SKN 12 5TEAM

E-COMMERCE 고객 이탈 분석

온라인 쇼핑몰 고객의 주요 이탈 원인 분석

소속 | SK Networks Family AI캠프 12기 2차 단위 프로젝트 5조

구성원 | 김도윤, 윤권, 이정민, 이준석, 허한결

CONTENTS

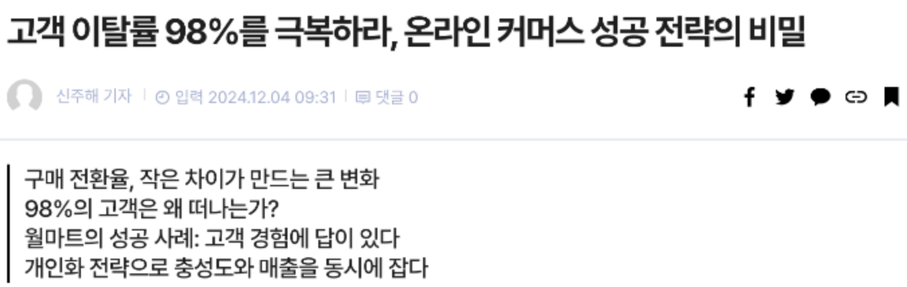
01 | 프로젝트 기획

PAYNT

Pain to Pay, Painted with Data

Pain → Pay → Paint": 고객의 문제(pain)를 발견하고, 수익화(pay)하며, 데이터를 바탕으로 고객 여정을 시각화(paint)한다

기획 의도 및 목표



고객 이탈률 98%를 극복하라, 온라인 커머스 성공 전략의 비밀

신주해 기자 | 입력 2024.12.04 09:31 | 댓글 0

구매 전환율, 작은 차이가 만드는 큰 변화
98%의 고객은 왜 떠나는가?
월마트의 성공 사례: 고객 경험에 답이 있다
개인화 전략으로 충성도와 매출을 동시에 잡다

프로젝트 목표

고객 데이터를 기본으로 이탈 가능성 사전 예측

관리자의 입장에서 머신러닝모델을 통해 이탈 원인 분석

기대효과

이탈 가능 고객에 대한 사전 대응

마케팅 비용 절감

데이터 기반 의사결정 강화

01 | 프로젝트 기획

기술 스택

언어	Python
분석	Pandas, Numpy, Scikit-Learn
모델링	XGBoost, Lightgbm
시각화	plotly, SHAP
UI 구현	Streamlit
버전 관리	Git

02 | 데이터 분석

데이터 소개 kaggle

사업 항목	상세 내용
데이터 소개	온라인 소매(E-commerce) 기업의 고객 이탈 분석 및 예측을 위한 것으로, 고객 행동 데이터를 기반으로 이탈 가능성을 분석하는 데 목적이 있다.
총 데이터 수	5,600개 이상치/결측치 처리 후 3,500개의 데이터 사용
주요 칼럼	One-hot 인코딩 후 칼럼 수 변화: 18개 -> 28개

컬럼(18개)

타킷 변수 | 이탈 여부

수치형

고객 번호 | 거래 기간 | 도시 등급 | 창고-집 거리

앱 사용 시간 | 등록 기기 수 | 등록 주소 수

사용한 쿠폰 수 | 불만 제기 여부 | 작년 대비 증가 수

만족도 | 주문 수 | 마지막 주문 일자 | 캐시백 금액

범주형

선호 로그인 기기 | 성별 | 선호 카테고리

기혼 여부 | 선호 결제 방식

02 | 데이터 분석

데이터 분석 1 : 데이터 처리 과정

원본 데이터셋	E Commerce Dataset2.xlsx
결측치 처리	수치형 컬럼은 평균값으로 대체 (7개 변수)
이상치 제거	IQR 기반 필터링
범주형 인코딩	get_dummies 적용 (drop_first = True)
클래스 불균형 처리	SMOTE 적용 (Train 데이터에 한정)
스케일링	StandardScaler 적용 (로지스틱 회귀, MLP 대응)

02 | 데이터 분석

데이터 분석 2 : 칼럼 간의 상관관계

1. Complain(불만 제기 여부) 이탈과 가장 강한 상관관계

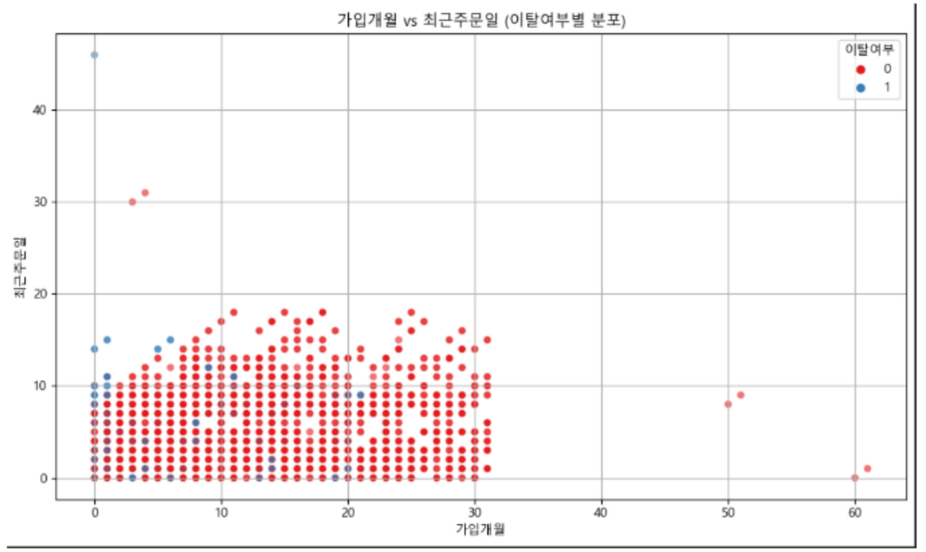
2. SatisfactionScore(만족도 점수, 1~5점) 점수가 낮을수록 이탈 확률 급격히 증가

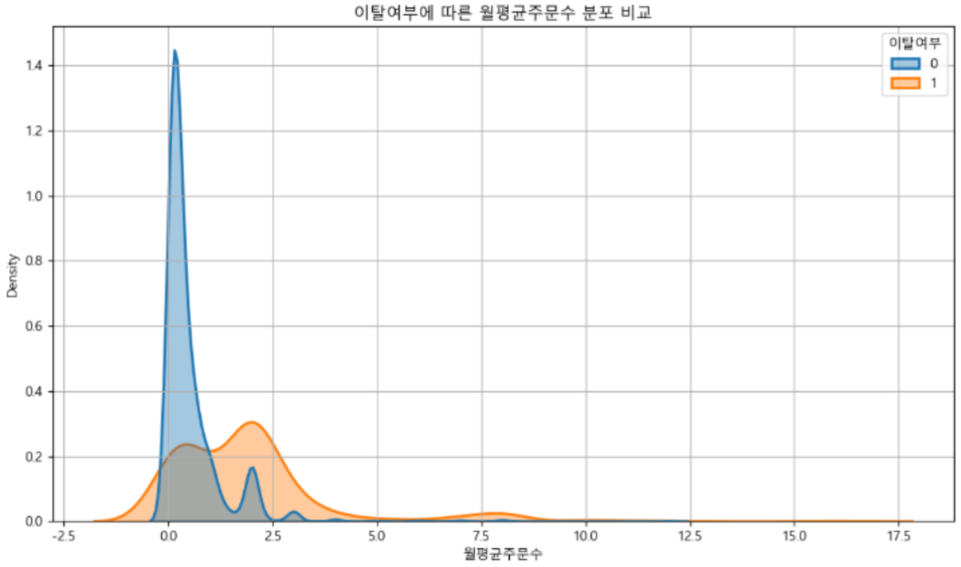
3. DaySinceLastOrder(마지막 주문일 경과) 최근 주문 이력이 없는 고객은 이탈 가능성 급증

4. HourSpendOnApp(앱 체류 시간) 앱에 머무는 시간이 짧을수록 이탈률 높음

5. OrderCount(주문 횟수) 주문 빈도는 충성도 및 만족도를 반영

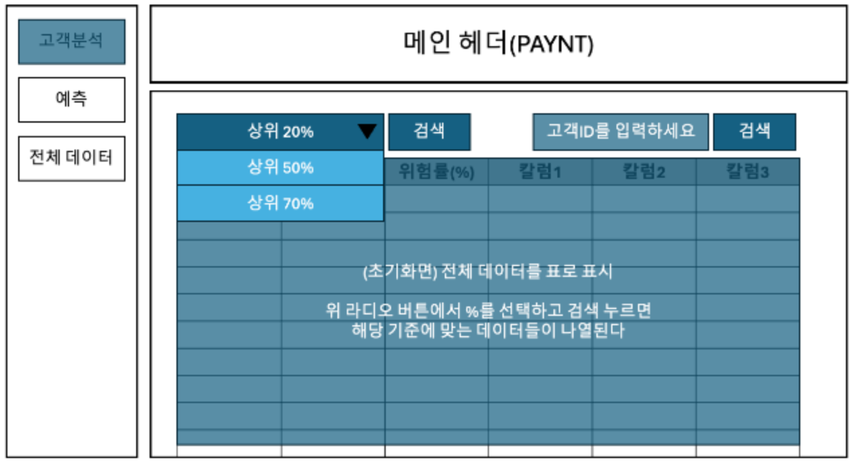
6. Tenure(가입 기간) 가입한 지 얼마 안 된 고객의 이탈률 높음

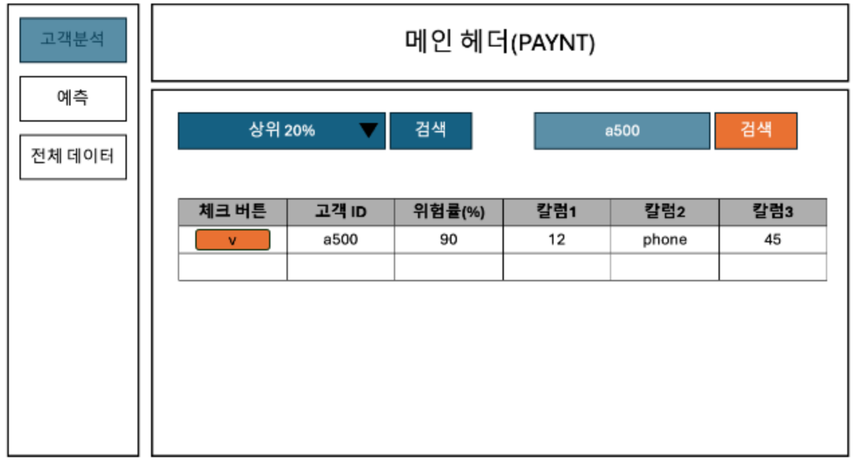


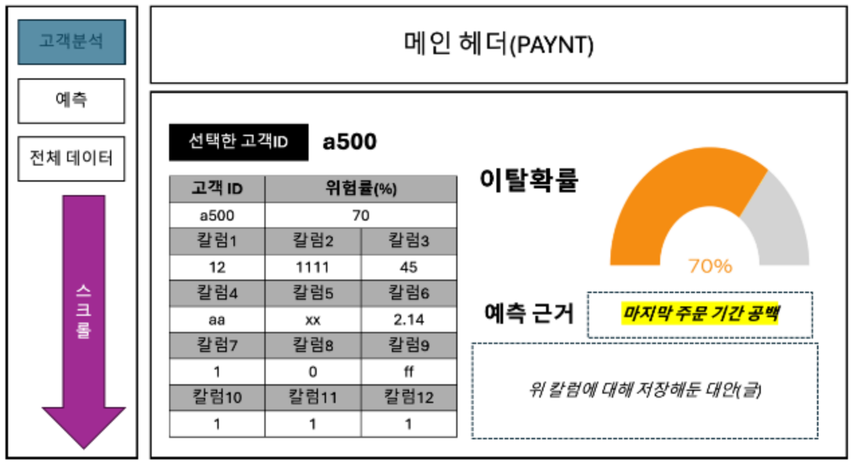


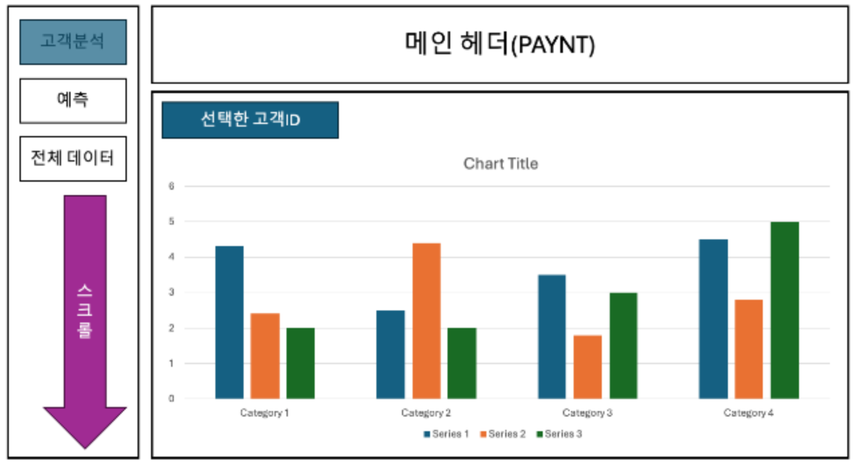
03 | 화면 설계서

1페이지: 개별 고객에 대한 이탈 여부 분석



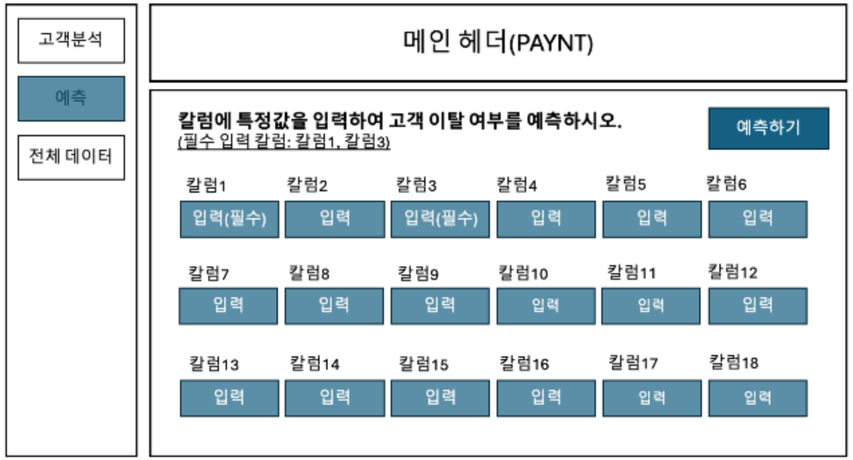


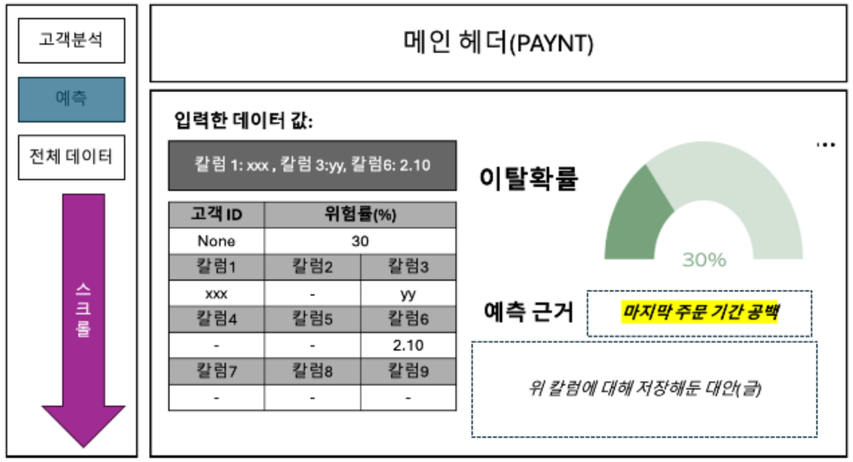


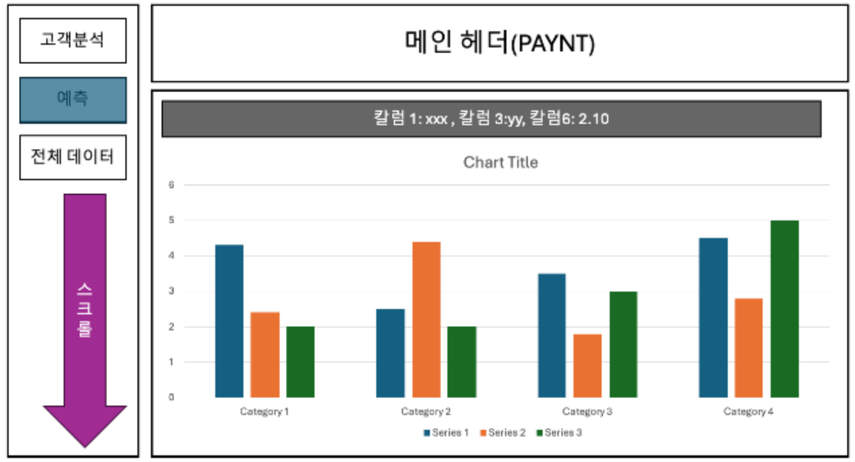


03 | 화면 설계서

2페이지: 임의의 데이터에 대한 이탈 여부 예측

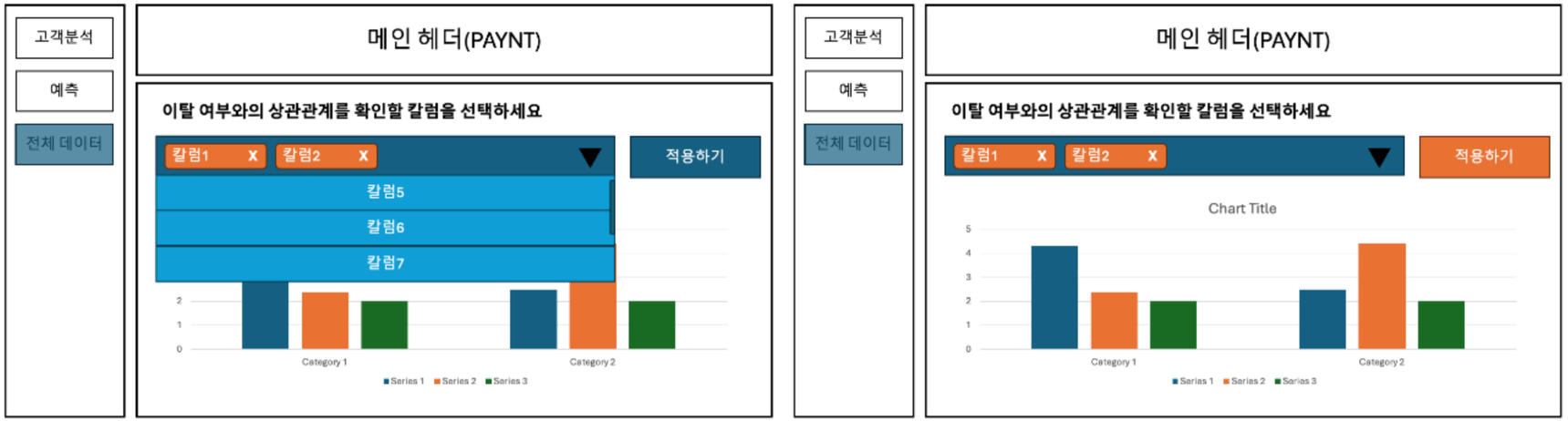


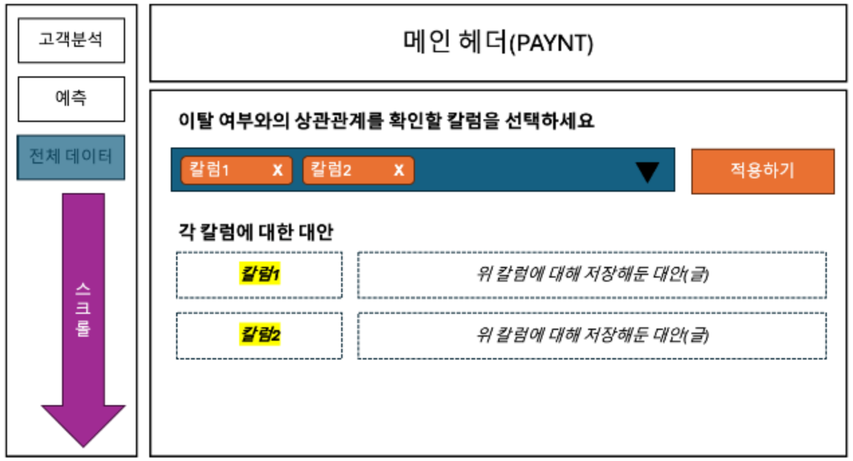




03 | 화면 설계서

3페이지: 전체 데이터 내 칼럼 별 상관관계





04 | 모델

모델 분석 방법

모델 비교 및 교차 검증	하이퍼파라미터 튜닝	최종 모델 저장	모델 평가	UI 구현
5-Fold 교차 검증을 통해 성능 파악 비교 모델: LogisticRegression, KNN, SVC, NaiveBayes 등	GridSarchCV로 최적의 하이퍼파라미터 조합 탐색	XGBoost으로 최종 선택 .pkl 형식으로 저장	테스트셋 기반으로 정확도, 정밀도, 재현율, F1 점수 등 평가	Streamlit 연동

04 | 모델

모델 분석 - 성능 비교

```
✅ Train Accuracy: 1.0
✅ Test Accuracy : 0.9746121297602257
✅ Precision     : 0.9838709677419355
✅ Recall        : 0.8840579710144928
✅ F1 Score      : 0.9312977099236642
✅ AUC Score     : 0.9846188989568264
⚠️ 과적합 여부 (Train - Test F1): 0.0687

[분류 리포트]
              precision    recall  f1-score   support

           0       0.97      1.00      0.98       571
           1       0.98      0.88      0.93       138

    accuracy                           0.97       709
   macro avg       0.98      0.94      0.96       709
weighted avg       0.97      0.97      0.97       709
```

Model	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score
LogisticRegression	0.8579	0.8293	0.5503	0.6739	0.6059
DecisionTree	1.0	0.8829	0.673	0.7754	0.7205
RandomForest	1.0	0.9394	0.9358	0.7391	0.8259
GradientBoosting	0.936	0.8942	0.7405	0.7029	0.7212
AdaBoost	0.91	0.8646	0.6438	0.6812	0.662
Bagging	1.0	0.9365	0.8661	0.7971	0.8302
ExtraTrees	1.0	0.9309	0.9238	0.7029	0.7984
KNN	0.9448	0.811	0.5119	0.6232	0.5621
SVC	0.9575	0.8942	0.7692	0.6522	0.7059
GaussianNB	0.551	0.2863	0.2125	0.9855	0.3496
XGBoost	1.0	0.969	0.9531	0.8841	0.9173
LightGBM	0.9978	0.9478	0.9106	0.8116	0.8582

최종 모델 선택: **XGBoost**

이유: XGBoost는 이커머스 고객 이탈 예측 문제에 있어
성능, 해석력, 실시간 적용성 모두에서 최적의 선택
과적합이 의심되지만 이탈자를 예민하게 선별해야 하는 프로그램의 취지와 잘 맞아서 선택

05 | UI 구현 및 시연

🚀 이탈 예측 대시보드

고객 이탈 예측 결과

고위험 고객 수

이탈 위험 70% 이상

696명

중위험 고객 수

이탈 위험 40~70%

713명

저위험 고객 수

이탈 위험 40% 미만

2133명

이탈률 필터

전체

고객 ID 검색

고객 ID를 입력하세요

총 3542명의 고객이 선택되었습니다.

이탈 위험도 표시:

- 🔴 70% 이상: 고위험 고객
- 🟡 40~70%: 중위험 고객
- 🟢 40% 미만: 저위험 고객

방향성 설명:

- (부정): 해당 특성이 이탈 확률을 높이는 방향으로 작용
- (긍정): 해당 특성이 이탈 확률을 낮추는 방향으로 작용
- 괄호 안의 %는 전체 예측에서 해당 특성이 차지하는 상대적 중요도입니다.

고객 ID	이탈 위험도	주요 영향 요인	2순위 영향 요인	3순위 영향 요인
CUST_000001	🔴 85.7%	마지막 주문 후 경과일 (⚠️ 부정) (13.1%)	불만 제기 (⚠️ 부정) (11.8%)	거래 기간 (⚠️ 부정) (10.3%)
CUST_000002	🔴 89.6%	마지막 주문 후 경과일 (⚠️ 부정) (14.1%)	배송 거리 (✅ 긍정) (12.6%)	거래 기간 (⚠️ 부정) (11.6%)

추후 업데이트 예정 항목

감사합니다

소속 | SK Networks Family AI캠프 12기 2차 단위 프로젝트 5조

구성원 | 김도윤, 윤권, 이정민, 이준석, 허한결