Project 03

공유오피스 체험 신청자의 유료 전환 예측 프로젝트

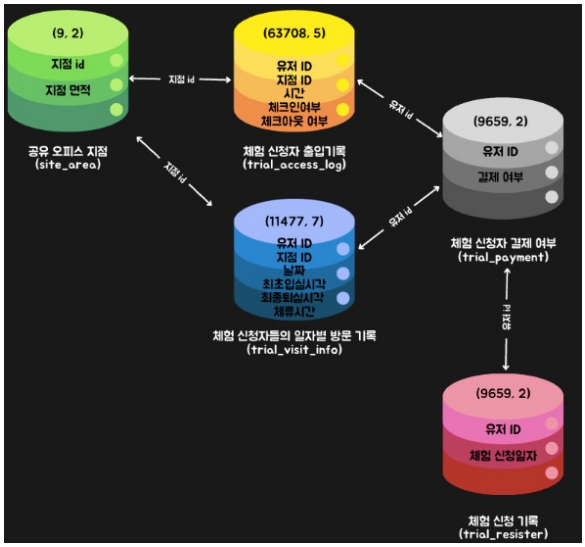
(2025.06.10 - 2025.06.19)

분석배경

- 공유오피스는 단순한 공간 임대에서 일하는 방식을 지원하는 플랫폼으로 확장하며 성숙기에 접어들었고, 그만큼 과열경쟁이 공존하는 상황
- 개인 고객을 대상으로 무료 체험권 제공하고 있으나, 유료 결제 전환율이 38%로 낮은 상황임
- 유료 전환 가능성이 높은 타깃군을 식별하여 맞춤형 마케팅과 리텐션 전략을 설계가 필요함

프로세스

활용 데이터 공유오피스 3일 체험 신청자의 로그 데이터



데이터 탐색을 통해 생성한 피처

구분	열	설명
체크인	Mean_checkins_level	일 평균 체크인 수 범주형 변수
	Mean_checkins_per_day	일 평균 체크인 수
	Frequent_checkins	일 평균 체크인 수가 3 이상인지 여부
	Total_checkins	총 체크인 수
체류시간	Mean_stay_time_hour	일 평균 체류시간
	Mean_stay_time_level	일 평균 체류시간 범주형 변수
	Log_stay	일 평균 체류시간이 4시간 이상인지 여부
	Stay_time_hour	총 체류 시간
	days	방문 일수
방문 시간	First_visit_hour	첫 방문 시간
	First_visit_month	첫 방문 월
	First_visit_q	첫 방문 분기
	First_visit_season	첫 방문 계절
	High_convert_month	첫 방문 월이 전환율 높은 월인지 여부 - 전환율 높은 월 : 5,6,10
	First_visit_dayofweek	첫 방문 요일
	First_visit_is_weekend	첫 방문 요일이 주말인지 여부
	high_convert_dayofweek	첫 방문 요일이 전환율 높은 요일인지 여부 - 전환율 높은 요일 : 화요일, 일요일
	Peak_top5	높은 혼잡대 시간 방문 여부
	Evening_visit	저녁 시간대 방문 여부
	Visit_delay_days	체험 신청 후 체크인 지연 일
타겟 변수	Is_payment	결제 여부

분석결과

LogisticRegression

RandomForestClssifier

GridSearchCV를 통한 하이퍼파라미터 튜닝

▪ Base Line Model : LogisticRegression

- 해석이 직관적이고 명확하여 이후 모델과 성능을 비교하기 위한 기준으로 적합하다고 판단되는 LogisticRegression 모델을 베이스 라인 모델로 선정

지표	값	비고
Accuracy	0.6337	
Precision	0.5545	
Recall	0.2563	양성 클래스 탐지 성능에 제한적
F1_Score	0.3506	양성 클래스 탐지 성능에 제한적
ROC_AUC	0.6019	

결과 및 한계점

Logistic Regression은 해석에는 유리하지만
선형적인 관계만 반영하는 한계 존재

▪ 2차 선정 모델 : RandomForestClassifier

- 변수 간 복잡한 상호작용과 비선형 구조를 반영할 수 있는 Random Forest 모델을 적용하여 보다 정교한 예측을 시도

지표	값	차이 (Base 모델 기준)
Accuracy	0.6251	- 0.0086
Precision	0.5195	- 0.0350
Recall	0.3739	+ 0.1176
F1_Score	0.4349	+ 0.0843
ROC_AUC	0.6308	+ 0.0289

결과

재현율, F1_Score, ROC_AUC가 모두 상승

▪ GridSearchCV를 통한 RandomForestClassifier 하이퍼파라미터 튜닝

지표	값	차이 (Base 모델 기준)	차이 (이전 모델 기준)
Accuracy	0.6288	- 0.0049	+ 0.0037
Precision	0.5223	- 0.0322	+ 0.0028
Recall	0.4426	+ 0.1863	+ 0.0687
F1_Score	0.4792	+ 0.1286	+ 0.0443
ROC_AUC	0.6385	+ 0.0366	+ 0.0077

결과

이전 모델 대비 모든 주요 성능 지표가 소폭 상승,
특히 재현율과 F1_Score의 지속적인 상승은 모델의 탐지력 및 안정성을 높일 수 있을 것으로 판단

Project 03

공유오피스 체험 신청자의 유료 전환 예측 프로젝트

(2025.06.10 - 2025.06.19)

분석결과

LogisticRegression

RandomForestClssifier

GridSearchCV를 통한 하이퍼파라미터 튜닝

XGBoost

XGBoost (Bayesian Optimization)

SMOTETomek

▪ 3차 선정 모델 : XGBClassifier

- Gradient Boosting 기반의 앙상블 모델인 XGBoost를 선정
- 개별 트리 간의 오차를 보완하는 방식으로 학습이 진행되어 **과적합 방지, 일반화 성능 향상에 강점**이 있고 Random Forest에 비해 파라미터 조정 **범위가 넓어** 성능 개선 여지가 크다고 판단

지표	값	차이 (Base 모델 기준)	차이 (이전 모델 기준)
Accuracy	0.6337	+ 0.0000	+ 0.0049
Precision	0.5419	- 0.0126	+ 0.0196
Recall	0.3263	+ 0.0700	- 0.1163
F1_Score	0.4073	+ 0.0567	- 0.0719
ROC_AUC	0.6435	+ 0.0416	+ 0.0050

결과 및 한계점

- 베이스 모델 대비 성능은 전반적으로 향상되었음
- 직전 모델과 비교 시 재현율, F1_Score 하락하여 양성 탐지 측면에서 아쉬움 존재하나, 정밀도 및 ROC_AUC는 소폭 상승하여 판별력 유지

▪ XGBoost 모델의 하이퍼파라미터 튜닝

- Hyperopt의 TPE 알고리즘 활용하여 베이지안 최적화 방식으로 튜닝

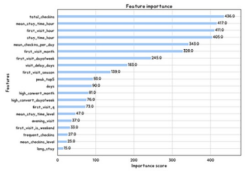
지표	값	차이 (Base 모델 기준)	차이 (이전 모델 기준)
Accuracy	0.6089	- 0.0248	- 0.0248
Precision	0.4937	- 0.0608	- 0.0482
Recall	0.5518	+ 0.2955	+ 0.2255
F1_Score	0.5512	+ 0.1706	+ 0.1139
ROC_AUC	0.6477	+ 0.0458	+ 0.0042

결과

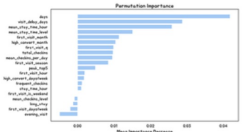
이전 모델 대비 재현율, F1_Score, ROC_AUC 상승함

▪ 피처 중요도 및 피처의 실제 기여도 확인

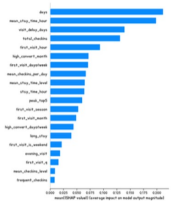
feature_importance



permutation_importance



SHAP



▪ 피처 조정 후 데이터를 재분리, XGBoost 모델 베이지안 최적화 기반 hyperopt 적용

지표	값	차이 (Base 모델 기준)	차이 (이전 모델 기준)
Accuracy	0.6202	- 0.0135	+ 0.0113
Precision	0.4859	- 0.0686	- 0.0078
Recall	0.5037	+ 0.2474	- 0.0481
F1_Score	0.4946	+ 0.1440	- 0.0266
ROC_AUC	0.6282	+ 0.0263	- 0.0195

결과 및 한계점

베이스 모델 대비 정확도 개선되어 균형형 모델 지향 시 더 적합할 수 있으나, 이전 모델 대비 재현율과 F1_score 소폭 하락하여 탐지 성능 다소 하락

▪ 클래스 불균형 문제 해결 : SMOTETomek 적용

- 기존 베이스라이 모델 대비 성능 향상 보이고 있으나 아직 소수 클래스에 대한 탐지 부족이라고 판단
- 클래스 문제 해결하고자 SMOTETomek 기법 적용

지표	값	차이 (Base 모델 기준)	차이 (이전 모델 기준)
Accuracy	0.6272	- 0.0065	+ 0.0070
Precision	0.4946	- 0.0599	+ 0.0087
Recall	0.4685	+ 0.2122	- 0.0352
F1_Score	0.4812	+ 0.1308	- 0.0134
ROC_AUC	0.6366	+ 0.0347	+ 0.0084

결과 및 한계점

- 클래스 불균형 완화를 통해 재현율과 ROC_AUC가 일정 수준 개선되었으나 정밀도가 낮아지면서 F1_Score, 정확도 등의 종합 성능은 XGBoost 단독 모델보다 다소 하락하였음
- 이는 SMOTE가 minority class 샘플을 인위적으로 생성하고 TomekLinks가 일부 majority 샘플을 제거함에 따라 데이터의 분포가 왜곡되거나 경계가 모호해져 모델의 결정 경계가 과적합되었을 가능성을 보임

Project 03

공유오피스 체험 신청자의 유료 전환 예측 프로젝트

(2025.06.10 - 2025.06.19)

분석결과

LogisticRegression

RandomForestClssifier

GridSearchCV를 통한 하이퍼파라미터 튜닝

XGBoost

XGBoost (Bayesian Optimization)

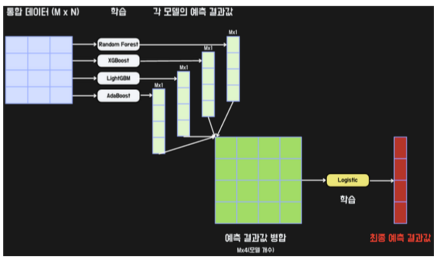
SMOTETomek

스태킹 앙상블

- 최종모델 구축 : 스태킹 앙상블

- 서로 다른 알고리즘 간의 보완적 특성을 활용하여 성능을 극대화할 수 있는 스태킹 앙상블기법 도입
- 개별 모델의 편향과 분산을 줄이고, 일관되면서도 강건한 예측 성능을 확보하기 위한 전략 수립

스태킹 모델 학습 경로



기반 모델

모델	선택 이유
XGBoost	강력한 성능과 정교한 분기 기준
RandomForest	앙상블 기반의 랜덤성으로 다양한 관점에서 예측 가능
LightBGM	빠르고 효율적이며, 다른 boosting 알고리즘과 상호 보완적
AdaBoost	가중치를 조절하며 오류에 민감하게 반응, 단순하지만 효과적인 부스팅

메타 모델 : Logistic Regression

복잡한 비선형 모델들의 예측 출력을 선형적으로 조합하여 안정적인 일반화 성능을 높이는 데 효과적인 Logistic Regression 선정

지표	값	차이 (Base 모델 기준)	차이 (이전 모델 기준)
Accuracy	0.6397	+ 0.0060	+ 0.0125
Precision	0.5116	- 0.0429	+ 0.0170
Recall	0.5168	+ 0.2605	+ 0.0483
F1_Score	0.5142	+ 0.1636	+ 0.0330
ROC_AUC	0.6341	+ 0.0322	- 0.0025

결과

- 베이스 모델 대비 모든 주요 지표 개선
- 정확도, 재현율, F1_Score 모두 최고 수준으로 정밀성과 탐지력 간 균형이 잘 잡힌 모델로 평가
- 모델의 실전 적용 가능성 높음

모델링 성능 비교 및 향상 과정 요약

베이스라인 모델 대비 약 2배의 유료 전환 고객 탐지 능력 향상을 달성하였고, 예측 모델이 실질적으로 의미 있는 피처를 바탕으로 학습함을 최종 검증함