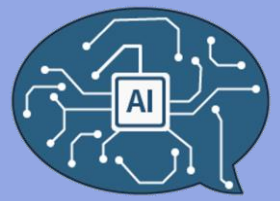
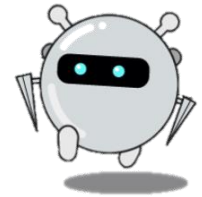


Regression



OUTTA

목차

01 지도학습 주요내용 Recap

OUTTA

| 01 지도학습 주요내용 Recap

분류 (Classification)

- 데이터를 기반으로 새로운 샘플의 범주형 클래스 레이블 예측이 목표

- 이진 분류(binary classification): 클래스 레이블 종류가 두 개인경우

- 다중 클래스 분류 (multi-class classification): 클래스 레이블 종류가 세 개 이상

회귀 (Regression)

- 연속적인 출력 값에 대한 예측이 목표

- 예) 시험 점수 예측,
  주식 가격 예측,
  부동산 가격 예측

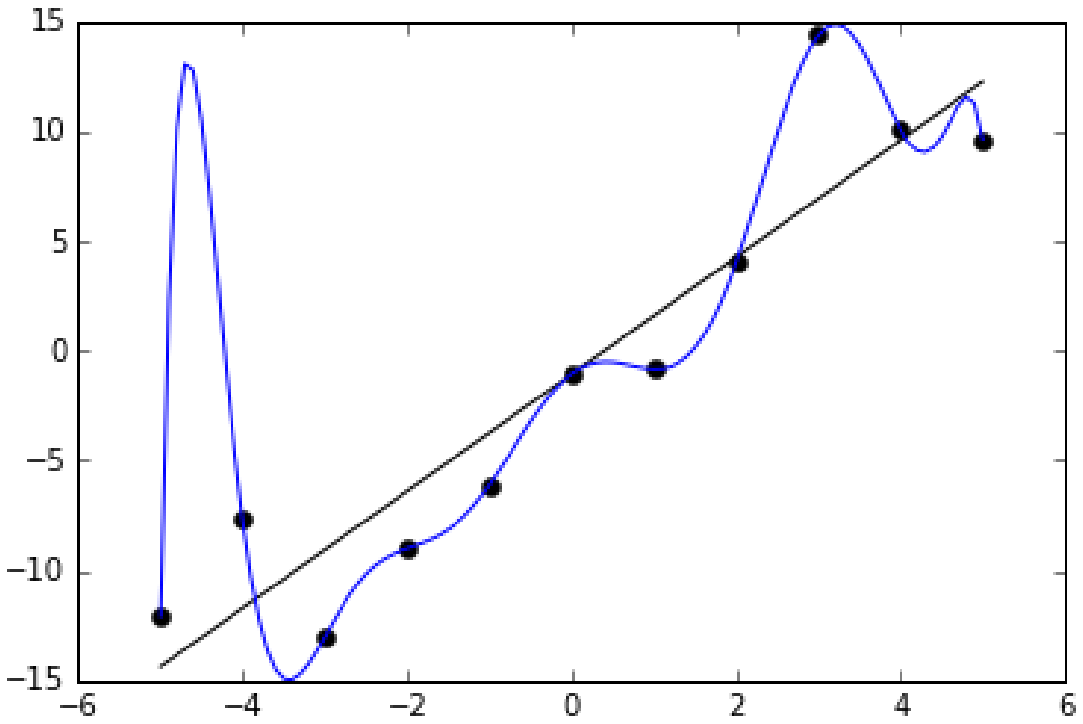
| 01 지도학습 주요내용 Recap

데이터 타입에 따라서 모델 구축/ 목적에 따라 선정하는 모델이 다름

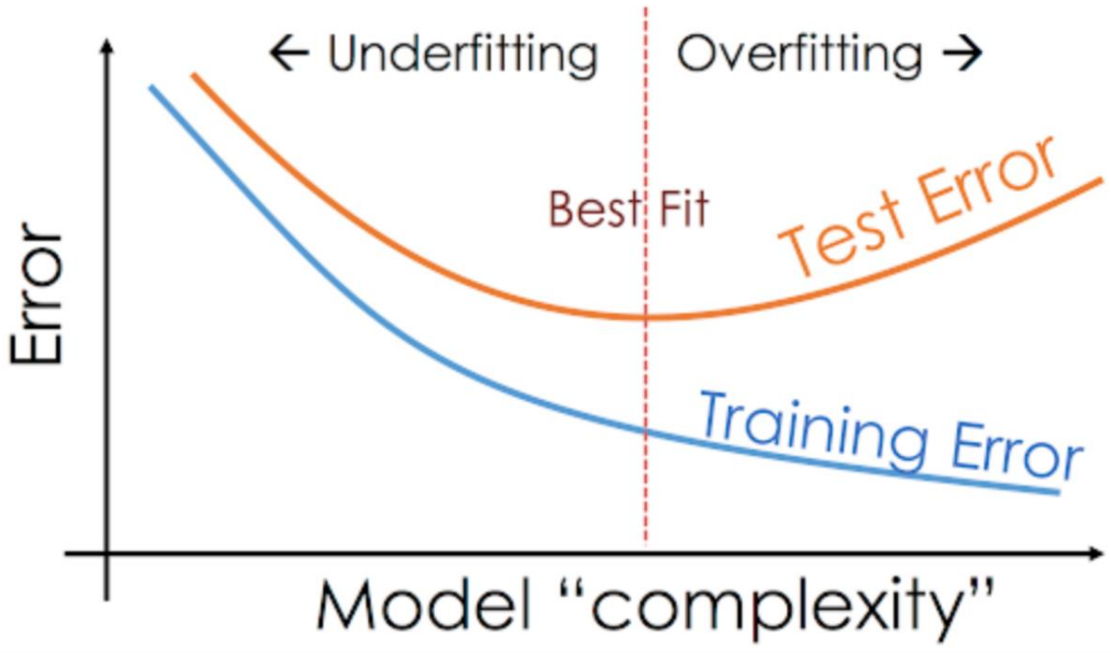
<모델에 대한 4가지 지표>

- 모델 정확도 (Accuracy): 모델의 예측/분류 등 성과에 대한 지표
- 모델 복잡도 (Complexity): 모델의 학습 과정에서 소요되는 계산량
- 모델 해석도 (Interpretability): 학습 결과에 대한 해석력이 높을수록 모델 피드백이 쉬움
- 모델 확장도 (Scalability): 데이터의 추가에 따른 모델의 정확도 상승이 어디까지 가능한지에 대한 지표

| 01 지도학습 주요내용 Recap



Overfitting & Underfitting

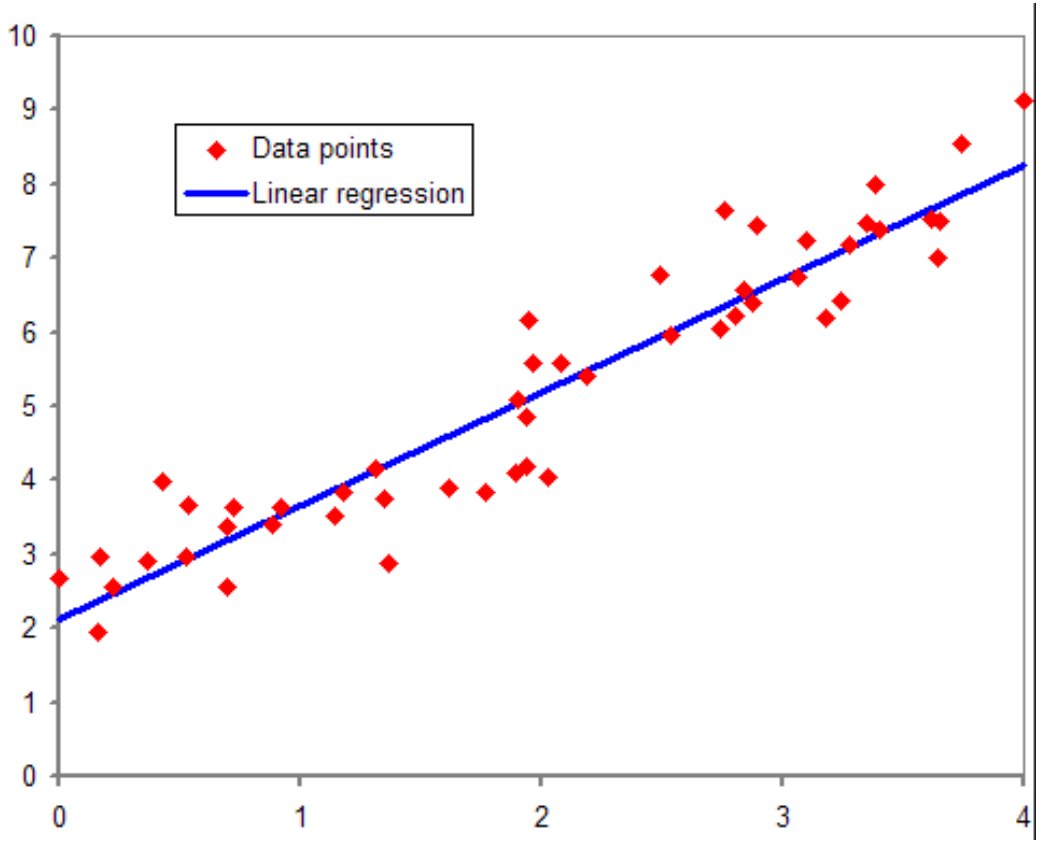


Bias-Variance Trade-off

02 회귀 모델

OUTTA

| 02 회귀모델: 선형 회귀



Simple Linear Regression

- w: weight, b: bias (parameter)
- x: 독립 변수
- y: 종속 변수
- 한 개의 독립 변수 x와 종속 변수 y의 선형 관계를 모델링

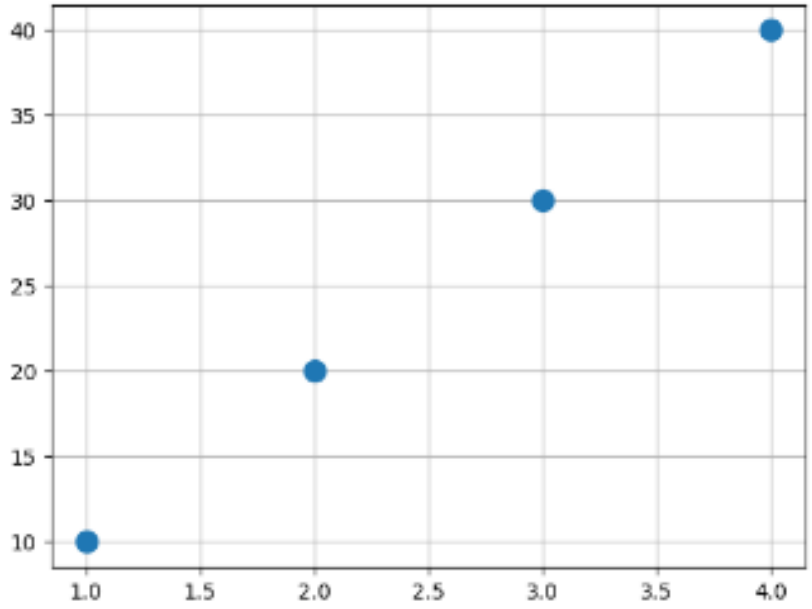
$$y = wx + b$$

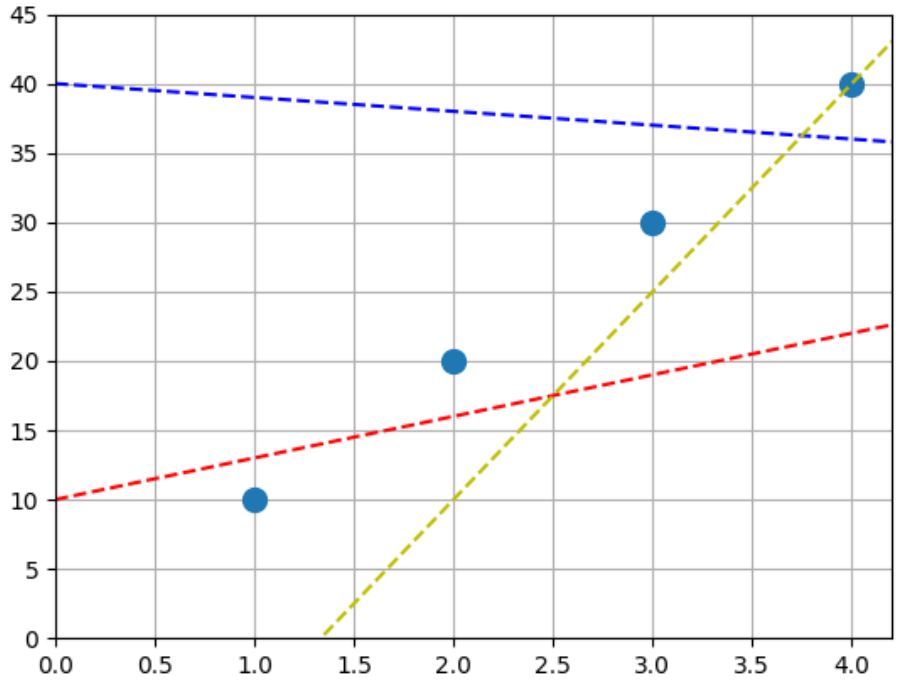
https://danawalab.github.io/machinelearning/2022/09/13/MachineLearning-LinearRegression.html

| 02 회귀모델: 선형 회귀

Simple Linear Regression

x	y
1	10
2	20
3	30
4	40



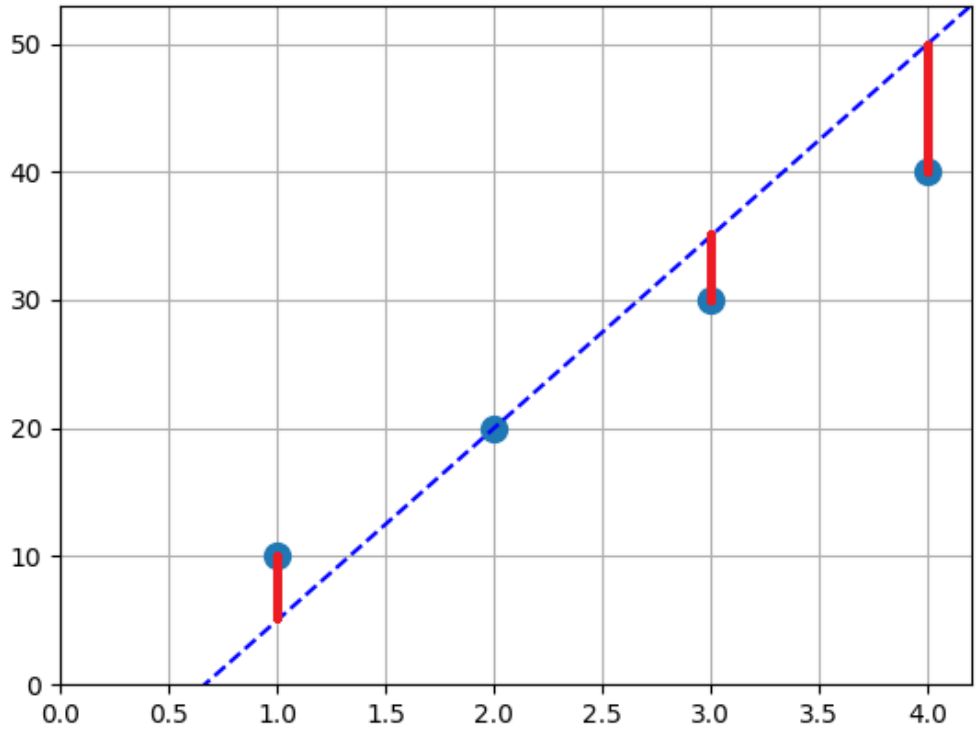


$$Find\ the\ best\ w, b\ such\ that\ y = wx + b$$

https://danawalab.github.io/machinelearning/2022/09/13/MachineLearning-LinearRegression.html

| 02 회귀모델: 선형 회귀

Simple Linear Regression



x	1	2	3	4
실제값	10	20	30	40
예측값	5	20	35	50
오차	5	0	-5	-10

$$\frac{1}{n}\sum_{i=1}^{n}[y_i - H(x_i)]^2 = \frac{5^2 + 0^2 + (-5)^2 + (-10)^2}{n} = \frac{150}{4} = 37.5$$

Find w, b such that minimizes MSE

https://danawalab.github.io/machinelearning/2022/09/13/MachineLearning-LinearRegression.html

OUTTA

| 02 회귀모델: 선형 회귀

Multiple Linear Regression

- $w_1, w_2, \ldots, w_n$: weight, b: bias (parameter)
- $x_1, x_2, \ldots, x_n$: 독립 변수
- y: 종속 변수
- n개의 독립 변수 $x_1, x_2, \ldots, x_n$와 종속 변수 y의 선형 관계를 모델링

$$y = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b$$

| 02 회귀모델: Ridge Regression

Ridge Regression

- 특정 계수의 weight에 패널티를 줌.
- RSS를 줄이려면 주어진 훈련 데이터에 적합을 잘하는 것 뿐만 아니라
- 계수들의 제곱의 합을 줄이는 것도 중요함.
- 결과적으로 특정 변수에 overfit되는 것을 방지한다고 볼 수 있음.

$$Find\ w, b\ such\ that\ minimizes\ \ldots$$

$$RSS_{ridge}(w,b) = \sum_{i=1}^{n}\left(y_i - (w_i x_i + b)\right)^2 + \alpha \sum_{j=1}^{p} w_j^2$$

| 02 회귀모델: Lasso Regression

Lasso Regression

- 특정 계수의 weight에 패널티를 줌.
- RSS를 줄이려면 주어진 훈련 데이터에 적합을 잘하는 것 뿐만 아니라
- 계수들의 절댓값을 줄이는 것도 중요함.
- 결과적으로 특정 변수에 overfit되는 것을 방지한다고 볼 수 있음.

$$Find\ w, b\ such\ that\ minimizes\ \ldots$$

$$RSS_{Lasso}(w,b) = \sum_{i=1}^{n}\left(y_i - (w_i x_i + b)\right)^2 + \alpha \sum_{j=1}^{p}|w_j|$$

| 02 회귀모델: Ridge vs Lasso Regression

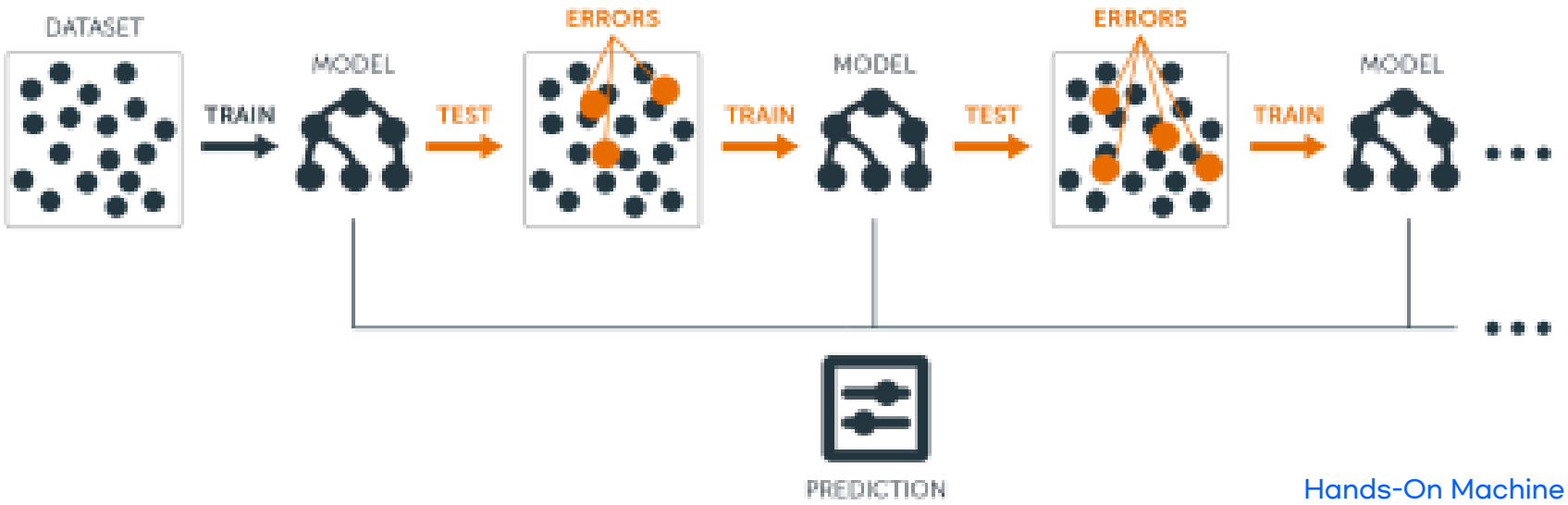
Ridge vs Lasso Regression

- 공통점: 계수에 대한 규제를 주어 특정 변수에 대한 overfit을 방지
- 차이점
  1) Ridge는 계수가 커질수록 패널티가 기하급수적으로 증가함 (제곱항)
  2) 반면 계수가 1보다 작을 때는 Lasso에 비해 덜 규제를 가함
  3) 또한 Lasso는 최적화 과정에서 계수가 0으로 수렴하는 경향이 있음.
     => 자연스럽게 Feature Selection 과정이 포함됨!

| 02 회귀모델: Gradient Boosting

Gradient Boosting

- 약한 모델이 순차적으로 적용되면서 앞선 모델의 에러를 개선해나감
- 분류와 회귀 문제 모두에 쓰일 수 있음
- 종속변수와 독립변수가 비선형적인 관계를 띨 때 특히 유용



Hands-On Machine Learning with R

| 02 회귀모델: XGBoost

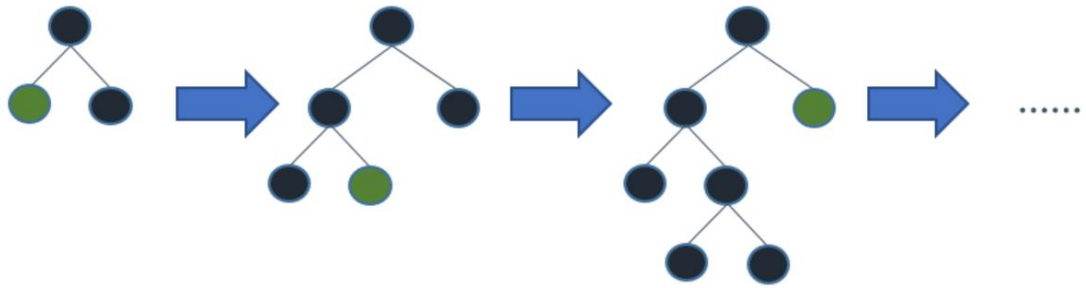
XGBoost (eXtreme Gradient Boosting)

- Gradient Boosting 알고리즘을 최적화한 것
- 알고리즘의 계산 속도 개선, 예측 성능 또한 개선
- Gradient Boosting에 규제를 추가해 모델 복잡성을 제어하고 과적합 방지
- 결측값 처리, 트리 가지치기, 내장된 교차 검증, 병렬 처리의 특징을 지님

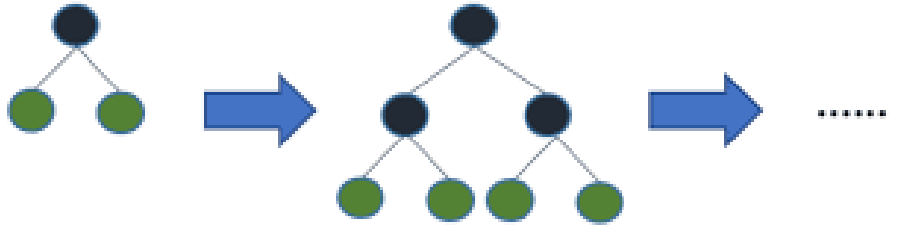
| 02 회귀모델: LightGBM

LightGBM (Light Gradient Boosting Machine)

- Microsoft에서 개발한 Gradient Boosting 프레임워크
- Leaf-wise Tree Growth가 특징. 좀 더 복잡한 트리를 만들 수 있음
- 범주형 특성을 자동으로 변환하는 기능 제공. 원-핫 인코딩할 필요 없음



Leaf-wise tree growth



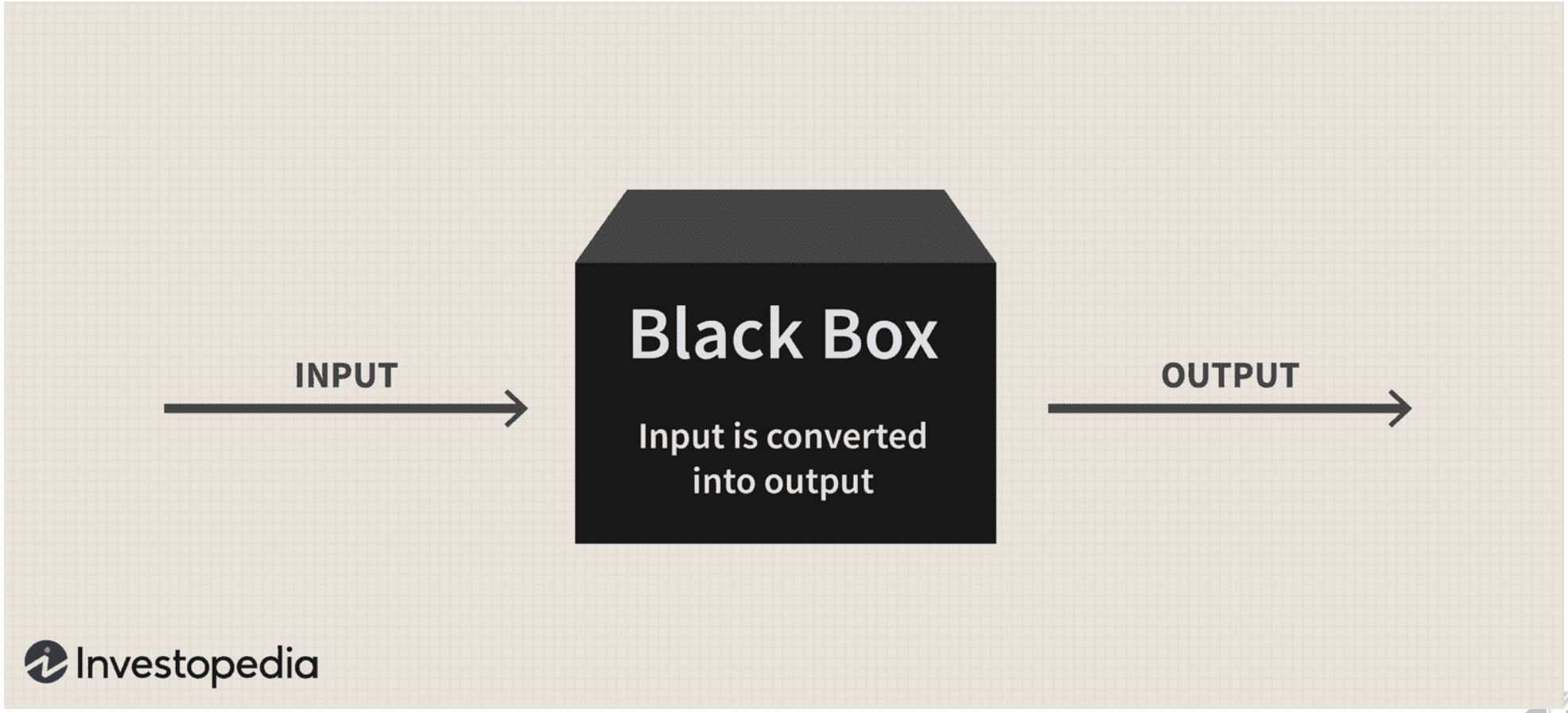
Level-wise tree growth

https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-it-how-to-fine-tune-the-parameters-60347819b7fc

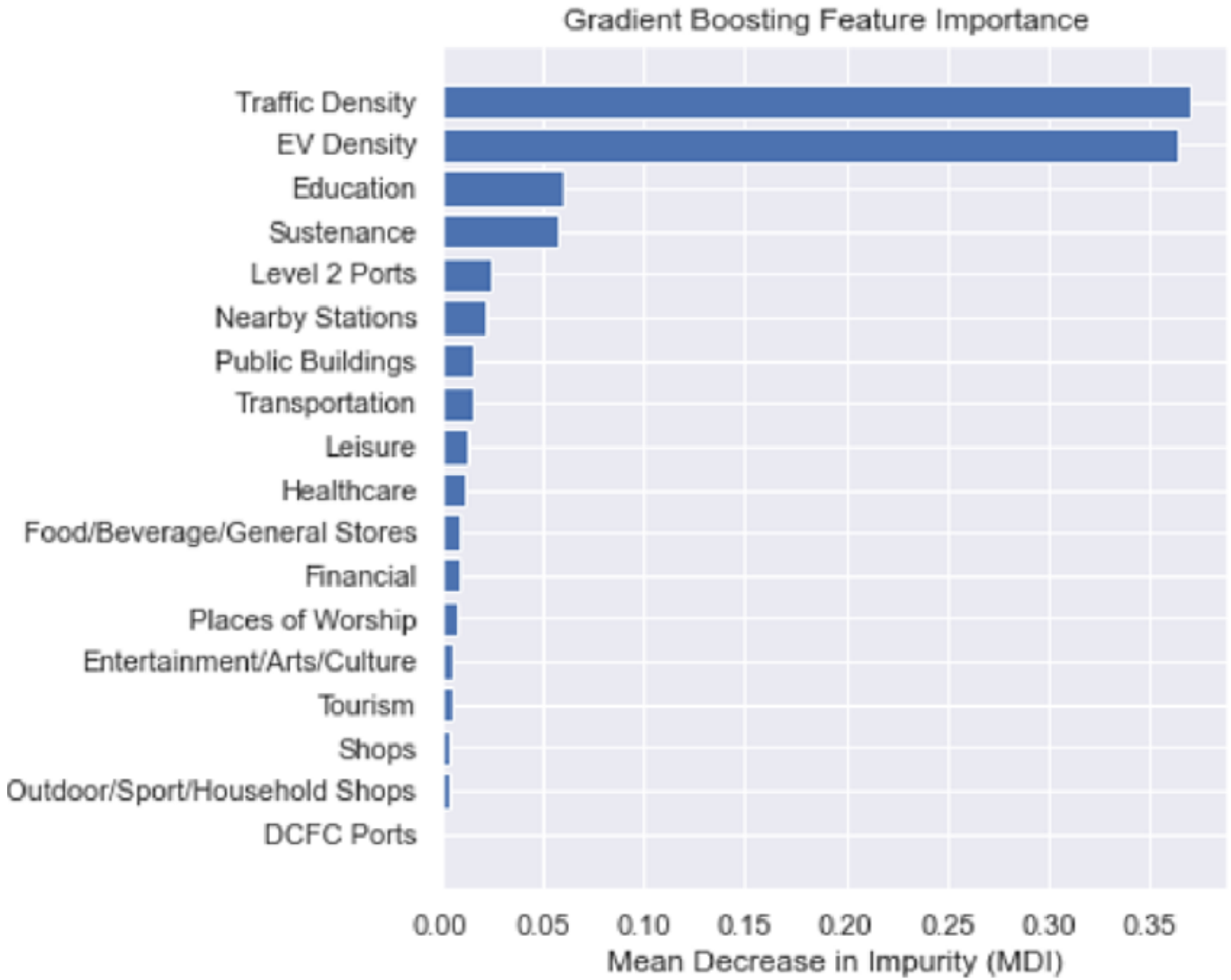
03 회귀 모델의 해석

OUTTA

| 03 회귀 모델의 해석: Gradient Boosting 기반 모델



| 03 회귀 모델의 해석: Gradient Boosting 기반 모델



OUTTA

| 03 회귀 모델의 해석: 선형 회귀 모델

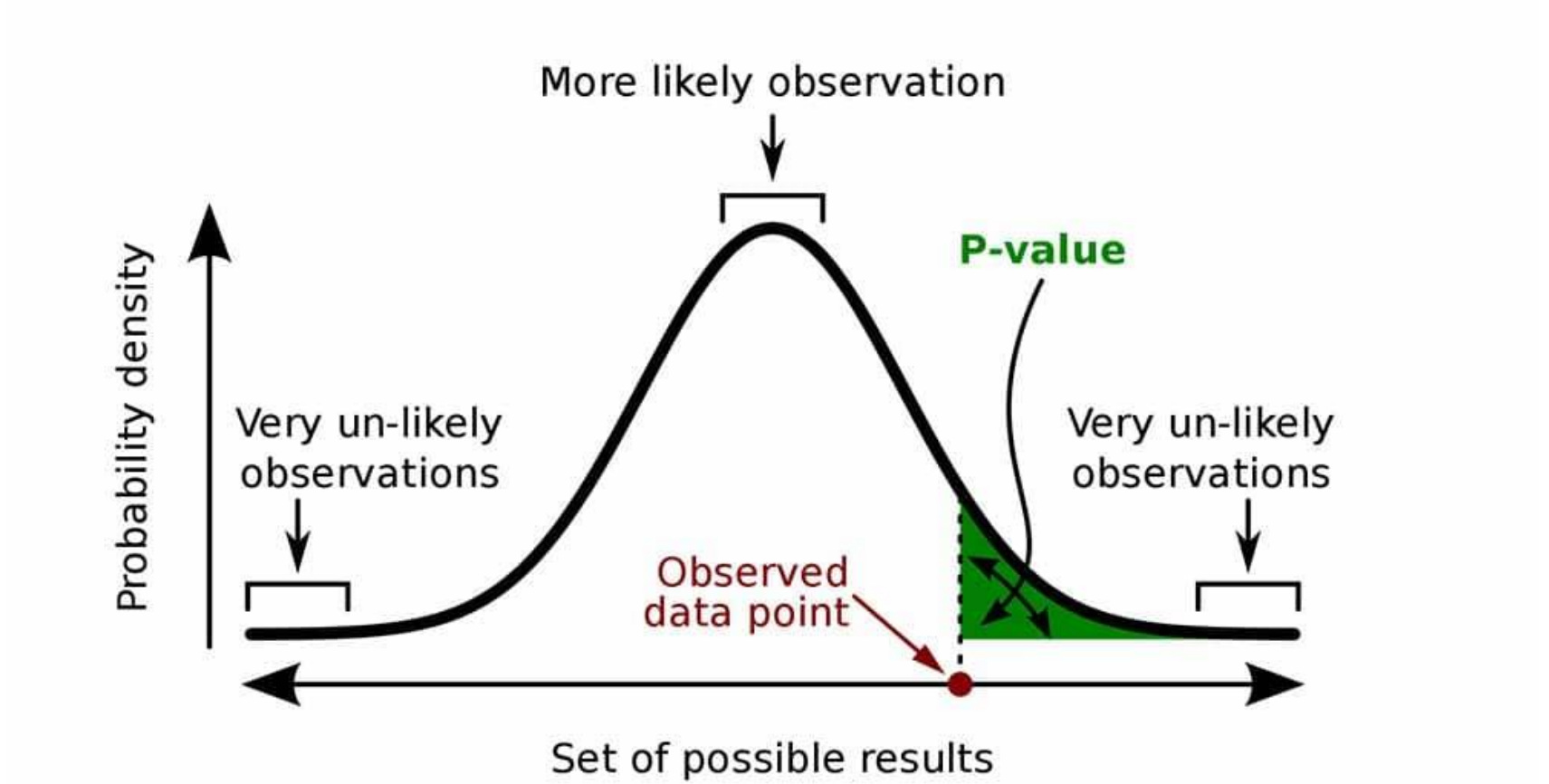
가설 검정

- 가설: 어떤 사실을 설명하거나 어떤 이론 체계를 연역하기 위하여 설정한 가정. 이로부터 이론적으로 도출된 결과가 관찰이나 실험에 의하여 검증되면, 가설의 위치를 벗어나 일정한 한계 안에서 타당한 진리가 된다. (네이버 지식백과)
  ex) “수능 날에는 날씨가 추워진다.”
  ex) “학생 평가가 자주 이뤄질수록 학생의 성적이 상승한다.”
  ex) “주식 가격 결정에는 기업의 내재 가치가 큰 영향을 준다.”
  ex) “이번에 개발한 신약이 기존의 약보다 효과가 좋다.”
  ex) “이번에 개발한 머신러닝 모델이 기존의 모델보다 효과가 좋다.”
- 가설 검정: 자료에 근거하여 자신이 세운 통계적 가설을 적정한 유의수준 하에 채택 또는 기각 하는 과정

| 03 회귀 모델의 해석: 선형 회귀 모델

귀무가설 (Null Hypothesis, H0)	대립가설 (Alternative Hypothesis, Ha)
• 가설검정의 대상이 되는 가설 • 일반적으로 기각될 것이 예상되어 세워진 가설 • 대립가설의 반대	• 귀무가설이 기각될 때 대체되는 가설 • 일반적으로 실험자의 믿음에 해당

| 03 회귀 모델의 해석: 선형 회귀 모델

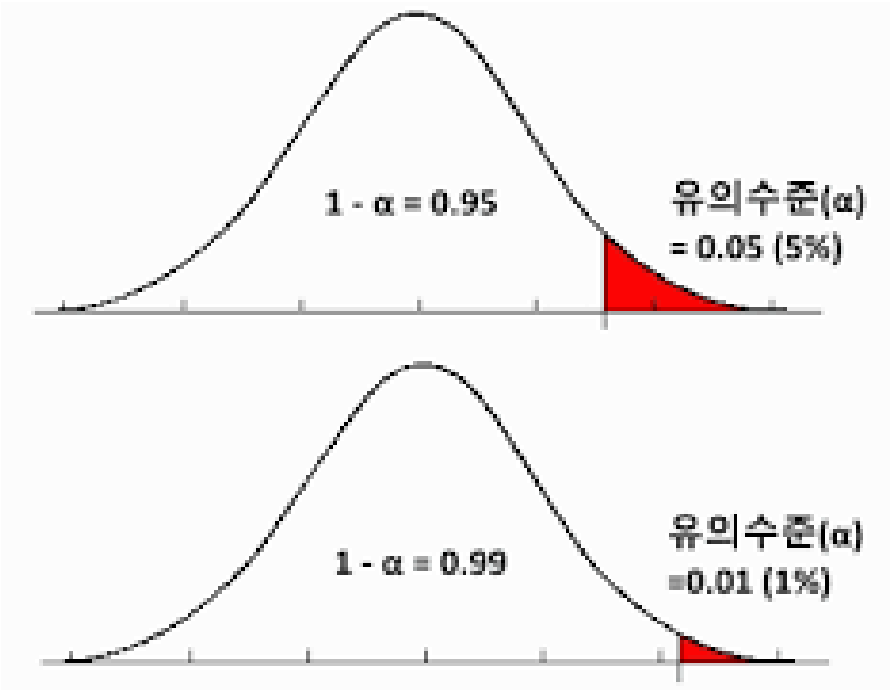


A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

| 03 회귀 모델의 해석: 선형 회귀 모델

유의 수준 (Significance Level)

- 통계적인 가설검정에서 사용되는 기준값
- 귀무가설이 참임에도 귀무가설을 기각할 확률
  => 낮을수록 해당 오차가 낮아짐
- 일반적으로 5%로 설정
- 연구의 특성에 따라서 유의수준은 바뀔 수 있음
  Ex) 사회과학적 문제: 10%, 공학적 문제: 1%, 0.5% 등



| 03 회귀 모델의 해석: 선형 회귀 모델

가설 검정을 통해 통계적 결론 내기

P-value < 유의수준	대립 가설에 대한 통계적으로 유의한 증거가 있다.
P-value >= 유의수준	대립 가설에 대한 통계적으로 유의한 증거가 없다.

| 03 회귀 모델의 해석: 선형 회귀 모델

Multiple Linear Regression

- $w_1, w_2, \ldots, w_n$: weight, b: bias (parameter)
- $x_1, x_2, \ldots, x_n$: 독립 변수
- y: 종속 변수
- n개의 독립 변수 $x_1, x_2, \ldots, x_n$와 종속 변수 y의 선형 관계를 모델링

$$y = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b$$

| 03 회귀 모델의 해석: 선형 회귀 모델

Multiple Linear Regression

$$y = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b$$

$$\frac{\partial y}{\partial x_i} = \frac{\partial (w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b)}{\partial x_i} = w_i$$

각 변수에 대한 계수

(다른 모든 것이 일정할 때) 변수 한 단위 증가에 따른 종속 변수의 변화량

| 03 회귀 모델의 해석: 선형 회귀 모델

Multiple Linear Regression

$$y = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b$$

$$y\Big|_{x_1=\cdots=x_n=0} = b$$

상수항

모든 독립변수의 값이 0일 때의 종속 변수의 값

| 03 회귀 모델의 해석: 선형 회귀 모델

다중 선형 회귀 모델 결과 해석

Dep. Variable
종속변수명

Coefficient
각 변수에 따른 계수

OLS Regression Results

Dep. Variable:	Head size	R-squared:	0.639
Model:	OLS	Adj. R-squared:	0.638
Method:	Least Squares	F-statistic:	416.5
Date:	Sun, 08 May 2022	Prob (F-statistic):	5.96e-54
Time:	21:40:40	Log-Likelihood:	-1613.4
No. Observations:	237	AIC:	3231.
Df Residuals:	235	BIC:	3238.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	520.6101	153.215	3.398	0.001	218.759	822.461
Brain_weight	2.4269	0.119	20.409	0.000	2.193	2.661

Omnibus:	2.687	Durbin-Watson:	1.726
Prob(Omnibus):	0.261	Jarque-Bera (JB):	2.321
Skew:	0.207	Prob(JB):	0.313
Kurtosis:	3.252	Cond. No.	1.38e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.38e+04. This might indicate that there are strong multicollinearity or other numerical problems.

R-Squared
회귀식이 원래의 자료를 얼마나 잘 설명하는지를 나타내는 수치

Adjusted R-Squared
R-Squared에 모델의 변수 증가에 따른 패널티를 준 수치

P>|t|
각 변수에 따른 p-value

OUTTA

| 03 회귀 모델의 해석: 선형 회귀 모델

인과관계 ≠ **상관관계**

<인과관계의 3가지 조건>

- 원인이 결과보다 시간적으로 앞설 것
- 원인과 결과는 서로 관련이 있을 것
- 원인이 변수만으로 설명이 되어야 하고 동시에 다른 변수에 의한 설명은 제거되어야 할 것

04 회귀 모델을 위한 데이터 처리

OUTTA

| 04 회귀 모델을 위한 데이터 처리

평균중심화 (mean-centering)

$$y = wx + b \qquad y\text{: 집값, } x\text{: 평수}$$

상수항: 평수가 0인 집의 가격 (의미가 없음)

$$y = w(x - \bar{x}) + b \qquad y\text{: 집값, } x\text{: 평수, } \bar{x}\text{: 평균 평수}$$

상수항: 평수가 평균인 집의 가격 (의미가 있음)

| 04 회귀 모델을 위한 데이터 처리

평균중심화 (mean-centering)

$$y = w_1x_1 + w_2x_2 + \cdots + w_ix_i + \cdots + w_nx_n + b$$

$$y = \sum_{i=1}^{n} w_ix_i + b$$

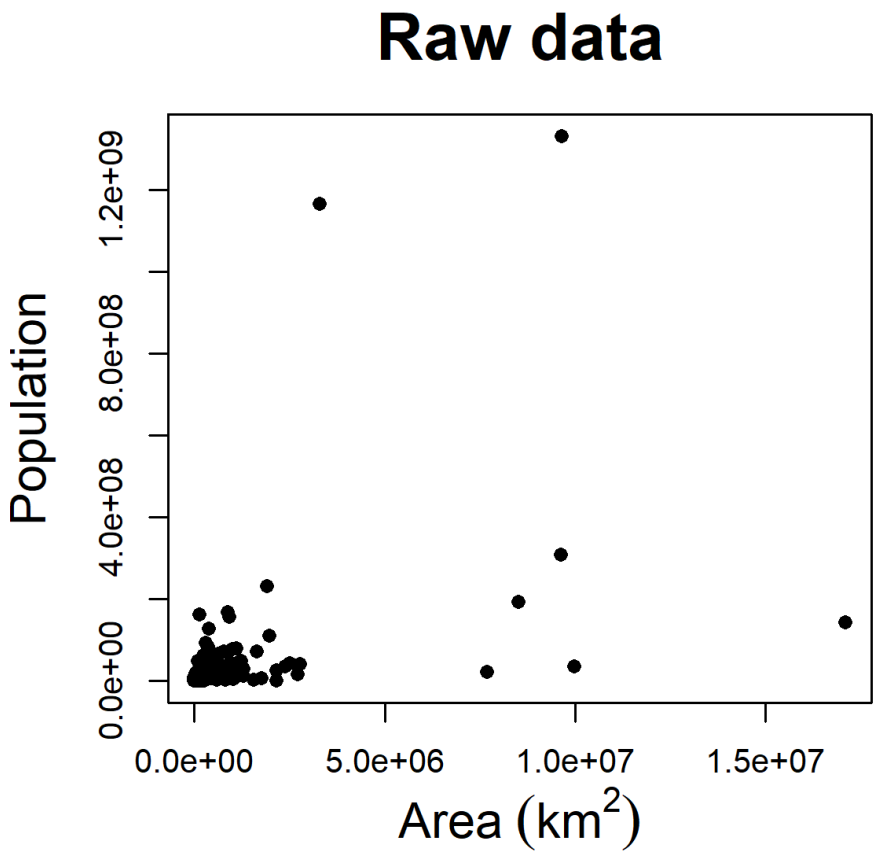
$$y = \sum_{i=1}^{n} w_i(x_i - \bar{x_i}) + b \; where \; \bar{x_i} : mean \; of \; x_i$$

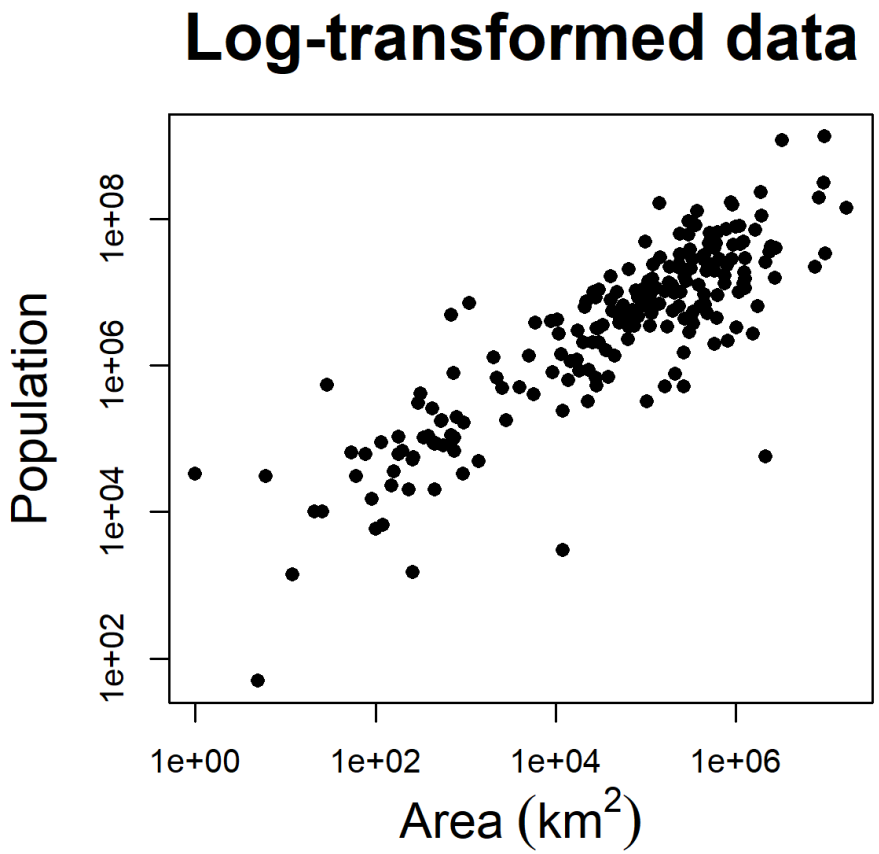
$$\frac{\partial y}{\partial x_i} = w_i$$

OUTTA

| 04 회귀 모델을 위한 데이터 처리

로그 변환 (log transformation)



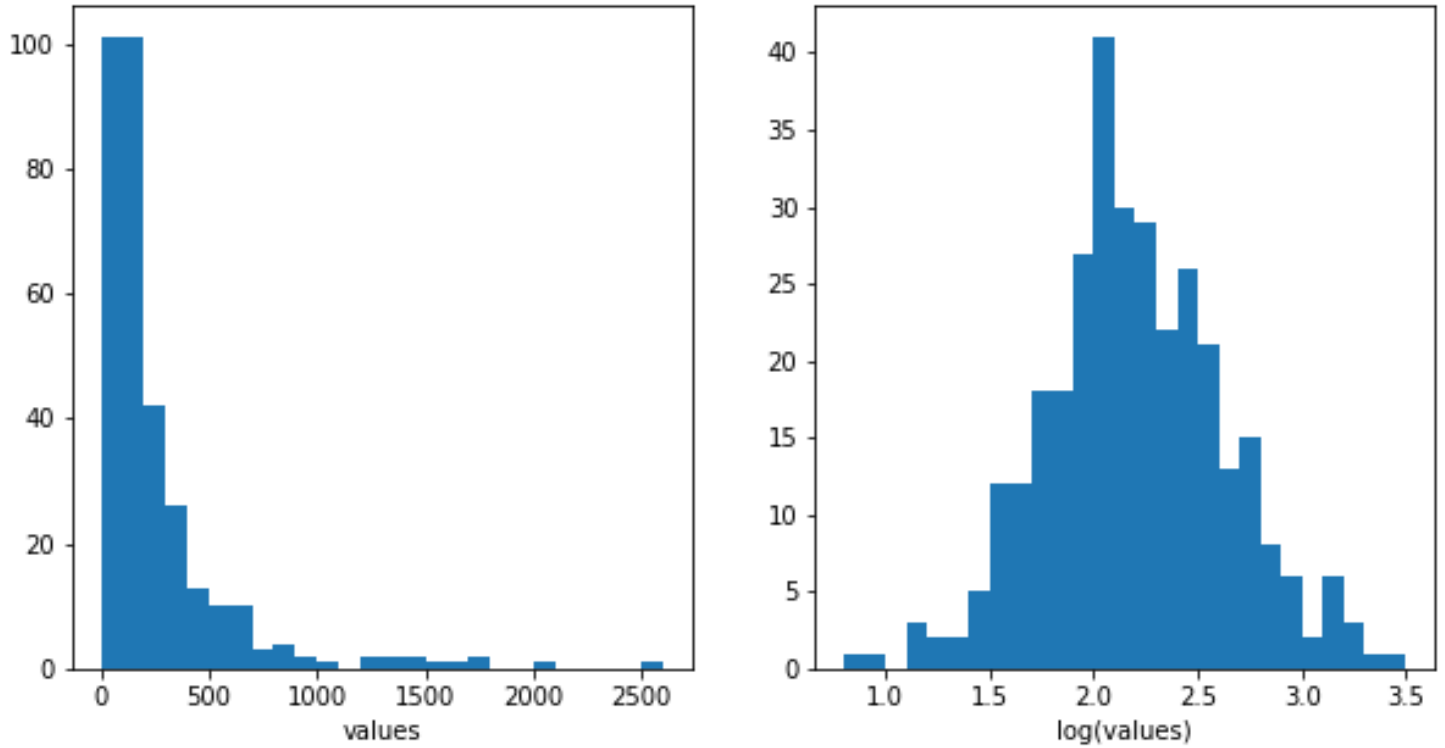


OUTTA

| 04 회귀 모델을 위한 데이터 처리

로그 변환 (log transformation)

- 선형 회귀 모델의 예측력을 높일 수 있음
- 변수가 right-skewed된 경우에 수행하면 좋음



| 04 회귀 모델을 위한 데이터 처리

로그 변환 (log transformation)

$$y = w\log(x) + b \qquad \frac{\partial y}{\partial x} = \frac{w}{x} \qquad w = \frac{\partial y}{\partial x}x$$

각 변수에 대한 계수

- (다른 모든 것이 일정할 때) 변수 한 단위 증가에 따른 종속 변수의 변화량과 그 때의 변수의 값의 곱
- (변수가 양수이므로) 계수가 양수이면 종속 변수와 독립 변수는 양의 상관관계
- 계수가 음수이면 종속 변수와 독립 변수는 음의 상관관계

| 04 회귀 모델을 위한 데이터 처리

로그 변환 (log transformation)

$$y = w\log(x) + b \qquad y\Big|_{x=1} = w\log(1) + b = b$$

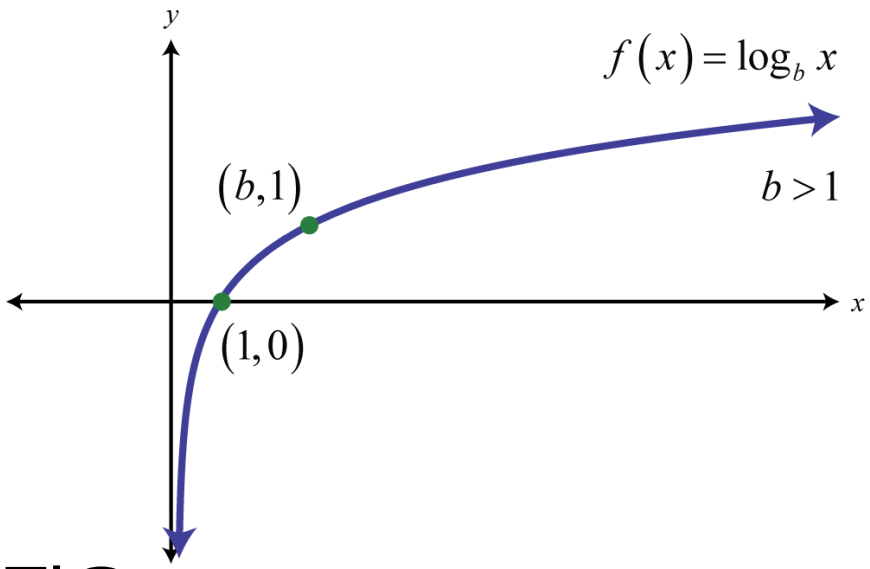
상수항

- 로그 변환된 변수가 1일 때의 종속 변수의 값

| 04 회귀 모델을 위한 데이터 처리

로그 변환 (log transformation)

$$y = w\log(x) + b$$



유의할 점

- Log 0은 정의되지 않음
- 따라서 변수에 0이 포함될 경우 별도의 처리 필요

$$y = w\log(x+1) + b$$

상수항

- 로그 변환된 변수가 0일 때의 종속 변수의 값

| 04 회귀 모델을 위한 데이터 처리

범주형 변수의 경우

순서형 변수	Label Encoding
명목형 변수	One-Hot Encoding

| 04 회귀 모델을 위한 데이터 처리

다중공선성

한 변수의 값이 다른 변수에 의해 온전히 결정될 때 발생

고객번호	국가명	이탈 여부
EA1243	한국	1
EA1244	중국	0
EA1245	한국	1
EA1246	일본	0
EA1247	일본	1

고객번호	한국	중국	일본	이탈 여부
EA1243	1	0	0	1
EA1244	0	1	0	0
EA1245	1	0	0	1
EA1246	0	0	1	0
EA1247	0	0	1	1

| 04 회귀 모델을 위한 데이터 처리

다중공선성

한 변수의 값이 다른 변수에 의해 온전히 결정될 때 발생
=> Dummy Variable을 범주의 개수보다 한 개 적게 설정하여 해결

고객번호	국가명	이탈 여부
EA1243	한국	1
EA1244	중국	0
EA1245	한국	1
EA1246	일본	0
EA1247	일본	1

고객번호	한국	중국	이탈 여부
EA1243	1	0	1
EA1244	0	1	0
EA1245	1	0	1
EA1246	0	0	0
EA1247	0	0	1

| 04 회귀 모델을 위한 데이터 처리

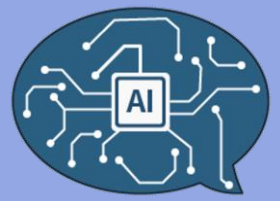
차원의 저주

Dummy Variable이 너무 많아질 때 발생 가능

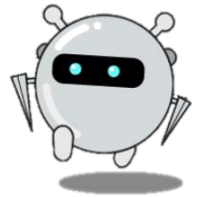
=> Feature Selection, Feature Engineering 통해 해결 가능

고객번호	국가명	이탈 여부
EA1243	한국	1
EA1244	중국	0
EA1245	미국	1
EA1246	일본	0
EA1247	영국	1

고객번호	동서양	이탈 여부
EA1243	동양권	1
EA1244	동양권	0
EA1245	서양권	1
EA1246	동양권	0
EA1247	서양권	1



감사합니다



OUTTA