Feature Engineering

17기 DA팀 김지훈

Content

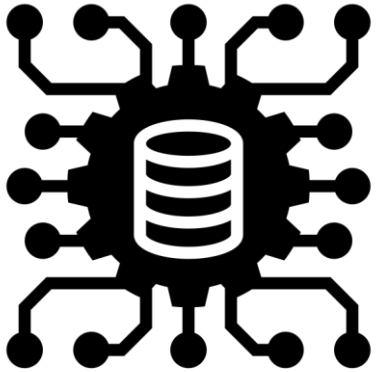
- ◆정의
- ◆Data structure

- ◆Missing Value
- ◆Outlier

- ◆Re-scaling
- ◆Non-numerical data

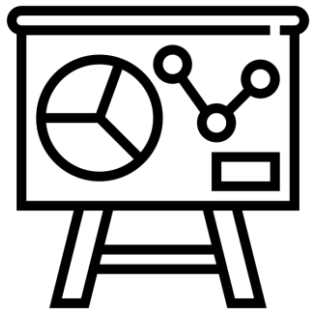
- ◆실습

Feature Engineering?



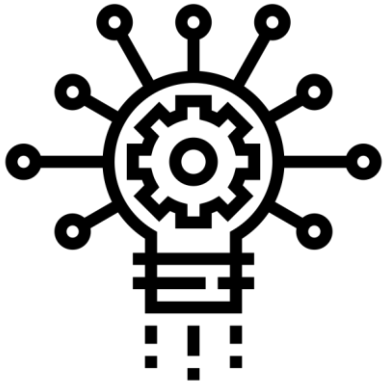
Created by Becris
from Noun Project

Data collecting



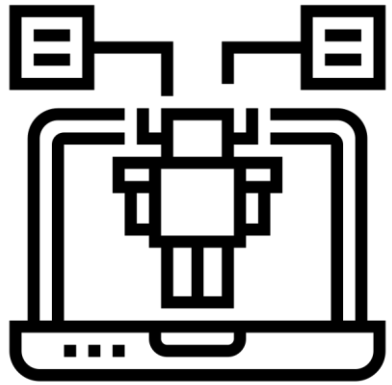
Created by Eucalyp
from Noun Project

EDA



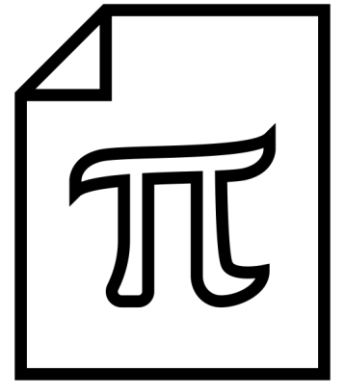
Created by Eucalyp
from Noun Project

Feature Engineering



Created by Eucalyp
from Noun Project

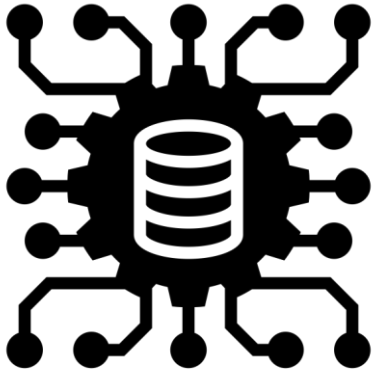
Modeling



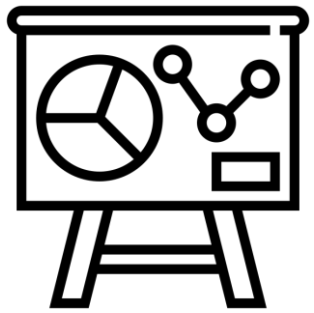
Created by Eucalyp
from Noun Project

Evaluation

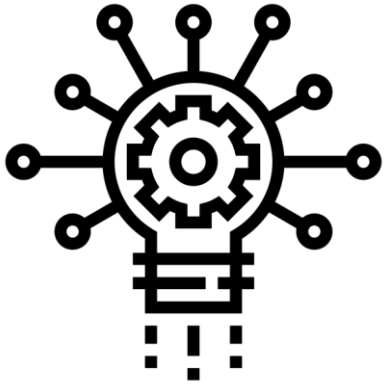
Feature Engineering?



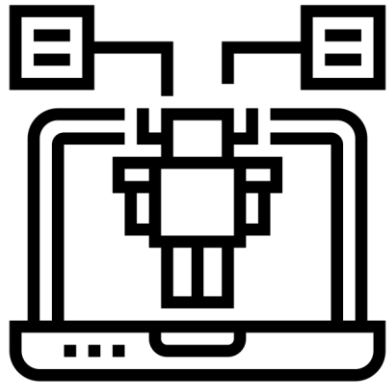
Created by Becris
from Noun Project



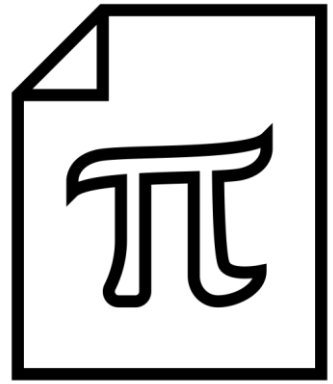
Created by Eucalyp
from Noun Project



Created by Eucalyp
from Noun Project



Created by Eucalyp
from Noun Project



Created by Eucalyp
from Noun Project

Data collecting

EDA **Feature Engineering**

Data Preprocessing

Modeling

Evaluation

Data Structure

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Target Variable

Data Structure

구분			예
수치적 자료	연속형 자료	비율 척도	키, 몸무게, 나이 등
		구간 척도	온도, 지수 등
	이산형 자료		주사위 결과, 사고 건수 등
범주형 자료	순위형 자료(순서척도)		평점, 선호도 등
	명목형 자료(명목척도)		혈액형, 성별 등

Data Structure

비율척도	간격에 대한 비율이 의미를 갖는 자료, 절대적 기준인 0이 존재하며 사칙연산이 가능하고, 가장 많은 정보를 갖고있는 척도 ex) 무게, 나이	양적 척도
구간척도	측정 대상이 갖고 있는 속성의 양을 측정하는 것으로 구간이나 구간 사이의 간격이 의미가 있는 자료, 곱셈과 나누기 불가 ex) 온도	
순서척도	측정 대상의 서열관계를 관측하는 척도 ex) 만족도, 학년, 신용등급	질적 척도
명목척도	측정 대상이 어느 집단에 속하는지 분류할 때 사용 ex) 성별, 출생지 구분	

Missing Value

Missing Value NA, NAN, 99999999, (공백), Unknown 등 다양한 형태로 표현	Missing Completely at Random (MCAR)	어떠한 변수의 결측치가 무작위로 발생한 경우	다른 변수와의 관계 X
	Missing at Random (MAR)	어떠한 변수의 결측 여부가 다른 변수와 관련이 있는 경우	Cabin의 결측치가 ticket class와 관련 있는 경우
	Missing not at Random (MNAR)	어떠한 변수의 결측 여부가 그 변수와 관련이 있는 경우	학업 성적이 좋지 않은 학생이 응답하지 않은 경우

Missing Value

결측치 처리

(1) Deletion

(2) Heuristic Imputation

(3) Mean/ Mode/ Median Imputation

(4) Prediction model

(5) KNN Imputation

Missing Value

결측치 처리

(1) Deletion

결측치가 있는 row(or column) 삭제

⇒ 완벽한 변수에 대해서만 분석을 진행

⇒ 결측치가 포함 된 row(or column)이 많을 경우 손실이 큼

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

Missing Value

결측치 처리

(1) Deletion

결측치가 있는 row(or column) 삭제

⇒ 완벽한 변수에 대해서만 분석을 진행

⇒ 결측치가 포함 된 row(or column)이 많을 경우 손실이 큼

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

Missing Value

결측치 처리

(1) Deletion

결측치가 있는 row(or column) 삭제

⇒ 완벽한 변수에 대해서만 분석을 진행

⇒ 결측치가 포함 된 row(or column)이 많을 경우 손실이 큼

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

Missing Value

결측치 처리

(2) Heuristic Imputation

상식이나 보편적인 사실로 채워 넣을 수 있는 경우

	Name	Sex
	Braund, Mr. Owen Harris	male
	Cumings, Mrs. John Bradley (Florence Briggs Th...	female
	Heikkinen, Miss. Laina	
	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female
	Allen, Mr. William Henry	

Missing Value

결측치 처리

(2) Heuristic Imputation

상식이나 보편적인 사실로 채워 넣을 수 있는 경우

	Name	Sex
	Braund, Mr. Owen Harris	male
	Cumings, Mrs. John Bradley (Florence Briggs Th...	female
	Heikkinen, Miss. Laina	female
	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female
	Allen, Mr. William Henry	male

Missing Value

결측치 처리

(3) Mean/ Mode/ Median Imputation

25	24	30	NaN	15	43	28	37	31	28

Missing Value

결측치 처리

(3) Mean/ Mode/ Median Imputation

25	24	30	NaN	15	43	28	37	31	28

Generalized Imputation : 모든 사람의 평균 29

Case Imputation : 특정 집단의 평균

Missing Value

결측치 처리

(4) Prediction model

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

Train set

Missing Value

결측치 처리

(5) KNN Imputation

K-Nearest Neighbor : 결측치 근처의 가장 가까운 k개의 값을 통해 분류 예측

⇒ K값?

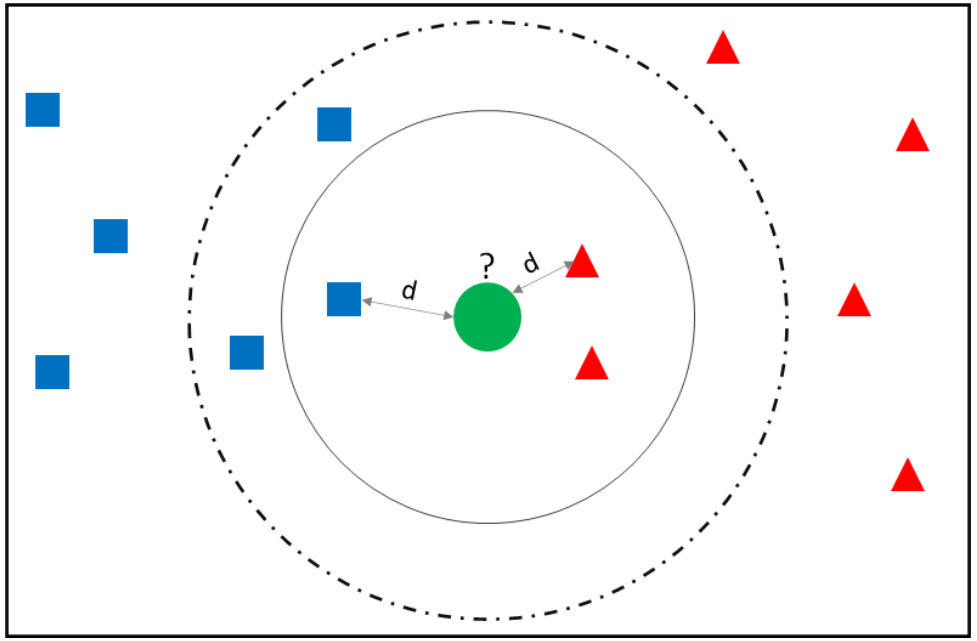
-Idea1:choosing hyperparameters that work best on the (training)data(:will always be K=1, Bad)

-Idea2:train, test split-> choose the best on test data(:no idea on performance of new data, Bad)

-Idea3:split to train, val, test -> choose hyperparameter on val and evaluate on test => better

-Idea4:Cross-validation split data into folds, and try each fold as val and average the results->expensive(for small datasets/ not usually for deep learning)

⇒ expensive



Outlier

Outlier?

해당하는 변수의 다른 값들과 비교해 지나치게 크거나 작은 값.

Linear regression, Adaboost 등의 모델은 outlier의 영향이 크지만, 트리기반의 모델은 큰 영향이 없음

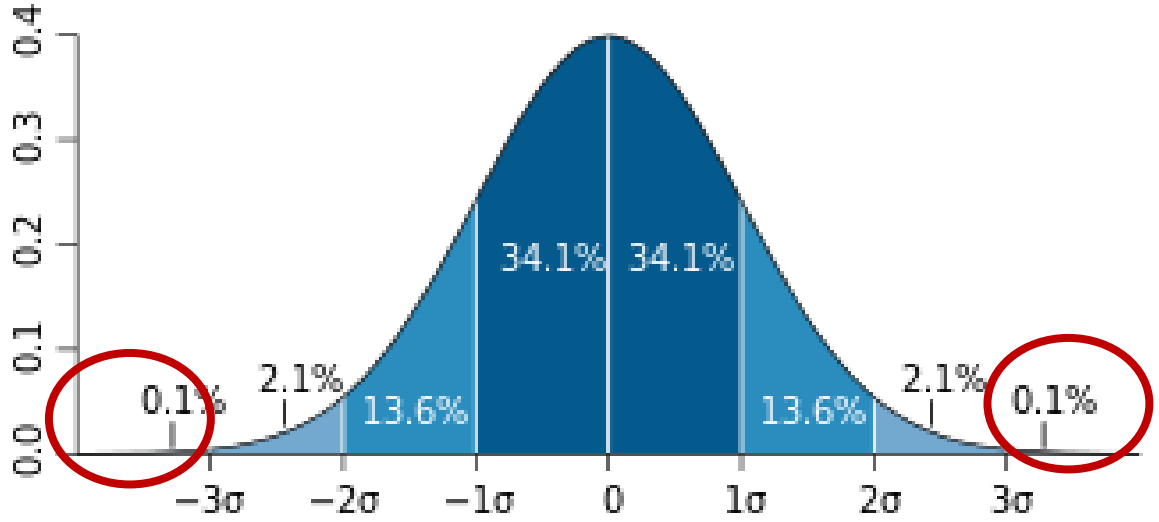
Outlier

이상값 인식

(1) ESD(Extreme Studentized Deviation) : 평균으로부터 3 표준편차 이상 떨어진 값(각 0.15%)

(2) 기하평균 – 2.5*표준편차 < DATA < 기하평균 + 2.5*표준편차, 외의 data

(3) 사분위수에서 상자 그림의 outer fence 밖의 값(Q1-1.5*(Q3-Q1) < DATA < Q3+1.5*(Q3-Q1) 밖의 값)



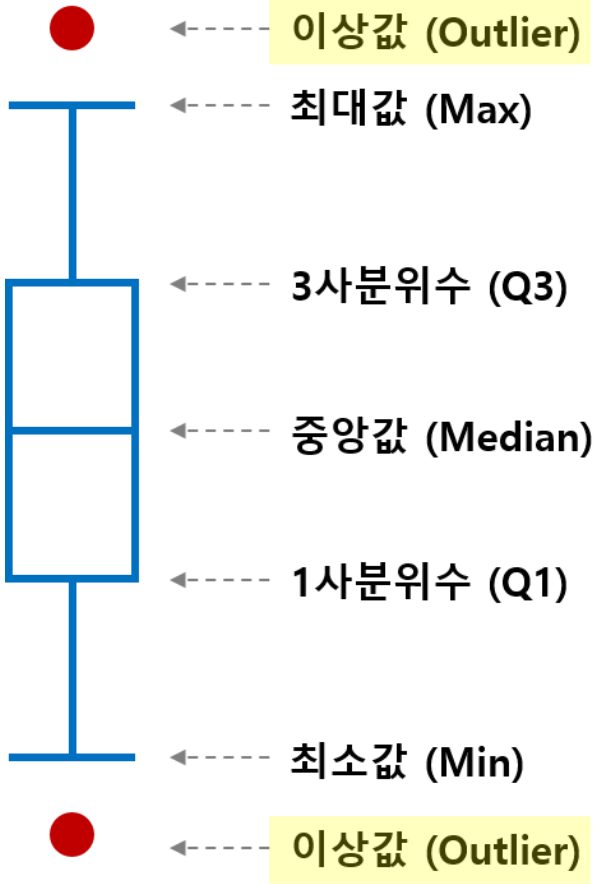
Outlier

이상값 인식

(1) ESD(Extreme Studentized Deviation) : 평균으로부터 3 표준편차 이상 떨어진 값(각 0.15%)

(2) 기하평균 – 2.5*표준편차 < DATA < 기하평균 + 2.5*표준편차, 외의 data

(3) 사분위수에서 상자 그림의 outer fence 밖의 값(Q1-1.5*(Q3-Q1) < DATA < Q3+1.5*(Q3-Q1) 밖의 값)



Outlier

이상값 처리

(1) Deletion

(2) Capping

(3) Change to other value

(4) Transformation

(5) Treating separately

Outlier

이상값 처리

(1) Deletion (극단값 제거)

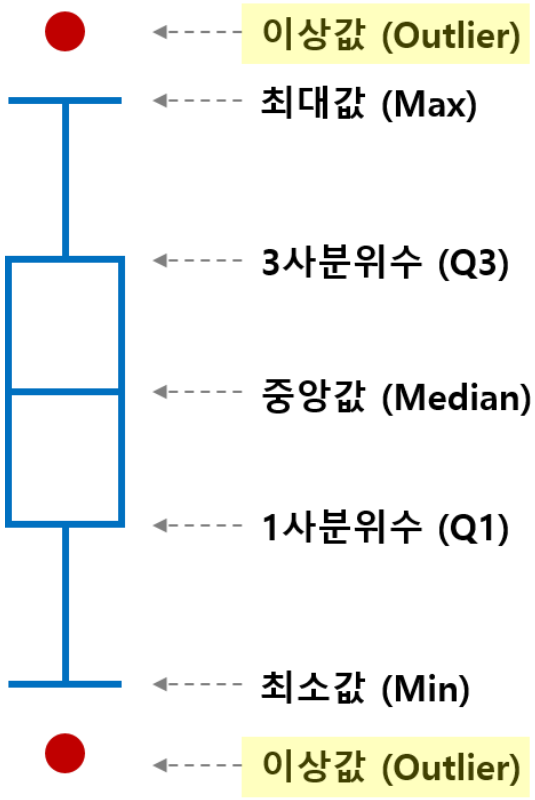
- Manually 제거

- 상단,하단 5%씩 제거

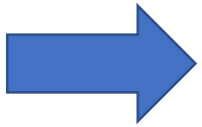
Outlier

이상값 처리

(2) Capping



가격 (₩)
5,000
4,000
4,000
3,700
6,900
7,200
168,200



가격 (₩)
5,000
4,000
4,000
3,700
6,900
7,200
10,000

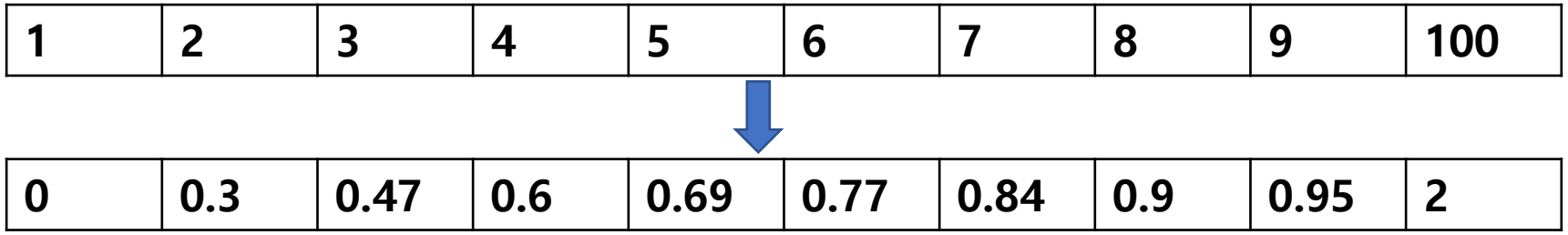
Outlier

이상값 처리

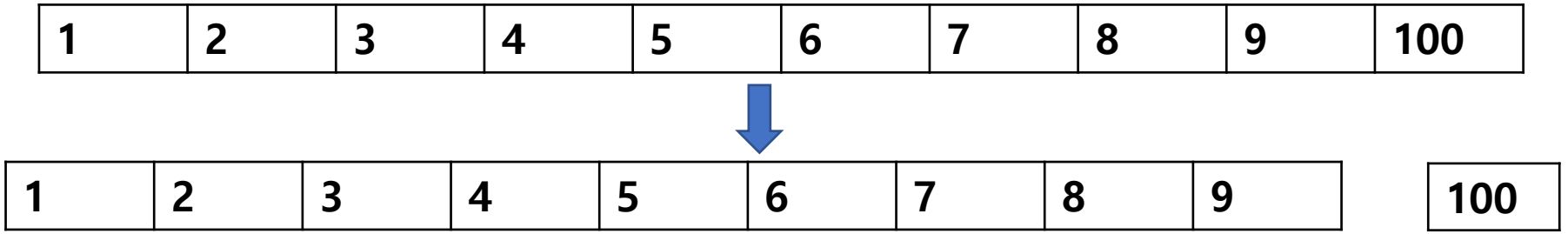
(3) Change to other value

'NA', Mean, Median, Mode 등

(4) Transformation



(5) Treating separately



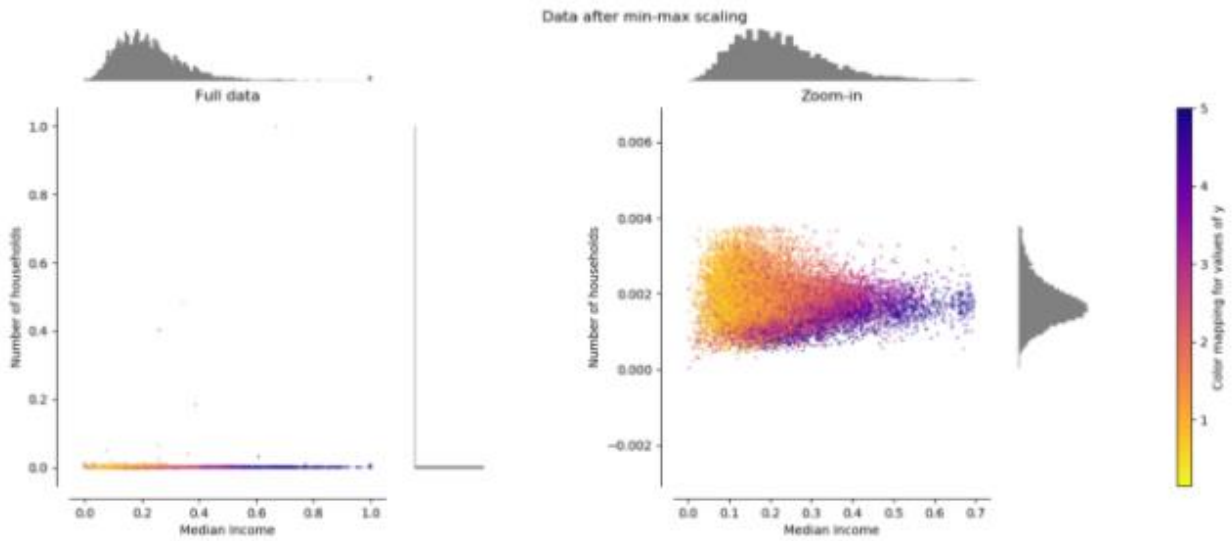
Re-scaling

MinMaxScaler (normalizing)

모델 학습 전에 normalizing을 해주는 이유는 서로 다른 Scale로 측정된 변수들이 모델에 학습 될 때, 스케일의 차이 때문에 bias가 발생 할 수 있기 때문!

⇒ Tree based model은 스케일에 크게 영향을 받지 않지만 SVM, LDA 같은 모델은 영향을 크게 받는다.

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)}$$



Re-scaling

normalizing

단순히 최대값으로 전체를 나눠줌

(0~255) => /255 => (0~1)

단순하지만 효과적인 방법

Re-scaling

StandardScaler (Standardization)

평균을 0, 분산을 1로 변경

⇒ 모든 특성이 같은 크기를 갖게 됨

⇒ 최대, 최소값이 제한되지 않음

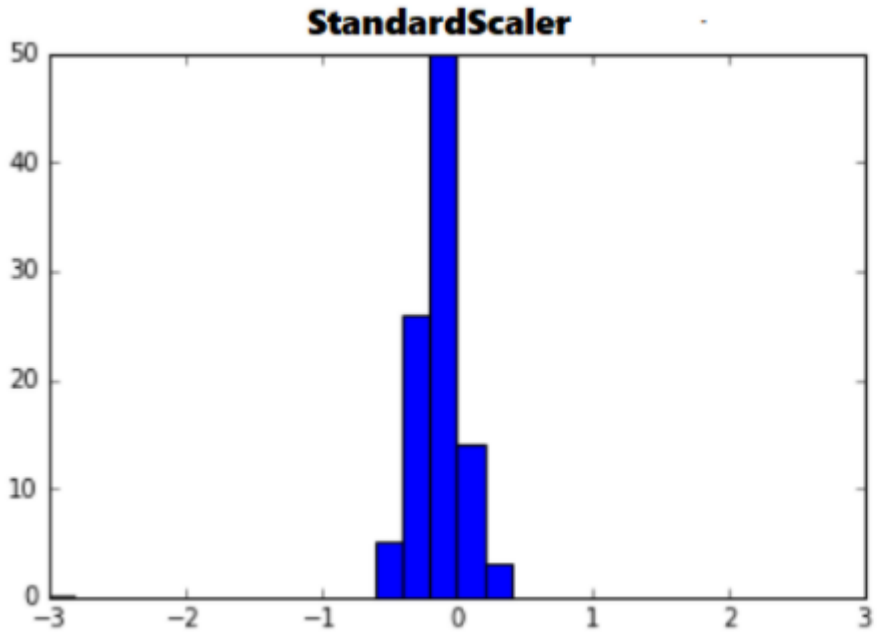
$$z = \frac{x - \bar{x}}{s}$$

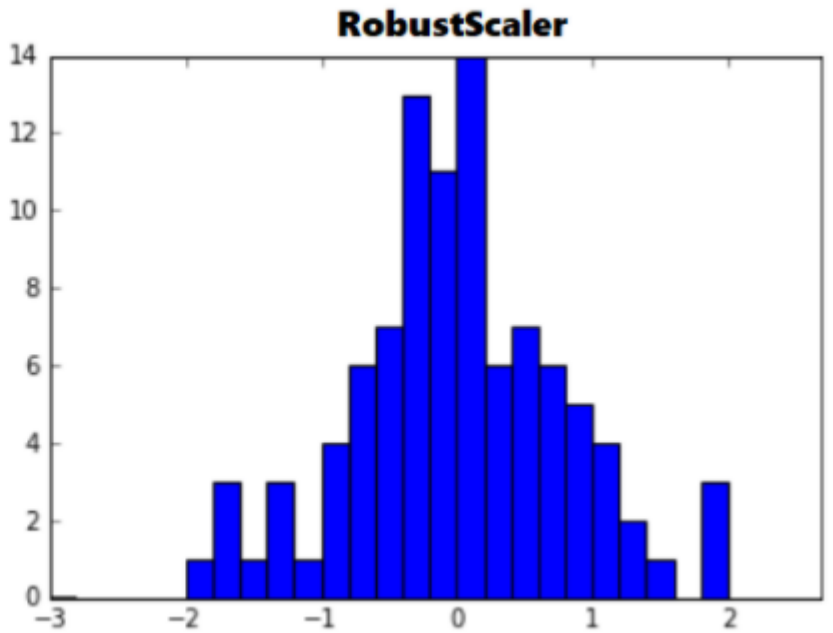
Re-scaling

RobustScaler

특성들이 같은 스케일을 갖게 되지만 평균대신 중앙값을 사용

=> 극단값에 영향을 적게 받음



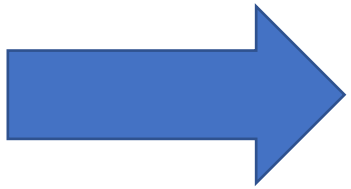


Non-numerical data

Variable creation

-

계약년월	계약일
201712	8
201712	22
201712	28
201712	13
201712	16
...	...
202003	11
202008	7
202007	10
202012	3
202009	28



transaction_year	transaction_month
2008	01
2008	01
2008	01
2008	01
2008	01
...	...
2017	11
2017	11
2017	11
2017	11
2017	11

Non-numerical data

One-hot encoding (Dummification)

id	color
1	red
2	blue
3	green
4	blue

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

Non-numerical data

Other Feature extractions

Age
23
42
NA
12
57
NA
49

Age	Age_isnull
23	0
42	0
NA	1
12	0
57	0
NA	1
49	0

감사합니다