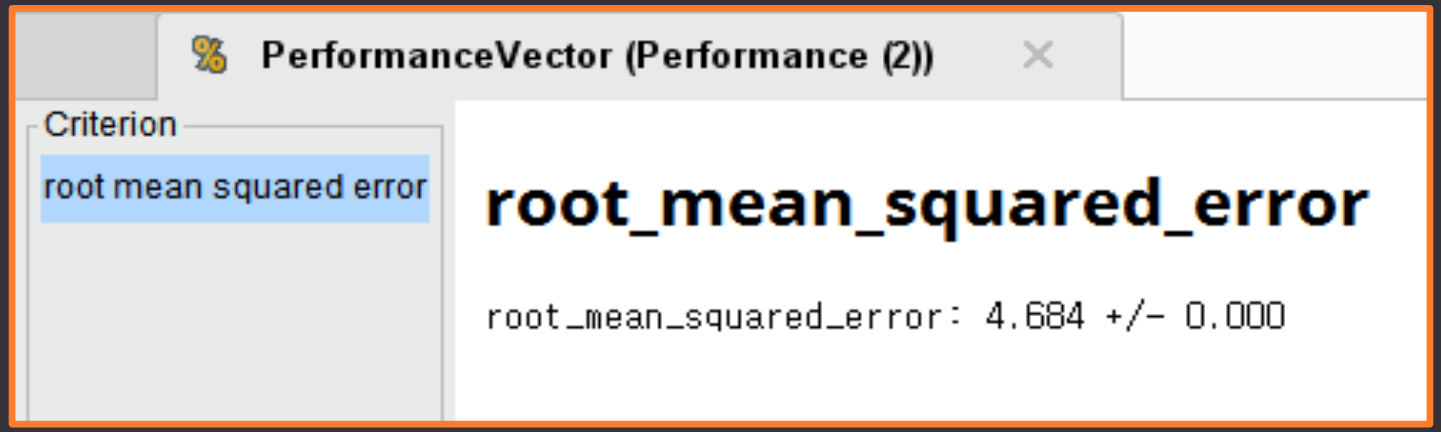
03
예측 모델



〈예측 모델의 관점에서 작성한 process〉

Shuffle과 Filter Example Range 오퍼레이터를 사용해서 데이터 셋을 나눈 process를 작성한다.

04
MSE



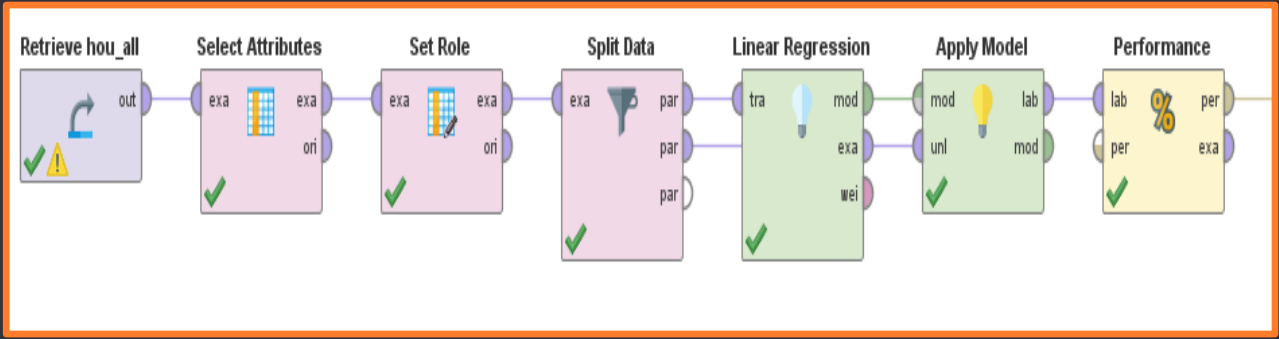
〈예측 모델의 시험용 집합에 대한 성능 측정 결과〉

모델의 시험용 데이터 셋에 performance 연산자를 추가해 성능을 측정해보았다.

MSE는 회귀예측 모델의 성능평가 척도 중 하나이며, 오차가 너무 클 경우를 대비해 root를 씌워준다.

그러므로 시험용 집합에 대한 MSE 값은 약 21.94이다.

05
설명 모델



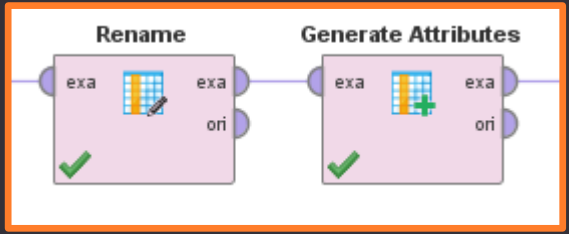
〈설명 모델의 관점에서 작성한 process〉

설명 모델의 관점에서 모델을 설명하기 위해 새로 process를 생성한다.

앞의 모델보다 더 간단한 형태의 process이다.

설명형 모델은 예측변수의 영향을 확인하고, 데이터의 적합도를 관찰해야한다.

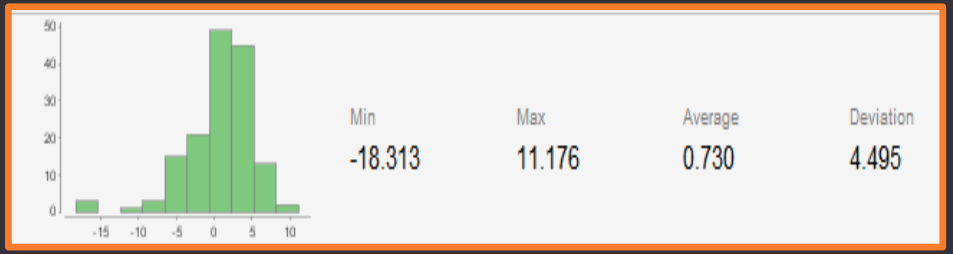
06
잔차 분석



Apply model 오퍼레이터 뒤에 rename과 generate attributes 추가해준다.
Rename은 예측한 MEDV이름을 PredMEDV로 바꿔줌

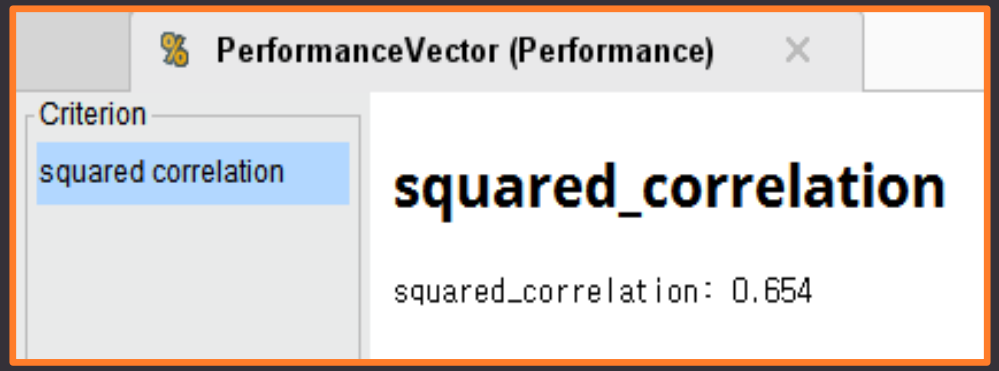
attribute name	function expressions
residual	PredMEDV-MEDV

새로 만든 속성의 이름을 residual(잔차)라 명명하고
함수 표현에 예측값과 실제값의 차이를 입력해준다.



〈잔차 분포도〉
모델의 statistic탭에서 residual의 히스토그램을 볼 수 있다.
평균= 0.730, 표준편차= 4.495로 오른쪽으로 치우친,
즉 정규분포를 따르지 않기 때문에 모델 개선을
계속해야한다는 것을 확인할 수 있다.

07
설명모델로서의 유의미



〈실제값과 예측값에 대한 상관계수 R의 제곱〉

설명모델로서의 유의미 여부는 상관계수의 제곱을 통해 알아 볼 수 있다.

Squared_correlation 즉 상관계수 R^2이 이론적으로 0.6, 실무적으로 0.4 이상이면 유의미하다 할 수 있다.

위와 같이 모델의 R^2 값이 0.654이며 0.6보다 크기 때문에 이 모델은 유의미하다는 것을 알 수 있다.

08
가격에 영향을 가장 많이 주는 변수

가격에 영향을 가장 많이 미치는 변수는, 후방제거법에서 가장 나중에 제거되는 변수를 뜻한다.

즉, 후방제거법을 연속으로 돌렸을 때 제거되는 변수들을 역순으로 나열하면, 영향을 많이 미치는 변수들의 순서가 된다.

Attribute	Code ↓	Std. Coefficient
NOX	****	-0.194
RM	****	0.301
DIS	****	-0.328
RAD	****	0.288
TAX	****	-0.286
PTRATIO	****	-0.221
LSTAT	****	-0.427
(Intercept)	****	?
ZN	***	0.119
B	***	0.104
CRIM		-0.060
INDUS		0.065
CHAS		0.042

또는 선형 회귀 모델의 Data 탭에서 변수들의 유의성을 확인할 수 있다.

Code열의 별의 개수는 예측 변수의 유의성을 표시한다.

상관계수 또한 절대값이 1에 가까울 수록 목표변수와 관련이 있다는 뜻이다.

부호는 음/양의 상관관계를 나타낸다.

모델에서 출력된 순서는

LSTAT-DIS-RM-RAD-TAX-PTRATIO-NOX-ZN-B-INDUS-CRIM-CHAS-AGE 순이며, 랜덤 시드 값에 따라 약간의 차이가 있지만,

대부분의 경우 LSTAT, DIS, RM은 의미있는 변수이며,

AGE, INDUS, CHAS는 의미가 없는 변수란 결과가 나온다.

09
회귀 계수

Q. 각 변수의 회귀 계수는 어떠한 의미를 가지는가?

우선 각 변수의 회귀계수는 설명력과는 무관하다. 회귀 계수의 크기는 같은 값을 가진다
예측변수의 Scale에 따라서 다르게 표현할 수 있으므로 **회귀계수의 크기가 변수의 중요도를 뜻하는 것은 아니다**.

그러나 회귀 계수의 **음/양에 따라서 예측변수와 목표변수의 관계를 알 수 있다.** 회귀 계수가 양의 값을 가지면
그 예측변수가 목표변수를 촉진한다는 것을 의미하며, 반대로 음의 값을 갖는다면 예측변수가 목표변수를
억제한다는 것을 알 수 있다.