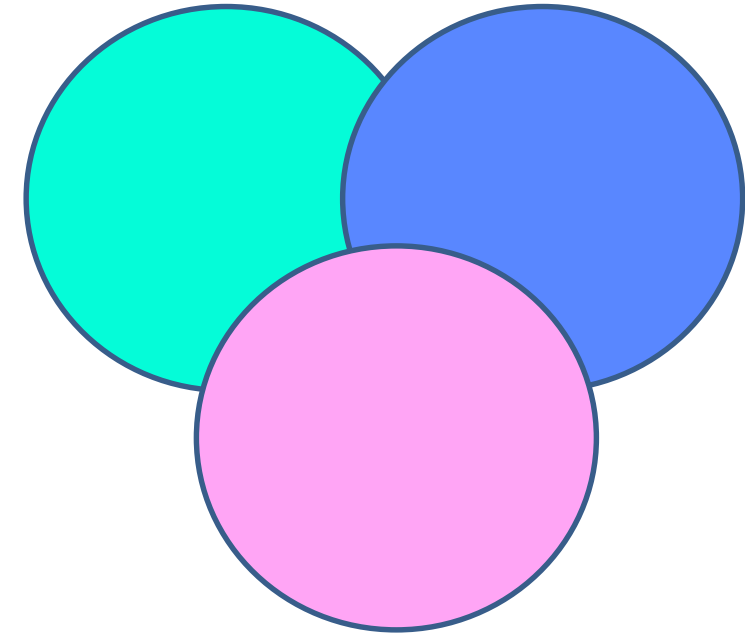


# Presentation For KOREN Data



# 목차

01	데이터 개요
02	데이터 EDA - ORIGINAL DATA
03	데이터 CDA - ORIGINAL DATA
04	데이터 전처리
05	데이터 EDA - PROCESSED DATA
06	데이터 CDA - PROCESSED DATA
07	모델링
08	모델 하이퍼파라미터 튜닝
09	모델링 - Sop_id 컬럼 제외
10	결론

# 01 데이터 개요

Feature 수(Column 수): 19개  
데이터 개수(Row 수): 7859개

컬럼명	설명	컬럼명	설명
sop_id	Standard Operating Procedure (SOP)의 고유 식별자	fault_detail_content	문제 또는 장애에 대한 자세한 내용을 포함하는 정보
ticket_id	티켓의 고유한 식별자	etc_content	기타 관련 내용을 나타내는 정보
ticket_type	티켓의 유형 또는 종류를 나타내는 정보	fault_type_content	문제 또는 장애 유형에 대한 자세한 내용을 포함하는 정보
ticket_result	티켓 처리 결과를 나타내는 정보	start_time	문제나 이벤트의 시작 시간
status	티켓의 상태를 나타내는 정보	end_time	문제나 이벤트의 종료 시간
request_time	티켓이 생성된 시각 또는 요청된 시각	handling_ack_user	문제나 이벤트 처리를 승인한 사용자를 식별하는 정보
receive_time	티켓을 수신한 시각	handling_ack_time	처리 승인 시간
detail	티켓에 대한 자세한 정보나 설명을 포함하는 정보	handling_fin_user	문제나 이벤트 처리를 완료한 사용자를 식별하는 정보
fault_classify	문제 또는 장애를 분류하는 정보	handling_fin_time	처리 완료 시간
fault_type	문제 또는 장애의 유형을 나타내는 정보	y	예측하기 위한 대상 변수를 나타냄

01 데이터 개요

컬럼명	데이터 타입	결측치	결측비율(%)
sop_id	int64	0	0.00
ticket_id	int64	0	0.00
ticket_type	object	1	0.01
ticket_result	object	2577	32.79
status	object	1	0.01
request_time	object	1	0.01
receive_time	object	7757	98.70
detail	object	7850	99.89
fault_classify	object	7839	99.75
fault_type	object	7839	99.75

컬럼명	데이터 타입	결측치	결측비율(%)
fault_detail_content	object	7839	99.75
etc_content	object	7840	99.76
fault_type_content	float64	7859	100.00
start_time	object	7857	99.97
end_time	object	7857	99.97
handling_ack_user	object	7856	99.96
handling_ack_time	object	7858	99.99
handling_fin_user	object	34	0.43
handling_fin_time	object	37	0.47
y	Object	0	0.00

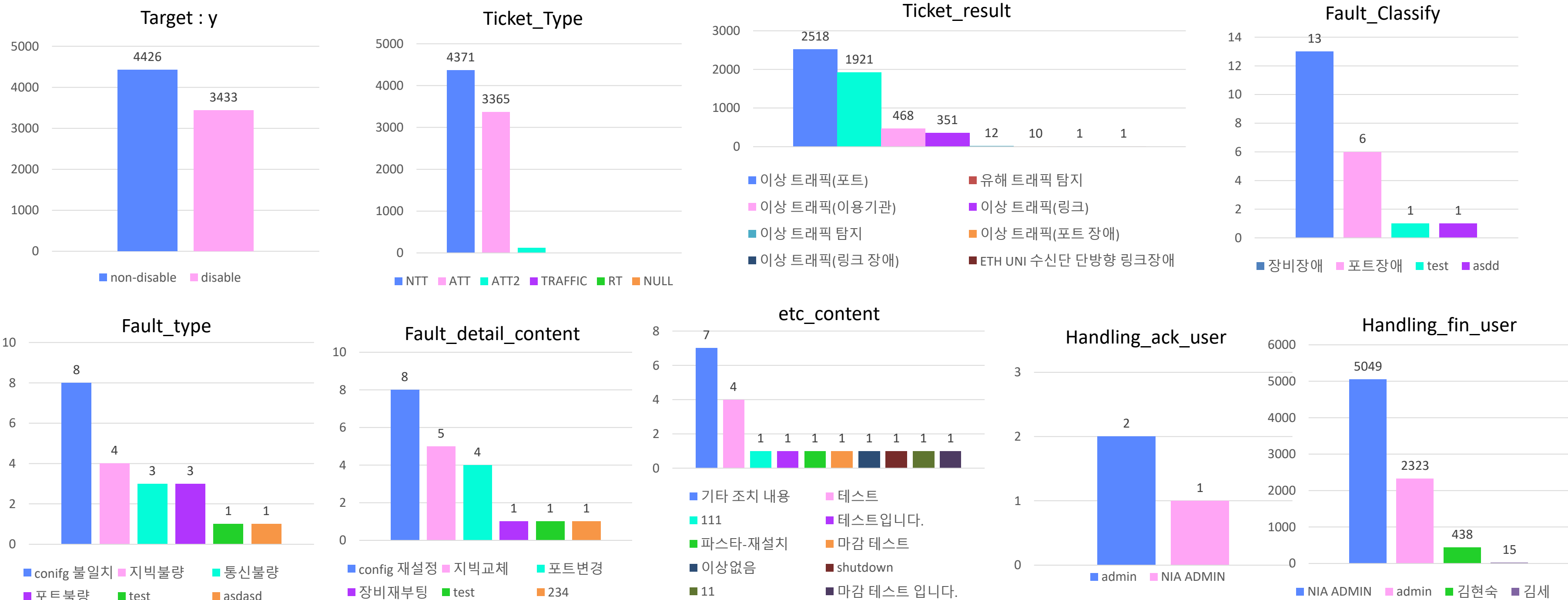
01 데이터 개요

컬럼명	nunique	unique	컬럼명	nunique	unique	컬럼명	nunique	unique
sop_id	7859	Unique	status	1	'FIN'	etc_content	10	'111', '테스트입니다.', '기타 조치 내용', '파스타-재설치', '마감 테스트', '이상없음', '테스트', 'shutdown'
ticket_id	7859	Unique	request_time	7858	unique1개 결측치	fault_type_content	0	null
ticket_type	5	'ATT','NTT','ATT2', 'TRAFFIC', 'RT'	receive_time	102	2023-08-25 11:07:14.643 +0900' '2022-07-28 15:55:46.349 +090, etc	start_time	2	'2021-12-14 13:06:23.754', '2022-04-05 12:28:50.376'
ticket_result	8	'이상 트래픽(포트)', '유해 트래픽 탐지', '이상 트래픽(링크)', '이상 트래픽(이용기관)', '이상 트래픽 탐지', '이상 트래픽(포트 장애)', '이상 트래픽(링크 장애)', 'ETH UNI 수신단 단방향 링크 장애 '	detail	1	'DETAIL'	end_time	2	'2021-12-14 14:06:23.754' '2022-04-05 13:28:50.376'
			fault_classify	4	'포트장애', ' 장비장애' , 'test', 'asdd'	handling_ack_user	2	'NIA ADMIN', 'admin'
			fault_type	6	'통신불량', ' config 불일치' , '포트불량', '지빅불량', 'test', 'asdd'	handling_ack_time	1	'2022-06-30 15:41:46.536 +0900'
			fault_detail_content	6	"포트변경', 'config 재설정', '장비재부팅' '지빅교체' , 'test', '234'	handling_fin_user	4	'admin' , 'NIA ADMIN', '김현숙' , 김세'
						handling_fin_time	7822	unique37개 결측치
						y	2	'non-disable', 'disable'

- 고유값 다수
- 컬럼전체 결측치
- Target 값

# 02 데이터 EDA - ORIGINAL DATA

※ Null 제외



# 03 데이터 CDA - ORIGINAL DATA

Target : y 에 대한 독립성 검증

P-value < 0.05 독립성을 띄지 않는다

P-value >= 0.05 독립성을 띈다

컬럼명	P.Value	독립성 여부
Sop_id	0.4946964925654553	독립
Ticket_id	0.4946964925654524	독립
Ticket_type	3.84295150256501e-53	독립X
Ticket_result	1.3636191849419942e-56	독립X
Status	1.0	독립
Request_time	0.49469615511240056	독립
Receive_time	0.45342248608678254	독립
Detail	1.0	독립
Fault_classify	0.5360000505482432	독립
Fault_type	0.04986599124408791	독립X

컬럼명	P.Value	독립성 여부
Fault_detail_content	0.24663415218605195	독립
Etc_content	0.07661320909916608	독립
Start_time	1.0	독립
End_time	1.0	독립
Handling_ack_user	1.0	독립
Handling_ack_time	1.0	독립
Handling_fin_user	5.3884745196930164e-86	독립X
Handling_fin_time	0.4946839637369011	독립

- y와 독립성을 가지지 않는 컬럼

04 데이터 전처리

01

→

02


→

03

→

04

→



Column 선별

Null 값 다수


- ticket\_result
- receive\_time
- detail
- fault\_classify
- fault\_type
- fault\_detail\_content
- etc\_content
- fault\_type\_content
- start\_time
- end\_time
- handling\_ack\_user
- handling\_ack\_time

전체 UNIQUE

- ticket\_id

UNIQUE 값 1개


- status



Request\_time / Handling\_fin\_time  
연도, 월, 시간으로 분리  
(파생변수 생성)

request_time	rf_year
	rf_month
	rf_hour

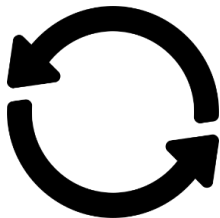
handling_fin_time	hf_year
	hf_month
	hf_hour



결측치 처리

rf_year	각 컬럼 내에 있는 값 중 random하게 결측치 대체
rf_month	
rf_hour	
ticket_type	

hf_year	각 컬럼의 최빈값으로 결측치 대체
hf_month	
hf_hour	
handling_fin_user	



Labeling

수동 라벨링

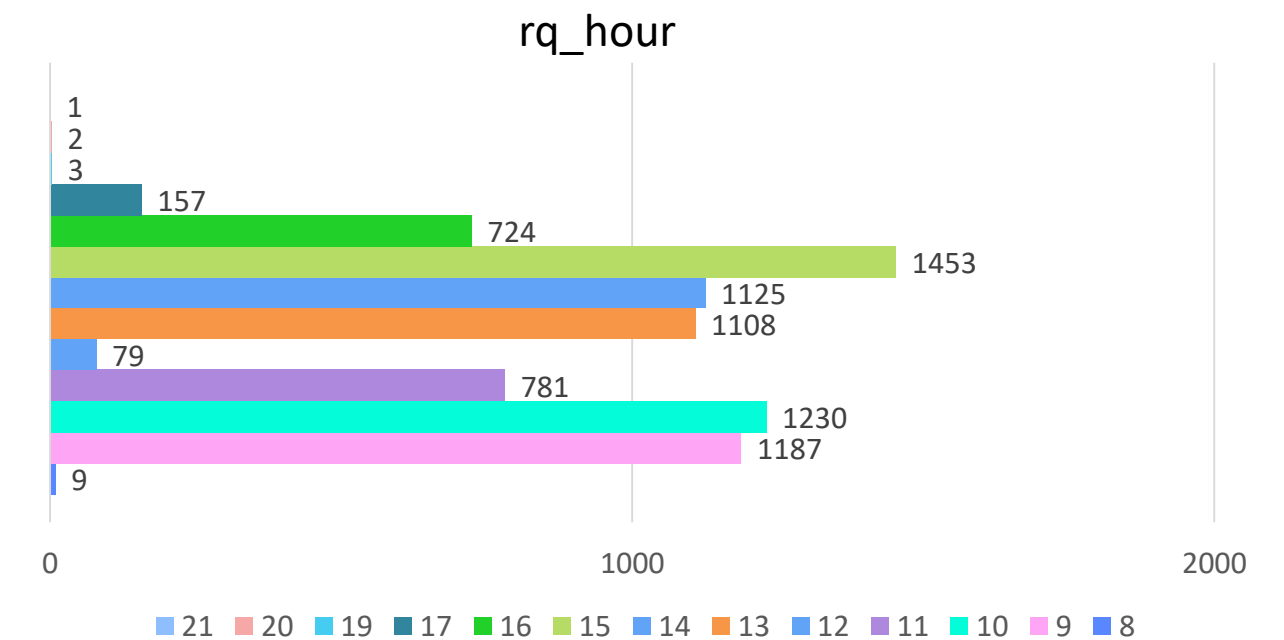
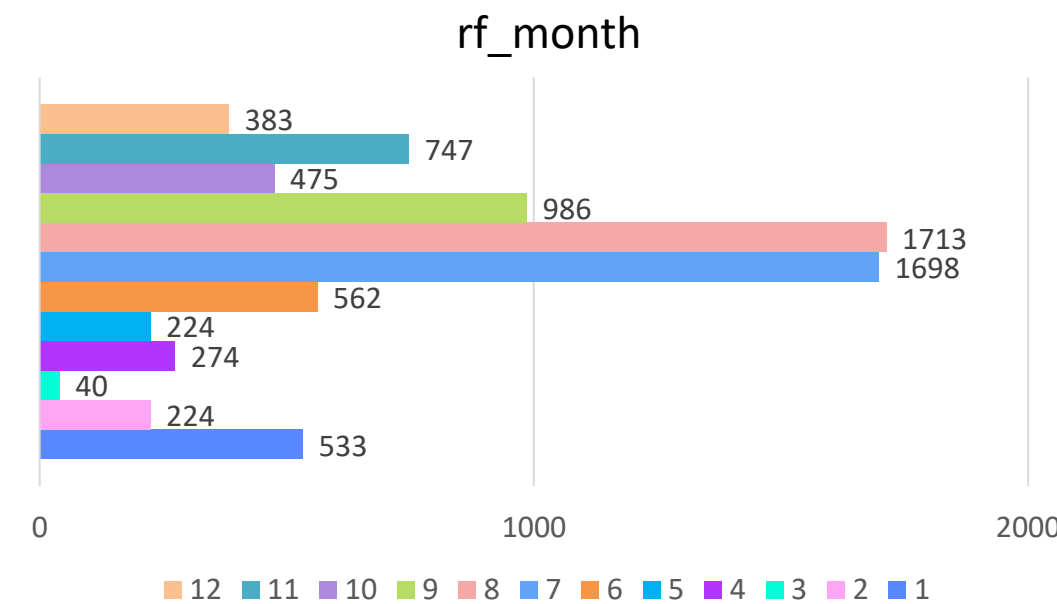
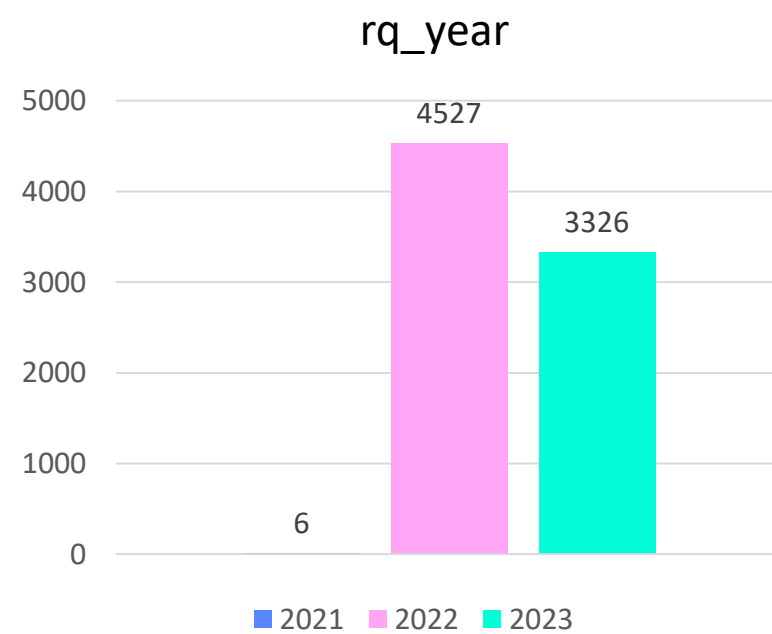
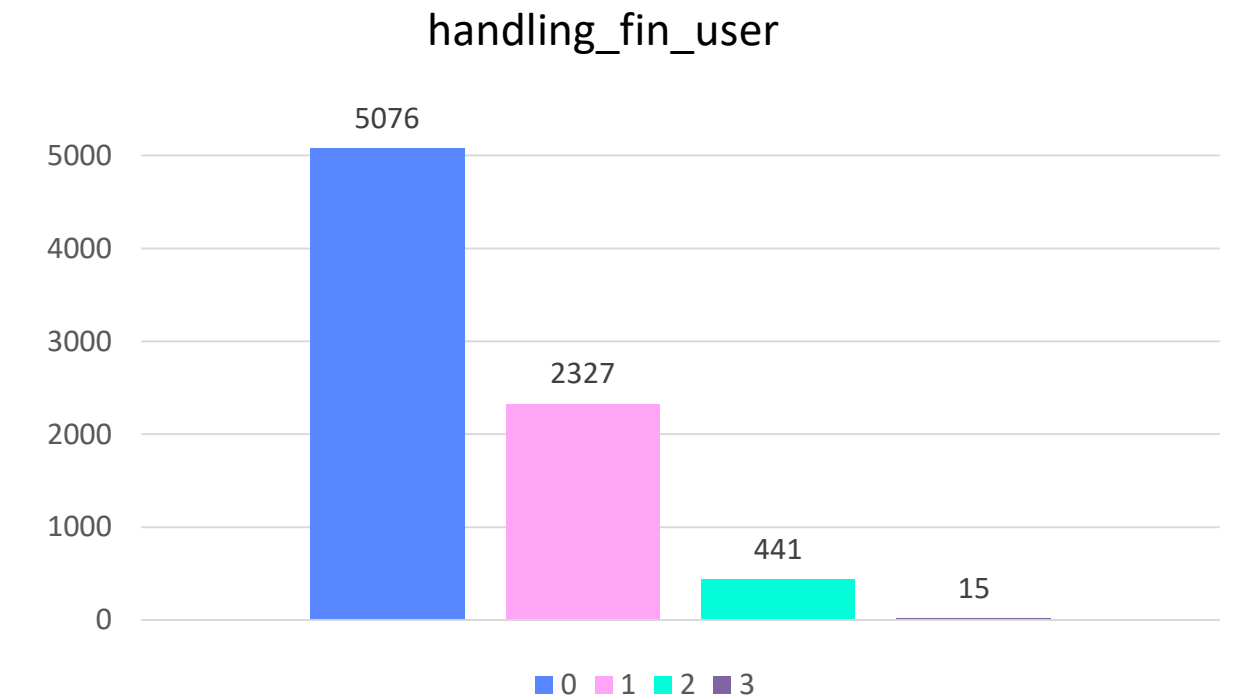
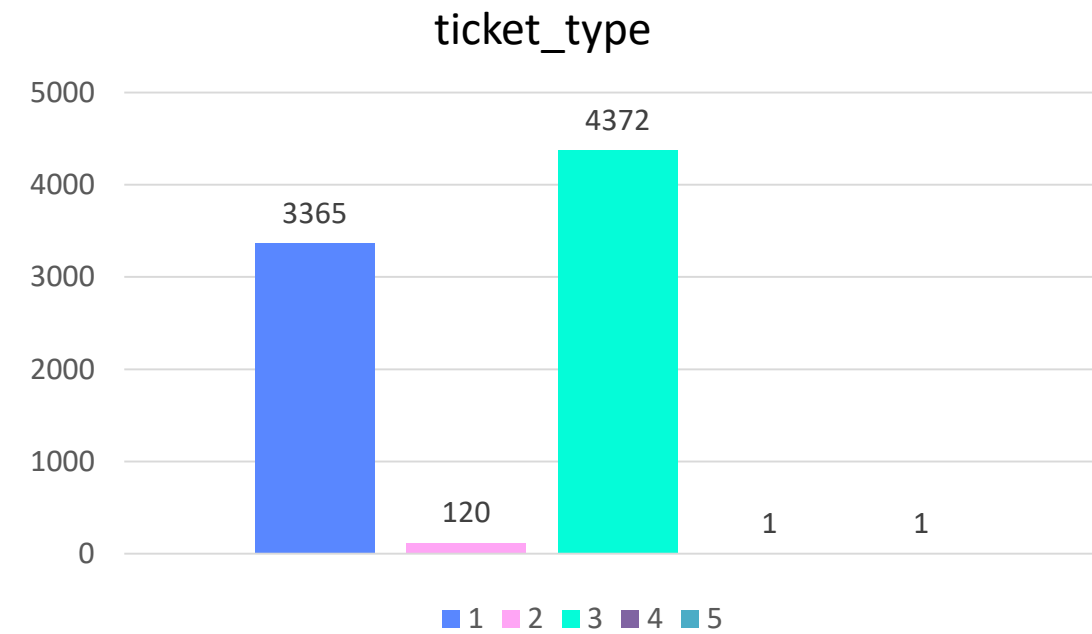
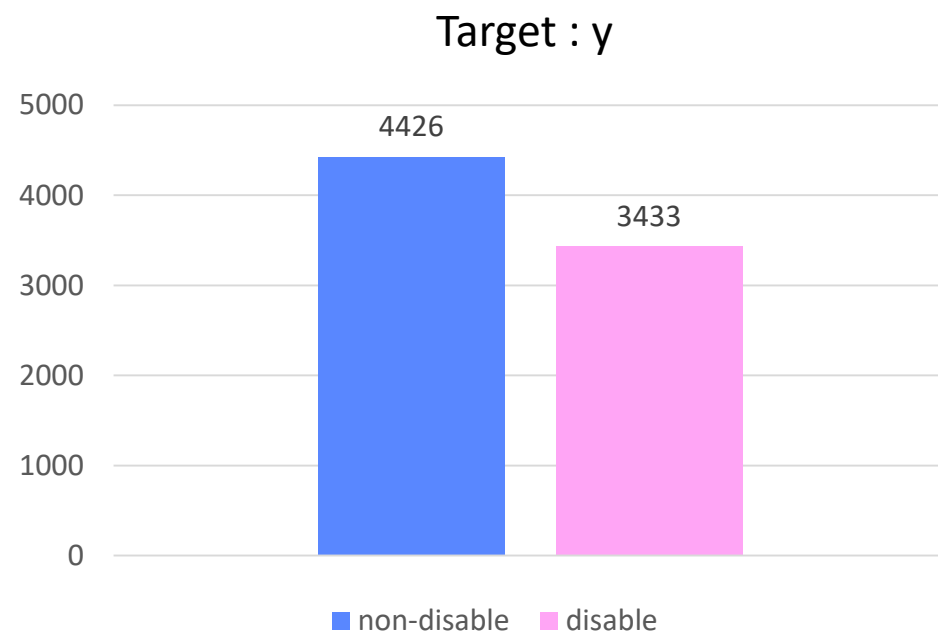
ticket_type	ATT	1
	ATT2	2
	NTT	3
	TRAFFIC	4
	RT	5

Label Encoder 사용

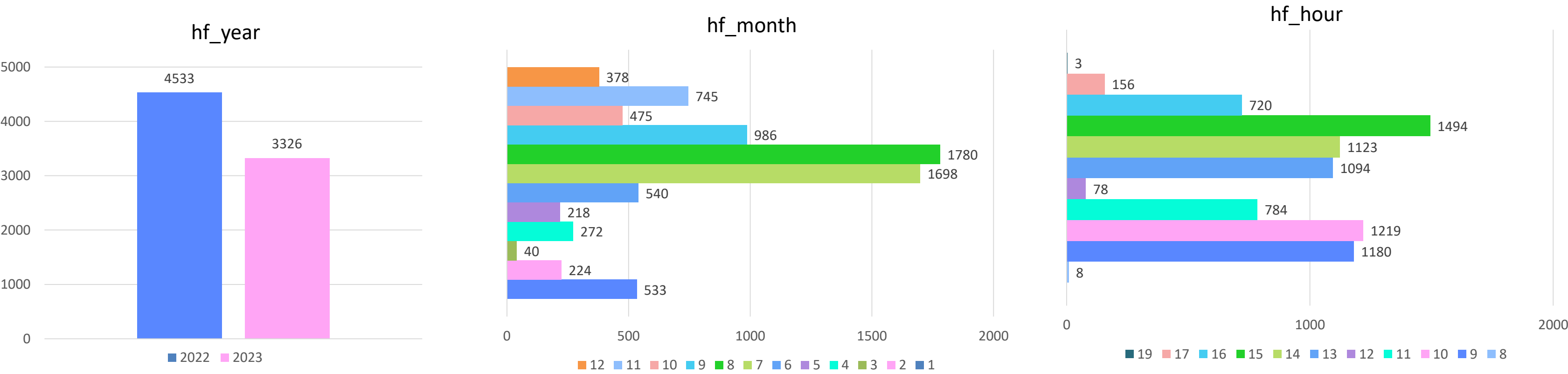
- handling\_fin\_user



# 05 데이터 EDA - PROCESSED DATA



05 데이터 EDA - PROCESSED DATA



06 데이터 CDA - PROCESSED DATA

Target : y 에 대한 독립성 검증

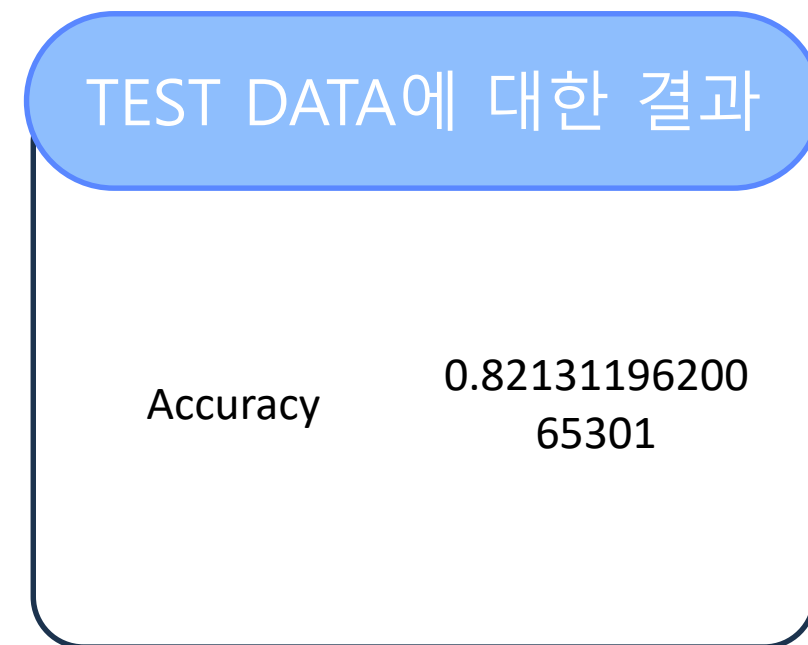
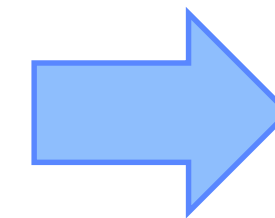
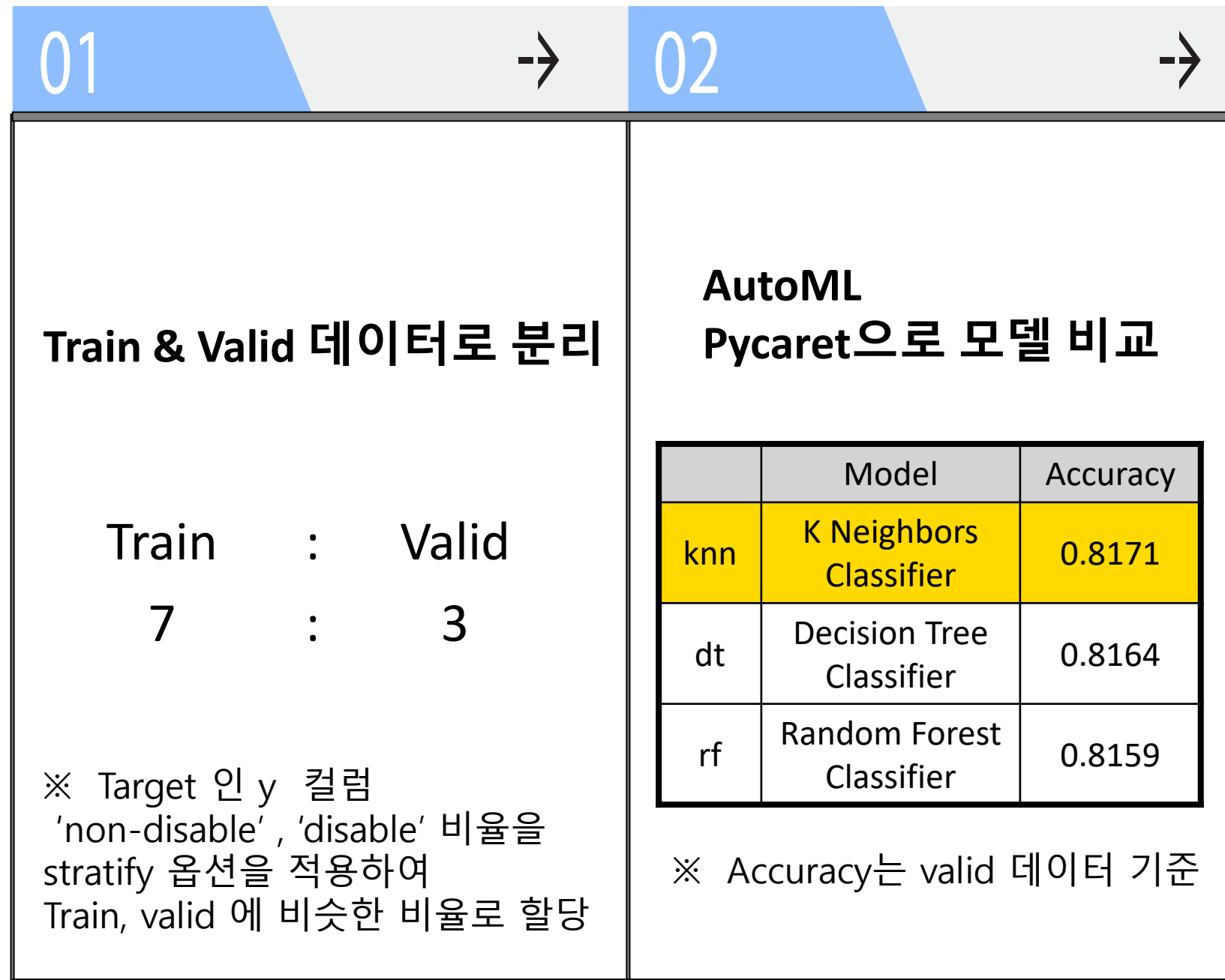
P-value < 0.05 독립성을 띄지 않는다

P-value >= 0.05 독립성을 띈다


컬럼명	P.Value	독립성 여부
Sop_id	0.4946964925654553	독립
Ticket_type	4.629968164630426e-53	독립x
handling_fin_user	4.7432416764471506e-86	독립x
rq_year	0.0	독립x
rq_month	0.0	독립x
rq_hour	1.9094715631868175e-77	독립x
hf_year	0.0	독립x
hf_month	0.0	독립x
hf_hour	1.9663931391617852e-80	독립x

- y와 독립성을 가지지 않는 컬럼

## 07 모델링

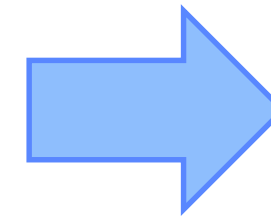


## 08 모델 하이퍼파라미터 튜닝



**AutoML**  
Pycaret으로 모델  
하이퍼파라미터 튜닝

KNeighborsClassifier	
algorithm	auto
leaf_size	30
metric	'manhattan'
metric_parameters	None
n_jobs	-1
n_neighbors	11
p	2
weights	distance



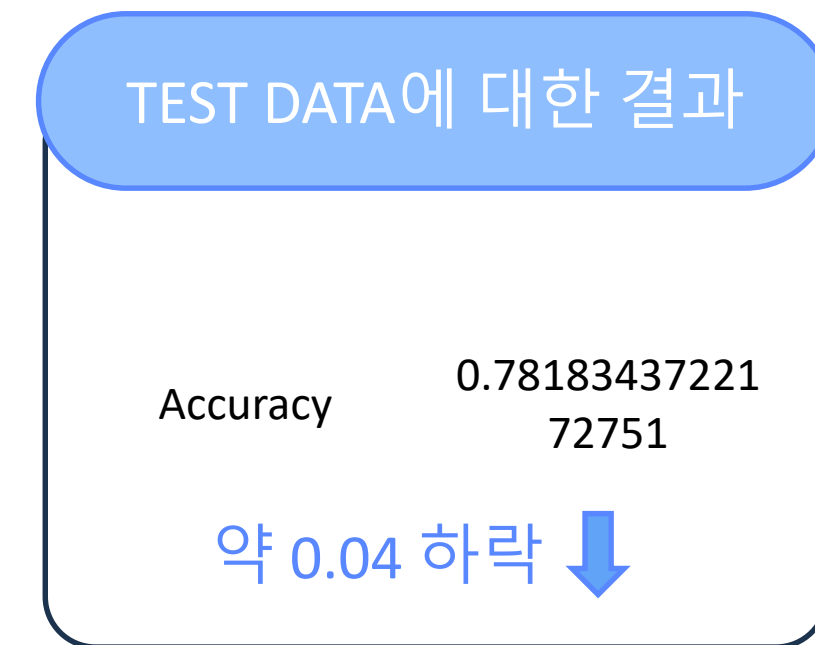
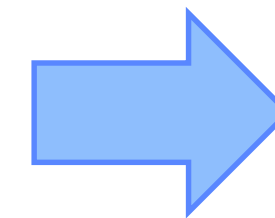
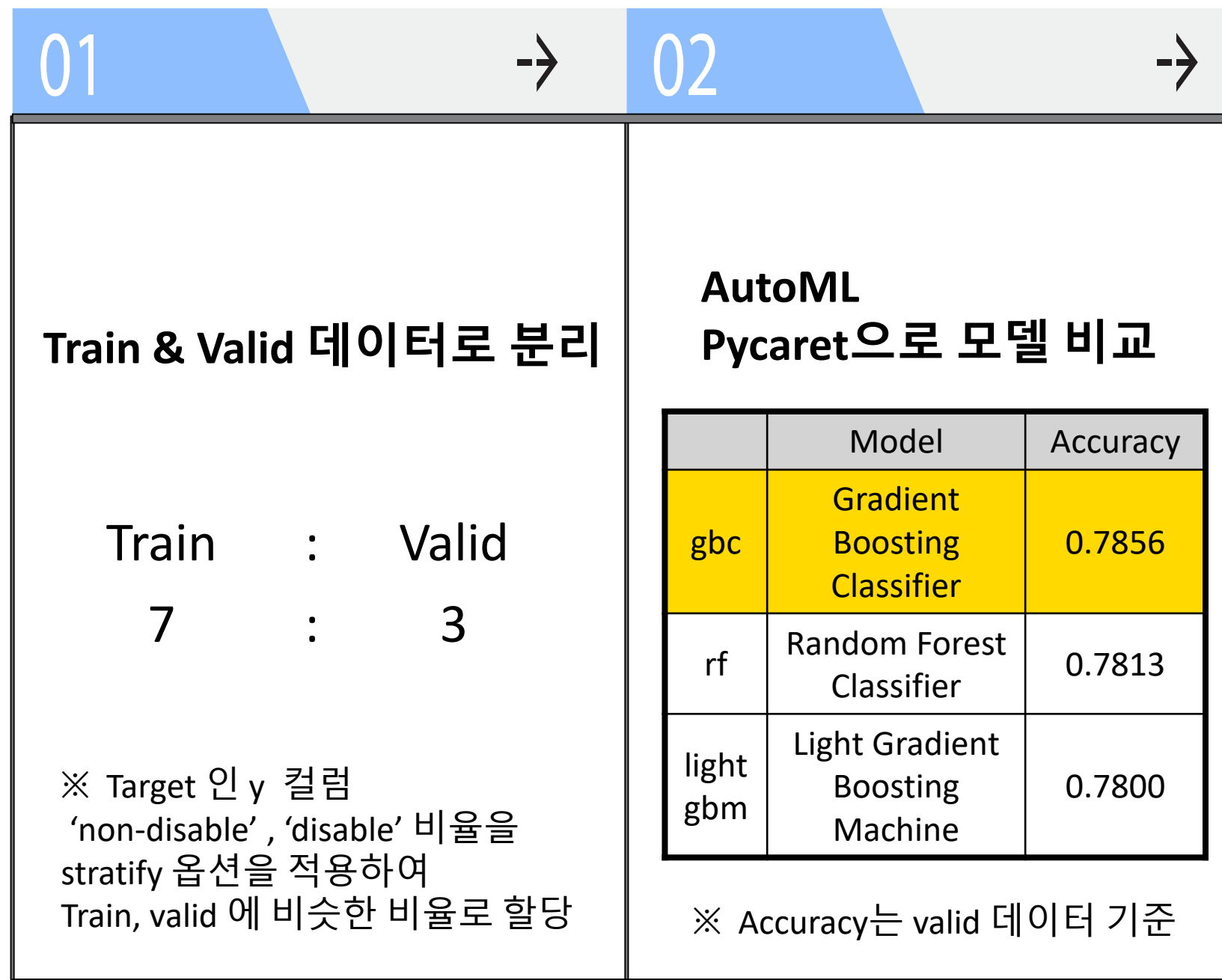
TEST DATA에 대한 결과

Accuracy 0.82962303354  
11101

약 0.008 상승 ↑

## 09 모델링

- sop\_id 컬럼 제외



## 10 결론

---

- 01    모델의 정확도는 0.83보다 낮음  
→ 하이퍼파라미터 튜닝 된 모델이 0.83에 가깝게 Accuracy가 나오기는 하지만 이 수치 이상으로 올라가지는 않음
  - 02    컬럼 'sop\_id' 제외하였을 때, 모델의 정확도는 다소 낮아졌음  
따라서 'sop\_id'가 중요한 예측 변수로 작용하고 있을 것으로 보임.  
→ 그러나 sop\_id는 전체가 unique한 값을 가진 컬럼이므로 왜 이렇게 작용하는지 이해하지 못함
  - 03    'y' 클래스에 대한 ' non-disable ' 와 ' disable ' 클래스의 분포가 균형을 이루는지 여부를 확인하고 데이터 분포가 균형을 이루지 않는다면,  
→ 각 클래스의 비율을 맞추고 모델링을 진행해볼 것
-

---

# 감사합니다

