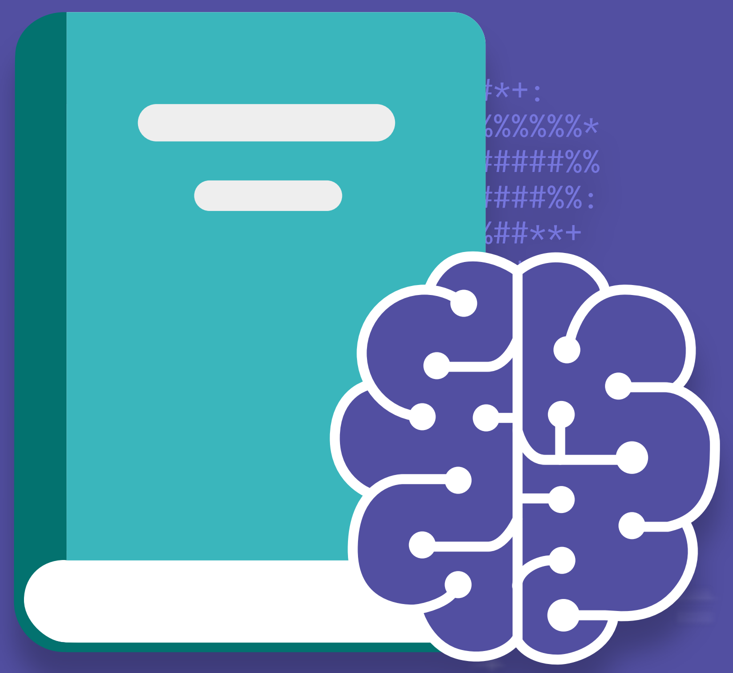
/* elice */

비전공자를 위한 머신러닝

4주차: 머신러닝 실무체험



Elice

목차

1. 머신러닝 업무 익히기
2. 타겟 마케팅을 위한 머신러닝 업무

1. 머신러닝 업무 익히기

머신러닝 업무 리뷰

< 데이터 과학의 목표 >

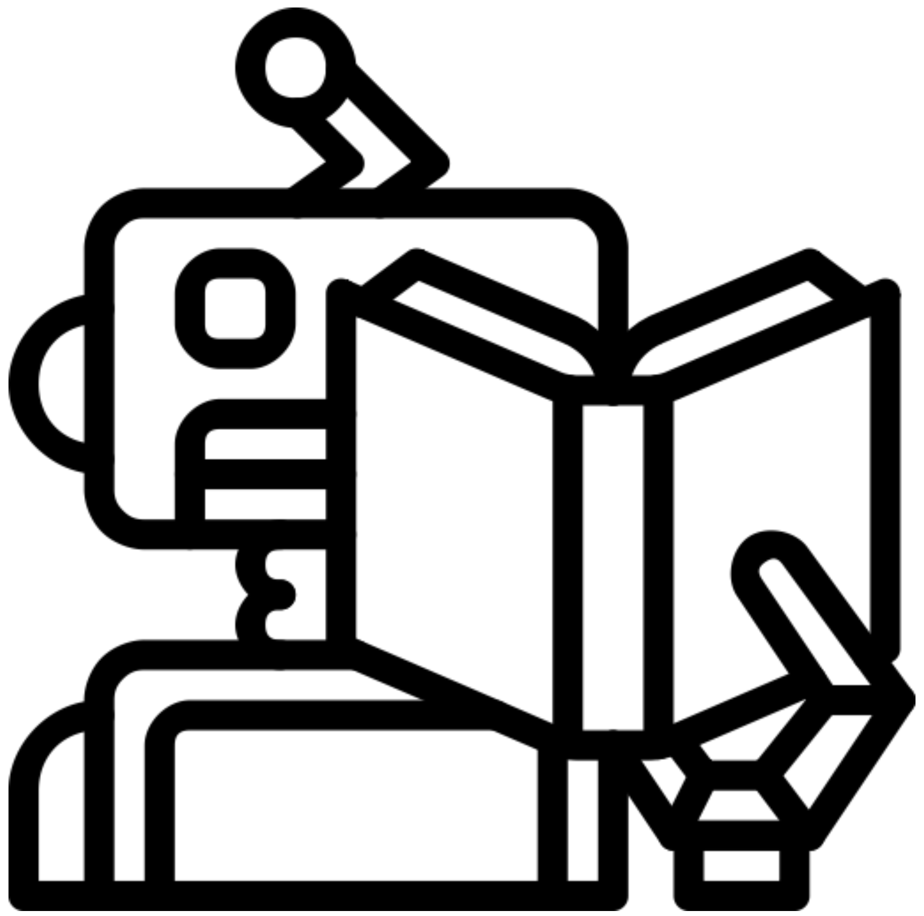
Decision Making

Monetization

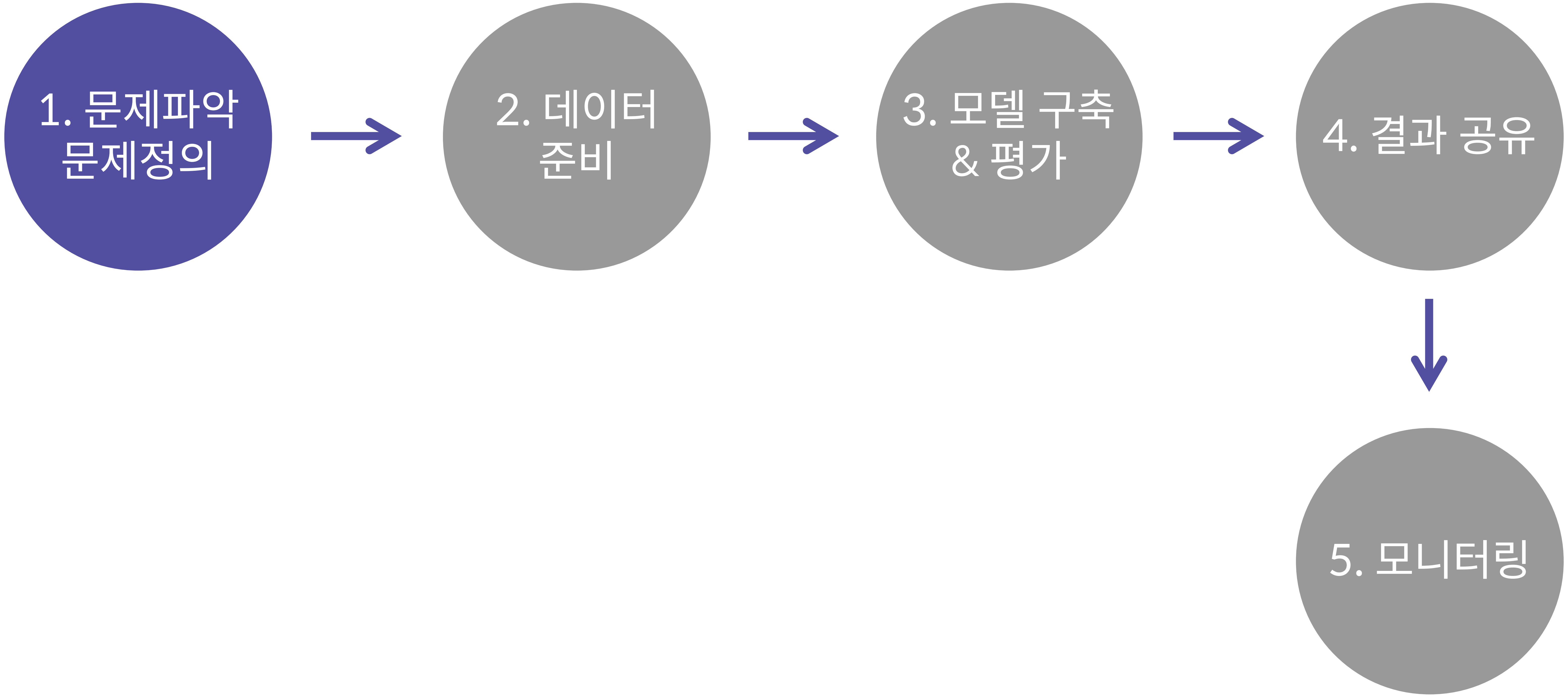
머신러닝 업무 리뷰

< 머신러닝의 목표 >

Prediction & Pattern Analysis



머신러닝 업무 프로세스



Define the Problem

머신러닝 프로젝트를 시작할 때

해결해야 하는 **비즈니스 문제**를 명확하게 먼저 정의

< 문제정의/문제파악을 위한 세부 프로세스 >

비즈니스 문제 파악 → 머신러닝 문제로 전환 → 머신러닝 도입 가능성/필요성 검토 → 효과검증 설계

Define the Problem

비즈니스 문제를 파악한 후에 이를 해결하기 위한

데이터 과학과 **머신러닝 문제**로 전환

Types of Machine Learning

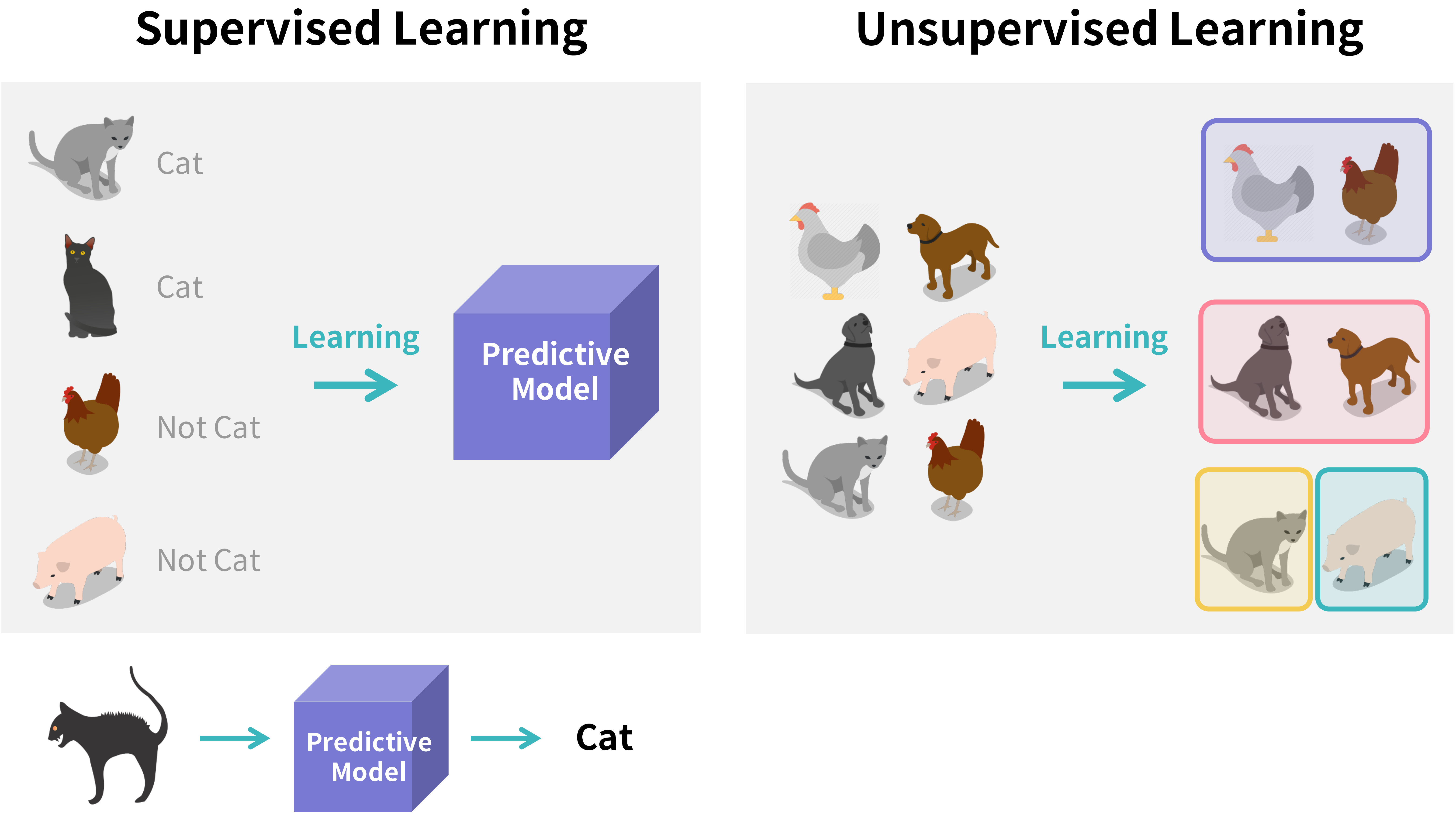
Supervised Learning

Develop Predictive Model
based on Input & Output Data

Unsupervised Learning

Group and Interpret Data
based on only Input Data

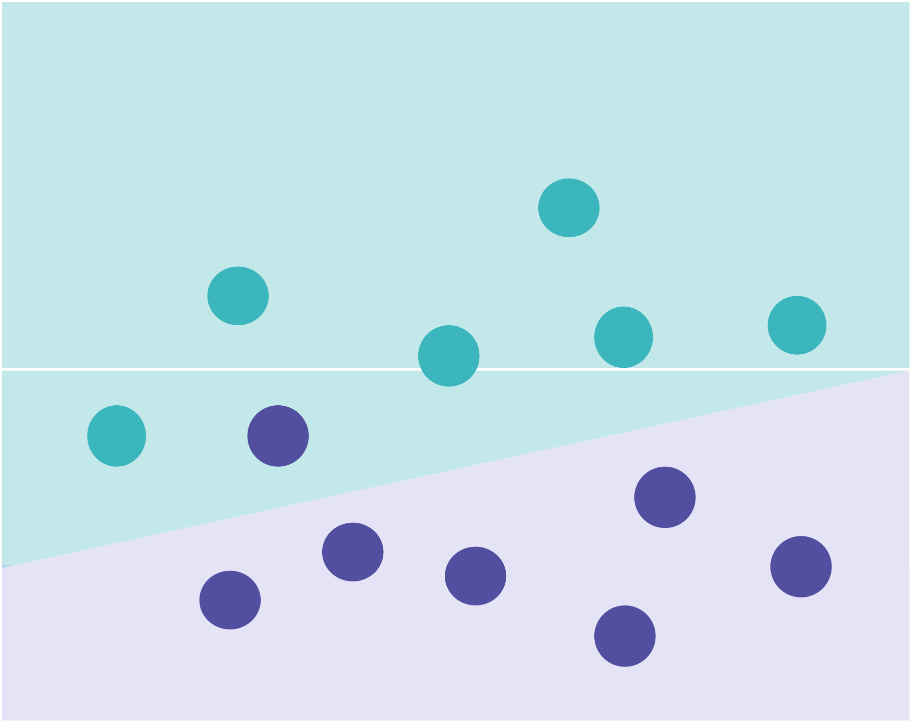
Types of Machine Learning



Supervised Learning

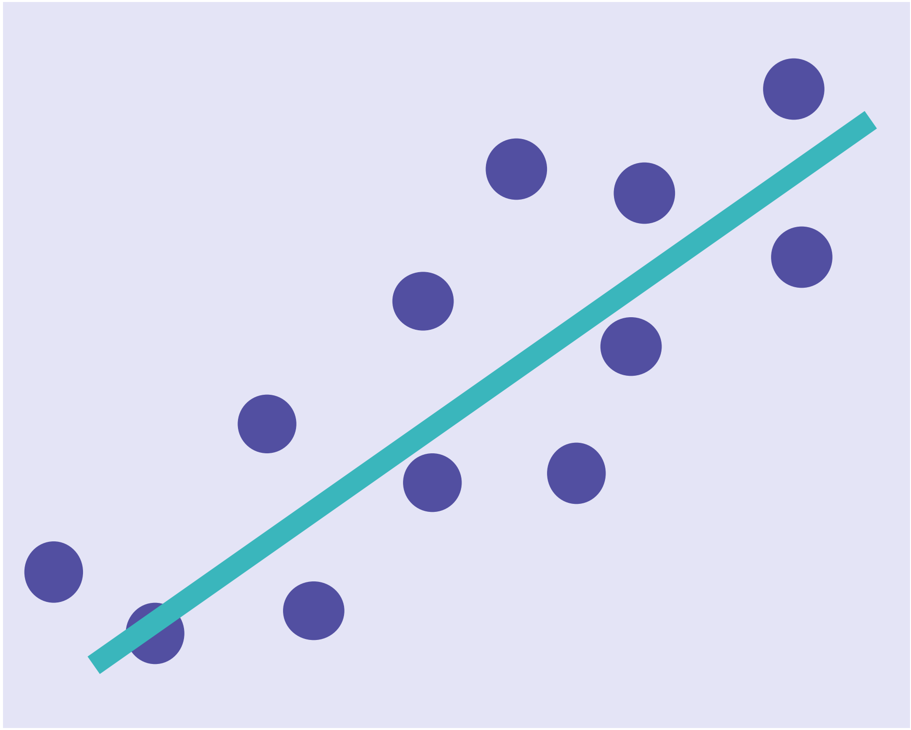
Classification

분류
범주를 예측



Regression

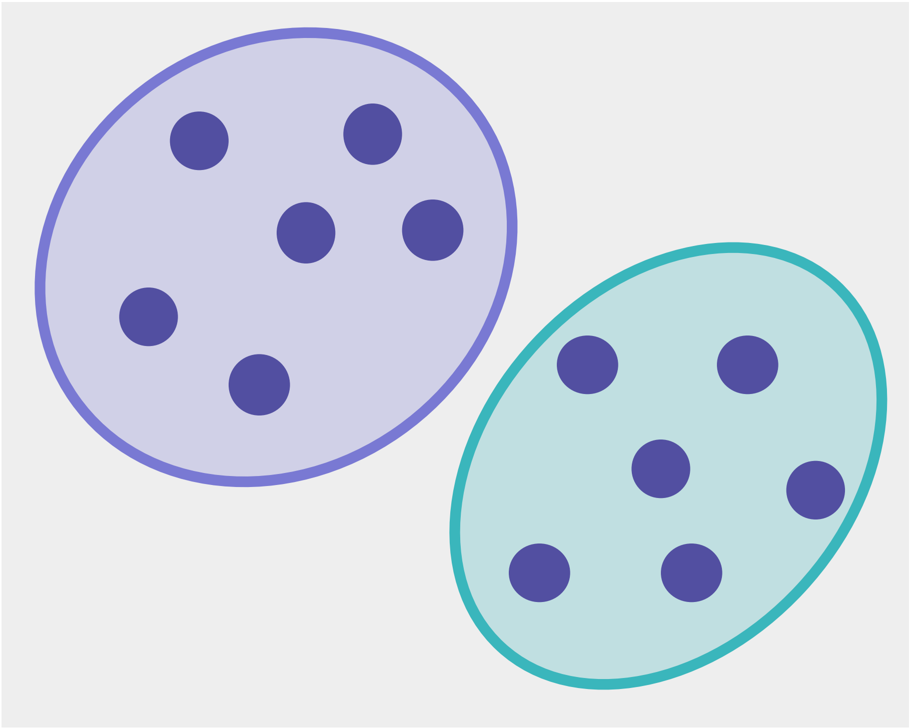
회귀
숫자를 예측



Unsupervised Learning

Clustering

유사한 그룹끼리 군집화



현실의 문제를 머신러닝 문제로

Business Problem	Target/Output	ML Problem
고객이 서비스를 이탈할 것인가	범주 : 이탈여부	Classification
내년도 서비스 예상 매출액은 얼마인가	숫자: 매출액	Regression
사용자 정보와 구매이력 기반 고객 세분화	-	Clustering

기타 머신러닝 문제

Recommender System

Anomaly Detection

Network Analysis

Dimensionality Reduction

Profiling

Time series Forecasting

효과검증 설계 예시

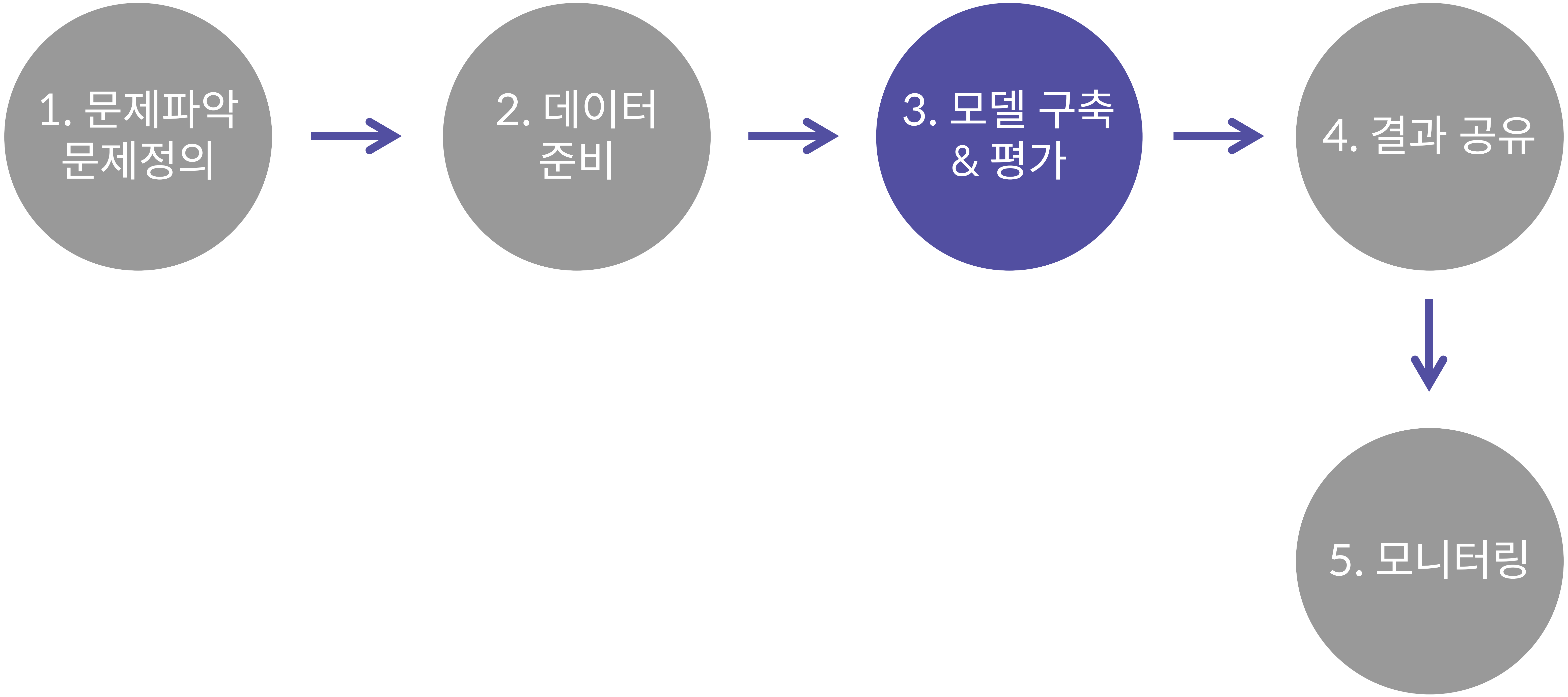
머신러닝 도입에 따른 효과 검증 프레임워크

효과검증 설계 예시

머신러닝 도입에 따른 효과 검증 프레임워크

문제정의	사용자의 서비스 **재방문율**(Retention)을 높이고 싶다
가설설정	서비스를 이탈할 것 같은 사용자에게 **프로모션**을 제공하여 재방문하게 한다
해결방안	서비스 **이탈 예측 모델**을 개발한다
효과검증	사용자의 **재방문율**(Retention)이 증가했는지 확인한다

머신러닝 업무 프로세스



Build Model & Evaluation

머신러닝 문제로 전환하고 데이터 준비를 마친 이후에는

적절한 머신러닝 모델 & 알고리즘을 선택하여 **모델을 구축하고 평가**

< 모델 구축 & 평가를 위한 세부 프로세스 >

모델 & 알고리즘 선택 → 실무적 제약사항 고려한 모델 적합 → 하이퍼파라미터 설정 → 모델 학습 → 모델 평가

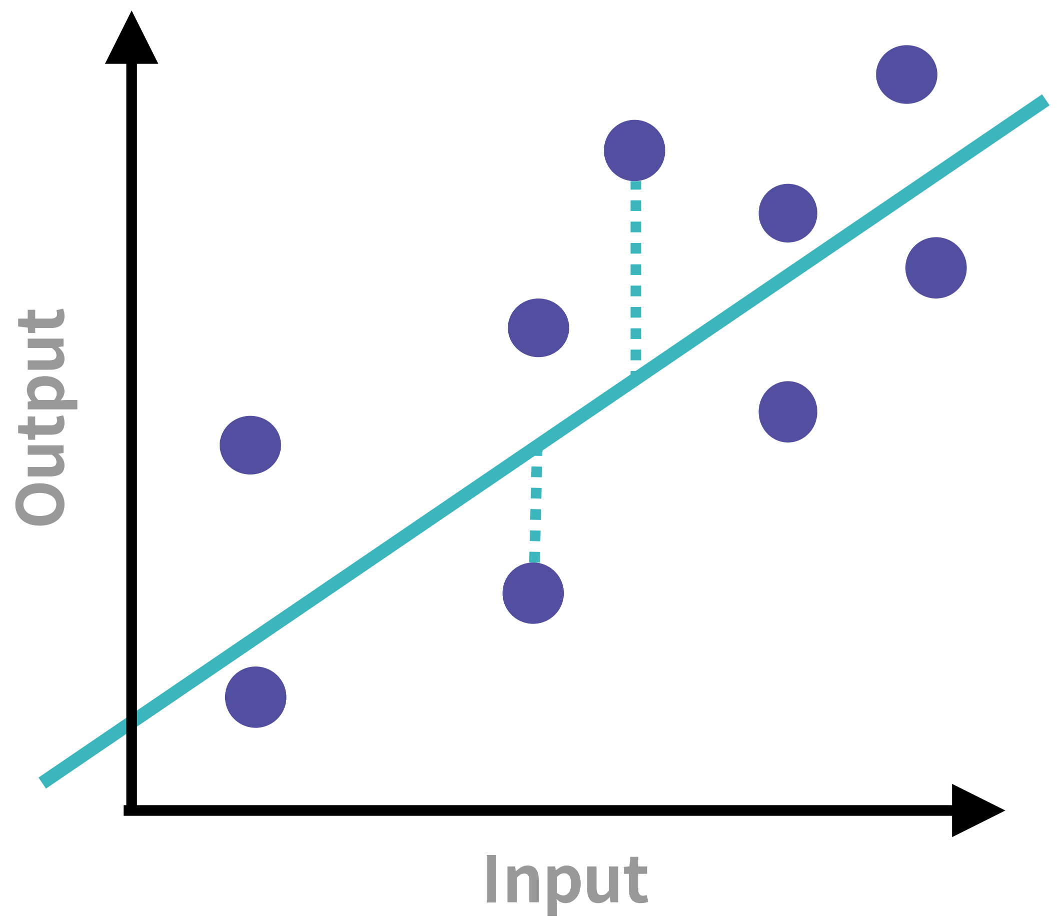
모델 & 알고리즘 선택

ML Model	Algorithm	Result
Classification	Logistic Regression Decision Tree Support Vector Machine	범주 예측
Regression	Linear Regression Ridge Regression Lasso Regression	숫자 예측
Clustering	K-means DBscan	군집

모델 평가

Regression은 **실제 값**(y)과 **예측한 값**$(\hat{y})$의 차이

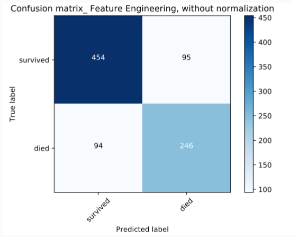
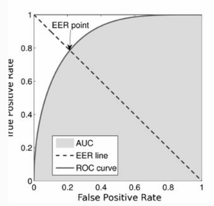
오차(Loss/Cost/Error)를 통해 모델의 성능 평가



Acroynm	Full Name	Description
MAE	Mean Absolute Error	$\frac{1}{n}\sum \lvert y-\hat{y}\rvert$
MSE	Mean Square Error	$\frac{1}{n}\sum (y-\hat{y})^2$
RMSE	Root Mean Square Error	$\sqrt{\frac{\Sigma(y-\hat{y})^2}{n}}$
MAPE	Mean Percentage Error	$\frac{100\%}{n}\sum \left\lvert\frac{y-\hat{y}}{y}\right\rvert$

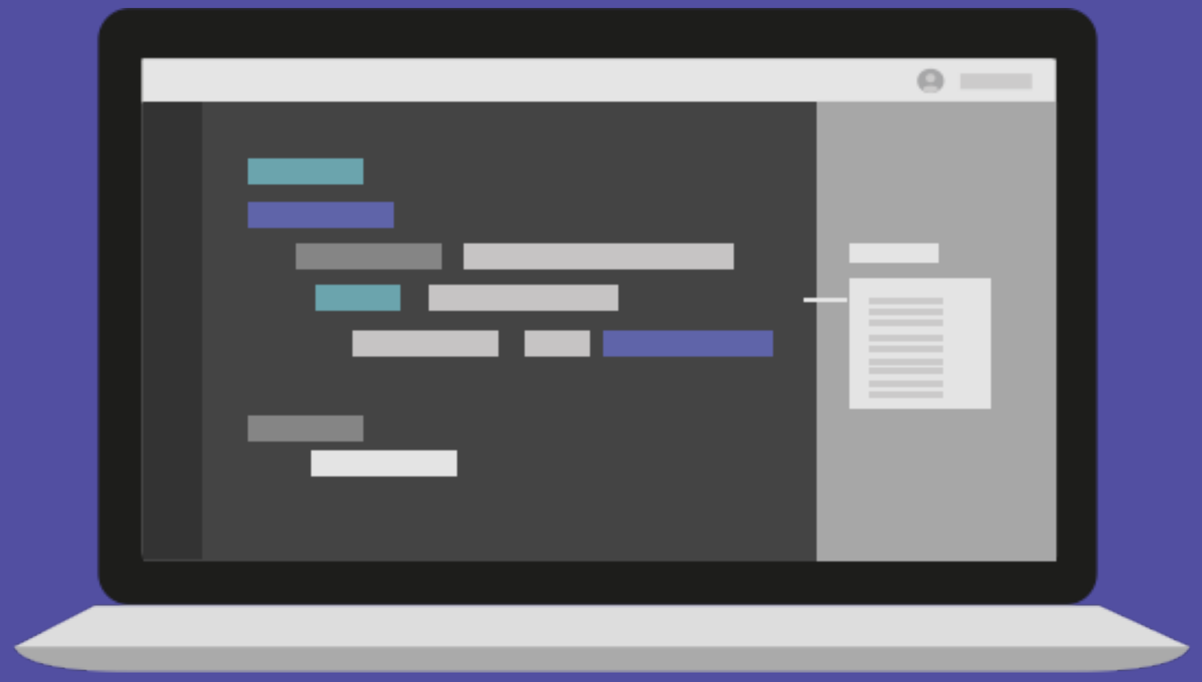
모델 평가

Classification은 **실제 범주**$(class)$와 **예측한 범주**$(class)$의 **정확도**(Accuracy)를 통해 모델의 성능 평가

Name	Description	Etc.
Accuracy	옳게 분류한 정확도	$\frac{correct\ prediction}{total\ data\ points} \times 100\%$
Confusion Matrix	분류 결과를 2x2의 표로 정리한 혼동행렬	
F-measure	precision과 recall의 조화평균 *precision: 예측한 범주에서 실제 True 범주 비율 *recall: 실제 범주에서 옳게 True라고 예측한 범주 비율	$F = \frac{precision \times recall}{precision + recall}$
AUC	TPR과 FPR을 각각 x축과 y축으로 했을 때의 생성되는 ROC curve 아래의 면적 *TPR: True Positive Rate 옳게 예측한 비율 *FPR: True라고 잘못 예측한 비율	

2. 타겟 마케팅을 위한 머신러닝 업무

[실습] 누구에게 프로모션을 제공해야 할까?



Bank Target Marketing 데이터 소개

1) 포르투갈의 기관 은행에서 진행한 마케팅 관련 데이터
2) Phone call을 통한 마케팅 프로모션 진행
3) Target은 마케팅을 통해 정기예금(term deposit)을 가입했는지 가입여부 의미

Source:
[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

실습 특징

1) 실습을 위해서 총 45,211명의 데이터에서 41,259명의 데이터를 사용하여 머신러닝 모델 학습
2) 학습한 모델을 이용하여 나머지 3,952명에 대한 정기예금 가입여부를 예측
3) 예측된 결과를 이용하여 기대손익 계산

스위치 소개

handling_missing_value

	결측치 비율(%)	결측치 수
age	0.00	0
job	0.80	330
marital	0.19	80
education	4.20	1731
default	20.87	8597
housing	2.40	990
loan	2.40	990
contact	0.00	0
month	0.00	0
day_of_week	0.00	0
duration	0.00	0
campaign	0.00	0
pdays	0.00	0
previous	0.00	0
poutcome	0.00	0
emp.var.rate	0.00	0
cons.price.idx	0.00	0
cons.conf.idx	0.00	0
euribor3m	0.00	0
nr.employed	0.00	0
y	0.00	0

handling_missing_value

False

age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y
56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	261	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	0
57	services	married	high.school	NaN	no	no	telephone	may	mon	149	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	0
37	services	married	high.school	no	yes	no	telephone	may	mon	226	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	0
40	admin.	married	basic.6y	no	no	no	telephone	may	mon	151	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	0
56	services	married	high.school	no	no	yes	telephone	may	mon	307	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	0
45	services	married	basic.9y	NaN	no	no	telephone	may	mon	198	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	0
59	admin.	married	professional.course	no	no	no	telephone	may	mon	139	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	0
41	blue-collar	married	NaN	NaN	no	no	telephone	may	mon	217	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	0
24	technician	single	professional.course	no	yes	no	telephone	may	mon	380	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	0
25	services	single	high.school	no	yes	no	telephone	may	mon	50	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	0

handling_missing_value

True

age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y
56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	261	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	0
57	services	married	high.school	no	no	no	telephone	may	mon	149	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	0
37	services	married	high.school	no	yes	no	telephone	may	mon	226	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	0
40	admin.	married	basic.6y	no	no	no	telephone	may	mon	151	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	0
56	services	married	high.school	no	no	yes	telephone	may	mon	307	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	0
45	services	married	basic.9y	no	대치됨 no	no	telephone	may	mon	198	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	0
59	admin.	married	professional.course	no	no	no	telephone	may	mon	139	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	0
										제거됨										
24	technician	single	professional.course	no	yes	no	telephone	may	mon	380	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	0
25	services	single	high.school	no	yes	no	telephone	may	mon	50	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	0
25	services	single	high.school	no	yes	no	telephone	may	mon	222	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	0

add_age_categorical

True

cons.conf.idx	euribor3m	nr.employed	y	age_cat
-36.4	4.857	5191.0	0	3
-36.4	4.857	5191.0	0	3
-36.4	4.857	5191.0	0	2
-36.4	4.857	5191.0	0	2
-36.4	4.857	5191.0	0	3
-36.4	4.857	5191.0	0	2
-36.4	4.857	5191.0	0	3
-36.4	4.857	5191.0	0	1
-36.4	4.857	5191.0	0	1
-36.4	4.857	5191.0	0	1

add_marketing_info

True

age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome
56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	261	1	999	0	nonexistent
57	services	married	high.school	no	no	no	telephone	may	mon	149	1	999	0	nonexistent
37	services	married	high.school	no	yes	no	telephone	may	mon	226	1	999	0	nonexistent
40	admin.	married	basic.6y	no	no	no	telephone	may	mon	151	1	999	0	nonexistent
56	services	married	high.school	no	no	yes	telephone	may	mon	[illegible]	1	999	0	nonexistent
45	services	married	basic.9y	no	no	no	telephone	may	mon	[illegible]	1	999	0	nonexistent
59	admin.	married	professional.course	no	no	no	telephone	may	mon	139	1	999	0	nonexistent
24	technician	single	professional.course	no	yes	no	telephone	may	mon	380	1	999	0	nonexistent
25	services	single	high.school	no	yes	no	telephone	may	mon	50	1	999	0	nonexistent
25	services	single	high.school	no	yes	no	telephone	may	mon	222	1	999	0	nonexistent

추가됨

add_social_economic_info

True

emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
-1.8	93.876	-40.0	0.699	5008.7
-3.4	92.379	-29.8	0.788	5017.5
1.1	93.994	-36.4	4.856	5191.0
-1.7	94.215	[illegible]	0.893	4991.6
1.4	93.918	[illegible]	4.968	5228.1
-1.7	94.215	-40.3	0.782	4991.6
1.1	93.994	-36.4	4.857	5191.0
1.1	93.994	-36.4	4.860	5191.0
1.4	94.465	-41.8	4.959	5228.1
-0.1	93.200	-42.0	4.021	5195.8

추가됨

add_marketing_info

True

emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
-1.8	93.876	-40.0	0.699	5008.7
-3.4	92.379	-29.8	0.788	5017.5
1.1	93.994	-36.4	4.856	5191.0
-1.7	94.215	-40.3	0.893	4991.6
1.4	93.918	-42.7	4.968	5228.1
-1.7	94.215	-40.3	0.782	4991.6
1.1	93.994	-36.4	4.857	5191.0
1.1	93.994	-36.4	4.860	5191.0
1.4	94.465	-41.8	4.959	5228.1
-0.1	93.200	-42.0	4.021	5195.8

			cation	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome
			sic.4y	no	no	no	telephone	may	mon	261	1	999	0	nonexistent
			school	no	no	no	telephone	may	mon	149	1	999	0	nonexistent
			school	no	yes	no	telephone	may	mon	226	1	999	0	nonexistent
			sic.6y	no	no	no	telephone	may	mon	151	1	999	0	nonexistent
			school	no	no	yes	telephone	may	mon		1	999	0	nonexistent
45	services	married	basic.9y	no	no	no	telephone	may	mon		1	999	0	nonexistent
59	admin.	married	professional.course	no	no	no	telephone	may	mon	139	1	999	0	nonexistent
24	technician	single	professional.course	no	yes	no	telephone	may	mon	380	1	999	0	nonexistent
25	services	single	high.school	no	yes	no	telephone	may	mon	50	1	999	0	nonexistent
25	services	single	high.school	no	yes	no	telephone	may	mon	222	1	999	0	nonexistent

추가됨

transform_pdays_to_categorical

True

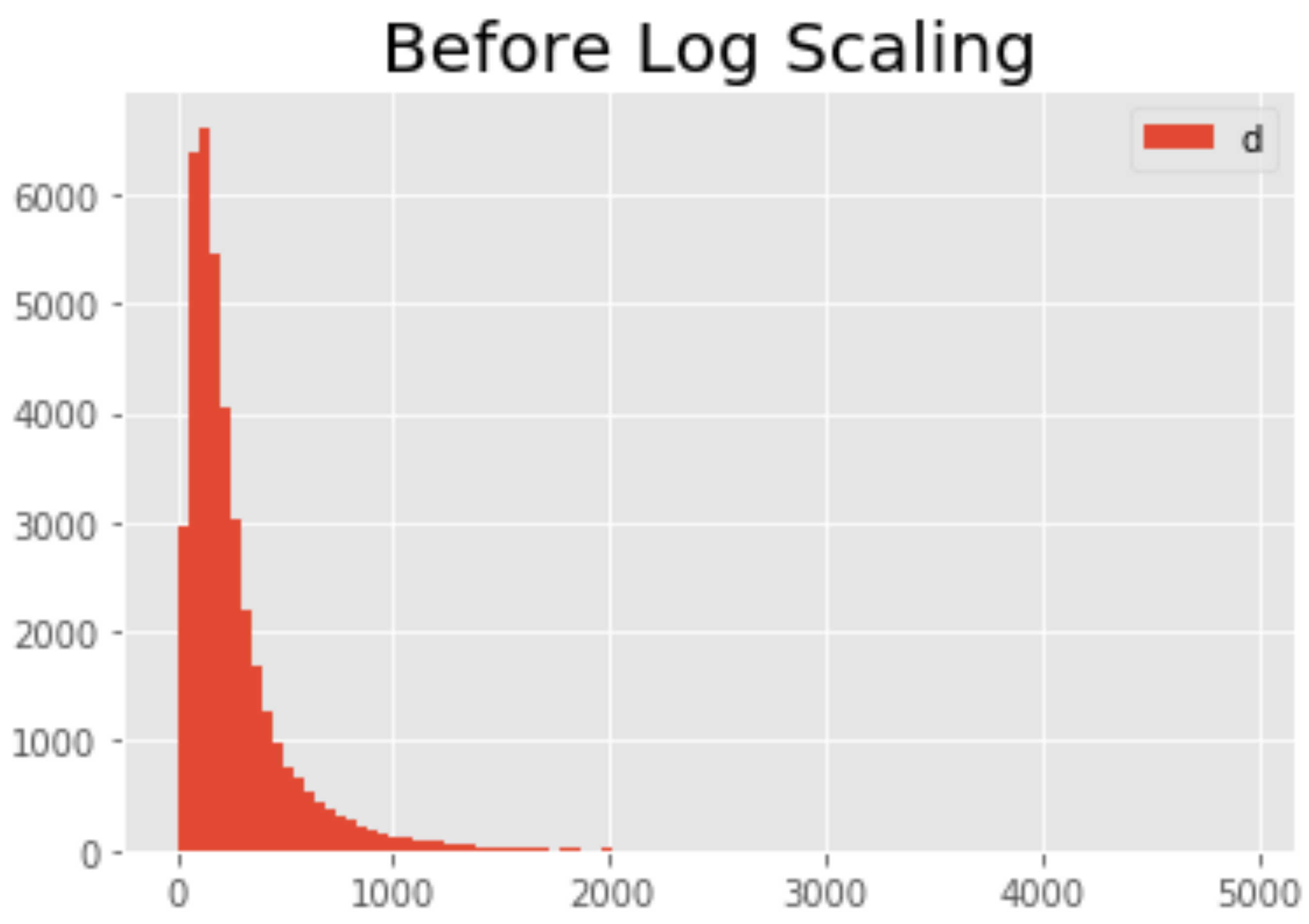
변환하지 않으면 해당 변수의 결측이 93%이므로 제거한다

- 0~4 → 1
- 5~9 → 2
- 10~14 → 3
- 15~19 → 4
- 20~24 → 5
- 25~ → 6
- 999 (missing value) → 7

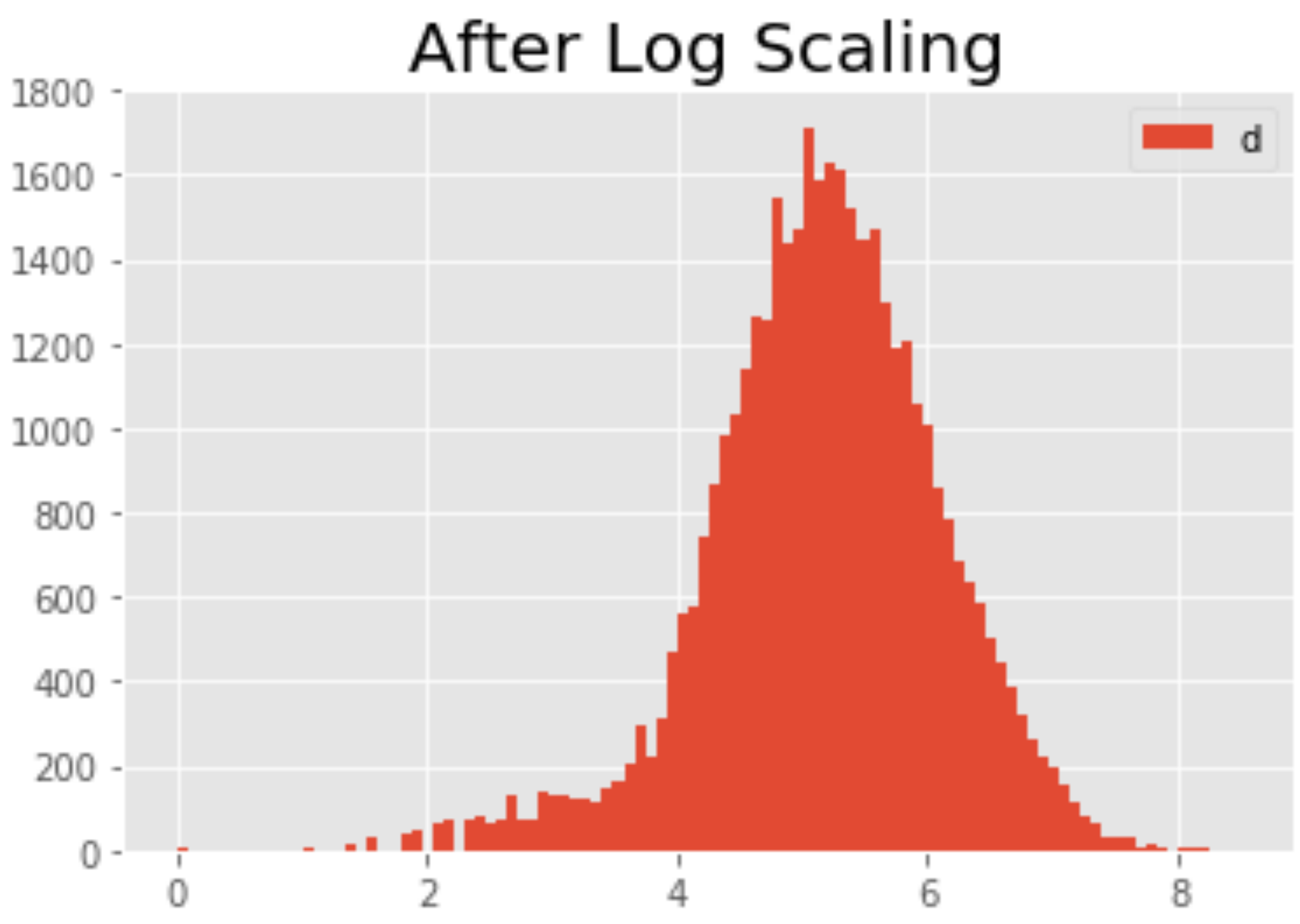
pdays
2
7
7
1
7
3
7
7
7
7

transform_duration_to_log_scale

False



True



feature_normalization

'minmax'

	duration
0	261
1	149
2	226
3	151
4	307
5	198
6	139
8	380
9	50
11	222
12	137

	duration
0	0.053070
1	0.030297
2	0.045954
3	0.030704
4	0.062424
5	0.040260
6	0.028264
8	0.077267
9	0.010167
11	0.045140
12	0.027857

'standard'

	duration
0	261
1	149
2	226
3	151
4	307
5	198
6	139
8	380
9	50
11	222
12	137

	duration
0	0.008676
1	-0.422568
2	-0.126088
3	-0.414867
4	0.185794
5	-0.233899
6	-0.461072
7	0.466873
8	-0.803757
9	-0.141489
10	-0.468773
11	0.131889
12	-0.434119

CREDIT

코스 매니저
손현곤

강사
오승우

콘텐츠 제작에 기여하신 분
오승우

영상 제작에 기여하신 분
박수광

검수와 자문에 도움주신 분
신현철

문의 및 연락처

academy.elice.io
contact@elice.io
facebook.com/elice.io
medium.com/elice