

실습 1

Linear Regression를 사용한
전복 순살 무게 예측

[Lecture] Dr. HeeSuk Kim

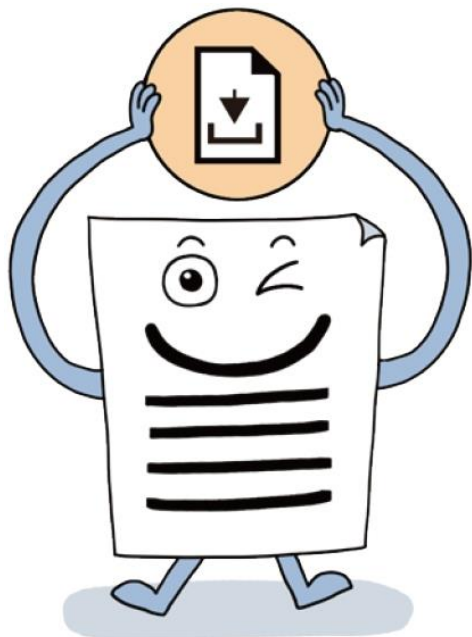
1

전복 순살의 무게를 맞춰 보!

Linear Regression를 사용하여
전복 순살의 무게를 예측해 보자.

데이터 종류:

정형 데이터



사용하는 모델:

Linear Regression



1 해결해야 할 문제는 무엇일까?

문제 상황

전통 시장이나 주변 마트 수산물 코너에 가면 전복을 흔히 찾아볼 수 있다.

전복은 크기와 가격대가 다양하다.

우리나라에서 전복을 판매할 때, 비슷한 크기끼리 묶어서 '미'라는 단위로 판매한다.

'미'는 1kg을 기준으로 몇 마리가 들어가느냐를 말한다.

'미'의 계수가 클수록 작은 전복이며, 계수가 작을수록 큰 전복이다.

전복은 크기가 작으면 살수율($\frac{\text{순살 무게}}{\text{전체 무게}}$)이 낮아 실제 먹을 수 있는 양이 줄어들고, 크기가 너무 크면 질겨지므로 구매 시 적당한 크기를 고려해야 한다.

전복 데이터를 분석하고
전복의 순살 무게를 예측할 수 있는
인공지능 모델을 만들어 보자.



2 데이터를 준비하자!

데이터 다운로드 링크

<https://bit.ly/3mBcZZb>

1 Orange3 데이터 다운로드

- **Data** 카테고리 - [Datasets] 위젯 더블 클릭 - Abalone(전복) 데이터 더블 클릭

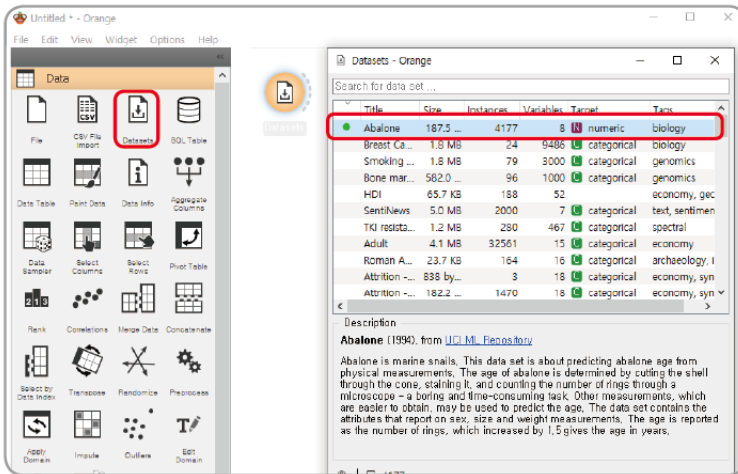


그림 1-1 Orange3에서 제공하는
Datasets의 다양한 데이터

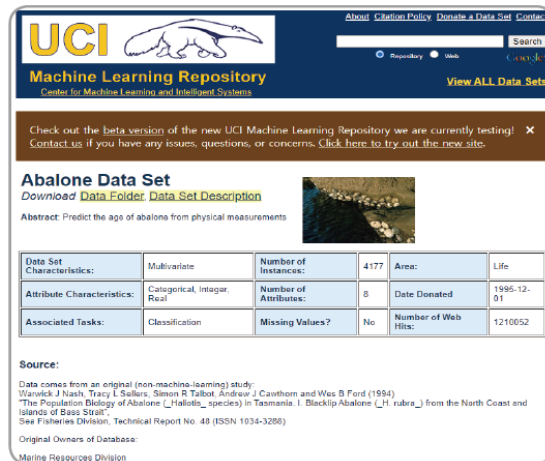
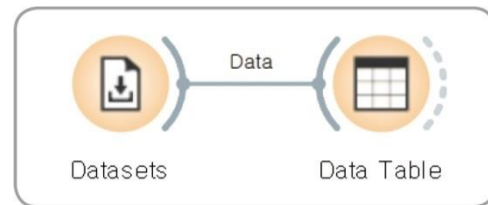


그림 1-2 UCI 전복 데이터

2 데이터 불러오기

- **Data** 카테고리 - [Data Table] 위젯을 가져와 [Datasets] 위젯에 연결 후 더블 클릭



	Rings	Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight
1	15	M	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.1500
2	7	M	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.0700
3	9	F	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.2100
4	10	M	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.1550
5	7	I	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.0550
6	8	I	0.425	0.300	0.095	0.3515	0.1410	0.0775	0.1200
7	20	F	0.530	0.415	0.150	0.7775	0.2370	0.1415	0.3300
8	16	F	0.545	0.425	0.125	0.7680	0.2940	0.1495	0.2600
9	9	M	0.475	0.370	0.125	0.5095	0.2165	0.1125	0.1650
10	19	F	0.550	0.440	0.150	0.8945	0.3145	0.1510	0.3200
11	14	F	0.525	0.380	0.140	0.6065	0.1940	0.1475	0.2100
12	10	M	0.430	0.350	0.110	0.4060	0.1675	0.0810	0.1350
4168	9	M	0.500	0.380	0.125	0.5770	0.2690	0.1265	0.1535
4169	8	F	0.515	0.400	0.125	0.6150	0.2865	0.1230	0.1765
4170	10	M	0.520	0.385	0.165	0.7910	0.3750	0.1800	0.1815
4171	10	M	0.550	0.430	0.130	0.8395	0.3155	0.1955	0.2405
4172	8	M	0.560	0.430	0.155	0.8675	0.4000	0.1720	0.2290
4173	11	F	0.565	0.450	0.165	0.8870	0.3700	0.2390	0.2490
4174	10	M	0.590	0.440	0.135	0.9660	0.4390	0.2145	0.2605
4175	9	M	0.600	0.475	0.205	1.1760	0.5255	0.2875	0.3080
4176	10	F	0.625	0.485	0.150	1.0945	0.5310	0.2610	0.2960
4177	12	M	0.710	0.555	0.195	1.9485	0.9455	0.3765	0.4950

→ 속성명

9개의 속성
4,177개의 전복 데이터

그림 1-3 [Data Table] 위젯으로
본 전복(Abalone) 데이터

3 데이터 속성 정보 확인하기

◆ Abalone Dataset: 4,177개 전복 데이터의 9개 속성

속성명	속성 정보
Rings	나이테: 연도를 나타냄.
Sex	성별: M(수컷), F(암컷), I(유아)
Length	길이: 최장 껍질 측정(mm)
Diameter	직경: 길이에 수직(mm)
Height	두께: 껍질과 살 포함(mm)

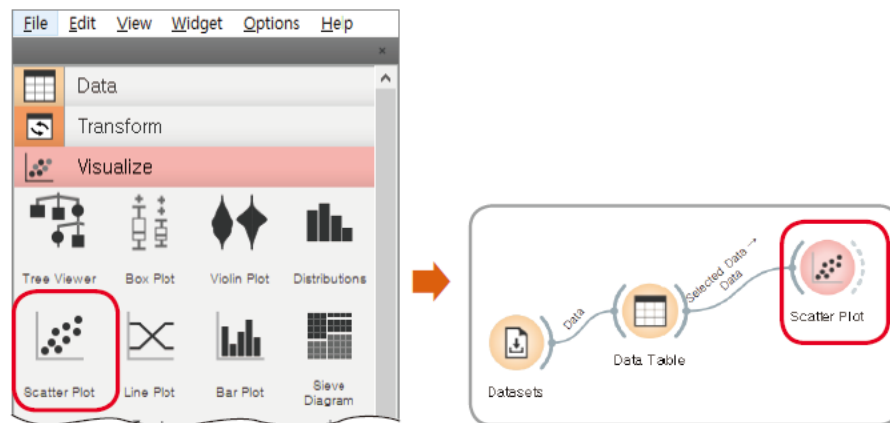
속성명	속성 정보
Whole weight	전체 무게: 그램 단위(g)
Shucked weight	순살 무게: 그램 단위(g)
Viscera weight	내장 무게: 피를 뺀 후 장 무게(g)
Shell weight	껍질 무게: 건조 후(g)

4 데이터 시각화하기

① [Scatter Plot] 위젯을 연결하여 산점도(Scatter Plot)로 시각화하기

Visualize 카테고리에는 산점도(Scatter Plot), 상자 그림(Box Plot), 도수분포표(Distributions Statistics) 등 데이터를 시각화할 수 있는 다양한 위젯 존재

- **Visualize** 카테고리 – [Scatter Plot] 위젯을 가져와 [Data Table] 위젯에 연결한 후 더블 클릭



② 전복 속성에 따른 산점도 확인하기

- **산점도**: 두 변수를 x축과 y축으로 설정하고 두 값이 만나는 곳에 점으로 나타낸 그래프
- **y축의 이름**: 전복 순살 무게 (Shucked weight)로 설정
- **x축의 이름**: 여러 가지 속성으로 변경
위 과정을 통해 각 속성에 따른 전복 순살 무게를 확인

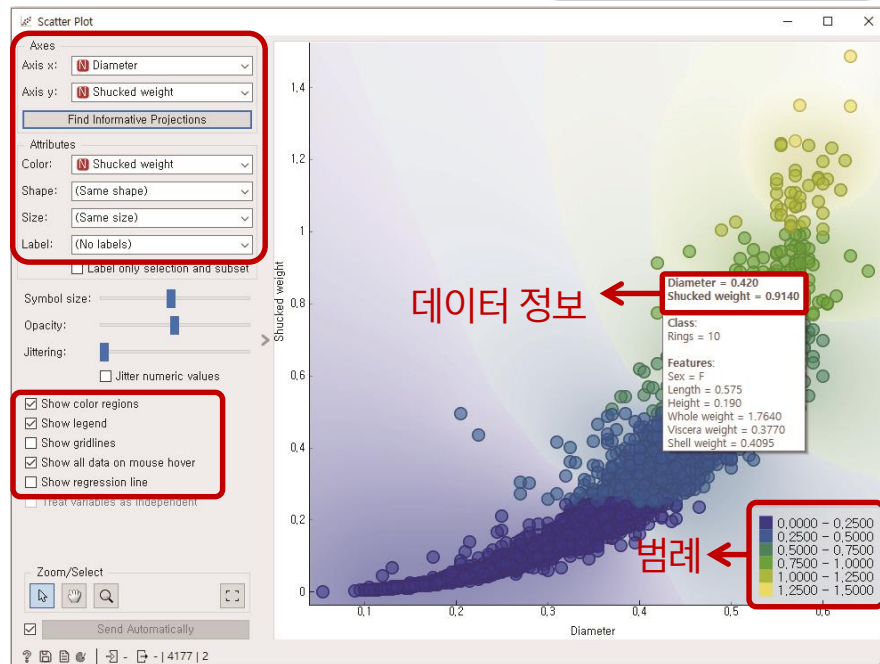
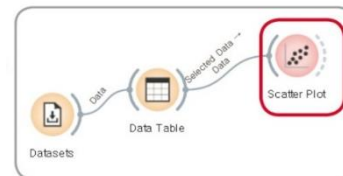
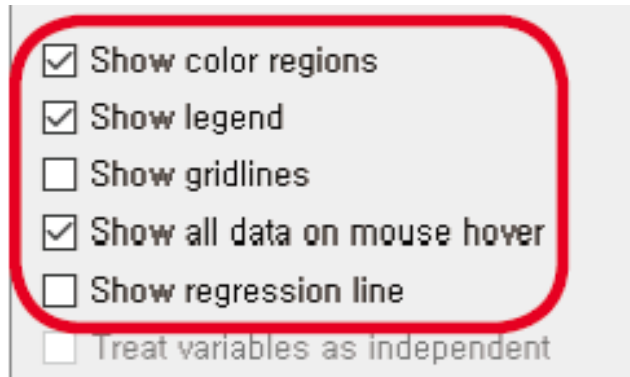


그림 1-4 [Scatter plot] 위젯 창과 전복 데이터 산점도 화면
(x가 Diameter 속성일 경우)

③ [Scatter Plot] 위젯의 다양한 설정을 이용하여 여러 가지 형태로 표현하기

Attribute의 Color, Shape, Size, Label을 설정하여 점의 색, 모양, 크기, 라벨 표현

- Show color regions를 체크하면 색이 지역별로 표시
- Show legend를 체크하면 범례 표시
- Show all data on mouse hover를 체크하고
산점도 점 위에 마우스를 올려놓으면
점에 해당하는 데이터 정보가 표시



5 데이터 전처리하기

① [Select Columns] 위젯을 이용하여 Target 정하기

- **Transform** 카테고리의 [Select Columns] 위젯을 [Datasets] 위젯에 연결하고, [Data Table] 위젯을 추가하여 [Select Columns] 위젯과 연결

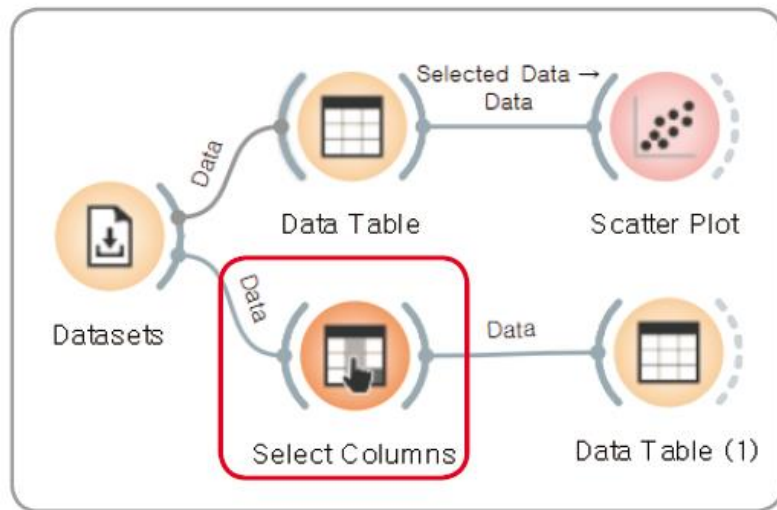


그림 1-5 [Select Columns]과 [Data Table(1)] 위젯 연결

① [Select Columns] 위젯을 이용하여 Target 정하기

- [Select Columns] 위젯을 더블 클릭하여 전복 순살 무게(Shucked weight)를 Features에서 Ignored로 옮긴 후(①), 다시 Target으로 옮기기(②).

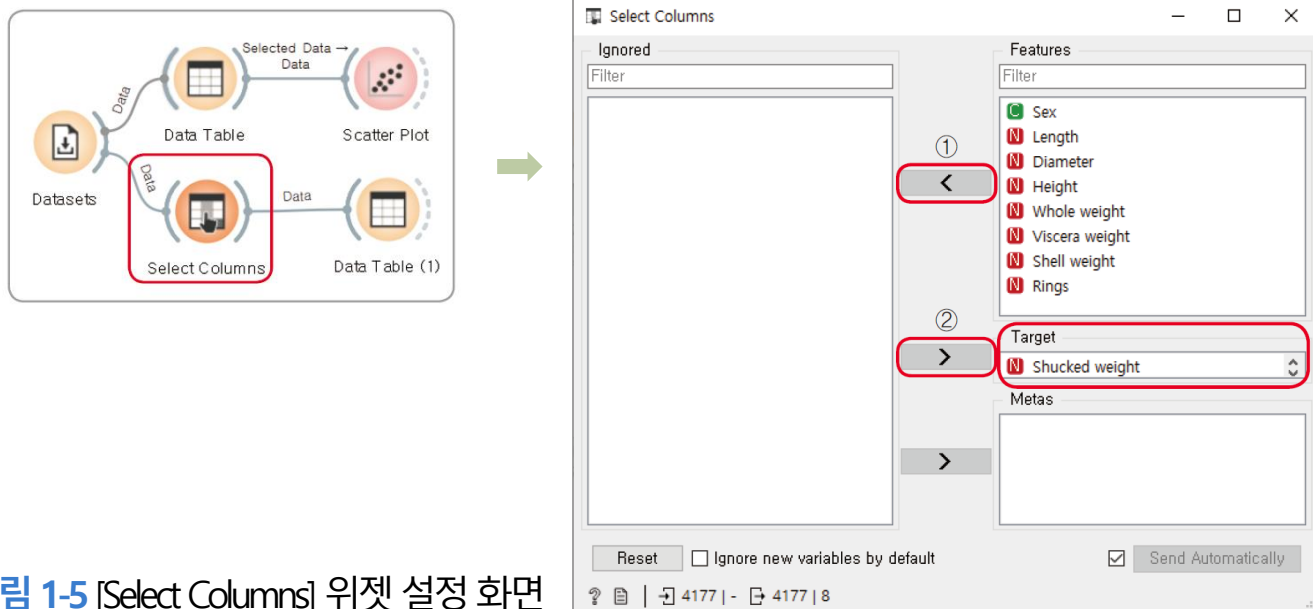






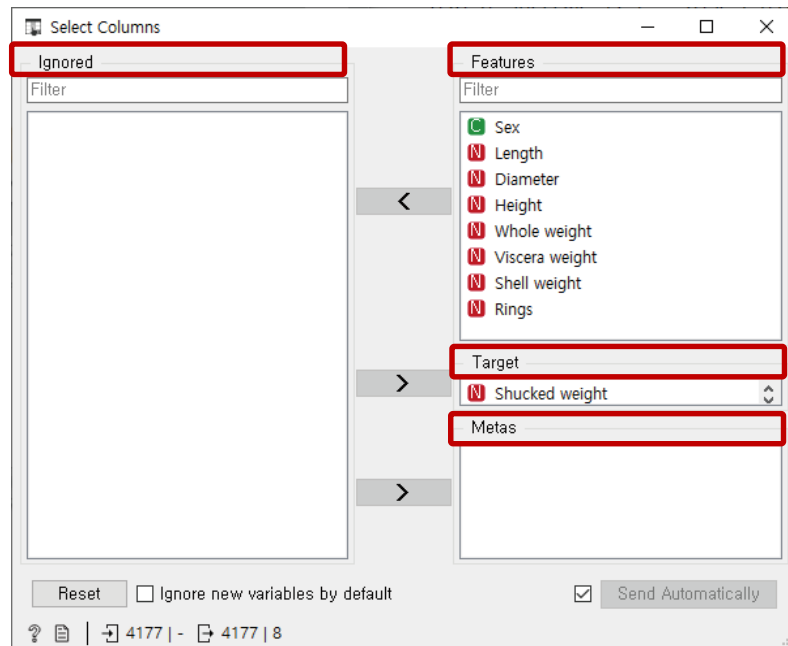
그림 1-5 [Select Columns] 위젯 설정 화면

② [Select Columns] 위젯을 이용하여 데이터 형식(Type) 알아보기

데이터 종류	형식(Type)	예시
범주형(같은 특성 부류)	categorical  categorical	있다/없다, 남자/여자, 식물/동물
수치형(정수, 소수)	numeric  numeric	1, -3, 0.5346
문자형	text  text	school, apple
날짜형	datetime  datetime	2020-06-01/ 2020-08-03 00:03:45

③ 속성과 결과를 정하기 위한 역할(Role) 설정하기

- **Feature**: 기계학습에 영향을 미치는 원인이 되는 속성(독립 변수)
- **Target**: 예측하고자 하는 결과가 되는 속성(종속 변수)
- **Meta**: 사용되지는 않지만, 참고만 하기 위해 보여지는 데이터
- **Ignored**: 분석에 사용되지 않아 무시할 데이터

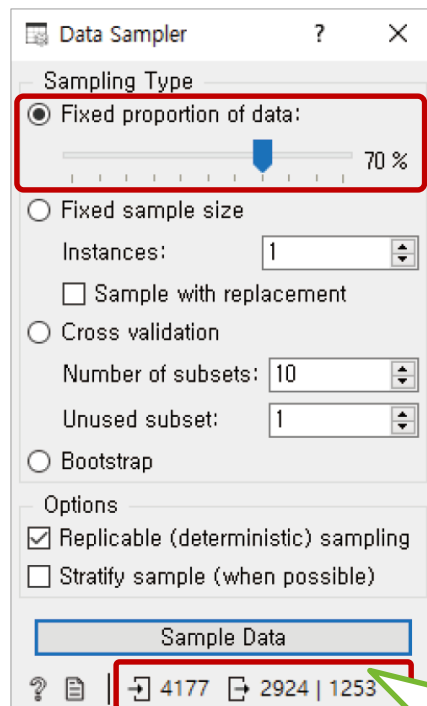
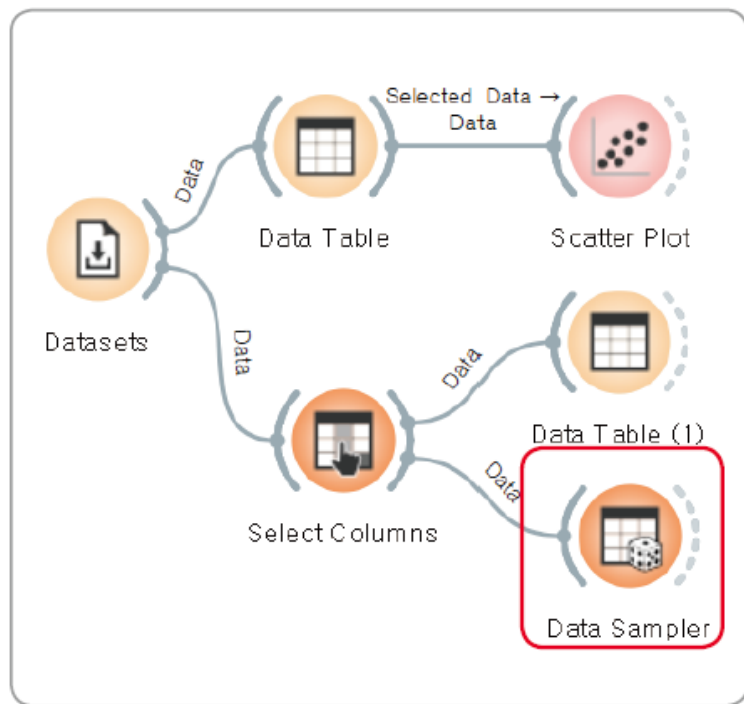


3 어떤 모델을 선택하고 학습시킬까?

1 학습 모델 선택하기

① 훈련 데이터와 테스트 데이터 나누기

- [Data Sampler] 위젯 이용하기
 - Transform 카테고리의 [Data Sampler] 위젯을 캔버스로 가져오기
- 훈련 데이터와 테스트 데이터 비율 정하기
 - [Data Sampler] 위젯을 더블 클릭하여 일반적으로 Fixed proportion of data를 70%로 설정
 - 훈련 데이터(70%), 테스트 데이터(30%)



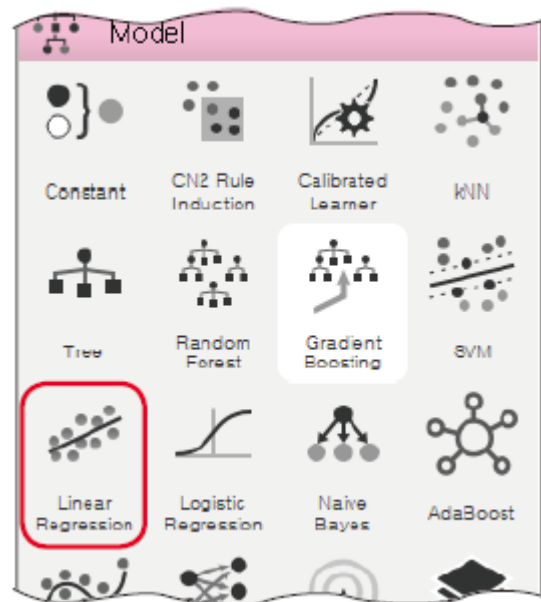
일반적으로 훈련 데이터와 테스트 데이터를 나누는 작업은 '데이터 준비하기' 과정에서 진행한다.

그림 1-7 [Data Sampler] 위젯으로 훈련 데이터와 테스트 데이터 분할

- 입력 데이터: 4,177개
- 훈련 데이터: 70%인 2,924개
- 테스트 데이터: 30%인 1,253개

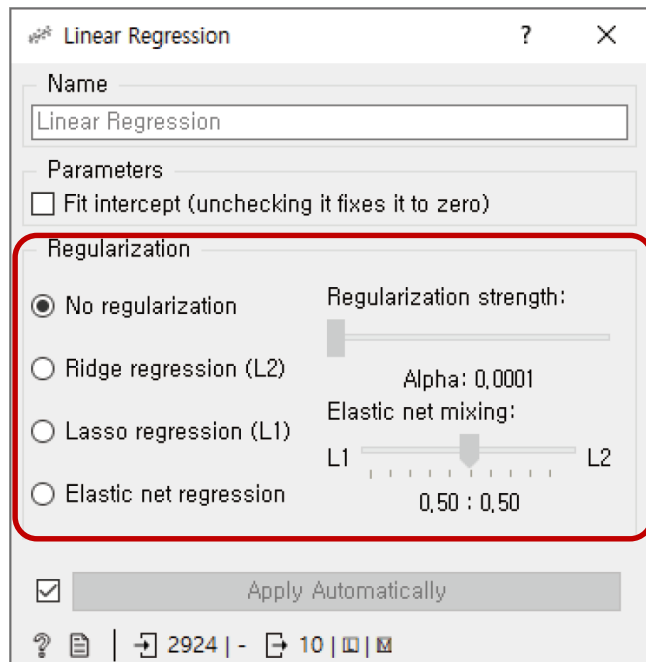
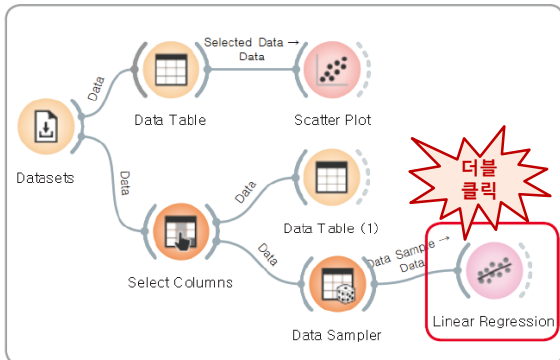
② 모델 선택하기

- Model 카테고리에서
[Linear Regression] 위젯 사용
- 연속적인 값을 예측하는 모델인
[Linear Regression] 위젯을
[Data Sampler] 위젯과 연결하면
전북 데이터(2,924개의 훈련 데이터)를
이용한 선형 회귀 모델 구현



2 학습시킴

[Data Sampler] 위젯에 [Linear Regression] 위젯을 연결하면 모델 위젯이 자동으로 실행되어 데이터 학습시킴.



Regularization(정규화, 과적합 방지)

- **Ridge regression(L2):** 각 계수의 제곱을 더하는 방식
- **Lasso regression(L1):** 각 계수의 절댓값을 더하는 방식
- **Elastic net regression:** L2와 L1 방식을 절충한 것

그림 1-8 훈련 데이터로 선형 회귀 모델 학습시킴

AI랑 친해지기

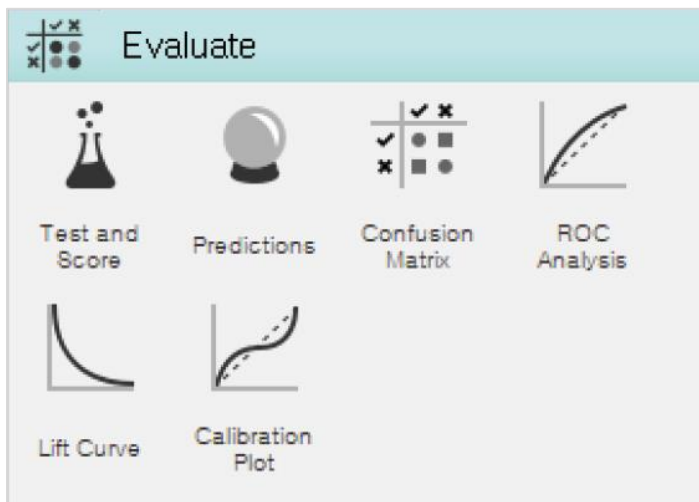
선형 회귀(Linear Regression)

- 직선의 방정식은 기본적으로 $y = ax + b$ 와 같다. 선형 회귀는 이러한 직선 형태의 함수를 이용하여 입력 데이터를 통해 오차가 가장 적은 방정식을 연속적으로 찾는다. 물론 이 방정식을 찾기 위해서는 a (기울기)와 b (절편)이 필요하다. 기계학습에서는 기울기는 가중치란 용어로 사용하고 절편은 바이어스라는 용어로 사용한다.
- x 와 y 간의 관계가 선형 관계라고 가정할 때, 데이터를 잘 나타내는 선을 찾는 것이 선형 회귀 모델의 핵심이다. 데이터를 가장 잘 나타내느냐의 판단은 실제 데이터와의 오차를 이용한다. 오차가 작은 선을 찾게 되고, 이 오차는 음수가 나올 수 있으므로 예측값과 실제값의 차이를 제곱하거나 절댓값을 취하는 등의 방법으로 평가한다.

4 모델의 성능을 확인해 보자!

1 학습 결과 확인하기

① Evaluate 카테고리 열기



- Evaluate 카테고리
인공지능 모델을 확인 및 평가할 수
있는 위젯 모음

② 위젯 연결하기

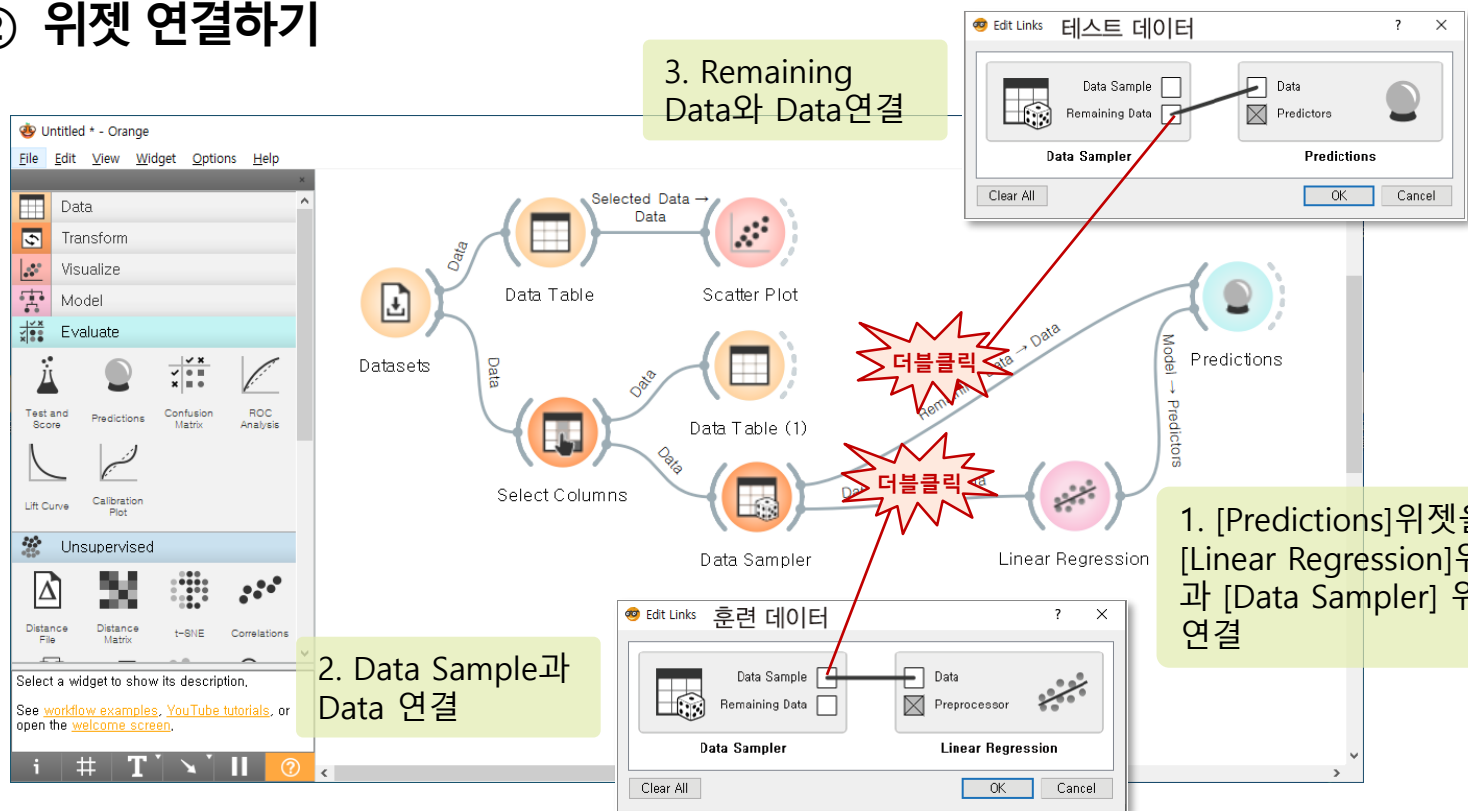


그림 1-9 선형 회귀 모델 성능 평가

2 학습시키기

① [Prediction] 위젯을 더블 클릭하여 성능 결과 확인하기

	Linear Regression	Shucked weight	Sex	Length	Diameter	Height	Whole weight	Viscera weight	Shell weight	Rings
1	0.5039	0.5105	F				1.2075	0.2620	0.3900	10
2	0.3789	0.3980	M				0.7575	0.1510	0.1750	8
3	0.5871	0.6000	F				1.3270	0.3015	0.3550	10
4	0.9299	1.0260	M	0.725	0.305	0.185	1.9780	0.4255	0.4505	12
5	0.2696	0.2580	F	0.515	0.400	0.170	0.7960	0.1755	0.2800	16
6	0.0215	0.0215	I	0.220	0.160	0.050	0.0490	0.0100	0.0150	4
7	0.4818	0.4835	F	0.610	0.465	0.160	1.0725	0.2515	0.2800	10
8	0.6183	0.6275	F	0.670	0.535	0.185	1.5970	0.3500	0.4700	21
9	0.5550	0.5560	M	0.675	0.515	0.150	1.3120	0.2845	0.4115	11
10	0.7901	0.8435	F	0.745	0.585	0.190	1.9660	0.4370	0.5855	18
11	0.1106	0.1080	I	0.365	0.275	0.135	0.2400	0.0445	0.0735	7
12	0.0775	0.0695	I	0.335	0.250	0.080	0.1695	0.0440	0.0495	6
13	0.1071	0.0930	I	0.370	0.275	0.080	0.2325	0.0560	0.0720	6
14	0.1160	0.1150	I	0.395	0.295	0.095	0.2725	0.0625	0.0850	8
15	0.3841	0.3685	M	0.545	0.435	0.145	0.9385	0.1245	0.3450	11
16	0.0171	0.0135	I	0.210	0.150	0.045	0.0400	0.0080	0.0105	4
17	0.3903	0.3895	F	0.560	0.430	0.145	0.8980	0.2325	0.2450	9
18	0.3750	0.3215	M	0.580	0.470	0.165	0.9270	0.1985	0.3150	11
19	0.5499	0.5320	M	0.670	0.540	0.195	1.2170	0.2735	0.3315	11
20	0.8133	0.8035	F	0.695	0.535	0.175	1.8385	0.3960	0.5030	10
21	0.3640	0.3135	F	0.565	0.445	0.125	0.8305	0.1785	0.2300	11
22	0.3234	0.2720	M	0.540	0.415	0.130	0.8245	0.2260	0.2400	13
23	0.5472	0.5785	M	0.625	0.500	0.130	1.0820	0.2045	0.2500	8

Model	MSE	RMSE	MAE	R2
Linear Regression	0.001	0.036	0.023	0.975

실제값과 예측값
유사함 확인 가능

- Shucked weight
테스트 데이터의 실제
순살 무게(실제값)
- Linear Regression
선형 회귀 모델로 예측한
순살 무게(예측값)

그림 1-10

선형 회귀 모델 평가하기 성능 확인

② 전복 순살 무게 예측 인공지능 모델 확인하기

어느 정도의 정확도로 예측하였는지 예측 모델의 성능을 평가하는 4개의 평가 지표

Model	MSE	RMSE	MAE	R2
Linear Regression	0.001	0.036	0.023	0.975

평가 지표

Linear Regression은
정확도가 매우 높은
모델 확인



MSE, RMSE, MAE는
0에 가까울수록,
R2는 1에 가까울수록
정확도가 높아요.

평가 지표

MSE(Mean Squared Error)

RMSE(Root Mean Squared Error)

MAE(Mean Absolute Error)

R2(R Squared, R2, 결정계수)

그림 1-11 예측 모델 평가 지표



Orange3 장점

Q Orange3의 Data 카테고리의 [Datasets] 위젯에 있는 데이터들은 사용하기 적절한가요?

A [Datasets] 위젯에는 70여 개의 데이터가 있으며, 붓꽃 품종 데이터, 타이타닉 생존자 데이터, 와인 품질 데이터, 보스턴 집값 예측 데이터 등 널리 알려진 데이터가 많습니다. 인공지능을 처음 시작하는 사람들이 사용하기 적합한 데이터로 구성되어 있습니다.

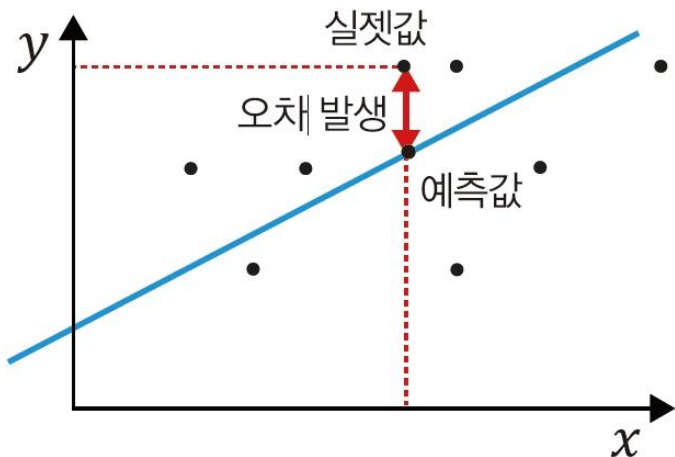
AI

전문가
되기

예측(회귀) 모델 평가 지표

지금까지 배운 내용을 바탕으로 성능 평가를 한 후 4개의 평가 지표를 정리해 보자.

$$H(x) = Wx + b$$



오차는 양수와 음수가 섞여 있으므로 정확도를 알기 위해서는 4개 평가 지표를 이용한다.

점들을 학습시켜서 모델 $H(x)$ 를 만든다.
 이때 어떤 x 상태일 때 실제값(거리값) y 와
 예측값 $H(x)$ 의 차이가 발생할 수 있다.
 이 차이가 작으면 작을수록 학습 모델이
 예측을 잘한다고 할 수 있다.

- ① **MSE(Mean Squared Error):** 예측값과 실젯값 차이(오차)의 제곱을 평균 낸 값으로, 0에 가까울수록 예측값과 실젯값의 차이가 없으므로 성능 우수

$$MSE = \frac{1}{n} \sum_{i=1}^n (H(x_i) - y_i)^2$$

- ② **RMSE(Root Mean Squared Error):** MSE를 Root로 처리한 값으로, 0에 가까울수록 성능 우수

$$RMSE = \sqrt{MSE}$$

- ③ **MAE(Mean Absolute Error):** 오차의 절댓값을 평균한 값으로, 0에 가까울수록 성능 우수

$$MAE = \frac{1}{n} \sum_{i=1}^n |H(x_i) - y_i|$$

- ④ **R₂(R Squared, 결정계수)**: 적합한 정도를 재는 척도로,
 1에 가까울수록 성능이 우수하고 0에 가까울수록 성능 저하

$$R^2 = \frac{\sum_{i=1}^n (H(x_i) - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

→ 예측값($H(x_i)$)과 실제값 평균(\bar{y})의 차이를 제공한 값들의 합

→ 실제값(y_i)과 실제값 평균(\bar{y})의 차이를 제공한 값들의 합

또는

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - H(x_i))^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

→ 실제값(y_i)과 예측값($H(x_i)$)의 차이를 제공한 값들의 합

→ 실제값(y_i)과 실제값 평균(\bar{y})의 차이를 제공한 값들의 합

정리하기

Orange3를 활용하면 별도의 프로그래밍 없이도 Orange3 데이터를 가져와 처리하고 인공지능 모델을 만들어 평가까지 할 수 있었다.

또한 Orange3 데이터로 학습하고 예측 모델의 평가 지표 수치를 확인해 보니 전반적으로 성능이 우수한 것으로 나왔다.

이처럼 Orange3에는 인공지능 모델을 학습하고 평가하는 데 유용한 데이터가 많다.

A large green circle with a thick border is centered on a background of repeating green chevrons. Inside the circle, the text "Q & A" is written in a bold, dark grey sans-serif font.

Q & A