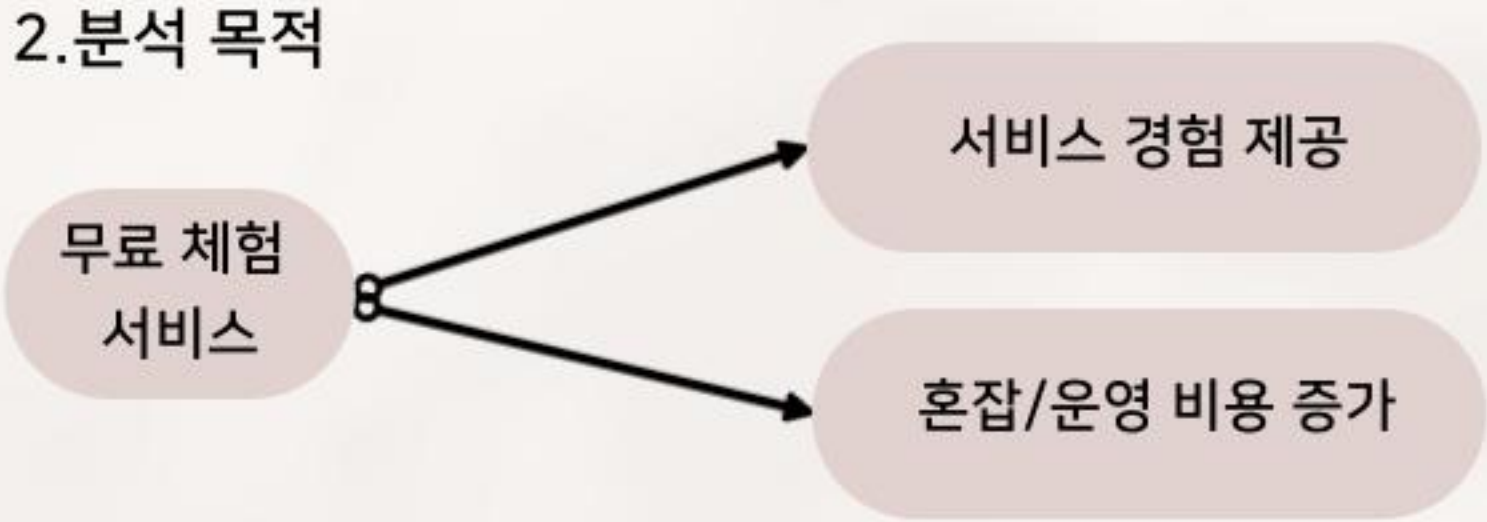
무료체험 사용자 수요 예측 모델링

1.데이터

무료체험 이용자 로그 데이터(등록, 방문, 출입 로그, 결제 여부) (3일)

2.분석 목적



유저에게 우리 서비스를 경험시켜 결제로 전환시키는 중간 단계

체험자의 수가 과도할 경우, 기존 고객의 경험을 저해할 수 있는 리스크

사용자 로그 데이터를 기반으로 익일 재방문 가능성을 예측하는 모델

체험자 유입 이후의 행동을 분석하고, 익일 재방문 가능성을 예측함으로써 효과적인 마케팅 및 운영 전략의 기준을 마련

3. 데이터 프레임 생성

방문자 로그 데이터를 기반으로 다음과 같은 피처를 생성하였습니다.

변수명	설명
days_from_register	등록일로부터 경과한 일수 (0, 1, 2 등)
is_weekend	기준 날짜가 주말인지 여부 (1: 주말, 0: 평일)
prev_1day_visited	하루 전 방문 여부 (1: 방문, 0: 미방문)
prev_2day_visited	이틀 전 방문 여부 (1: 방문, 0: 미방문)
visited_today	해당 날짜에 방문 여부 (1: 방문, 0: 미방문)
stay_duration_minutes	해당 날짜의 총 체류 시간 (분 단위, 없으면 0)
visit_next_day	익일 방문 여부 (타겟 변수, 1: 방문, 0: 미방문)

트라이얼 수요 예측 모델링

4. 모델링

베이스 모델

BASELINE 모델로 로지스틱 회귀 분석을 적용하였습니다.
입력 피처들과 익일 재방문 확률 간의 관계를 로지스틱 함수를 통해 설명하는 선형 모델

A 안

체류 시간을 계산할 때, 가장 이른 로그가 입실이고, 가장 마지막 로그가 퇴실인 사용자만을 필터링한 후, 퇴실 시각에서 입실 시각을 뺀 값으로 계산하는 방식
외출이나 중간 로그 누락 여부는 고려하지 않았습니다.

B 안

체류 시간의 정확도를 높이기 위해 입실과 퇴실 로그가 정확히 순서대로 짝을 이루는 경우만 필터링하여 계산하는 방식
이 경우 외출 및 재입실 패턴까지 고려할 수 있어 보다 정교한 체류 시간 계산이 가능
로그 쌍이 완벽하게 일치하는 사용자만을 대상으로 하기 때문에 데이터 손실이 발생

5,745명 중 3,715명
분석에 포함

항목	A안	B안
데이터 수	1,423명	1,011명
Accuracy	0.64	0.72
Precision (클래스 1)	0.37	0.44
Recall (클래스 1)	0.3	0.2
F1-score (클래스 1)	0.33	0.27
ROC AUC Score	0.703	0.713

- 전체 정확도(ACCURACY)가 더 높음 (0.72 VS. 0.64)
- ROC AUC 점수도 더 우수 (0.713 VS. 0.703)
- 무엇보다 데이터 정합성과 신뢰성이 높다는 점에서
- 모델이 예측하는 "재방문 가능성"의 품질이 더 우수하다고 판단됩니다.
- 이에 따라, 정확한 체류 시간 기반의 신뢰성 있는 예측을 위해 B안을 최종 선택

트라이얼 수요 예측 모델링

5. 피처 확장

날씨 정보: 해당 날짜의 기온, 강수량 등 외부 환경 요인을 반영
사용자 행동 클러스터링: 체류 패턴 및 방문 이력을 기반으로 유사 행동 유형 그룹화
지점별 구분: 공간적 특성을 고려한 지점 단위 피처 추가

6. 다양한 알고리즘 성능 비교

- 로지스틱 회귀 분석 (Baseline)
  - 가장 단순하고 해석이 쉬운 선형 모델로, 입력 피처와 재방문 확률 간의 관계를 로지스틱 함수로 설명
- 서포트 벡터 머신 (SVM)
  - N차원 공간에서 각 클래스 간의 경계를 최대화하는 초평면을 찾아 분류하며, 이진 분류 문제에 강점을 가짐
- 랜덤 포레스트
  - 다수의 결정 트리를 무작위로 구성해 예측을 앙상블하는 방식으로, 변수 간 비선형 관계와 상호작용을 효과적으로 포착
- 부스팅 모델 (LightGBM, XGBoost)
  - 이전 모델의 오차를 보완하며 트리를 순차적으로 학습시키는 방식으로, 높은 예측 성능과 정교한 학습이 가능함

모델	Test ROC AUC	Accuracy	Positive(1) Precision	Positive Recall	Positive F1	Negative (0) Precision	Negative Recall	Negative F1
로지스틱 회귀	0.7777	0.7061	0.5621	0.8359	0.6722	0.8725	0.6329	0.7336
랜덤 포레스트	0.8077	0.7335	0.5926	0.8342	0.693	0.8786	0.6767	0.7646
LightGBM	0.8248	0.7347	0.6114	0.7254	0.6635	0.827	0.74	0.7811
서포트 벡터 머신	0.7765	0.7098	0.5574	0.9482	0.702	0.9517	0.5755	0.7172

7. 결론

- 랜덤 포레스트
  - F1(Positive) 0.6930으로 불균형 데이터에서 양호한 균형을 보였음.
  - 안정적인 성능과 피처 중요도(예: days_from_register, prev_1day_visited)를 통해 해석 가능성이 용이했음.
- 서포트 벡터 머신
  - 이진분류에 강한 알고리즘인 만큼 Positive 리콜이 0.9482로 매우 높아 재방문자를 놓치는 비율이 낮았음.
  - 그러나 Precision(0.5574)이 낮아, 많은 오탐(False Positive)이 발생할 수 있으므로 모델 조정이 필요.
- LightGBM
  - 전체 ROC-AUC가 가장 높아 전반적인 순위 결정 능력이 우수하므로 최종 모델로 선택함.
  - 다만 Positive 클래스 리콜(0.7254)이 SVM, 랜덤 포레스트 대비 다소 낮아, 재방문자(Positive) 탐지율을 높이기 위해 Recall을 높이는 모델 검토 필요.