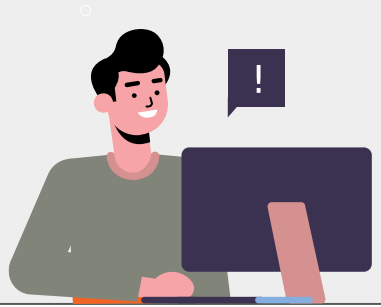
SECTION 2 PROJECT

인도 E-Commerce 플랫폼의 고객 이탈 예측 및 분석

AI 18TH 정희재

목차



01 개요 (Overview)

- 프로젝트의 수립 배경 및 목표

개요(Overview)

What's the 'Problem'?

1. 매출 하락

- 고객 이탈 발생은 곧 매출 감소로 이어짐

2. 마케팅 비용 증가

- 새로운 고객을 유치하기 위해 추가적인 마케팅 비용이 발생

3. 시장 점유율 하락

- 경쟁 업체로의 고객 이동으로 시장 점유율 하락

개요(Overview)

따라서,

- '기업'은 '고객'이 더 이상 서비스를 이용하지 않거나 '이탈'할 가능성이 높은 경우를 파악하고 예측해야 한다.

02 데이터 확인 및 정리(EDA)

(1) 데이터 확인 및 분석 방향

(2) 데이터 탐색 및 정리(EDA)

(1) 데이터 확인 및 분석 방향

데이터 출처

캐글(kaggle)의 2021년 E-commerce Customer 데이터셋 (QR 코드 참조)

데이터 선정 근거

1. 이커머스 분야에서 고객 이탈은 많은 기업에서 큰 문제

2. 고객 이탈 예측을 위한 다양한 변수들이 기재됨

3. 인도는 현재 전 세계에서 가장 빠르게 성장하는 이커머스 시장 중 하나*

- 2021년부터 2026년까지 연평균 25% 증가해 10조3290억 루피(약 1296억 달러) 규모로 성장할 것으로 예상
- 2020/21 회계연도에 1억 5000만 명의 온라인 쇼핑객을 보유하고 있으며 2025/26 회계연도에 3억 5000만 명으로 상승할 것으로 예상

* 코트라(KOTRA) 보도자료(2022.09.22) 참조
(https://dream.kotra.or.kr/kotranews/cms/news/actionKotraBoardDetail.do?SITE_NO=3&MENU_ID=180&CONTENTS_NO=1&bbsGbn=243&bbsSn=243&pNttSn=196900)

(1) 데이터 확인 및 분석 방향

데이터 확인

구성 : 5630 (행) x 20 (열)

: 온라인 쇼핑몰을 사용하는 5630명의 20개의 특성에 관해 기재된 데이터

주요 정보 (특성)

: 이탈 여부, 불만 여부 ,서비스 이용기간,
선호 결제 방법, 만족도 점수, 마지막 주문 이후 경과일 수,
지난 달에 사용한 쿠폰 수, 지난 달에 주문한 총 주문 수,
지난 달에 받은 캐시백 금액 등

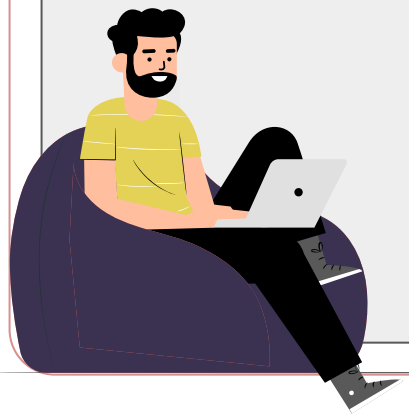
(1) 데이터 확인 및 분석 방향

데이터 분석 방향

1. 데이터셋을 살펴보면서 '고객 이탈'을 예측할 수 있는 중요한 특성이 무엇인지 알아보고자 한다.

2. 이를 위해, 타겟(종속 변수)을 'Churn' 이라는 특성으로 설정하고, 각 특성(독립 변수)간의 상관관계를 분석하면서 모델을 만들고자 한다.

3. 최종적으로 고객 이탈 여부를 예측하는 '분류'모델을 만들어서 성능을 비교 및 평가한 뒤 인사이트를 도출하고자 한다.



(2) 데이터 탐색 및 정리 (EDA)

전처리 ① 활용도가 없다고 판단한 컬럼('CustomerID') 삭제

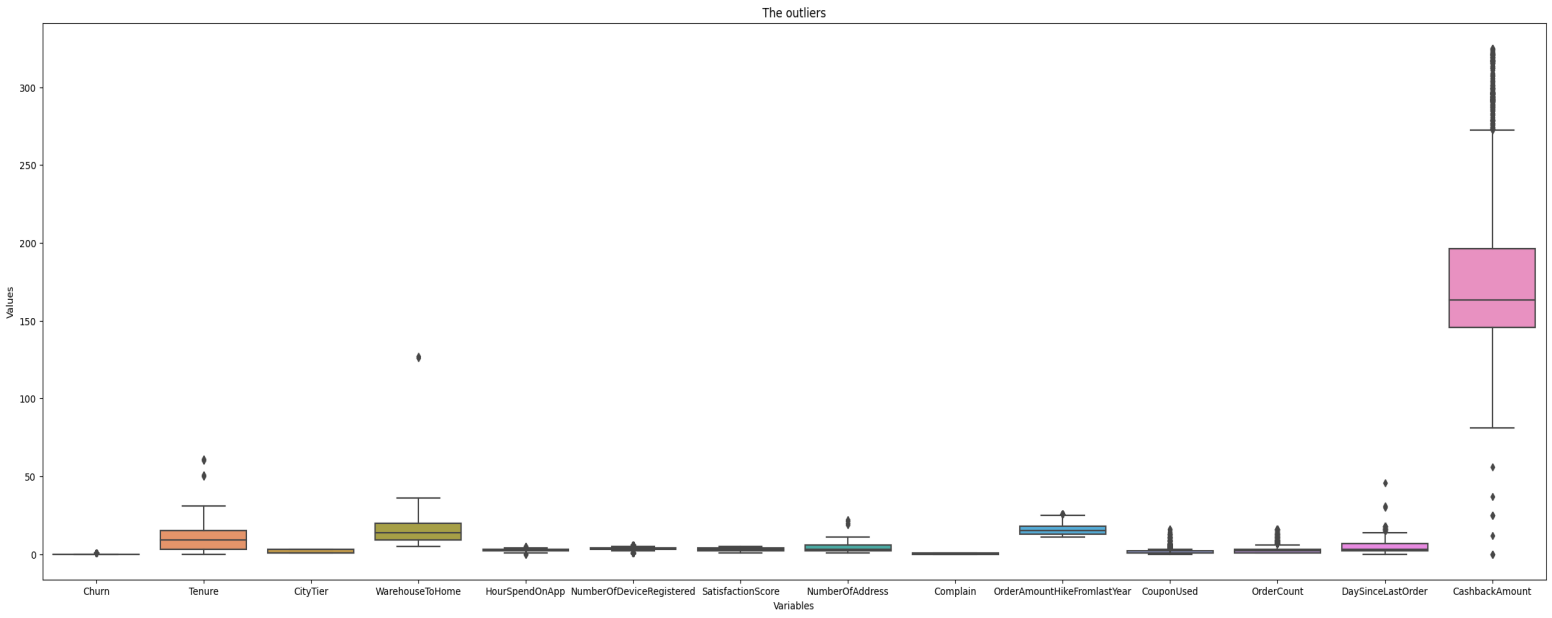
CustomerID
50001
50002
50003
50004
50005
...
55626
55627
55628
55629
55630

전처리 ② 결측값이 있는 특성 확인 후 중앙값(Median) 값으로 채움

	count	mean	std	min	25%	50%	75%	max
Tenure	5366.0	10.189899	8.557241	0.0	2.00	9.00	16.0000	61.00
CityTier	5630.0	1.654707	0.915389	1.0	1.00	1.00	3.0000	3.00
WarehouseToHome	5379.0	15.639896	8.531475	5.0	9.00	14.00	20.0000	127.00
HourSpendOnApp	5375.0	2.931535	0.721926	0.0	2.00	3.00	3.0000	5.00
NumberOfDeviceRegistered	5630.0	3.688988	1.023999	1.0	3.00	4.00	4.0000	6.00
SatisfactionScore	5630.0	3.066785	1.380194	1.0	2.00	3.00	4.0000	5.00
NumberOfAddress	5630.0	4.214032	2.583586	1.0	2.00	3.00	6.0000	22.00
Complain	5630.0	0.284902	0.451408	0.0	0.00	0.00	1.0000	1.00
OrderAmountHikeFromlastYear	5365.0	15.707922	3.675485	11.0	13.00	15.00	18.0000	26.00
CouponUsed	5374.0	1.751023	1.894621	0.0	1.00	1.00	2.0000	16.00
OrderCount	5372.0	3.008004	2.939680	1.0	1.00	2.00	3.0000	16.00
DaySinceLastOrder	5323.0	4.543491	3.654433	0.0	2.00	3.00	7.0000	46.00
CashbackAmount	5630.0	177.223030	49.207036	0.0	145.77	163.28	196.3925	324.99

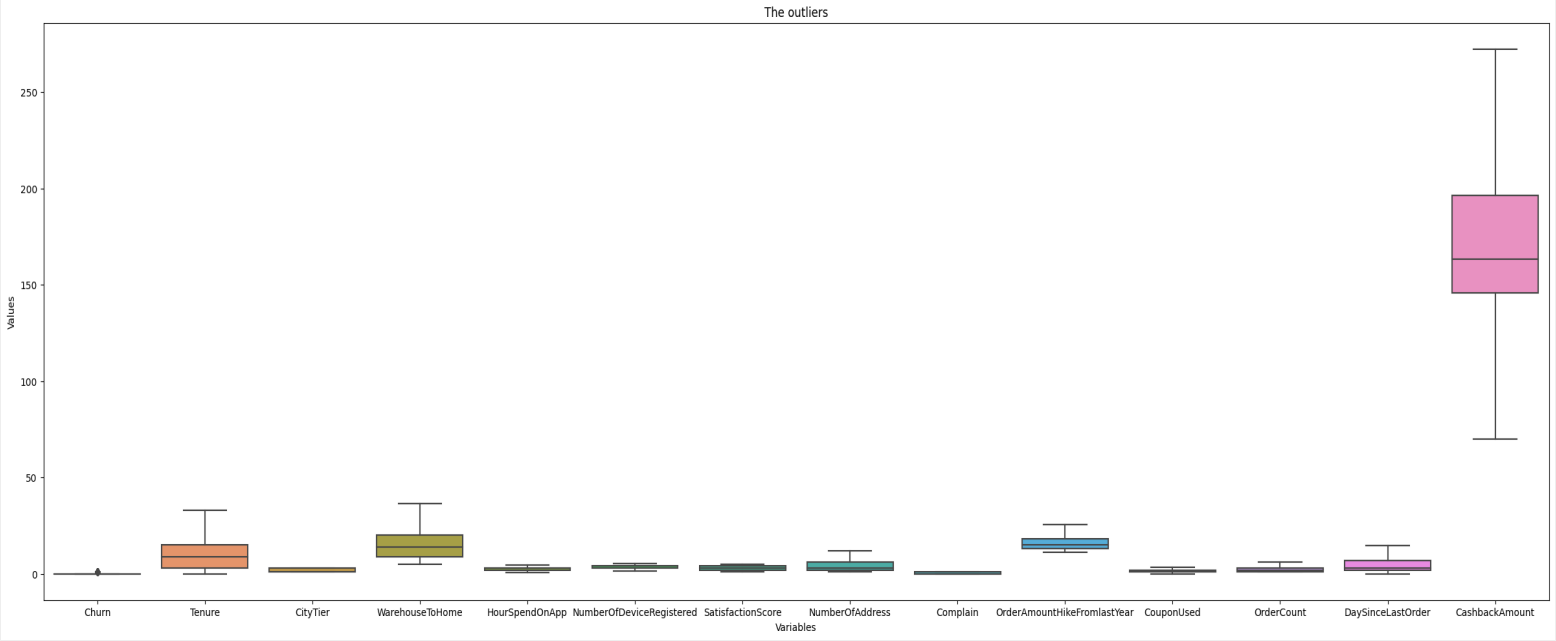
(2) 데이터 탐색 및 정리 (EDA)

전처리 ③ 이상치 확인 후 추후 모델링을 위해 삭제



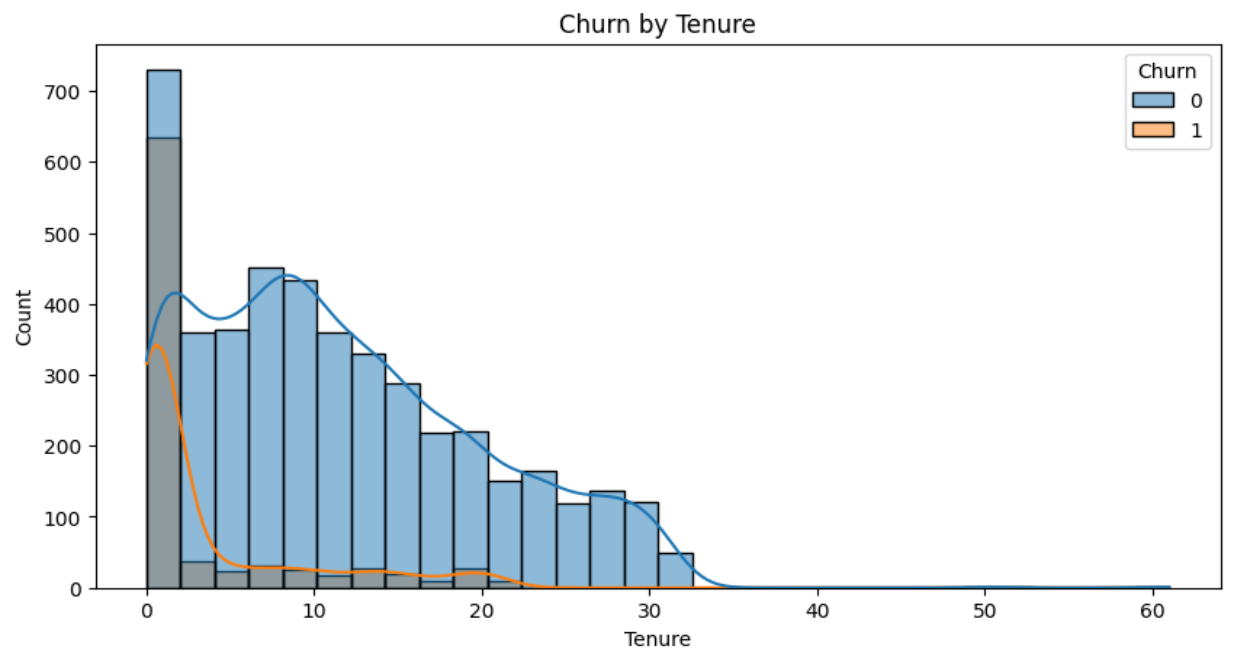
(2) 데이터 탐색 및 정리 (EDA)

전처리 ③ 이상치 확인 후 추후 모델링을 위해 삭제



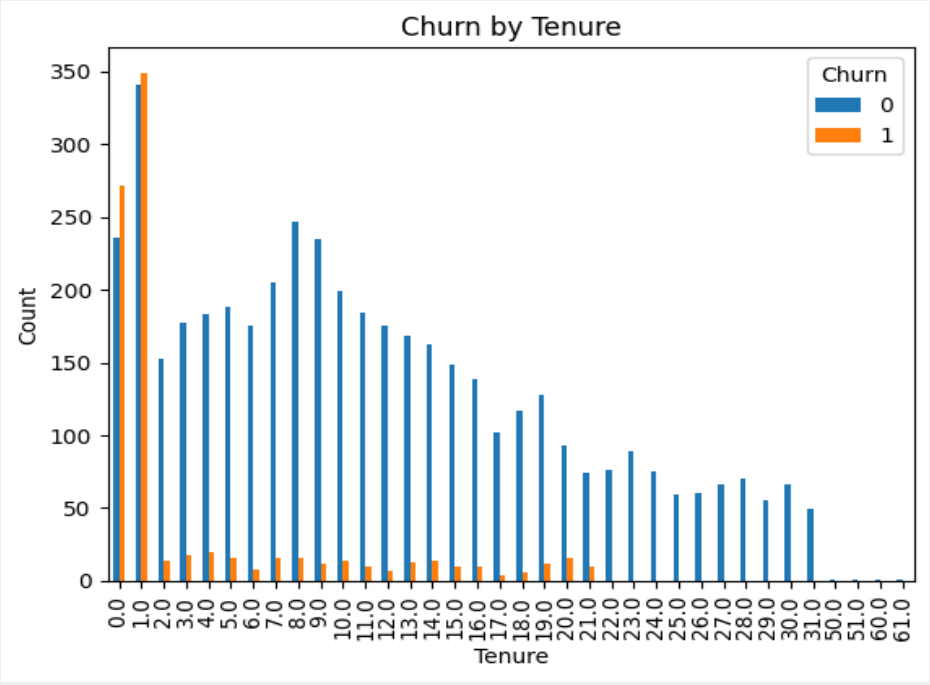
(2) 데이터 탐색 및 정리 (EDA)

가설 1. 플랫폼 이용 기간이 짧을 수록('Tenure') 고객 이탈 가능성이 높을 것이다.



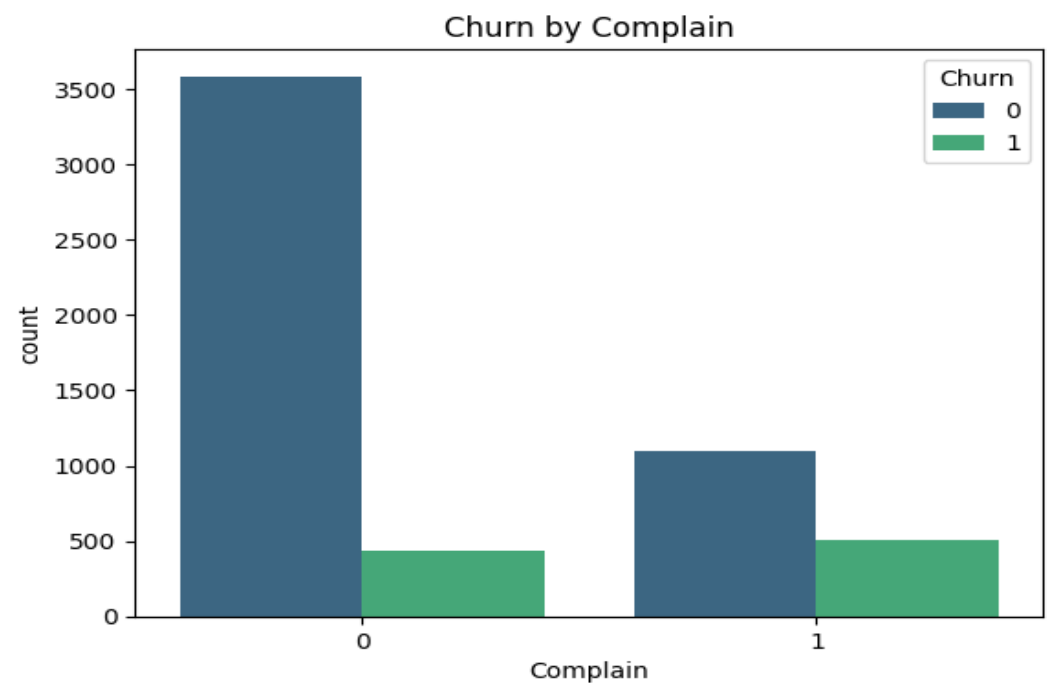
(2) 데이터 탐색 및 정리 (EDA)

가설 1. 플랫폼 이용 기간이 짧을 수록('Tenure') 고객 이탈 가능성이 높을 것이다.



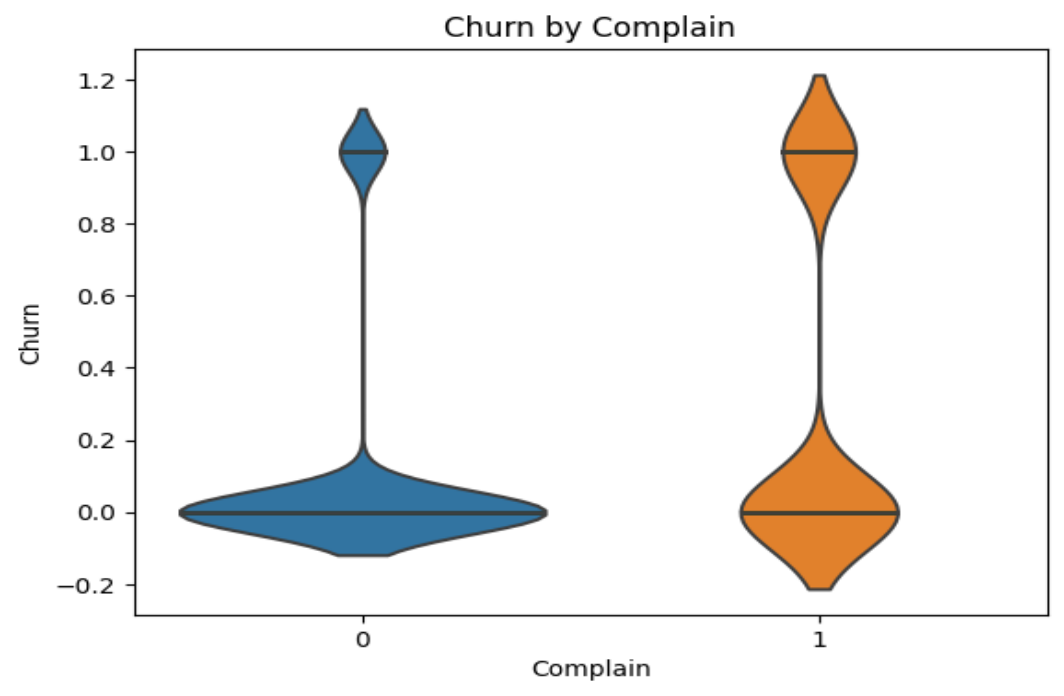
(2) 데이터 탐색 및 정리 (EDA)

가설 2. 불만('Complain')을 제기한 고객일수록 고객 이탈 가능성이 높을 것이다.



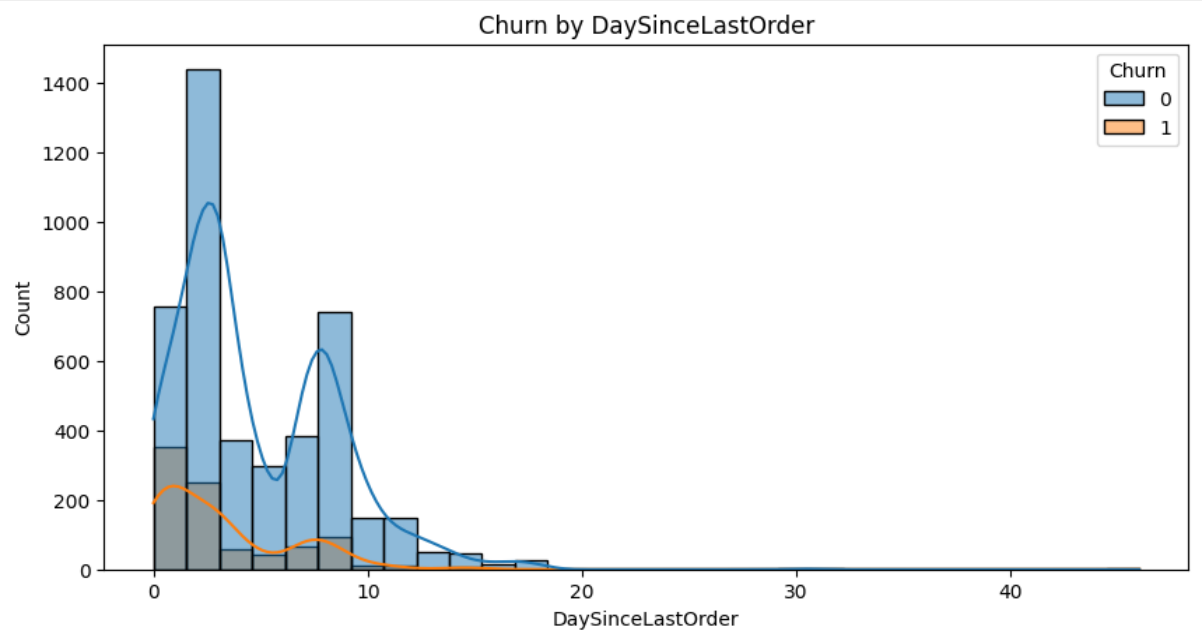
(2) 데이터 탐색 및 정리 (EDA)

가설 2. 불만('Complain')을 제기한 고객일수록 고객 이탈 가능성이 높을 것이다.



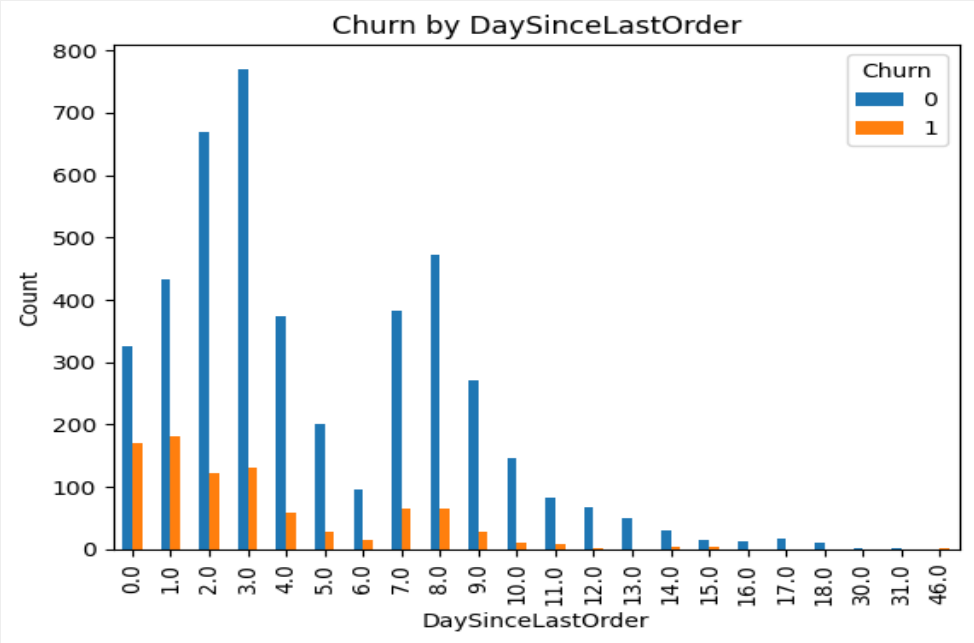
(2) 데이터 탐색 및 정리 (EDA)

가설 3. 마지막 주문 후 경과일 수('DaySinceLastOrder')가 길수록 고객 이탈 가능성이 높을 것이다.



(2) 데이터 탐색 및 정리 (EDA)

가설 3. 마지막 주문 후 경과일 수('DaySinceLastOrder')가 길수록 고객 이탈 가능성이 높을 것이다.



03 모델링 (ML Modeling)

(1) 기준 모델 선정

(2) 머신 러닝 (Machine Learning) 모델 비교

(3) 모델 성능 확인 및 평가

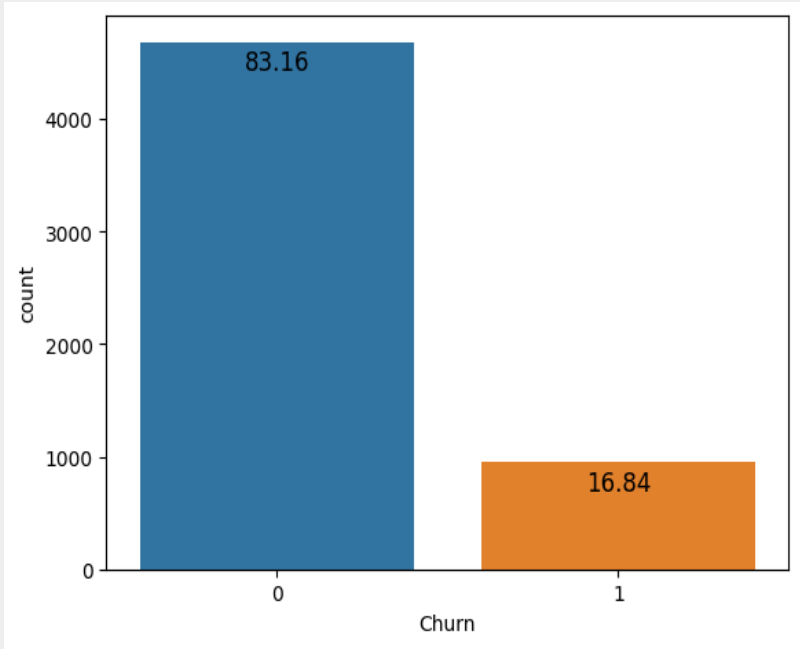
(1) 기준 모델 선정

❶ Feature Engineering

앱 사용시간 별로 주문량을 파악하기 위해
'OrderFrequencyOnApp' 특성 추가

OrderFrequencyOnApp
0.333333
0.333333
0.500000
0.500000
0.333333
...
0.666667
0.666667
0.666667
0.500000
0.666667

❷ 타겟('Churn') 분포 확인



(1) 기준 모델 선정

❸ 모델 성능 평가지표 선정

ⅰ. Recall (재현율)

- 실제 이탈한 고객 중에서 모델이 이탈을 제대로 예측한 비율

ⅱ. F1 Score

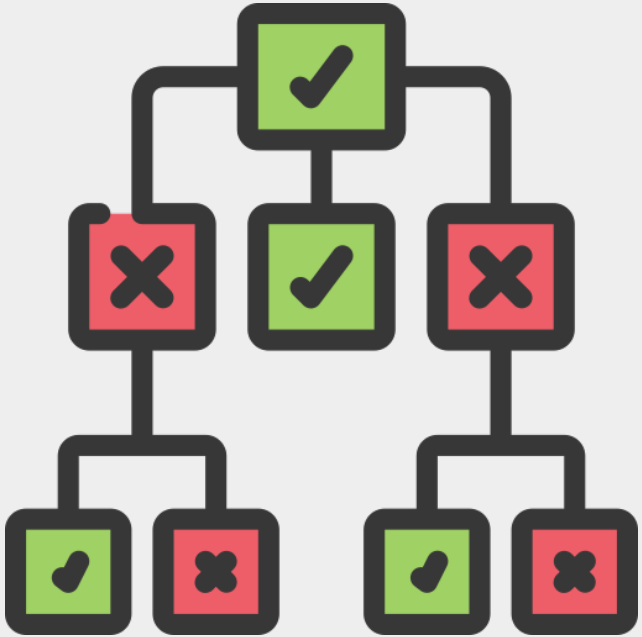
- 정밀도(Precision)와 재현율(Recall)을 모두 고려한 지표

ⅲ. Precision Recall Curve(Average Precision)

- X축은 Recall, y축은 Precision

- AP는 Precision-Recall Curve 아래 면적을 나타내는데 모델의 성능을 종합적으로 나타내며, 값이 높을수록 성능이 좋다고 할 수 있음

(1) 기준 모델 선정 : Decision Tree

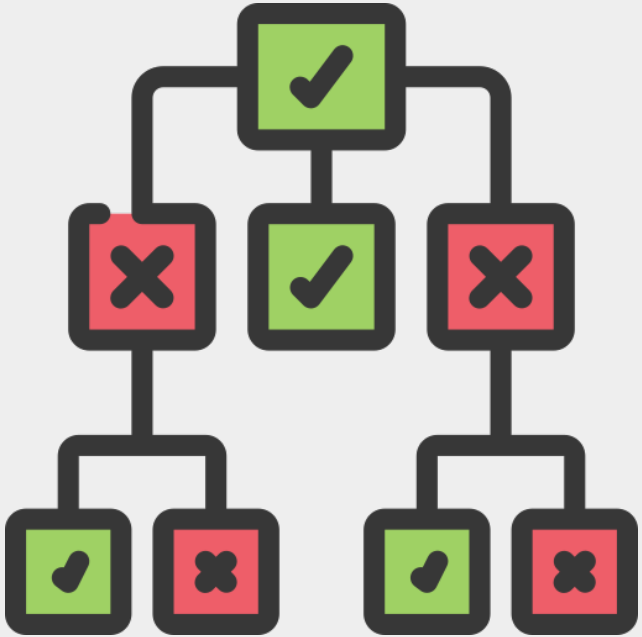


❹ 기준 모델 선정 (모델 성능 평가의 기준)

'Decision Tree' 모델 선택

- 트리(Tree) 형태의 의사결정 규칙 생성
- 맨 위에 위치한 노드는 'Root node', 맨 아래 위치한 노드는 'Leaf node'라고 함
- 최종적으로 'Leaf node'에 도달하면 해당 'Leaf node'에 속하는 클래스를 예측값으로 출력
- 트리가 깊어지면 규칙이 많아져 과적합에 취약

(1) 기준 모델 선정 : Decision Tree



❺ 기준 모델 최적화

'Decision Tree' 모델 'Tuning'

- 하이퍼파라미터(모델이 학습되기 전에 지정해주는 파라미터) 튜닝
- Max_depth(트리의 최대 깊이) 조정, Class_weight(타겟 불균형 조절) 조정 등
- 여러 가지 시도 후 최적의 파라미터 선정

(1) 기준 모델 선정 : Decision Tree

❻ 기준 모델 성능 평가 지표 결과

i . Recall (재현율)

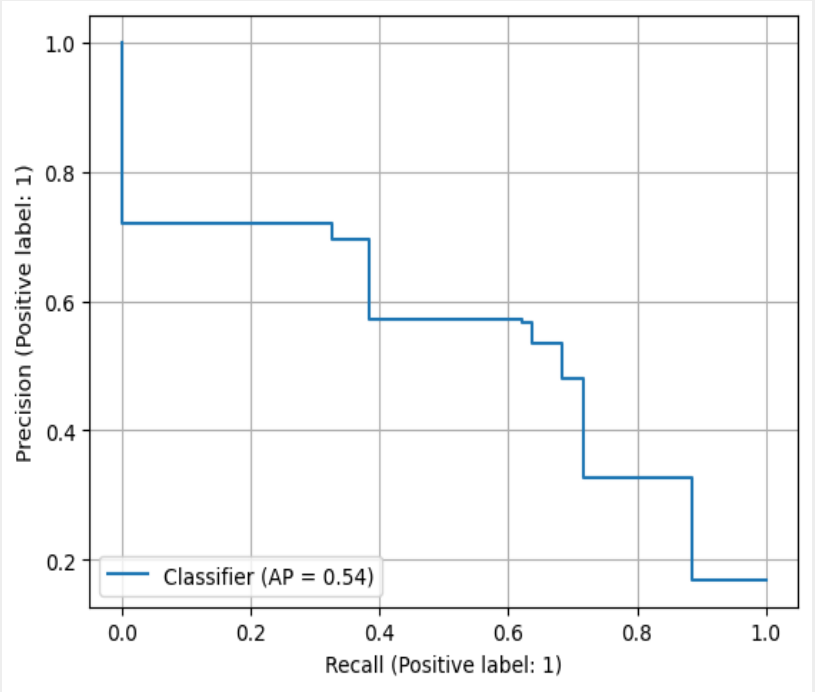
검증 Recall: 0.74

ii . F1 Score

검증 F1: 0.58

iii . Precision Recall Curve(AP)

검증 AP : 0.54



(2) 머신 러닝(ML) 모델 비교 : Logistic Regression

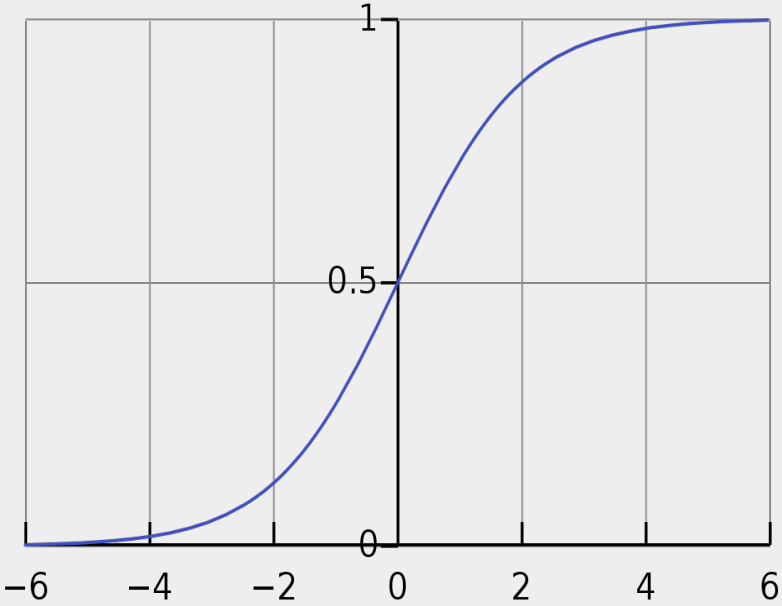
❶ 로지스틱 회귀모델 (Logistic Regression)

- 로지스틱 함수를 통해 확률값을 예측하는 모델

- 로지스틱 함수는 S자 모양의 곡선, 0과 1 사이의 값을 가짐

- 이진 분류에서는 0.5를 기준으로 확률값이 0.5보다 크면 1, 작으면 0으로 분류

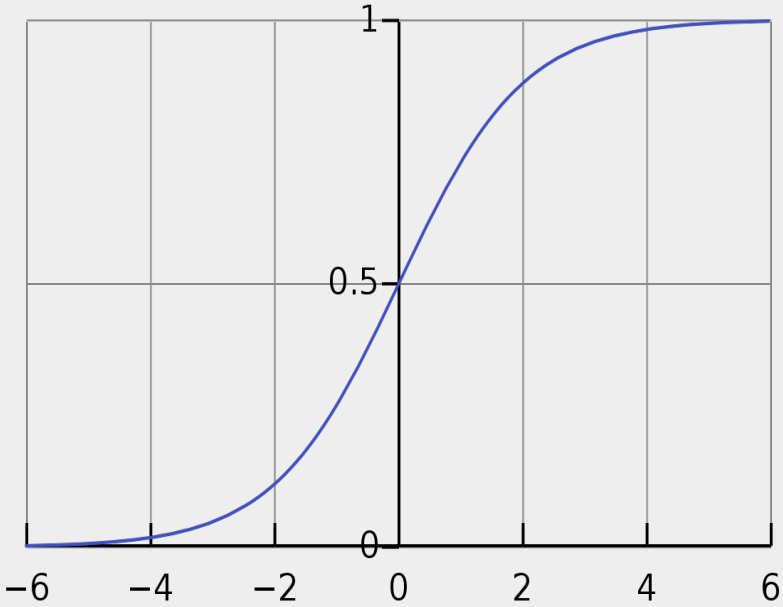
- 독립변수 간에 다중공선성(강한 상관관계)가 있다면 예측력 저하



(2) 머신 러닝(ML) 모델 비교 : Logistic Regression

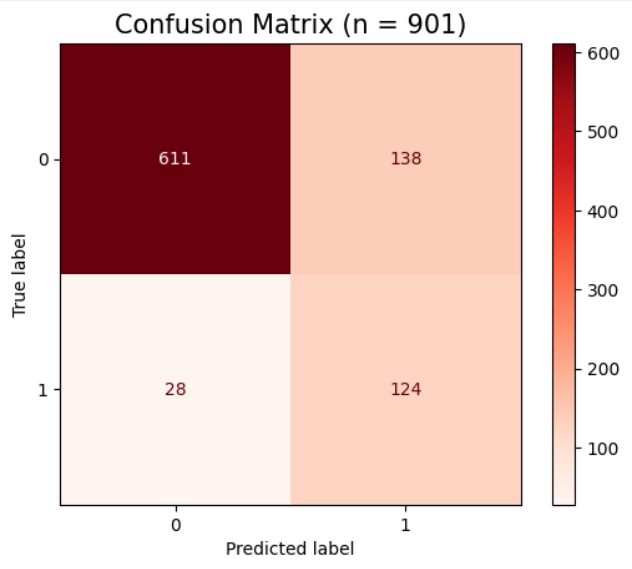
❷ 로지스틱 회귀모델 최적화 'Tuning'

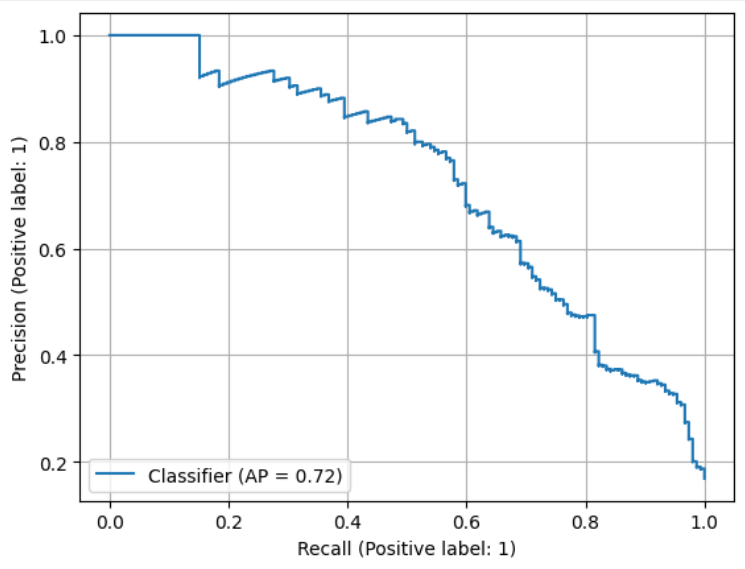
- 수치형 특성의 스케일링(표준화) 작업 실시
- 범주형 특성의 OneHot 인코딩 실시
- 클래스 불균형으로 인한 Class_weight 설정
- Grid Search CV를 통해 최적의 파라미터 조합 탐색 시도



(2) 머신 러닝(ML) 모델 비교 : Logistic Regression

❸ 로지스틱 회귀모델 성능 평가 지표 결과



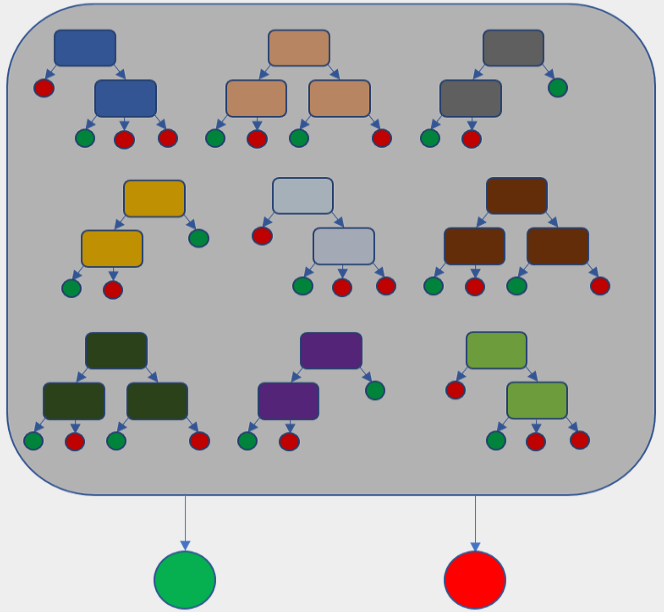


i . Recall (재현율) : 0.81, ii . F1 Score : 0.59, iii . Precision Recall Curve(AP) : 0.72

(2) 머신 러닝(ML) 모델 비교 : Random Forest

❶ 랜덤 포레스트 (Random Forest)

- 여러 개의 Decision Tree를 결합한 앙상블 기법 중 하나
- 전체 데이터셋에서 샘플을 무작위로 추출하여 다양한 Tree를 생성 후 결합
- Tree의 개수와 변수의 수가 많을 수록 계산 시간이 증가하여 대용량 데이터셋에서는 적합하지 않을 수 있다.

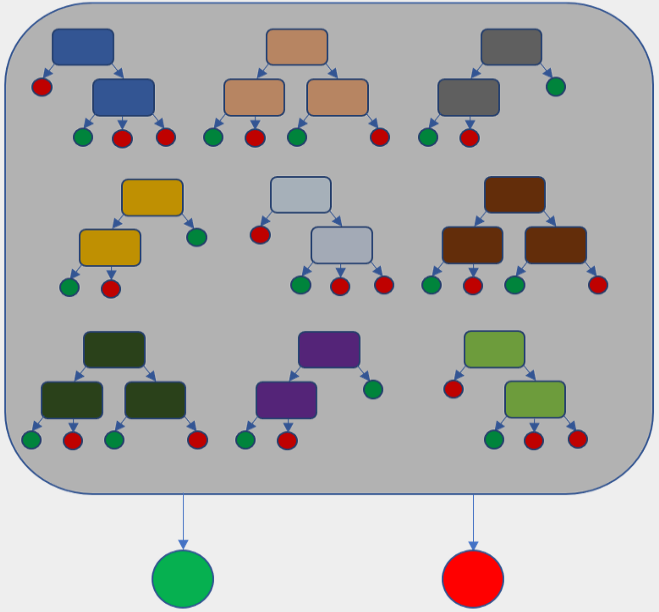


Random Forest

(2) 머신 러닝(ML) 모델 비교 : Random Forest

❷ 랜덤 포레스트 최적화 'Tuning'

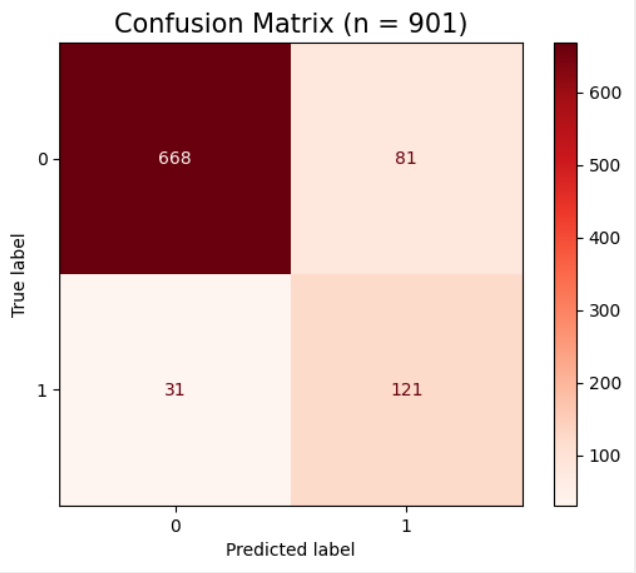
- 특성의 차원을 줄이기 위해 Ordinal 인코딩 실시
- 클래스 불균형으로 인한 Class_weight 설정
- Grid Search CV를 통해 최적의 파라미터 조합 탐색 시도

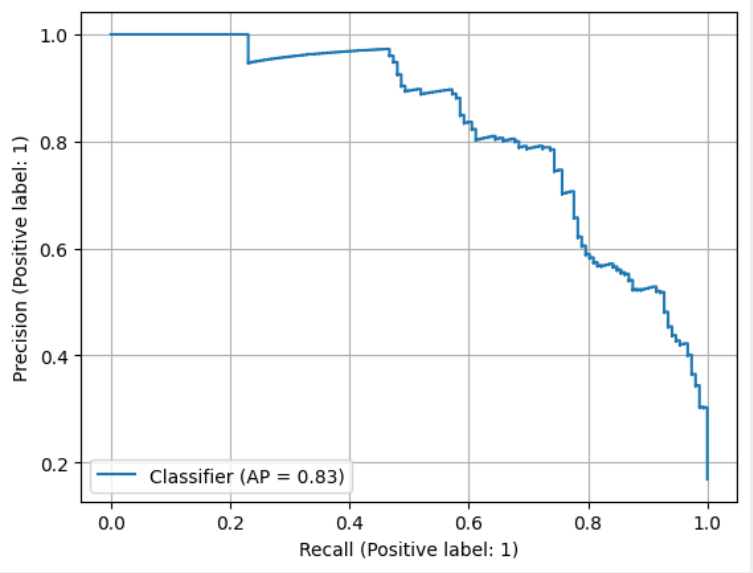


Random Forest

(2) 머신 러닝(ML) 모델 비교 : Random Forest

❸ 랜덤 포레스트 성능 평가 지표 결과



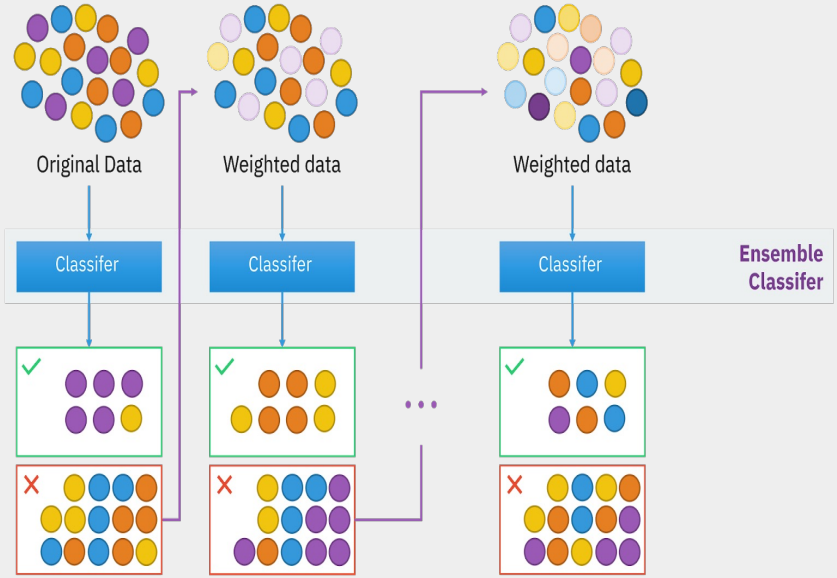


i . Recall (재현율) : 0.79, ii . F1 Score : 0.68, iii . Precision Recall Curve(AP) : 0.83

(2) 머신 러닝(ML) 모델 비교 : XGBoost

❶ XGBoost

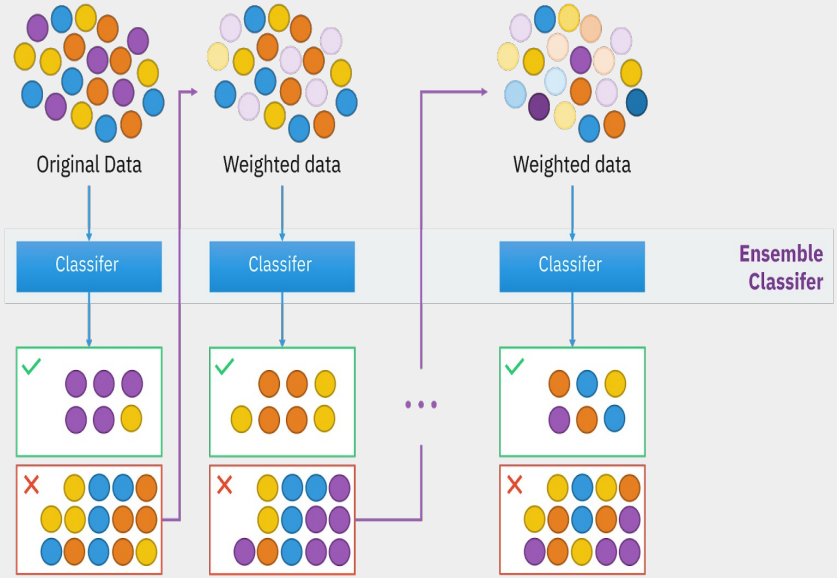
- Gradient Boosting 알고리즘을 기반으로 과적합 문제를 방지하는 기능(Early Stopping)들이 내장됨
- 예측값과 실제값의 잔차를 업데이트하며 Decision Tree 학습 과정 반복
- Gradient Boosting 보다 더 높은 예측 성능을 보이지만, 데이터가 충분히 크지 않은 경우 성능 차이가 크지 않을 수 있음



(2) 머신 러닝(ML) 모델 비교 : XGBoost

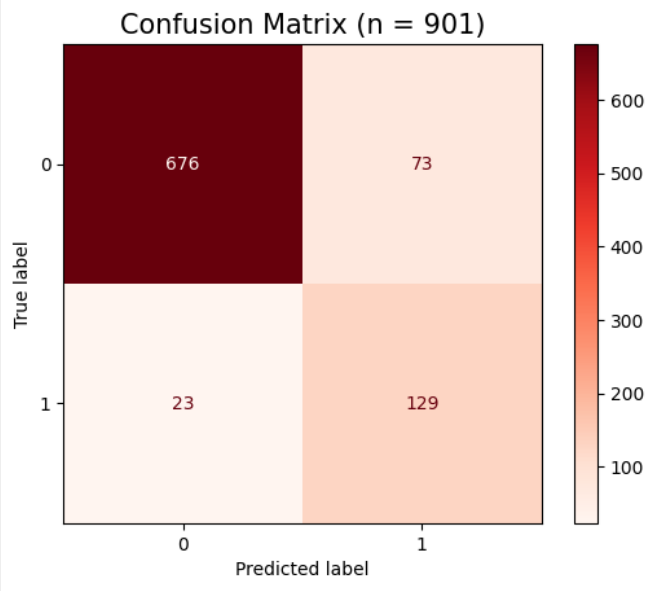
❷ XGBoost 최적화 'Tuning'

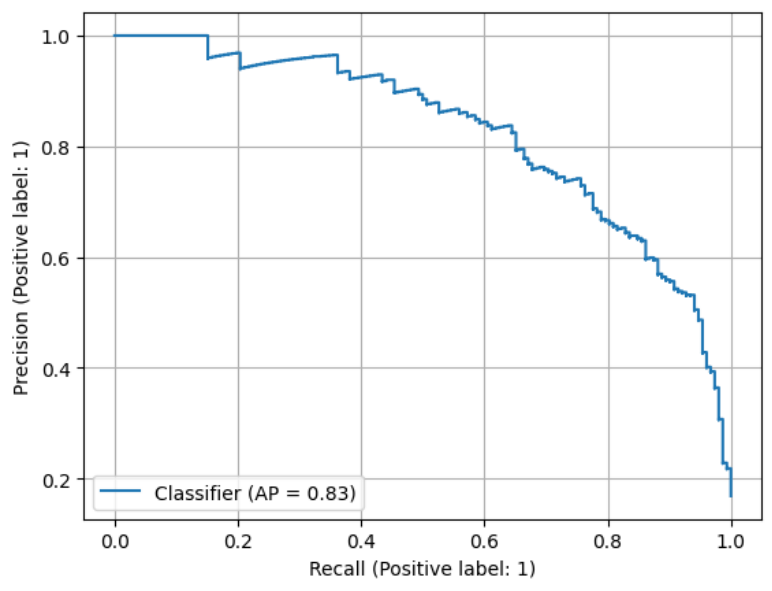
- 특성의 차원을 줄이기 위해 Ordinal 인코딩 실시
- 클래스 불균형으로 인한 Scale_pos_weight 설정
- Grid Search CV를 통해 최적의 파라미터 조합 탐색 시도



(2) 머신 러닝(ML) 모델 비교 : XGBoost

❸ XGBoost 성능 평가 지표 결과





i . Recall (재현율) : 0.84, ii . F1 Score : 0.72, iii . Precision Recall Curve(AP) : 0.83

(2) 머신 러닝(ML) 모델 비교 : 정리 및 평가

Ⅰ. 모델 성능 비교 정리

① 기준모델 (Decision Tree)

Recall: 0.74, F1: 0.58, AP : 0.54

② 로지스틱 회귀 모델

Recall: 0.81, F1: 0.59, AP : 0.72

③ 랜덤 포레스트 (Random Forest)

Recall: 0.79, F1: 0.68, AP : 0.83

④ XGBoost

Recall: 0.84, F1: 0.72, AP : 0.83

Ⅱ. 최종 모델 성능 평가

- XGBoost 선택

- # 기준 모델에 비해
- Recall : 13.5% (10/74) 향상
- F1 : 24.1% (14/58) 향상
- AP : 53.7% (29/54) 향상

Ⅲ. TEST SET 일반화 성능 평가

Recall: 0.87, F1: 0.72, AP : 0.81

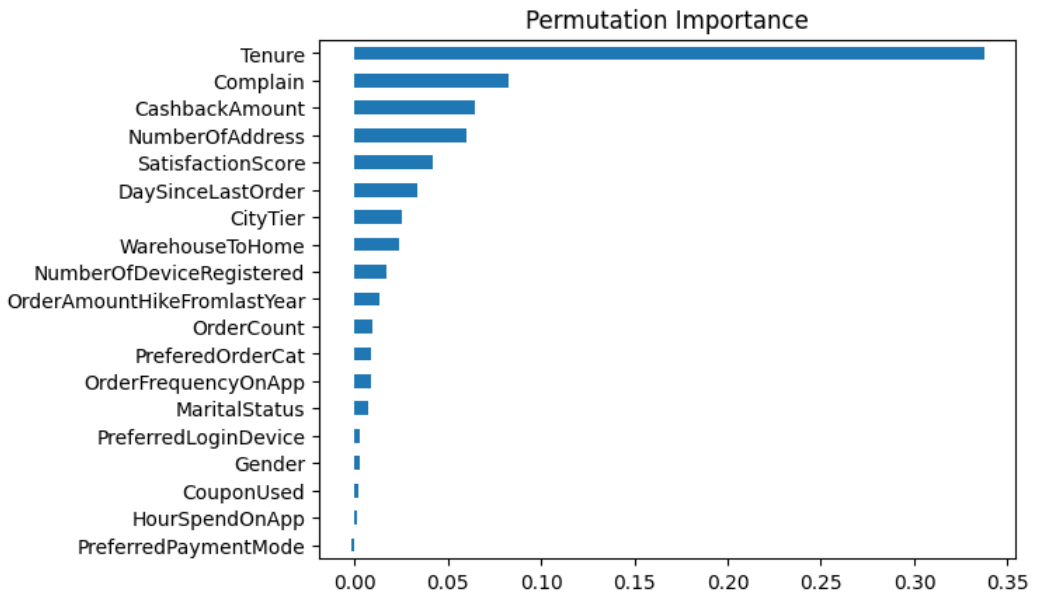
04 모델 해석

- Permutation Importance, PDP 활용

모델 해석 : Permutation Importance

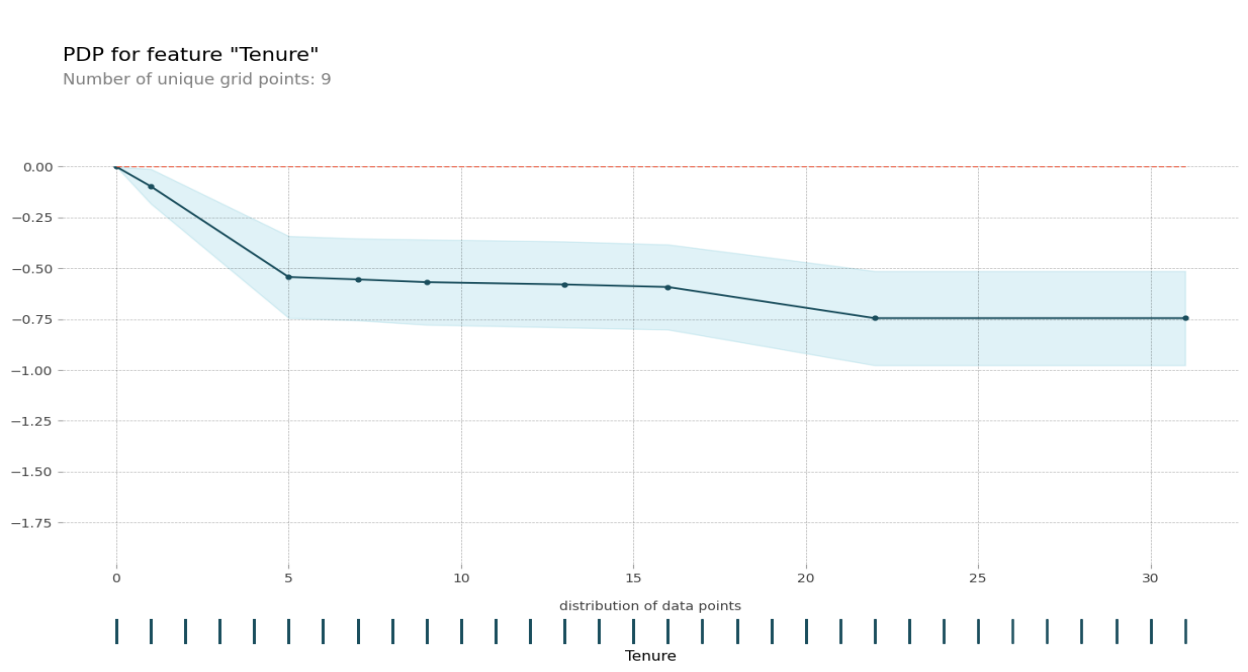
❶ Permutation Importance를 활용하여 특성 중요도 파악

Weight	Feature
0.3373 ± 0.0379	Tenure
0.0826 ± 0.0169	Complain
0.0650 ± 0.0209	CashbackAmount
0.0604 ± 0.0309	NumberOfAddress
0.0417 ± 0.0231	SatisfactionScore
0.0341 ± 0.0182	DaySinceLastOrder
0.0252 ± 0.0226	CityTier
0.0244 ± 0.0125	WarehouseToHome
0.0175 ± 0.0137	NumberOfDeviceRegistered
0.0133 ± 0.0141	OrderAmountHikeFromlastYear
0.0101 ± 0.0103	OrderCount
0.0093 ± 0.0158	PreferedOrderCat
0.0088 ± 0.0079	OrderFrequencyOnApp
0.0076 ± 0.0141	MaritalStatus
0.0032 ± 0.0101	PreferredLoginDevice
0.0032 ± 0.0063	Gender
0.0019 ± 0.0066	CouponUsed
0.0014 ± 0.0040	HourSpendOnApp
-0.0018 ± 0.0063	PreferredPaymentMode



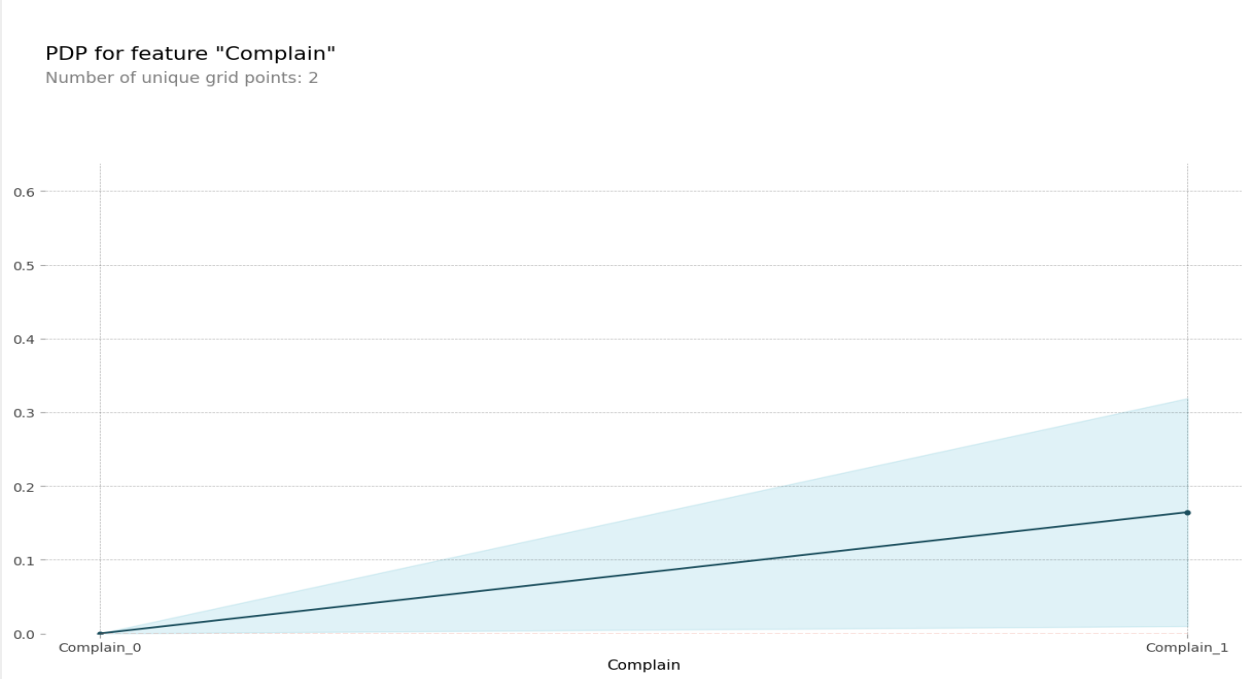
모델 해석 : Partial Dependence Plot

❷ 가설 1. 플랫폼 이용 기간이 짧을 수록('Tenure') 고객 이탈 가능성이 높을 것이다.



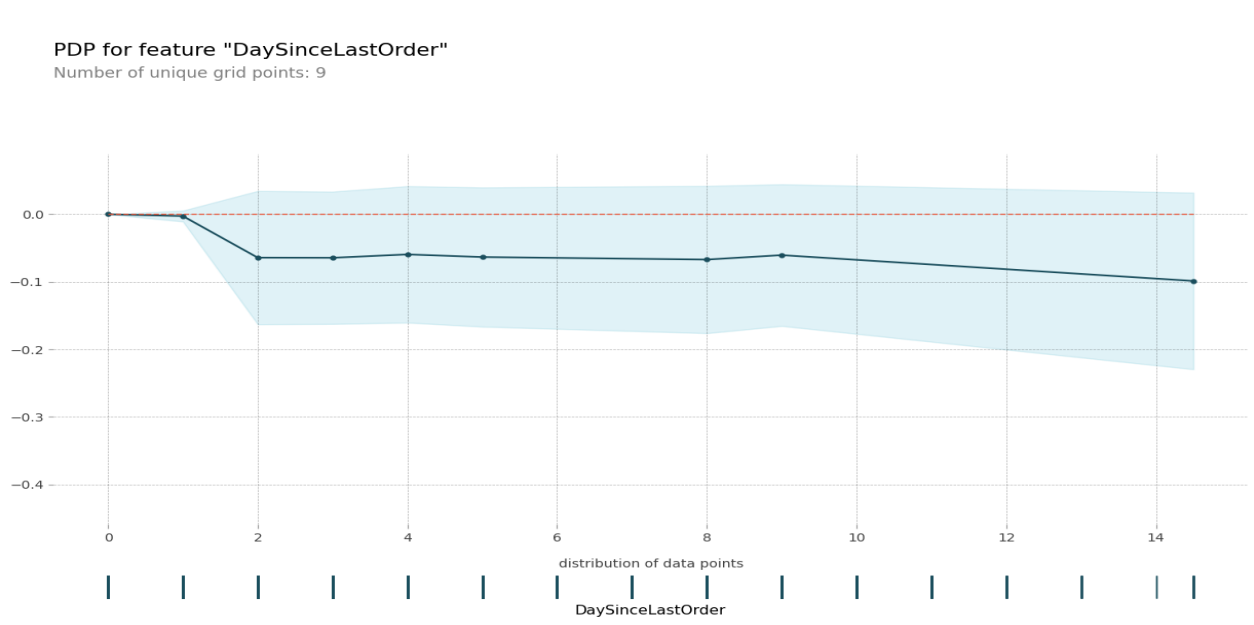
모델 해석 : Partial Dependence Plot

❸ 가설 2. 불만('Complain')을 제기한 고객일수록 고객 이탈 가능성이 높을 것이다.



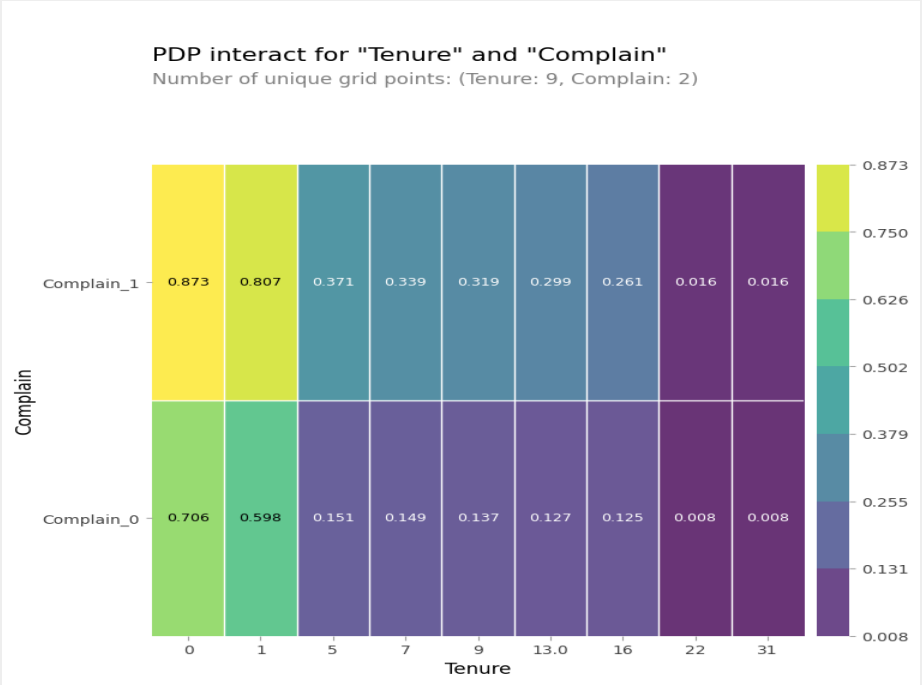
모델 해석 : Partial Dependence Plot

❹ 가설 3. 마지막 주문 후 경과일 수('DaySinceLastOrder')가 길수록 고객 이탈 가능성이 높을 것이다.



모델 해석 : Partial Dependence Plot

❺ 특성 중요도가 가장 높은 두 특성('Tenure', 'Complain')의 Interact 파악



05 결론 (Conclusion)

- 관점별 분석 및 한계 제시 I

결론 : 관점별 분석 및 한계

Ⅰ. 데이터 관점

① 데이터셋이 수집된 기간과 정확한 날짜를 파악할 수 있는 데이터 변수 추가 필요
② 고객의 연령대 데이터 변수 추가 필요
③ 특성 중요도가 비교적 높은 'CashbackAmount', 'NumberOfAddress' 컬럼을 분석하지 못함

Ⅱ. 모델링 관점

① 클래스 불균형 문제를 해결하기 위한 여러 방법을 시도해보지 못함 (Class_weight만 시도)
② 모델에 맞는 다양한 파라미터를 사용해보지 못함
③ 최적의 파라미터 조합을 찾기 위한 여러 방법을 시도해보지 못함 (Grid Search만 시도)

Ⅲ. 비즈니스 관점

① 정확히 어느 온라인 쇼핑몰인지 알지 못해 현재 기업에 맞는 비즈니스 전략을 세우는데 한계가 있음
② 가설 1, 가설 2는 EDA를 통한 시각화와 특성 중요도 및 PDP 분석을 통해 유의미한 결과를 얻었다고 판단함
③ 예측 모델의 비즈니스 가치를 어떻게 나타낼지 보완 필요

감사합니다.