PORT
FOLIO

정 수 연 JUNG SOO YEON	010-9314-2905	tndus29290@gmail.com

PROJECT 1

이탈 예측 모델링을 통한 고객 이탈 최소화 및 내부 고객 관리 강화

프로젝트 개요

기획 의도

엔데믹 영향과 고물가로 인해,이커머스 소비시장이 분산되고, 최저가 경쟁의 이유등으로 기업들은 경쟁적인 상황에 직면하고 있다.
⇒ 이로 인해, 고객 이탈율 상승 및 내부 고객 관리의 필요성이 대두되고 있다.

문제 정의

CRM 팀 : 성별,디바이스,결혼 여부 등의 다양한 고객 특성별 세분화된 마케팅 전략 수립 및 이탈 예측 모델링이 필요하다.
도시등급 ,결제 방식등의 데이터를 활용하여, 비즈니스 전략을 수립하고 장기적인 가치를 창출해야 한다.
Product 팀: 앱/웹 사용 시간, 디바이스 선호도 등의 데이터를 활용하여, 사용자 경험 개선 및 더나은 서비스를 제공해야 한다.

분석 목표

1. 고객 이탈 예측 모델 개발
2. 이탈에 영향을 주는 특성 중요도 분석
3. CRM 팀,Product팀 별 내부 고객 관리 강화 및 이탈 위험 고객 관리 액션 플랜 도출

Language

Python

Tool

Pandas

PROJECT 01. 이탈 예측 모델링을 통한 고객 이탈 최소화 및 내부 고객 관리 강화

분석 프로세스

데이터 탐색

- EDA 데이터 시각화 ⇒ 범주형 변수 ⇒ 수치형 변수

데이터 전처리

- 결측치 처리 ⇒ 이상치 처리 ⇒ 라벨 인코딩 ⇒ MinMaxScaling

모델 학습

- 모델링 준비 : SMOTETomke 샘플링 ⇒ 파라미터 튜닝 (Grid Search)

성과 평가

- 알고리즘
  - Decision Tree
  - Random Forest
  - LBGM Classifier
  - XGB Classifier ...
- 성과 비교
  - Accuracy
  - Precision
  - Recall
  - F1 score

모델 해석

- 변수 중요도
  - Feature Importance
  - SHAP Value
- 활용방안
  - CRM 팀 액션 플랜
  - Product 팀 액션 플랜

데이터 준비

데이터 소개

Ecommerce Customer Churn Analysis and Prediction

Predict customer churn and make suggestions

데이터 출처: Kaggle
5630 rows X 20 columns
고객의 행동 데이터 및 이탈 여부 포함

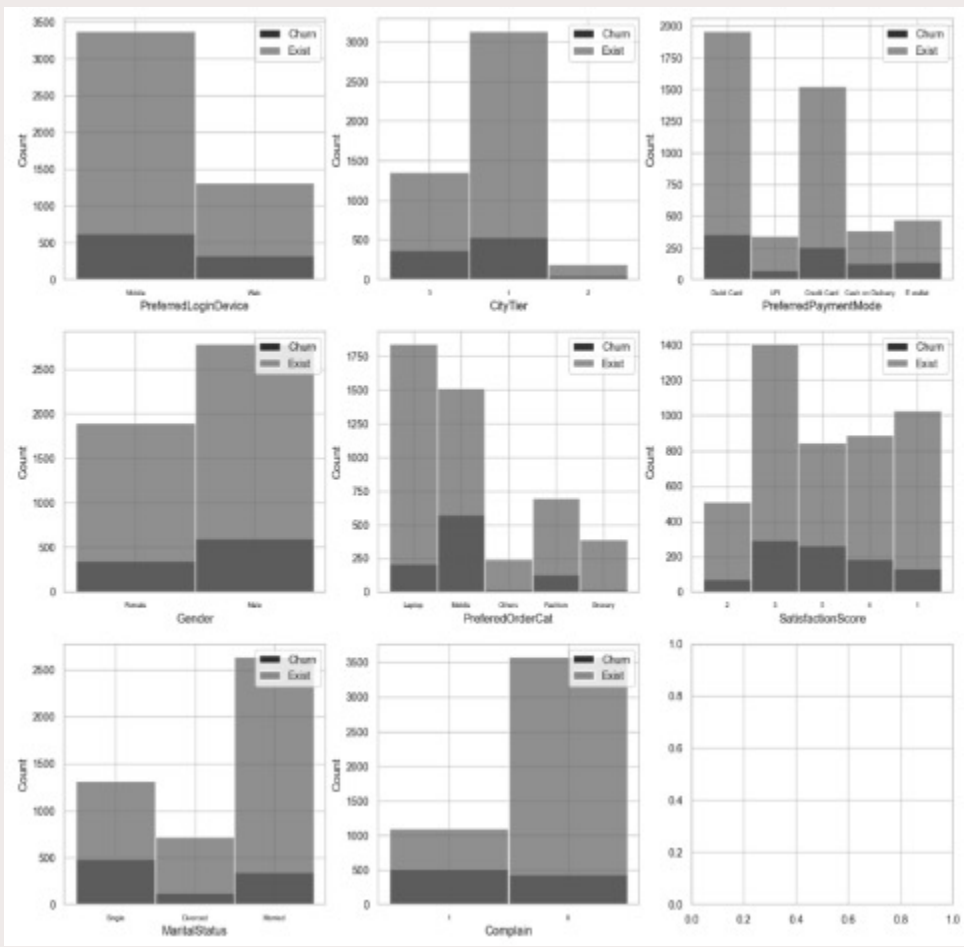
변수명	변수 설명	변수 유형	변수명	변수 설명	변수 유형
CustomerID	고객 고유아이디	Index	PreferedOrderCat	지난달 고객이 주문한 가장 선호하는 카테고리	Categorical
Churn	이탈 여부	Categorical	SatisfactionScore	만족도 점수	Categorical
Tenure	고객 사용 기간	Numerical	MaritalStatus	결혼 여부	Categorical
PreferredLoginDevice	선호하는 로그인 디바이스	Categorical	NumberOfAddress	등록한 주소 갯수	Numerical
CityTier	도시 등급 (1이 가장 발전된 도시)	Categorical	Complain	지난달 고객의 컴플레인 여부	Categorical
WarehouseToHome	창고에서 고객의 집까지 거리	Numerical	OrderAmountHikeFromlastYear	지난해 대비 주문 증가율	Numerical
PreferredPaymentMode	선호하는 결제 방식	Categorical	CouponUsed	지난달 사용한 쿠폰 갯수	Numerical
Gender	성별	Categorical	OrderCount	지난달 총 주문횟수	Numerical
HourSpendOnApp	앱 혹은 웹 사용 시간	Numerical	DaySinceLastOrder	마지막 주문 후 지난 일수	Numerical
NumberOfDeviceRegistered	고객별 등록한 총 디바이스 갯수	Numerical	CashbackAmount	지난달 평균 캐시백 금액	Numerical

데이터 전처리

변수명	변수 설명
결측치	결측치는 4~5%로 미비한 수치형 데이터로, 평균값으로 대체
이상치	이상치는 IQR 방식을 활용하여, 이상치 제거
인코딩	'LabelEncoder'을 통해, 범주형 변수 인코딩

EDA

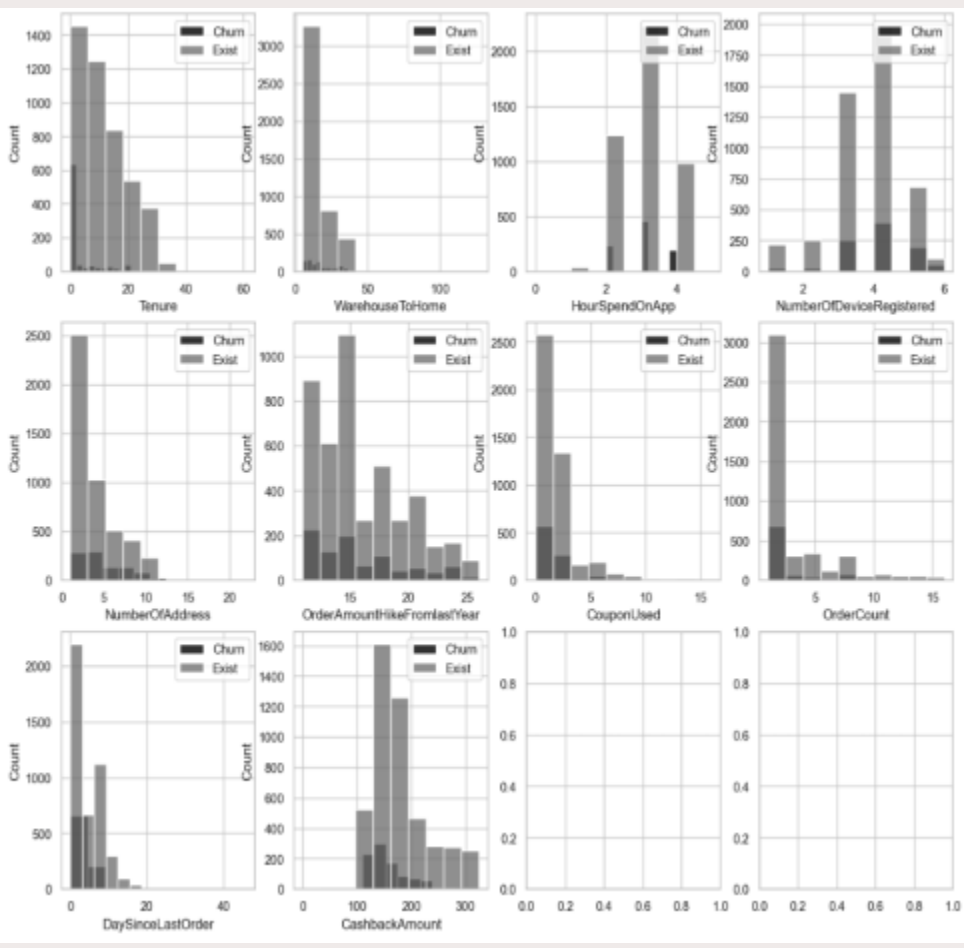
범주형 변수



변수명	변수 특징
PreferredLoginDevice (선호하는 로그인 디바이스)	- 71% 비율로,Mobile로 로그인하는 것을 선호함 - Web 기기는 모바일 대비 이탈율이 3.2% 더높다.
CityTier (도시 등급)	- 도시등급 3등급의 이탈율이 21.4%로 가장 높다. ➡ *도시등급이 낮을수록, 이탈율이 증가한다.*
PreferredPaymentMode (선호하는 결제 방식)	- Debit Card(직불카드) 결제방식을 가장 선호한다. - 이탈율이 가장 높은 결제 방식은 Cash on Delivery(배송 후 대금 지급)방식
Gender (성별)	- 남성 고객은 전체 주문의 59%,여성 고객은 40% 점유 ➡ 남성고객이 더 많다.
PreferedOrderCat (선호하는 카테고리)	- 고객이 가장 선호하는 카테고리는 Mobile 카테고리로, 약 37% 점유 - 이탈율이 가장 많은 카테고리는 Mobile (모바일 관련 상품) 카테고리 - 이탈율이 가장 적은 카테고리는 Grocery (식품) 카테고리
SatisfactionScore (만족도 점수)	- 대부분의 고객은 만족도 3점 부여 - 만족도와 이탈간의 관계는 없다. - 앱/웹 사용시간 혹은 주문건수와도 관계가 없는것으로 확인된다.
MaritalStatus (결혼 여부)	- Married(결혼) 고객이 가장 많다. - Single(싱글)고객의 이탈율이 26.7%로 가장 높다.
Complain (컴플레인 여부)	- 컴플레인이 있는 고객의 이탈율이 31% 더 높다

EDA

수치형 변수

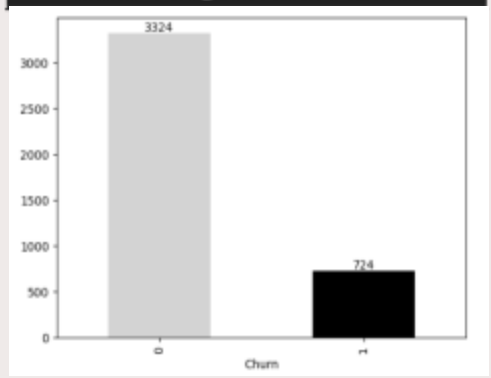


변수명	변수 특징
Tenure (사용 기간)	- 사용 기간이 10일 이내인 고객의 이탈율이 24.4%로 가장 높다. ➡ *사용기간이 짧을수록 이탈율이 높다.*
WarehouseToHome (창고에서 고객 집까지의 거리)	- 고객의 집에서 창고까지의 거리는 대부분 20km 이내에 있다.
HourSpendOnApp (앱/웹 사용 시간)	- 대부분의 고객은 3-4시간 앱/웹 디바이스를 사용한다. - 사용 시간과 이탈과는 무관하다.
NumberOfDeviceRegistered (디바이스 등록 갯수)	- 디바이스 6개 등록한 고객의 이탈율은 37.1%로 가장 높다. ➡*디바이스 등록 갯수가 많을수록,이탈율이 높아진다.*
NumberOfAddress (주소 등록 갯수)	- 이탈율이 가장 높은 고객은 8개~11개이하의 주소를 등록했다. ➡*주소 등록 갯수가 많을수록,이탈율이 높아진다.*
OrderAmountHikeFromlastYear (지난해 주문금액)	- 21$~ 26$이하의 주문금액을 소비한 고객의 이탈율이 가장높다, - 지난해 주문금액과 이탈과는 무관하다.
CouponUsed (쿠폰 사용갯수)	- 대부분의 고객은 1~2개 쿠폰을 사용한다. - 쿠폰 사용과 이탈과는 무관하다.
OrderCount (주문 건수)	- 대부분의 고객은 2-3번 주문한 것을 알 수 있다. - 주문 건수와 이탈은 무관하다.
DaySinceLastOrder (마지막 주문 후 경과일)	- 마지막 주문 후 3일이내 경과한 고객의 이탈율이 가장 높다. ➡*경과일이 짧아질수록,이탈 가능성이 높다.*
CashbackAmount (캐시백 금액)	- 캐시백 받은 금액이 적을수록, 이탈 가능성이 높다.

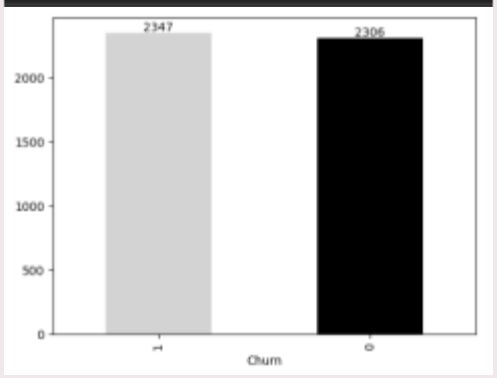
모델링

1차 시도

원본 Target 데이터 분포



SMOTETomek 후



target 데이터 (Churn 여부)의 클래스 불균형 해결을 위해,
"SMOTETomek" 기법 적용

⇒ 정보 손실 감소 및 과적합 가능성 감소

베이스라인 평가 지표

	Model	Accuarcy	Precision	Recall	F1 Score	AUC Score
0	Logistic Regression	0.762350	0.807645	0.685146	0.741370	0.761911
1	Support Vector Machine	0.855954	0.891580	0.808577	0.848053	0.855685
2	Decision Tree	0.835673	0.874707	0.781381	0.825414	0.835365
3	Random Forest	0.907956	0.947187	0.862971	0.903120	0.907700
4	XGBClassifier	0.931877	0.966102	0.894351	0.928843	0.931664
5	LGBMClassifier	0.921997	0.952809	0.887029	0.918743	0.921798
6	GradientBoostingClassifier	0.890796	0.924829	0.849372	0.885496	0.890560

베이스라인 평가지표 해석

1. 정확도: XGB Classifier은 0.934로 가장 정확도가 높다.
2. 정밀도: XGB Classifier Classifier는 0.968로 정밀도가 가장 높다.
3. 재현율:XGB Classifier는 0.896으로 재현율이 가장 높다 .
4. F1_score:XGB Classifier 0.931로 가장 높은 F1_score

⇒ 공통적으로 'XGB Classifier'의 평가지표가 가장 우수하다.

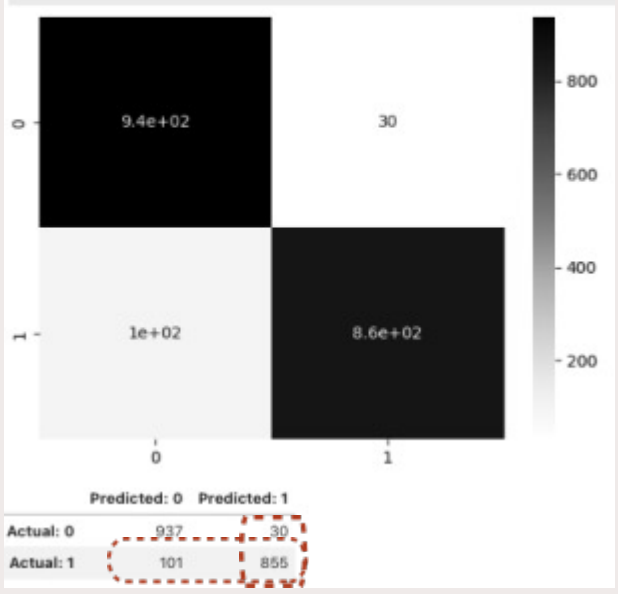
모델링

2차 시도

	Model	Accuarcy	Precision	Recall	F1 Score	AUC Score
XGBClassifier		0.931877	0.966102	0.894351	0.928843	0.931664

하이퍼파라미터 튜닝 전

	precision	recall	f1-score	support
0	0.90	0.97	0.93	967
1	0.97	0.89	0.93	956
accuracy			0.93	1923
macro avg	0.93	0.93	0.93	1923
weighted avg	0.93	0.93	0.93	1923

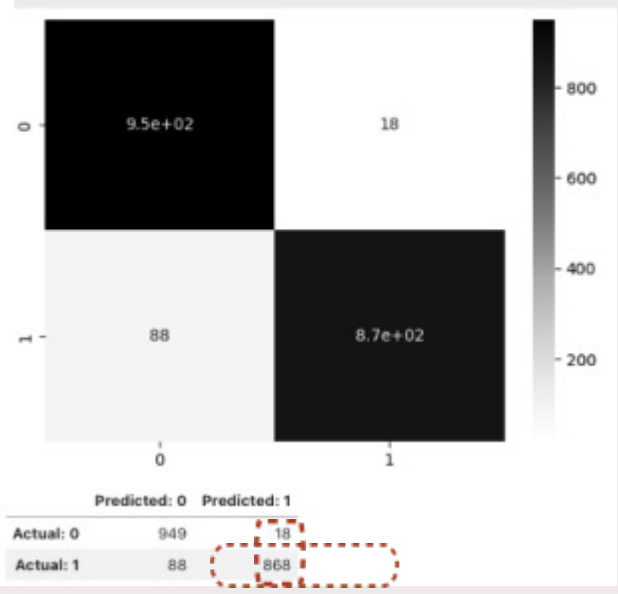


	Predicted: 0	Predicted: 1
Actual: 0	937	30
Actual: 1	101	855

	Model	Accuarcy	Precision	Recall	F1 Score	AUC Score
XGBClassifier		0.944878	0.979684	0.90795	0.942454	0.944668

하이퍼파라미터 튜닝 후

	precision	recall	f1-score	support
0	0.92	0.98	0.95	967
1	0.98	0.91	0.94	956
accuracy			0.94	1923
macro avg	0.95	0.94	0.94	1923
weighted avg	0.95	0.94	0.94	1923



	Predicted: 0	Predicted: 1
Actual: 0	949	18
Actual: 1	88	868

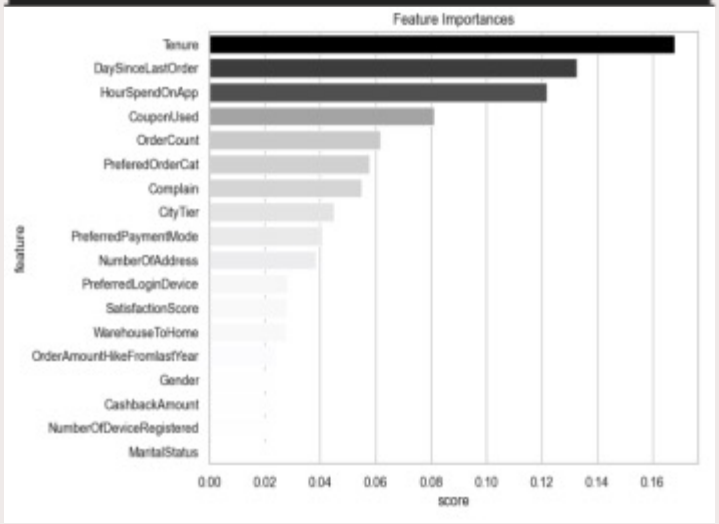
최적의 파라미터	
max_depth	[5,7,9]
min_child_weight	[1,3,5]
colsample_bytree	[0.5,0.75]
n_estimators	[100,200,300,400,500]

정확도 0.01 상승,
정밀도 0.01 상승,
재현율 0.02상승,
F1 score 0.01상승

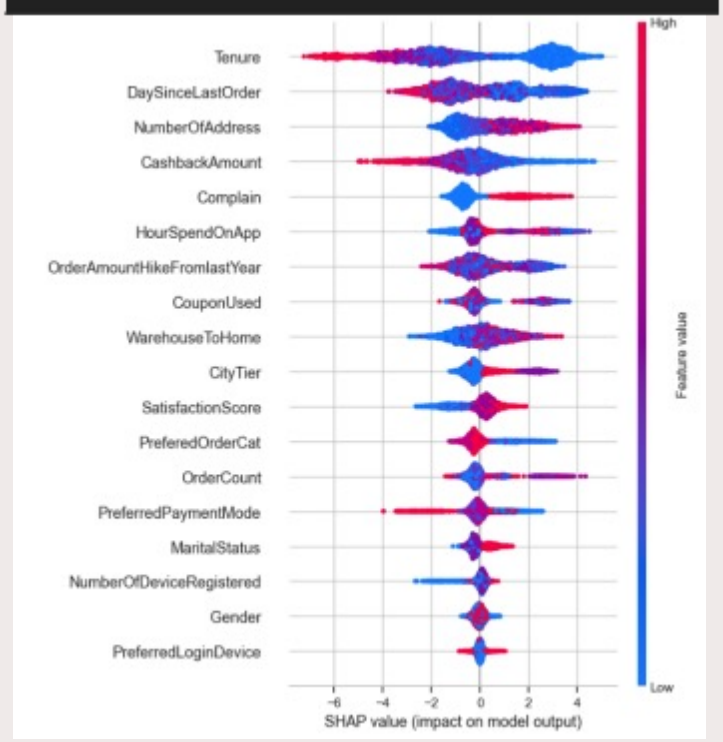
모델 해석

Feature Importance, SHAP Value

Feature Importance



SHAP Value



1. Tenure(고객 사용 기간)

⇒사용 기간이 짧을수록 이탈 가능성이 높다.

2. DaySinceLastOrder (마지막 주문 후 소요기간)

⇒마지막 주문 후 경과일이 짧을수록, 이탈 가능성이 높다

3.NumberOfAddress(등록한 주소 갯수)

⇒등록한 주소 갯수가 많을수록 ,이탈 가능성이 높다.

4.CashbackAmount(캐시백 받은 금액)

⇒캐시백 받은 금액이 적을수록, 이탈 가능성이 높다.

4.Complain(컴플레인 여부)

⇒컴플레인 경험이 있는 고객의 이탈 가능성이 높다.

5.HourSpendOnApp(디바이스 사용 시간)

⇒디바이스 사용기간이 길어질수록 이탈 가능성이 높다.

활용 방안

이탈 예측 모델

	CustomerID	Predict	Predcit_proba	rank
0	55400	1	100.000000	1
1	53212	1	100.000000	1
2	52921	1	100.000000	1
3	50469	1	100.000000	1
4	53214	1	100.000000	1

High risk 이탈 위험 고객
이탈 예상률 60%~90%⇒rank 1

+

고객 데이터 분석을 통한 Insight

High risk 이탈 위험 고객 식별 데이터 + 고객 데이터 분석을 통한 인사이트

⇒ CRM 팀,프로덕트 팀 액션 플랜 제안

액션 플랜

CRM 팀_액션플랜

1. 고객 별 세분화된 마케팅 전략 수립
   a. 내부 고객 : 남성 고객에게 'Mobile' 상품 할인 쿠폰 제공 /여성 고객에게 'Laptop'상품 할인 쿠폰
      결혼 고객에게 "가족 회원 추가 할인" 혜택 제공 / 싱글 및 이혼 고객에게 독립 생활을 위한 할인 혜택 제공
   b. 이탈 위험 고객 : 이탈 위험이 있는 성별,디바이스,결혼 여부등의 특성에 따라 특별 혜택 제공
2. 서비스 이용 기간에 따른 멤버십 혜택 부여
   a. 내부 고객 : 장기 이용 고객 대상 정기적인 감사 이벤트를 통해 충성도 강화
   b. 이탈 위험 고객 : 이용 기간이 짧은 신규 고객에게 멤버쉽 혜택 소개 및 특별 리워드 증정
3. 마지막 주문 후,경과기간에 따른 서비스 제공
   a. 내부 고객:마지막 주문 후, 경과기간이 늘어나는 고객에게 맞춤형 알림 서비스를 제공하여,재주문 유도
   b. 이탈 위험 고객: 마지막 주문 후,경과 기간이 짧은 고객을 대상으로 특별 혜택을 제공하여 재구매 유도
4. 캐시백 프로그램 개선 및 활용
   a. 내부 고객:적립된 캐시백 포인트 사용 권장 앱 푸시 메세지를 통해 , 고객의 지속적인 거래 유도
   b. 이탈 위험 고객:캐시백 금액 재조정을 통해,이탈 위험 고객에게 더 많은 캐시백 혜택을 제공하여,이탈 방지
5. 컴플레인 해결 및 고객 응대 강화
   a. 내부 고객:컴플레인 고객 응대 및 문제 해결을 위한 컴플레인 해결 프로세스를 개선하고,
      컴플레인 피드백을 반영하여 제품 및 서비스의 개선 방향 도출
   b. 이탈 위험 고객: 이탈 위험 고객 전용 서비스 팀을 구성하여,불만사항에 대해 우선적으로 대응, 주요 이슈에 대해 지속적인 모니터링
6. 도시 등급 별 선호하는 결제 방식에 따른 프로모션 진행
   a. 내부 고객 : 도시 등급1은 "Debit card"(직불 카드), 도시 등급 2는 "UPI"(은행 간 계좌이체) , 도시 등급 3은 'E wallet'(전자 지갑 시스템)에 맞춰,
      맞춤형 혜택 제공
   b. 이탈 위험 고객 : 도시 등급 3의 높은 이탈율에 주목하여, 해당 지역에 특화된 혜택 및 프로모션 진행을 통해, 이탈 고객 유지 및 고객 경험 개선

Product팀_액션플랜

1. 디바이스 등록 및 주소 등록 정보 활용
   a. 내부 고객 : 고객별 자주 이용하는 디바이스와 주소 정보에 따라 ,맞춤형 상품 추천 및 할인 혜택 제공
   b. 이탈 위험 고객 : 디바이스 및 주소 등록이 빈번한 경우,이탈 시 놓치게 될 특별 혜택에 대한 경고 및 안내 메시지를 전송하여 이탈을 방지하고 이용 동기 부여
2. 앱/웹 사용 시간에 따른 전략
   a. 내부 고객 : 고객의 앱/웹 사용시간에 따라 사용자 행동 데이터를 분석하고, 이를 기반으로 한 맞춤형 콘텐츠 및 할인 혜택을 제공하여 개인화된 경험 강화
   b. 이탈 위험 고객 : 사용자의 디바이스 선호도에 따라 최적화된 UI/UX 디자인을 개선하여 사용자의 편의성 및 만족도 향상
3. 로케이션 기능 개선
   a. 내부 고객:고객 집 주소와 창고 사이의 거리를 실시간으로 확인하여, 고객 만족도 향상

한계점

1. 최적의 파라미터를 찾기 위해, 전 파라미터의 조합을 찾기에는 시간적인 한계가 있었다.
2. 고객의 만족도 점수 외에, 피드백 내용에 대한 변수가 있었다면 추가적인 개선사항을 제안할 수 있었을 것 같다.
3. 상품 카테고리뿐만 아니라 개별 상품에 대한 구매 선호도를 알 수 있었다면, 추가적인 개인화 추천시스템까지 도전할 수 있을 것 같다.

THANK YOU.

정 수 연 JUNG SOO YEON	010-9314-2905	tndus29290@gmail.com