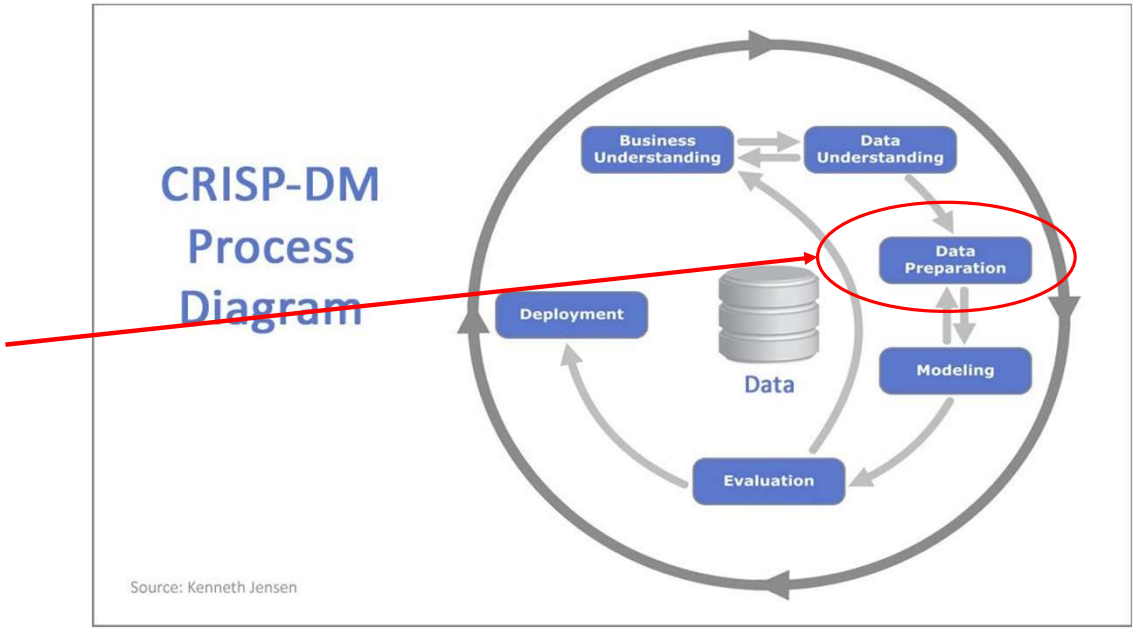
Feature Engineering

Outline

- Feature Engineering
- Introduction
- Scaling
- Encoding
- Missing Value
- Outlier
- Binning
- Generating new features
- Feature Selection

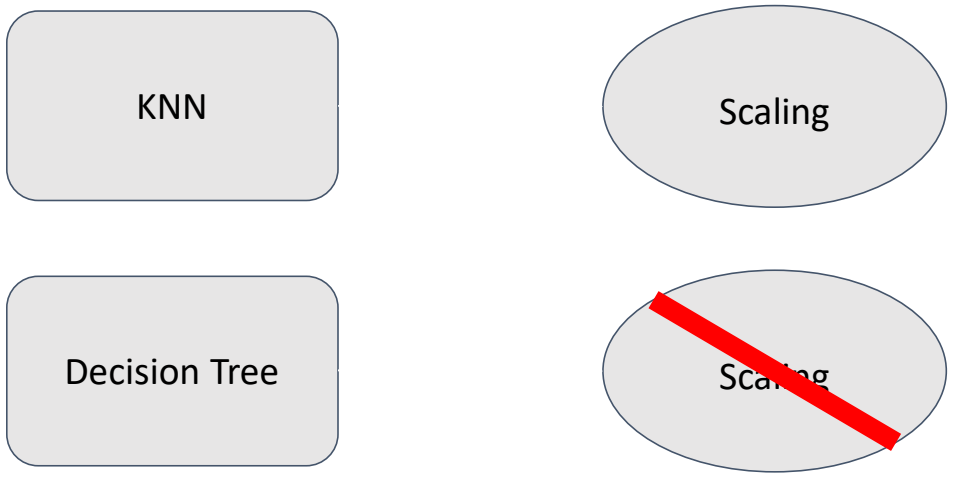


Feature Engineering

Your machine learning only as good as your data
With Feature Engineering, you can provide better input
In real word practice data is not clean:

- missing value
- outlier
- unreliable and invalid data
- covariate/lurking variable, etc

Each Model Optimize Differently



- For each model, some variable work best when we give certain treatment

Why Does It Matter ?

Scaling

What is Scaling ?

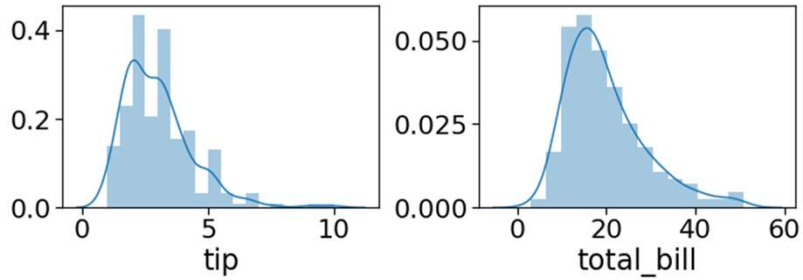
Transform numerical data into same range (typically small)

Scaling:

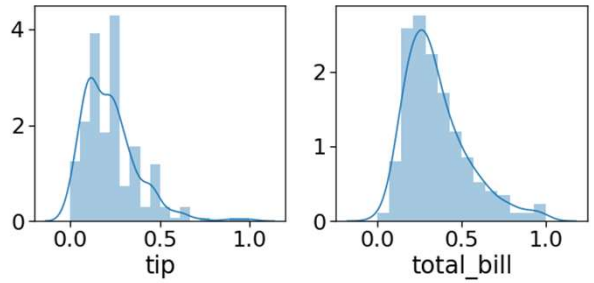
- MinMax Scaler
- Standard Scaler
- Robust Scaler

Some method may work best with scaling

- ex. KNN, Neural Network, Linear Model

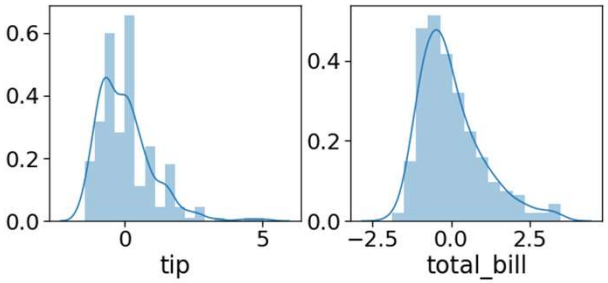


Default Distribution



Transform To Range
0 - 1

$$y = \frac{x - \min x_i}{\max x_i - \min x_i}$$



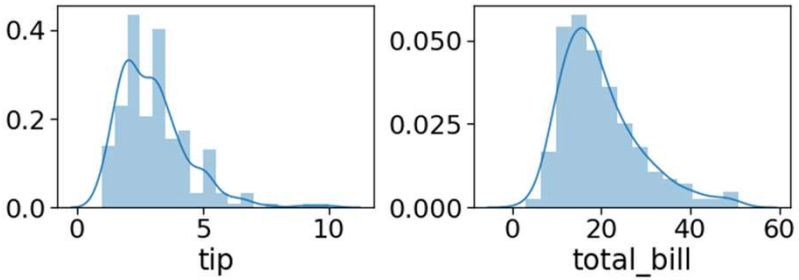
Transform To mean =
0 and sd =1

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

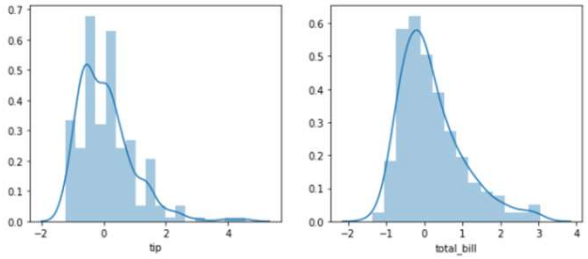
Where

$\bar{x}$ = mean

s = Standard deviation



Default Distribution



Transform To Small Range

$$z_i = \frac{x_i - Q_1(x_i)}{Q_3(x_i) - Q_1(x_i)}$$

Where:

$Q_1(x_i)$ = first quartile

$Q_3(x_i)$ = third quartile

Encoding

What is Encoding ?

- Encoding is used as our way to represent categorical variable in Machine Learning
- In Python , there are available some method such as one hot encoding, ordinal encoding dan binary encoding.
- Which method to use depend on the variable's scale of measurement

Scale of Measurement	Suggested Method		
	One Hot Encoding	Ordinal Encoding	Binary Encoding
Nominal	v	x	v
Ordinal	v	v	x

One Hot Encoding

Gender
Male
Female
Female
Male
Female

Male	Female
1	0
0	1
0	1
1	0
0	1

City
Jakarta
Bogor
Bogor
Bekasi
Bekasi

Jakarta	Bogor	Bekasi
0	1	0
1	0	0
1	0	0
0	0	1
0	0	1

Work best for nominal variable and can used for ordinal variable as well

One Hot Encoding For Linear Model

Gender
Male
Female
Female
Male
Female

Male
1
0
0
1
0

City
Jakarta
Bogor
Bogor
Bekasi
Bekasi

Jakarta	Bogor
0	1
1	0
1	0
0	0
0	0

- Only need k-1 variable from k category
- k variable will cause multicollinearity

Ordinal Encoding

Education
SD
SMP
SD
SMA
S1
S1

Education Encode
1
2
1
3
4
4

Value	Mapping
Other/None	0
SD	1
SMP	2
SMA	3
S1	4
Post-Grad	5

- Work best for ordinal variable
- can mislead if you use this method for nominal variable

Binary Encoding

CAR
Avanza
Xenia
Xenia
CR-V
Avanza
Calya
City
Calya
Jazz

Order
1
2
2
3
1
4
5
4
6

Binary Num
001
010
010
011
001
100
101
100
110

C1	C2	C3
0	0	1
0	1	0
0	1	0
0	1	1
0	0	1
1	0	0
1	0	1
1	0	0
1	1	0

Work best for nominal categorical variable that has too many categories

Binary Number

Number	Binary Number	Binary Number(alt.)
1	1	0001
2	10	0010
3	11	0011
4	100	0100
5	101	0101
6	110	0110
7	111	0111
8	1000	1000
9	1001	1001

Follow the largest digit

EXAMPLE :

3 :
3 = 11
3 = 2**1 x (1) + 2**0 x (1)

5 :
5 = 101
5 = 2**2 x (1) + 2**1 x (0) + 2**0 x (1)

6 :
6 = 110
6 = 2**2 x (1) + 2**1 x (1) + 2**0 x (0)

10 :
10 = 1010
10 = 2**3 x (1) + 2**2 x (0) + 2**1 x (1) + 2**0 x (0)

100 ?

.fit and .transform Method in preprocessing

Method	training set	test set or validation set
.fit	V	X
.transform	V	V

```
scaler = MinMaxScaler()
scaler.fit(X_train)
X_train_scaled = scaler.transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

- .fit method only applied to training set to avoid many problem such as information leakage (overly optimistic score in test set or validation set)
- some method simply require it. For example, binary encoding and tf-idf.

Apply Several Preprocessing Method to Modeling at once

Part 1a : Ridge

data : tips

target : tip

preprocess:

1. one hot encoding : sex, smoker, time
2. binary encoding : day
3. robust scaler : total_bill
4. no treatment : size

Random state 10, data splitting 70:30 model Ridge default

Apply Several Preprocessing Method to Modeling at once

Part 1b : Tree

data : tips

target : tip

preprocess v1:

1. one hot encoding : sex, smoker, time
2. ordinal encoding : day
3. no treatment : size, total_bill

Random state 10, data splitting 70:30 model

Tree(max depth 3)

data : tips

target : tip

preprocess v2:

1. one hot encoding : sex, smoker
2. ordinal encoding : time, day
3. no treatment : size, total_bill

Random state 10, data splitting 70:30 model

Tree(max depth 3)

Missing Value

What is Missing Value ?

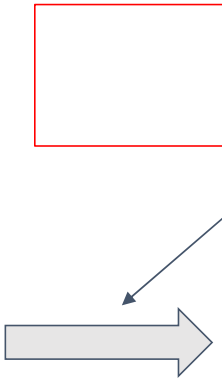
Gender	City	Income(IDR)
Male	Jakarta	-1
Female	Bogor	5,000,000
NaN	Unknown	2,500,000
Male	Bekasi	7,000,000
Female	Bekasi	12,000,000

Missing Value

Another value that might represent missing value :
“?”, 999999, “miss”, etc

Missing Value

	x1	x2	x3	x4	x5	x6
0	4.0	3.0	10	A	X	M
1	5.0	5.0	11	A	Y	M
2	NaN	6.0	12	C	X	NaN
3	6.0	5.0	9	C	X	M
4	7.0	NaN	8	D	NaN	N
5	9.0	5.0	11	NaN	Y	NaN



	x1	x2	x3	x4	x5	x6
0	4.0	3.0	10	A	X	M
1	5.0	5.0	11	A	Y	M
3	6.0	5.0	9	C	X	M

Simple Technique:

- Drop Column
- Drop Row
- Substitution with mean, median or mode.

Advance Technique for Handling Missing Data

- Regression imputation
- Last observation carried forward (Time Series Data)
- Maximum Likelihood
- Expectation-Maximization (Regression imputation done iteratively until stable)
- Multivariate Feature Imputation

Another Ways Of Handling Missing Value

- Track back where is the data coming from and find the real value
- Just let it be missing. Some method is able to automatically handle missing value

Missing Value Imputation in Python

Pandas:

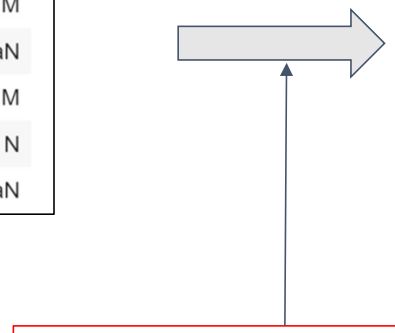
- fillna

Scikit-Learn:

- Mean
- Median
- Mode or new constant
- Multivariate feature imputation (equivalent to Expectation-Maximization)
- KNN-Imputer

Simple Imputer : Mean or Median

	x1	x2	x3	x4	x5	x6
0	4.0	3.0	10	A	X	M
1	5.0	5.0	11	A	Y	M
2	NaN	6.0	12	C	X	NaN
3	6.0	5.0	9	C	X	M
4	7.0	NaN	8	D	NaN	N
5	9.0	5.0	11	NaN	Y	NaN



	x1	x2	x3	x4	x5	x6
0	4.0	3.0	10.0	A	X	M
1	5.0	5.0	11.0	A	Y	M
2	6.2	6.0	12.0	C	X	NaN
3	6.0	5.0	9.0	C	X	M
4	7.0	4.8	8.0	D	NaN	N
5	9.0	5.0	11.0	NaN	Y	NaN

Mean Imputation

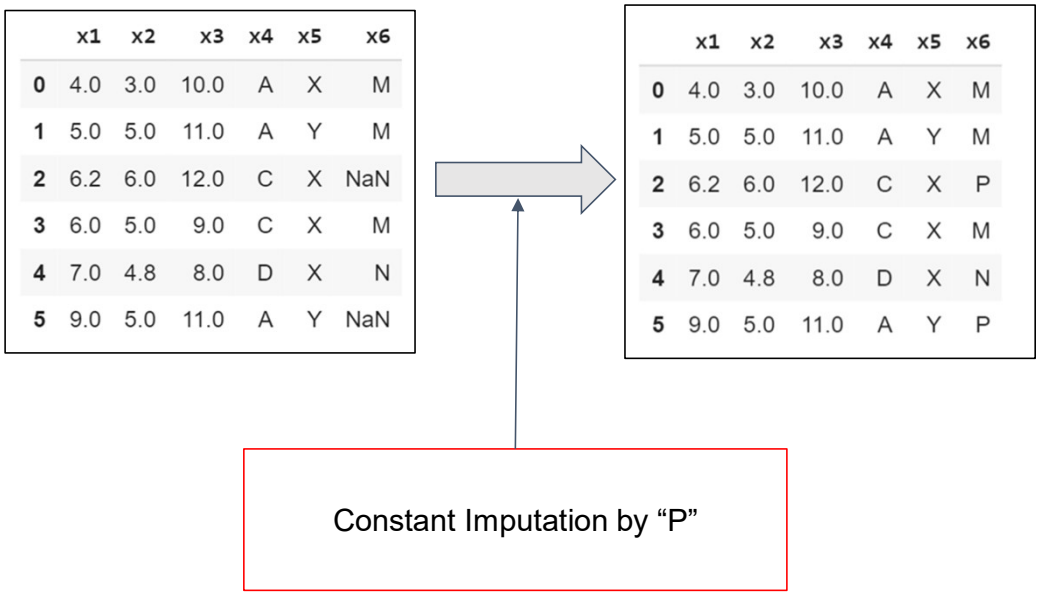
Simple Imputer : Mode

	x1	x2	x3	x4	x5	x6
0	4.0	3.0	10.0	A	X	M
1	5.0	5.0	11.0	A	Y	M
2	6.2	6.0	12.0	C	X	NaN
3	6.0	5.0	9.0	C	X	M
4	7.0	4.8	8.0	D	NaN	N
5	9.0	5.0	11.0	NaN	Y	NaN

	x1	x2	x3	x4	x5	x6
0	4.0	3.0	10.0	A	X	M
1	5.0	5.0	11.0	A	Y	M
2	6.2	6.0	12.0	C	X	NaN
3	6.0	5.0	9.0	C	X	M
4	7.0	4.8	8.0	D	X	N
5	9.0	5.0	11.0	A	Y	NaN

Mode Imputation “A” and “X”

Simple Imputer : Constant



	x1	x2	x3	x4	x5	x6
0	4.0	3.0	10.0	A	X	M
1	5.0	5.0	11.0	A	Y	M
2	6.2	6.0	12.0	C	X	NaN
3	6.0	5.0	9.0	C	X	M
4	7.0	4.8	8.0	D	X	N
5	9.0	5.0	11.0	A	Y	NaN

	x1	x2	x3	x4	x5	x6
0	4.0	3.0	10.0	A	X	M
1	5.0	5.0	11.0	A	Y	M
2	6.2	6.0	12.0	C	X	P
3	6.0	5.0	9.0	C	X	M
4	7.0	4.8	8.0	D	X	N
5	9.0	5.0	11.0	A	Y	P

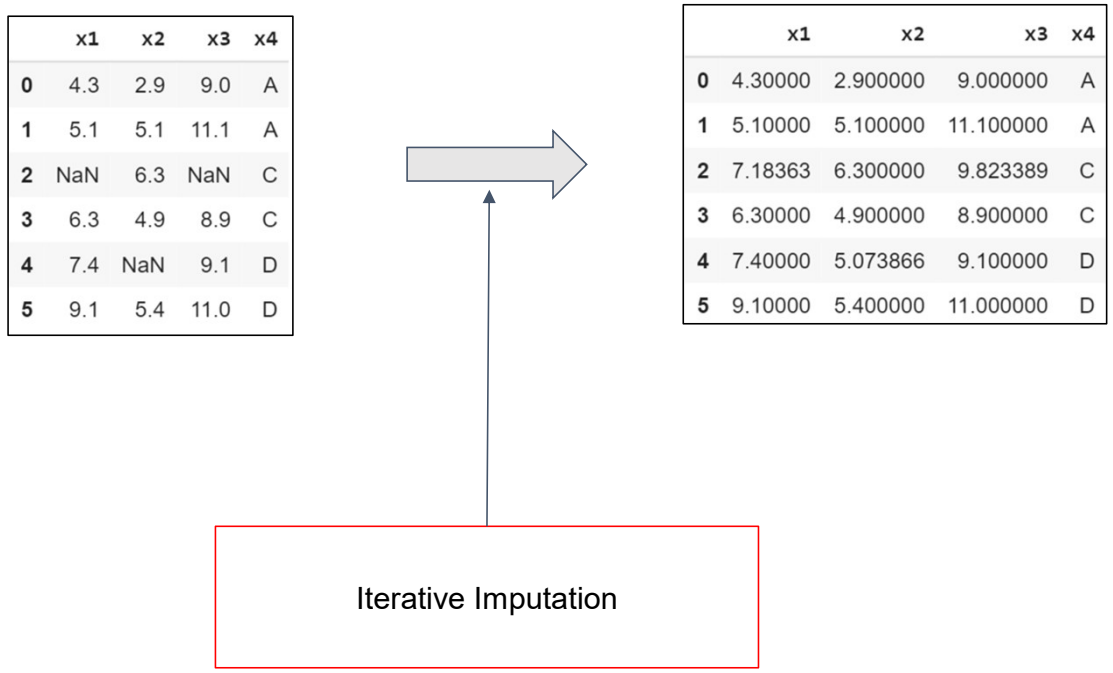
Iterative Imputer

work for multiple variable at once

In Sklearn, work only for numerical

How does it work:

- Predict missing value using regression
- Update the predicted missing value using regression until certain changes in from previous iteration



	x1	x2	x3	x4
0	4.3	2.9	9.0	A
1	5.1	5.1	11.1	A
2	NaN	6.3	NaN	C
3	6.3	4.9	8.9	C
4	7.4	NaN	9.1	D
5	9.1	5.4	11.0	D

	x1	x2	x3	x4
0	4.30000	2.900000	9.000000	A
1	5.10000	5.100000	11.100000	A
2	7.18363	6.300000	9.823389	C
3	6.30000	4.900000	8.900000	C
4	7.40000	5.073866	9.100000	D
5	9.10000	5.400000	11.000000	D

KNN Imputer

In Sklearn, work only for numerical

work for multivariable at once

How does it work:

- Predict missing value using KNN algorithm

	x1	x2	x3	x4
0	4.3	2.9	9.0	A
1	5.1	5.1	11.1	A
2	NaN	6.3	NaN	C
3	6.3	4.9	8.9	C
4	7.4	NaN	9.1	D
5	9.1	5.4	11.0	D

	x1	x2	x3	x4
0	4.3	2.90	9.00	A
1	5.1	5.10	11.10	A
2	7.1	6.30	11.05	C
3	6.3	4.90	8.90	C
4	7.4	5.15	9.10	D
5	9.1	5.40	11.00	D

KNN Imputation

Outlier

Outlier

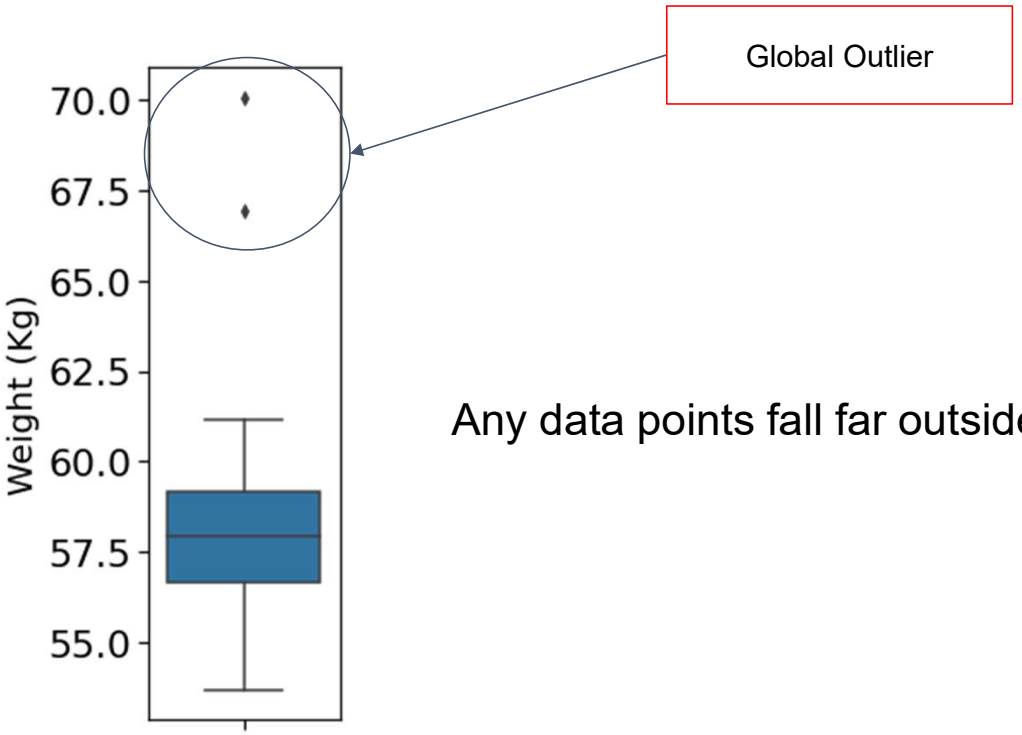
Outlier is an observation point that is distant from other observations

An outlier may indicate an experimental error, or it may be due to variability in the measurement

Outlier type:

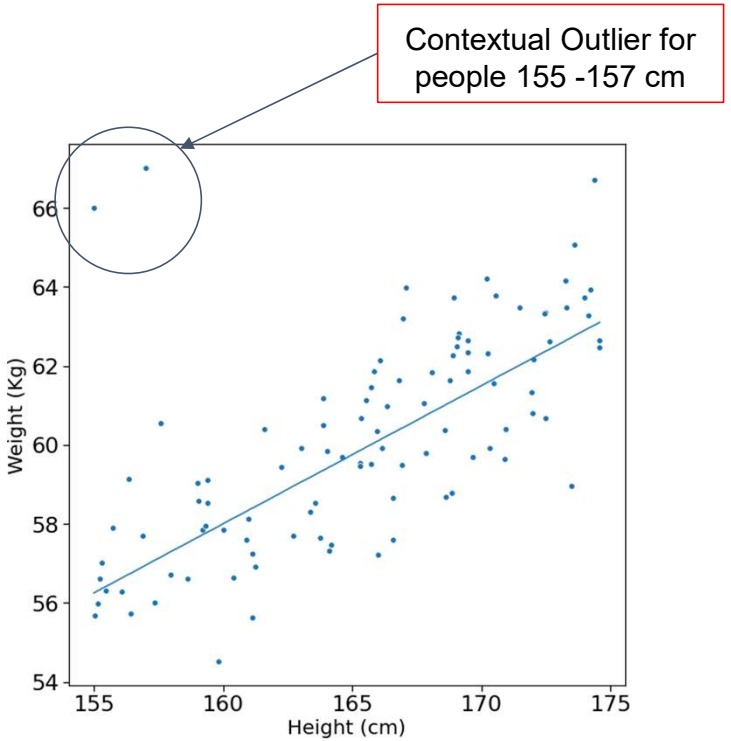
- Global Outlier
- Contextual Outlier
- Collective Outlier

Global Outlier



Any data points fall far outside the entirely data

Contextual Outlier

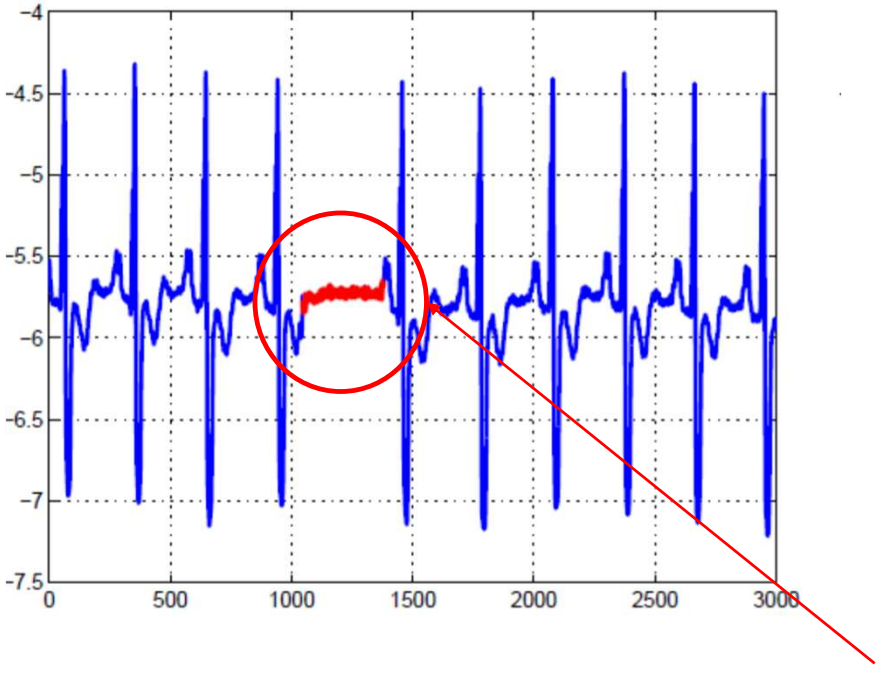


Any data points fall far outside the data points within the same context.

- When we talked about people who has low height, it is rare if those people has heavy weight
  - it is rare that people with height around 155 - 157 cm weighted above 66
  - but for people around 175 - 177 it is common
- Another example, For American it is common thing if the height fall around 180 but not for Asian

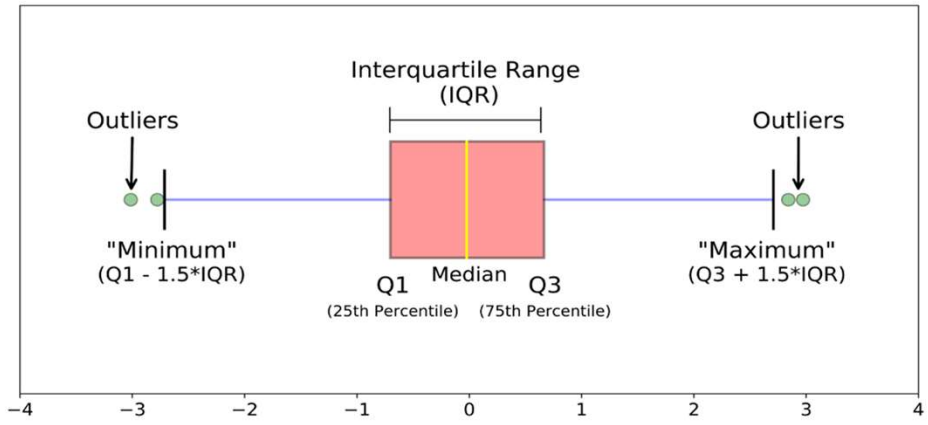
Collective Outlier

Any data point deviate significantly from the entire dataset but neither considered as either global outlier nor contextual outlier.



- The individual data instances in a collective outlier may not be outliers by themselves, but their occurrence together as a collection is anomalous.
- Only happened in data sets where data instances are related
- Often happened in sequence data, graph data, spatial data.
- Collective outlier can also appeared contextually
- Example : Human electro diagram output

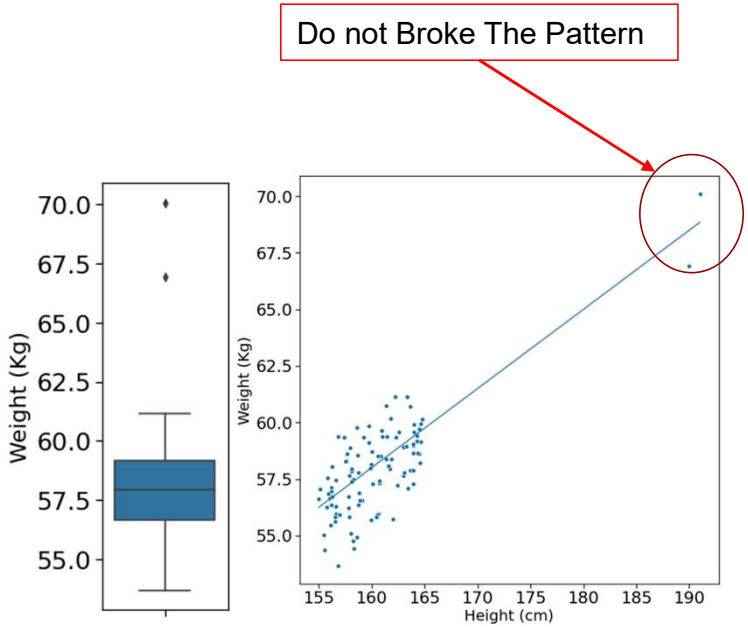
Outlier in Univariate Variable



A method can be used to detect outlier :

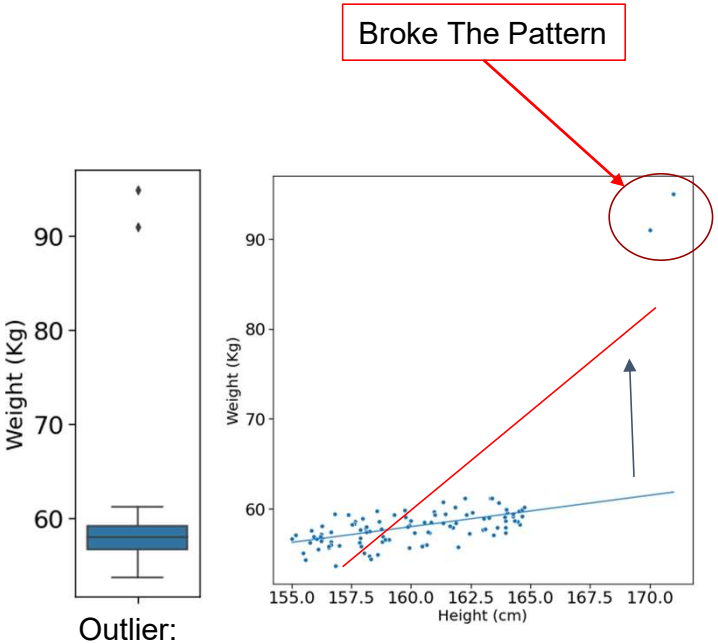
1. The data value > Q3 + 1.5 IQR, or
2. The data value < Q1 - 1.5 IQR

Outlier in Linear Regression



Outlier :

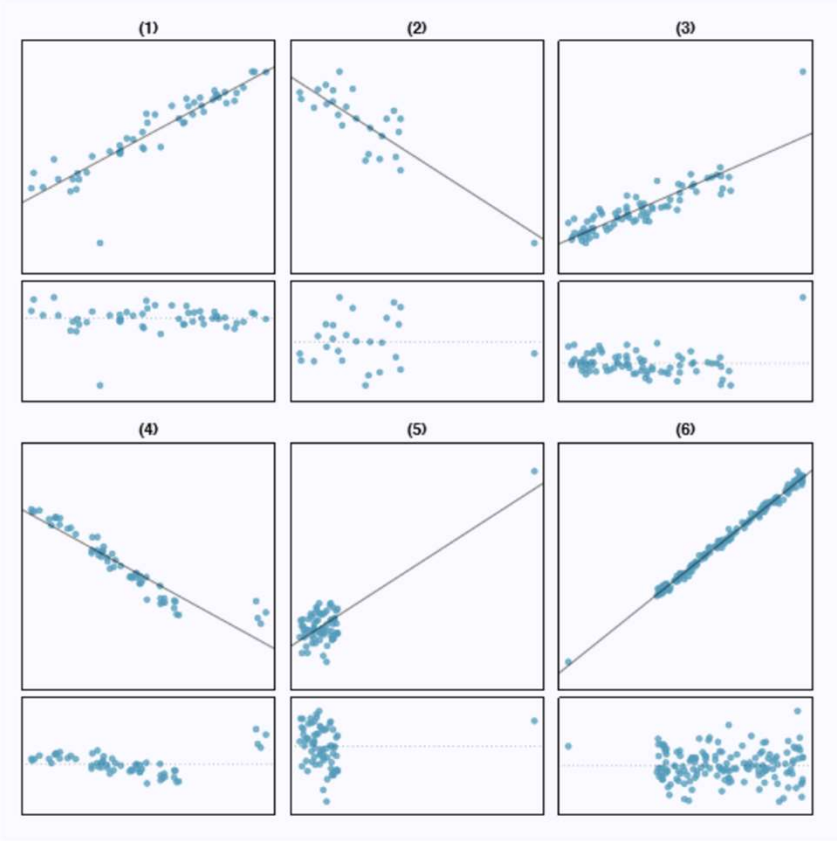
- Do not broke the pattern



Outlier:

- Broke the pattern
- also known as influential observation

Outlier In Linear Regression



1. Outlier slightly influence the line
2. Outlier do not much influence the line
3. Outlier slightly influence the line
4. Line badly fitted because outlier slightly influence the line and each of the cluster data points may have interesting explanation
5. Actually there is no certain pattern but the line appeared to be linearly positive because of the outlier
6. Outlier do not much influence the line

Outlier Cases

outlier detection aims to find patterns in data that do not conform to expected behavior. It is extensively used in many application domains such as

- Fraud detection for credit cards,
- Insurance,
- Healthcare
- Telecom fraud detection

Binning

What is Binning ?

Transform numerical variable into interval or categorical variable.

Tip Binning	Name
0 <= Tip <=1	Very Low
1 < Tip <= 2.5	Low
2.5 < Tip <= 4	Medium
4 < Tip <= 5.5	High
Tip > 5.5	Very High

Tip ($)
1.3
1.89
4.5
2.4
4.1
3.8
4.9
13

Tip Binning
1 - 2.5
1 - 2.5
4 - 5.5
1 - 2.5
4 - 5.5
2.5 - 4
4 - 5.5
> 5.5

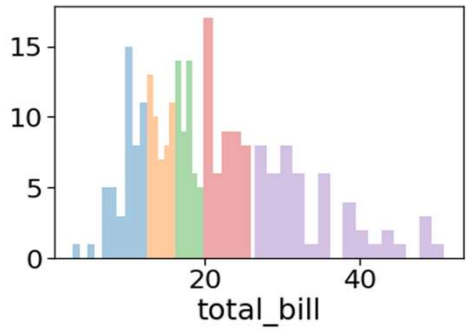
Binning Method

	total_bill	total bill eqfreq	total bill eqintv
0	16.99	(16.222, 19.818]	(12.618, 22.166]
1	10.34	(3.069, 12.636]	(3.022, 12.618]
2	21.01	(19.818, 26.098]	(12.618, 22.166]
3	23.68	(19.818, 26.098]	(22.166, 31.714]
4	24.59	(19.818, 26.098]	(22.166, 31.714]
...	...	...	...
239	29.03	(26.098, 50.81]	(22.166, 31.714]
240	27.18	(26.098, 50.81]	(22.166, 31.714]
241	22.67	(19.818, 26.098]	(22.166, 31.714]
242	17.82	(16.222, 19.818]	(12.618, 22.166]
243	18.78	(16.222, 19.818]	(12.618, 22.166]

244 rows × 3 columns

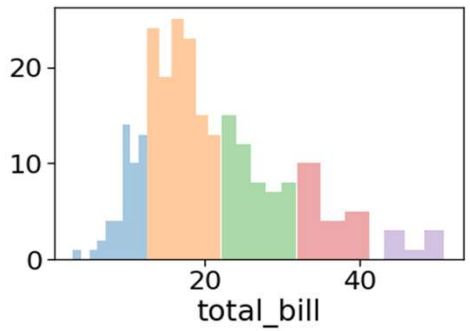
Equal Frequencies

total bill eqfreq	freq
(3.069, 12.636]	49.0
(12.636, 16.222]	49.0
(16.222, 19.818]	48.0
(19.818, 26.098]	49.0
(26.098, 50.81]	49.0

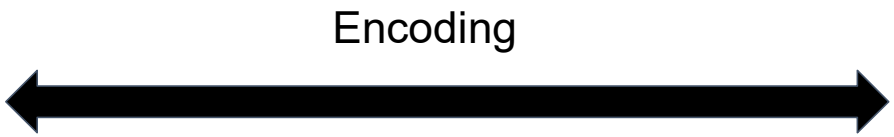


Equal Interval

total bill eqintv	freq
(3.022, 12.618]	49.0
(12.618, 22.166]	119.0
(22.166, 31.714]	50.0
(31.714, 41.262]	19.0
(41.262, 50.81]	7.0



Binning



Outlier

Tip ($)
1.3
1.89
4.5
2.4
4.1
3.8
4.9
13

Tip Binning
1 - 2.5
1 - 2.5
4 - 5.5
1 - 2.5
4 - 5.5
2.5 - 4
4 - 5.5
> 5.5

For Linear Relationship

Tip Encode
1
1
3
1
3
2
3
4

OR

For Nonlinear Relationship

K1	K2	K3	K4
1	0	0	0
1	0	0	0
0	0	1	0
0	0	0	0
0	0	1	0
0	1	0	0
0	0	1	0
0	0	0	1

Generating New Features : Polynomial

What is Polynomial Features ?

X
3
4
6
7
6

X	X**2
3	9
4	16
6	36
7	49
6	36

Second Order

Or

X	X**2	X**3
3	9	27
4	16	64
6	36	216
7	49	343
6	36	216

Third Order

OR

X	X**2	....	X**k
3	9	....	3**k
4	16	....	4**k
6	36	....	6**k
7	49	....	7**k
6	36	....	6**k

k-th Order

- model performance will increase significantly If the right order chosen
- too low : underfitting
- too high : overfitting

Polynomial Features for several variables

X1	X2
3	10
4	13
6	12
7	11
6	10

X1	X2	X1**2	X2**2
3	10	9	100
4	13	16	168
6	12	36	144
7	11	49	121
6	10	36	100

OR

X1	X2	X1**2	X2**2	X1**3	X2**3
3	10	9	100	27	1000
4	13	16	168	64	2197
6	12	36	144	216	1728
7	11	49	121	343	1331
6	10	36	100	216	1000

Apply Several Preprocessing Method to Modeling at once

Part 2 : Decision Tree

data : adult.csv

target : income

preprocess:

1. missing value : simple imputer with constant
2. one hot encoding : relationship, race, sex
3. binary encoding : workclass, marital status, occupation, native country
4. ordinal encoding : education (already encoded)
5. no treatment : numerical
6. out : fnlwgt

Random state 10, data splitting 70:30 model Tree(max depth 5, criterion entropy)

Feature Selection

What is Feature Selection ?

- Feature selection is a method to choose feature that actually have significant impact or important in the modeling
- Feature selection can be used as generalization method because too many feature may cause overfitting too little feature may cause undefittng
- Fewer feature can make interpretation easier (but beware of underfitting)

X1	X2	X3	X4	X5	X6	X7	Y
3	10	11	32	0.5	100	54	12
4	13	12	30	0.5	99	56	10
6	12	15	33	0.1	87	57	13
…	…	…	…	…	…	…	…
6	10	12	12	1.9	81	78	16

X1	X4	X6	Y
3	32	100	12
4	30	99	10
6	33	87	13
…	…	…	…
6	12	81	16

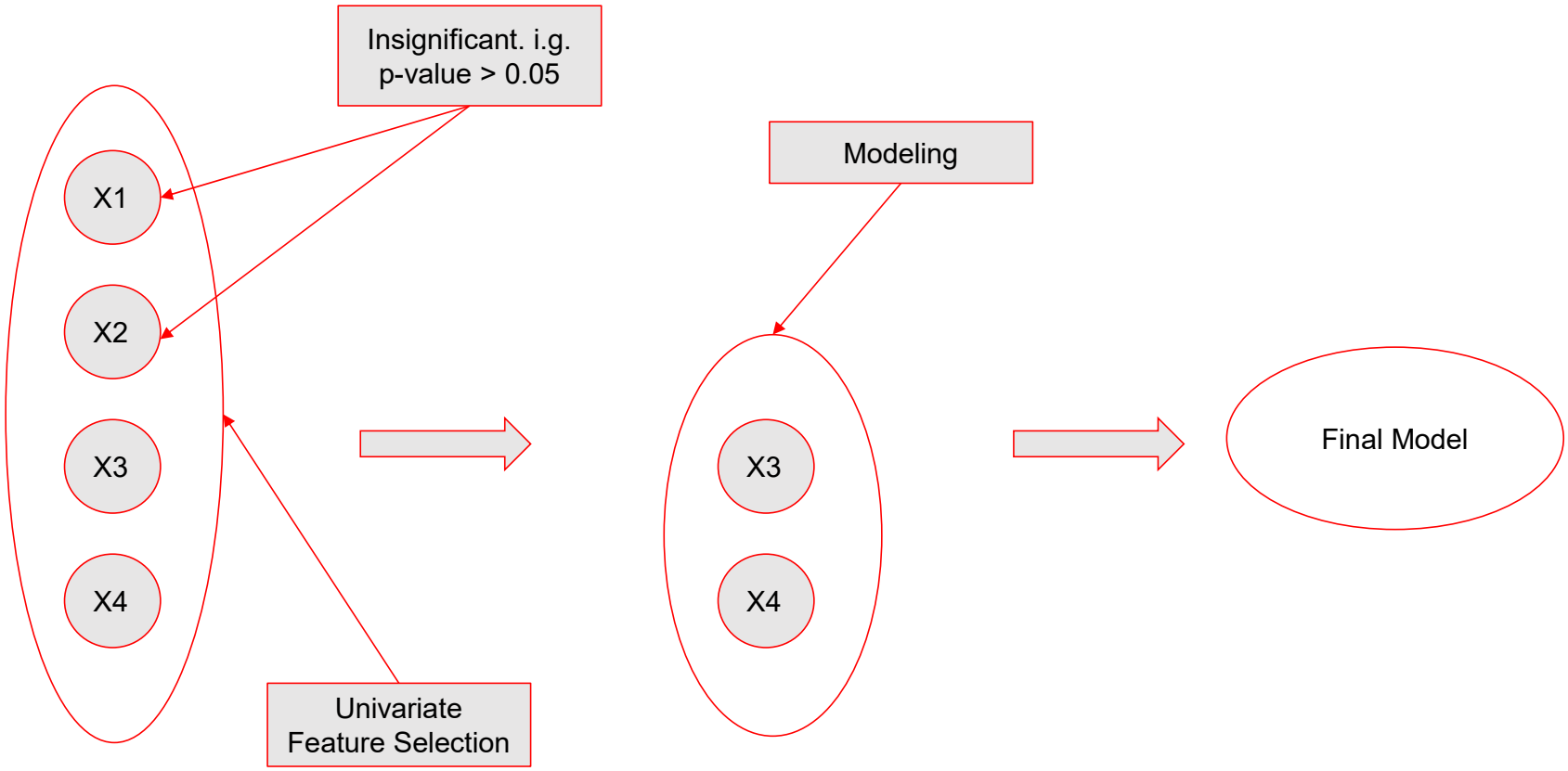
Feature Selection Method

1. Univariate Statistics Feature Selection
2. Model Based Feature Selection
3. Iterative Feature Selection

Univariate Statistics Feature Selection

- Chose feature that has a statistically significant (based on F-Statistics or Log Likelihood) relationship with the target
  - SelectKBest : selects a fixed number k of features
  - SelectPercentile : selects a fixed percentage of features
- Do not need to build any model
  - pros : fast to compute
  - cons : the result completely independent of the model that you might use (potentially less optimal)
- Only consider feature individually
  - cons : some feature might be useful after combined with another feature (can’t capture interaction)

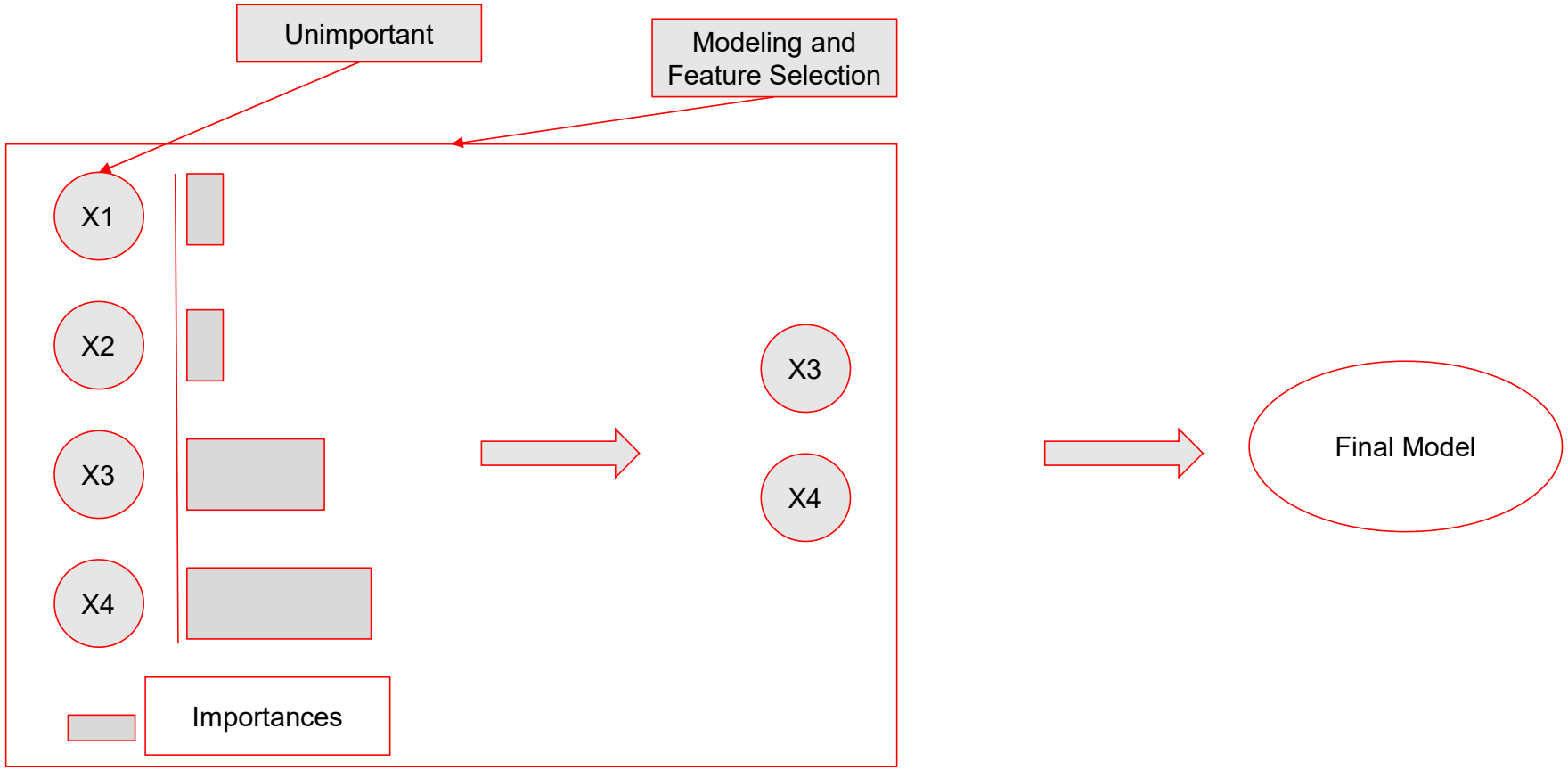
Univariate Statistics Feature Selection



Model Based Feature Selection

- Judge the importance of each feature using a supervised machine learning (a single model)
  - Decision Tree and Tree based models : feature importances
  - Linear model : coefficient's abs. value can be seen as feature importance (feature must have same scale or standardized feature)
- Need to build the model first
  - pros : the result depend on the model that you used (potentially more optimal)
  - cons : might make whole modeling process take longer time
- Selection consider all feature at once
  - pros : can capture interaction

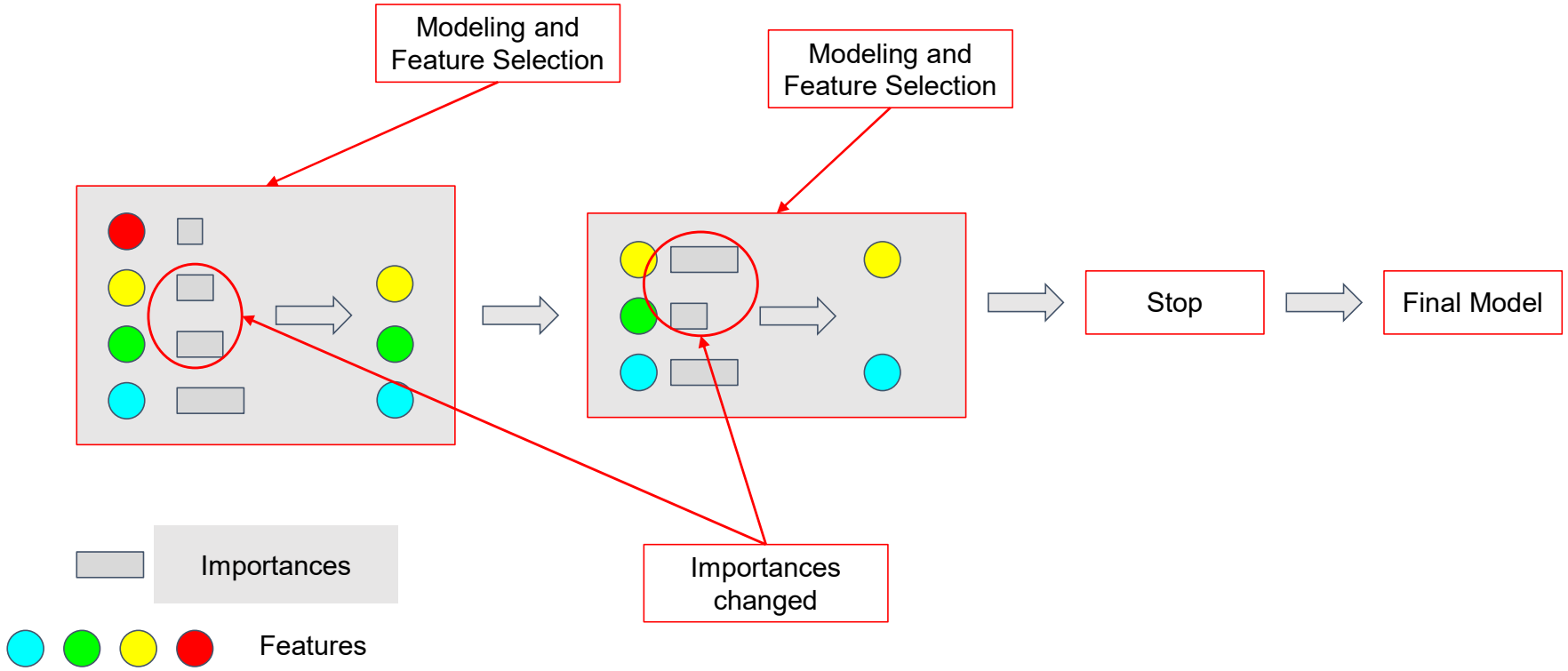
Model Based Feature Selection



Iterative Feature Selection

- Building a series of model with varying number of features
  - backward : start with using all features and keep removing it one by one until some criterion is reached (RFE(Recursive Feature Elimination))
  - forward : start with no feature and keep adding it one by one until some criterion is reached
- Need to build many model
  - pros : tried many possible combination and often outperform univariate and model based
  - cons : take significantly longer time than univariate and model based one
- Selection consider all feature at once
  - pros : can capture interaction

Iterative Feature Selection (backward)



Apply Several Preprocessing Method to Modeling at once

Part 3 : Logistic Regression

data : adult.csv

target : income

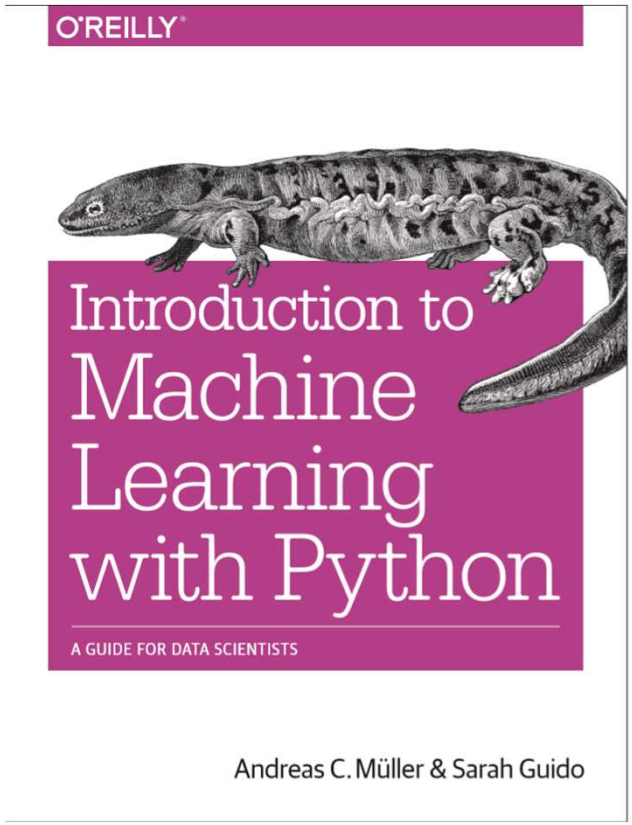
preprocess:

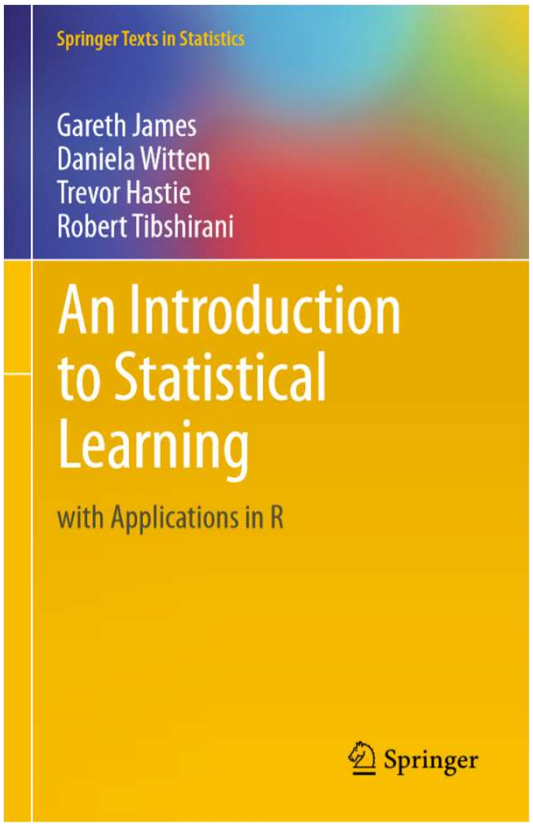
- missing value : simple imputer with constant
- one hot encoding : relationship, race, sex
- binary encoding : workclass, marital status, occupation, native country
- ordinal encoding : education (already encoded)
- no treatment : numerical
- out : fnlwgt

Random state 10, data splitting 70:30

- feature selection : select percentile
- model : logistic regression(max iter 1000, solver liblinear, C 10)

References


O'REILLY
Introduction to Machine Learning with Python
A GUIDE FOR DATA SCIENTISTS
Andreas C. Müller & Sarah Guido


Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Springer

References

https://medium.com/@danberdov/types-of-missing-data-902120fa4248’

https://www.real-statistics.com/handling-missing-data/types-of-missing-data/

https://www.anblicks.com/resources/insights-blogs/an-introduction-to-outliers/#:~:text=Outliers%20can%20be%20classified%20into,Intrusion%20detection%20in%20computer%20networks.&text=outlier%20classes.

https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_OpenIntro_Statistics_(Diez_et_al)./07%3A_Introduction_to_Linear_Regression/7.04%3A_Types_of_Outliers_in_Linear_Regression

https://towardsdatascience.com/all-about-categorical-variable-encoding-305f3361fd02

http://contrib.scikit-learn.org/category_encoders/index.html

https://www.researchgate.net/publication/267964435_Outlier_Detection_Applications_And_Techniques

https://www.the-modeling-agency.com/crisp-dm.pdf

https://scikit-learn.org/stable/

https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114