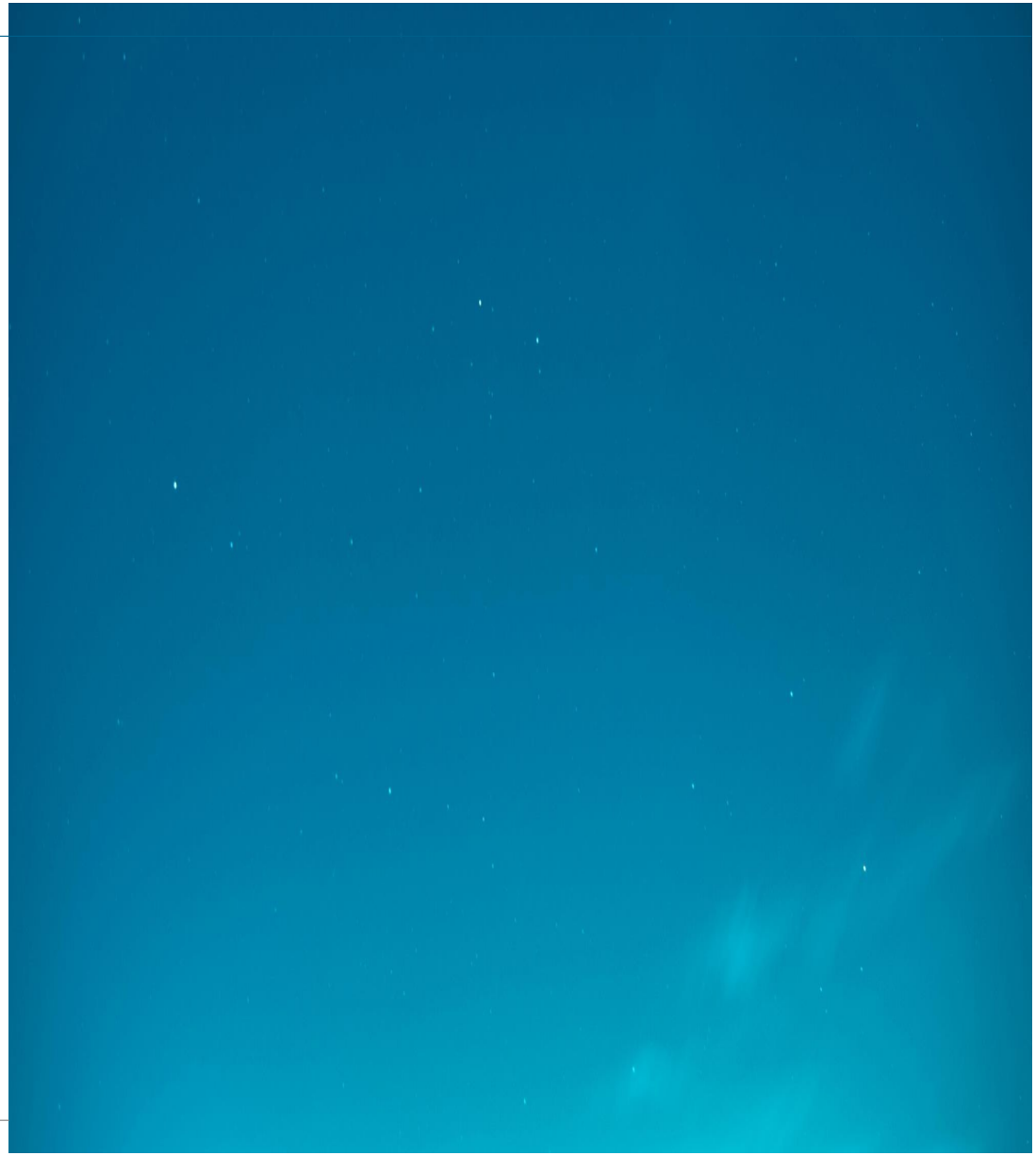
이커머스 플랫폼 문제진단 프로젝트

자스민 팀(18기_3팀)
경동연/ 이주연/ 방은혜

목차 Table of Contents



Part 1

프로젝트 개요

Part1 프로젝트 개요

1 프로젝트 주제 및 목표

1 주제	• 인도네시아 패션 이커머스 플랫폼의 데이터를 활용 • 해당 기업이 가지고 있는 각종 문제점을 정의 내려 원인 진단
2 목표	• 문제를 정의하고 해당 문제 해결을 위한 지표 개선 액션 도출 • 이탈 모형을 통한 이탈방지 액션 도출
3 역량	• 도메인 지식 • 분석 기술 • 커뮤니케이션 능력

Part 2

프로젝트 팀 구성 및 역할

Part2

2 프로젝트 팀 구성 및 역할

훈련생	역할	담당업무
경동연	팀장	고객데이터 분석 RFM 분석 이탈 예측 모델 만들기
이주연	팀원	클릭 데이터 분석 리텐션 분석
방은혜	팀원	거래 데이터 분석 문제점 진단 및 데이터 분석 시각화 대시보드 만들기

Part 3

프로젝트 수행 절차 및 방법

Part3

3 프로젝트 수행 절차 및 방법

구분	기간	활동
사전기획	8/9 – 8/10	✓ 프로젝트 기획 및 주제 선정 ✓ 데이터 탐색
데이터 전처리	8/11 – 8/20	✓ 데이터 정제 및 특성공학 ✓ 도메인 지식 및 인도네시아 배경지식 수집
모델링	8/21 – 8/26	✓ 이탈 방지 모델 구현
인사이트 도출	8/27 – 9/1	✓ 시각화 및 인사이트 도출
서비스 구축	9/2 - 9/4	✓ 대시보드 작업
총 개발 기간	8/9 – 9/5(총4주)	

Part 4

프로젝트 수행 결과

4 Data set

Part4 프로젝트 수행결과

001

- 고객들의 온라인 상호 작용을 기록한 데이터
- 사용자의 경로 추적
- 주요 column
  - 페이지 클릭수
  - 제품 인기도
  - 프로모션코드
- Size: 12,833,602 x 12

df_click_stream

002

- 고객관련 정보 데이터
- 주요 column
  - 가입날짜
  - RFM 등급
  - 고객 Rank
  - 이탈여부
- Size : 100,000 x 23

df_customer_chrun

003

- 제품 특징 관련 데이터
- 주요 column
  - 제품 종류
  - 브랜드
- Size: 44,424 x 10

df_product

004

- 거래관련 정보 데이터
- 주요 column
  - 가격
  - 수량
  - 프로모션
  - 배송지
- Size: 1,254,585 x 16

df_transaction

1 문제 정의

Part4 프로젝트 수행결과

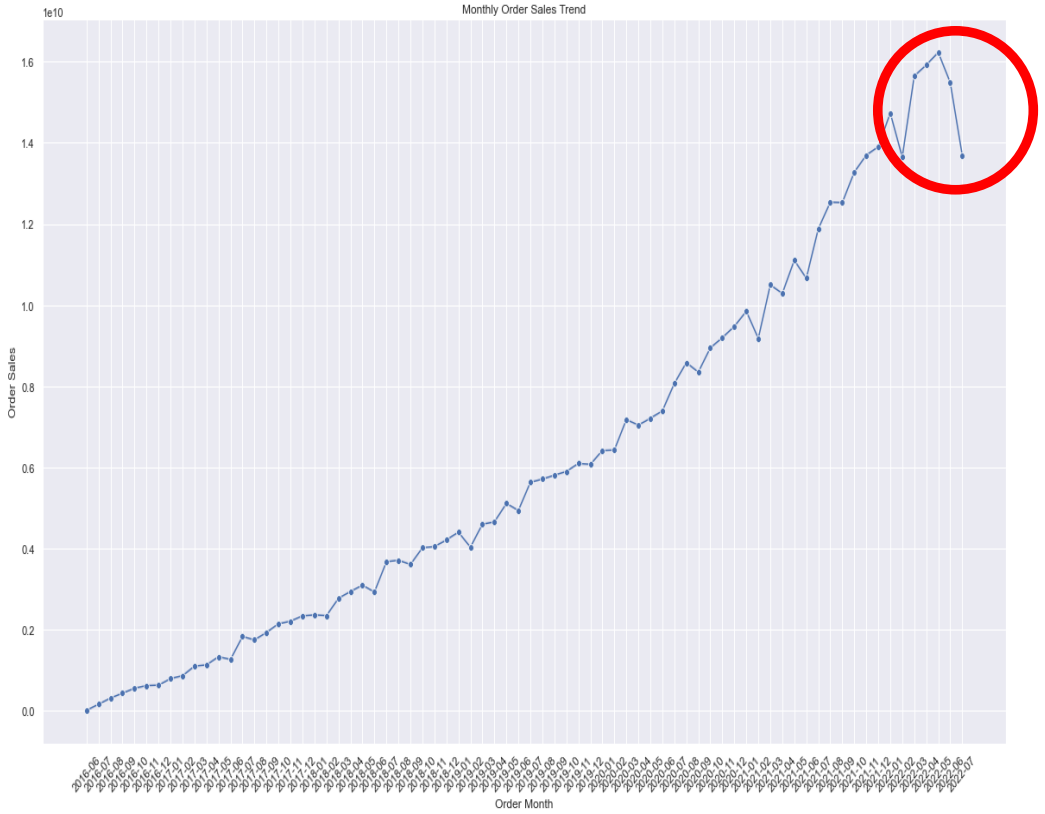
step1 **문제정의** | step2 가설수립 | step3 가설검정 | step4 지표도출 | step5 시각화

문제의 정의 및 목표

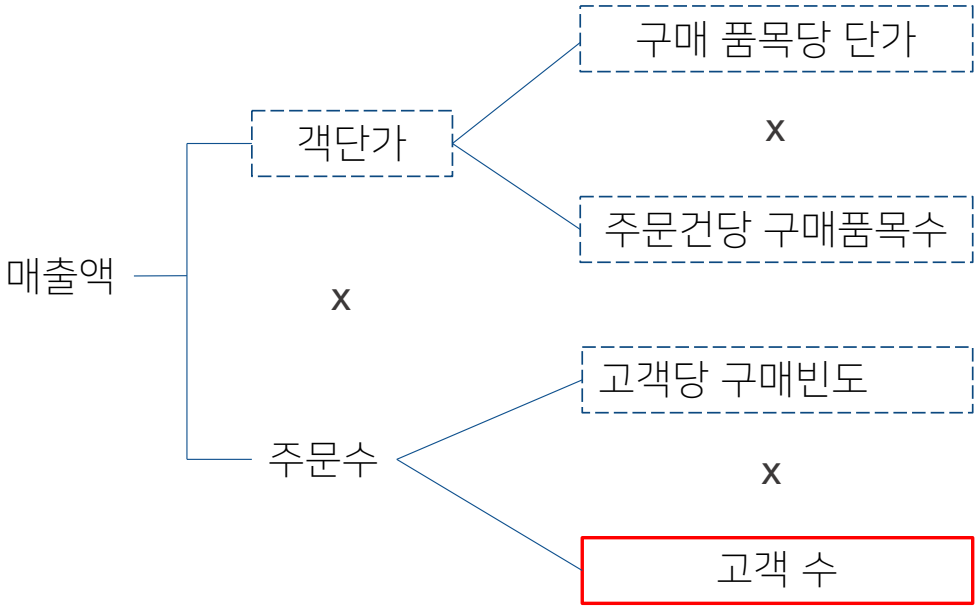
최근 연속 2개월(22.06-07) 매출액이 큰 폭으로 감소

2021년도 매출액 평균 증감률은 3.5% 였으나,
최근 2개월 -6.7% 로 감소함

KPI 를 전년도 평균 증감률의 80%인 2.9%로 설정
이를 달성하기 위한 최근 매출액 급감원인에 대한
문제원인 분석 및 해결액션 아이템 필요

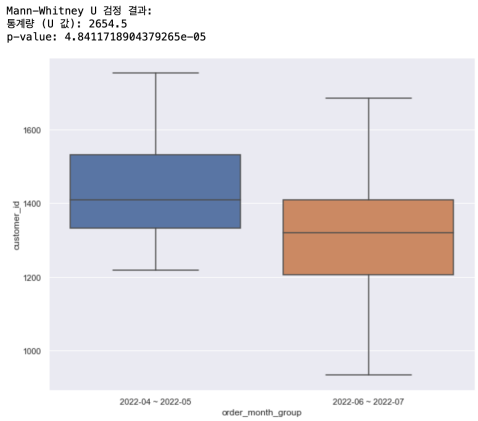


e-commerce이슈트리



이슈 트리를 활용한 가설수립 및 검정

최근 2개월 간,
점선(- - - -) 박스 들은 매출액 급감에 대한 통계적 유의미성을 입증 할 수 있는 변화가 없었음



반면. 붉은실선(——)
박스에 있는 고객 수의 급감을 확인

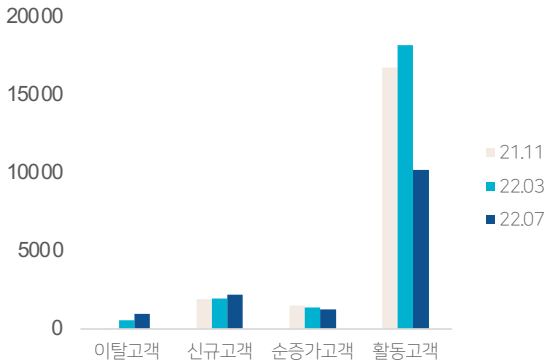
따라서 " 고객 수"에 집중하여 매출액 증가방안 모색

4 지표도출

Part4 프로젝트 수행결과

step1 문제정의 | step2 가설검정 | step3 가설검정 | **step4 지표도출** | step5 시각화

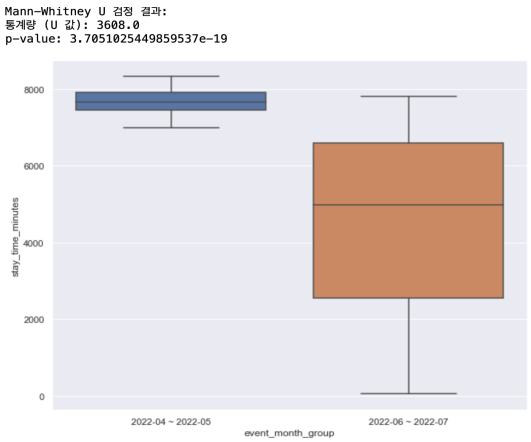
1 고객 구성 변화



최근 이탈고객 증가 및 활동고객 급감
순증가고객(신규가입-이탈) 감소

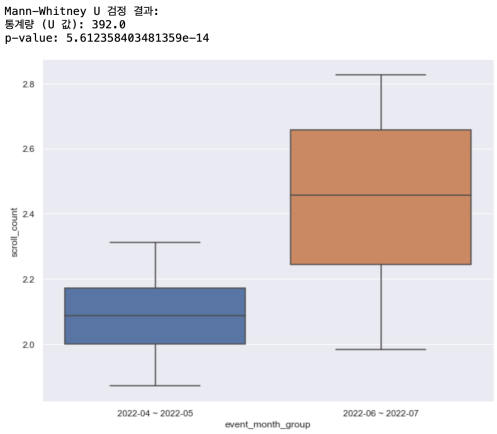
∴ 관리지표화 하여 개선 필요

2 주문까지 소요시간



22.4~5 vs 22.6~7
구매결정까지 걸린 시간이 현저히 짧아짐

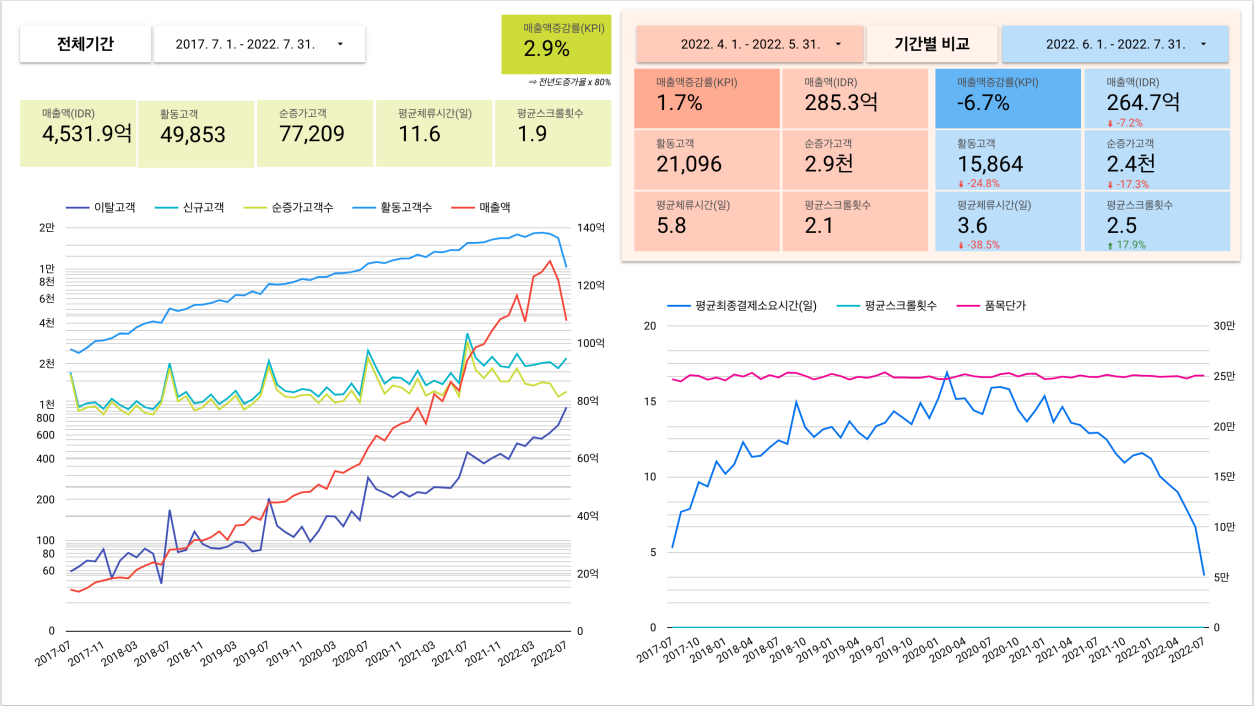
3 스크롤 횟수



22.4~5 vs 22.6~7
스크롤 횟수가 유의미하게 증가함

5 시각화

Part4 프로젝트 수행결과



대시보드

https://lookerstudio.google.com/reporting/e7efb167-777b-4b75-854d-bd932657ab4e/page/p_51ph0bkj9c

목표 매출액 증가율을 기준으로
핵심지표를 고객 구성차원에서 활동고객수,
순 증가고객수, 이탈고객, 신규 고객로 지표를 관리하고
고객 행동분석 차원에서 평균체류시간,
평균 스크롤횟수로 구분함
화면의 좌측 상단에는 기간별 비교를 할 수 있게
스코어카드를 구성하여, 해당 지표들을 모니터링
할 수 있음

이탈방지모형

1

Part4 프로젝트 수행결과

RFM 분석

랭크	Recency	Frequency (횟수)	Monetary(달러)
5	90일 이내	20 이상	5000 이상
4	180일 이내	13~19	2000~4999
3	365일 이내	7~12	800~1999
2	730일 이내	2~6	300~799
1	730일 초과	1	300미만

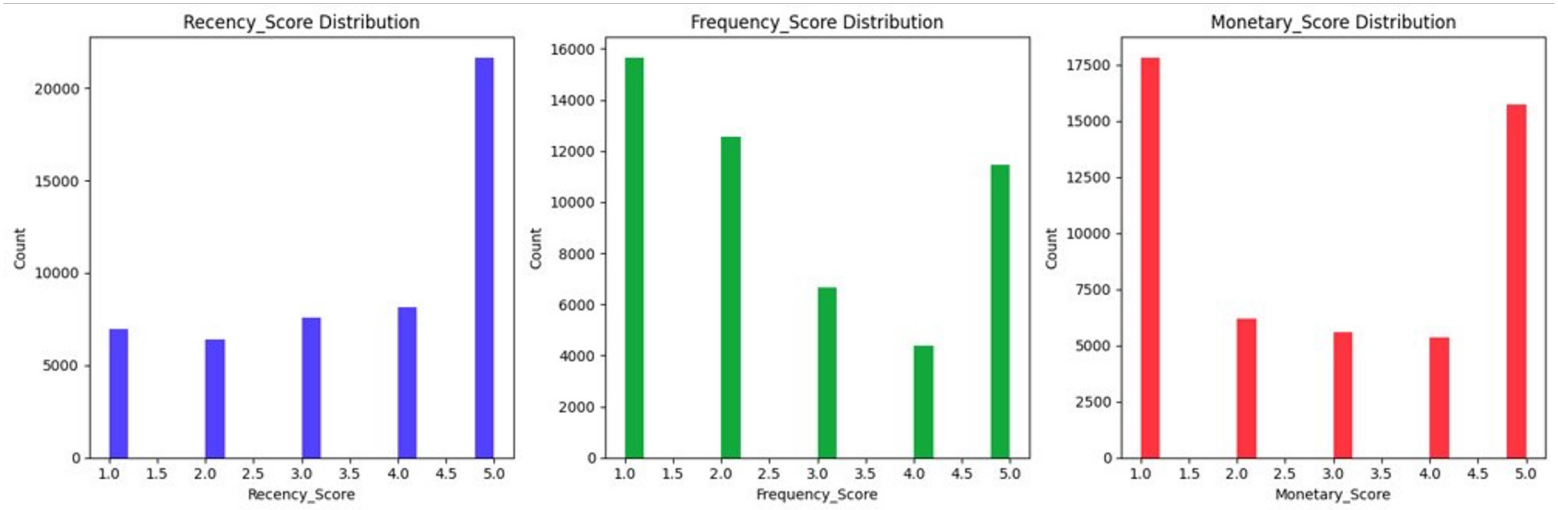
RFM 분석 표

Recency, Frequency, Monetary 칼럼을 생성
2022년 7월 31일을 최근 날짜로 표의 기준으로
고객별로 RFM 각각의 점수 데이터를 생성

2

Part4 프로젝트 수행결과

RFM 분석으로 이탈기준 정의



고객들의 RFM 시각화

개인정보 관리를 위해 1년간 접속 되지 않은 계정은 휴면 상태로 분리되기 때문에
최근에 주문한 기록이 1년이 넘어가는 Recency_Score가 1, 2인 고객들과 주문을 하지 않은
고객을 이탈 고객으로 선정

Part4 프로젝트 수행결과

3 RFM 분석으로 이용자 정의

Frequency_Score	All	5	4	3	2	1
Recency_Score						
All	50704	11444	4369	6660	12565	15666
5	21660	11077	2993	3133	3199	1258
4	8126	366	1351	2339	2682	1388
3	7561	1	25	1179	3780	2576
2	6411	0	0	9	2647	3755
1	6946	0	0	0	257	6689

이용자 정의

고객들의 등급을 나눠서 RANK 칼럼을 생성

regular, stable_customer, churn_precursor, new_customer, churn_customer 다음과 같은 5가지 등급으로 고객들을 분류

단골 고객 (regular)	안정적으로 유지되는 고객 (stable_customer)	이탈 전조가 있는 고객 (churn_precursor)	신규 고객 (new_customer)	이탈 고객 수 (churn_customer)
14070	11259	7561	4457	13357

Part4 프로젝트 수행결과

4 예측 모형에 사용할 feature

- customer_id
- gender
- device_type
- device_version
- home_location_lat
- home_location_long
- home_location
- age
- first_join_date

- booking_id
- session_id
- payment_method
- payment_status
- promo_amount
- promo_code
- created_at
- shipment_date_limit
- shipment_fee
- shipment_location_lat
- shipment_location_long

- total_amount
- product_id
- quantity
- item_price
- masterCategory
- subcategory
- articleType
- baseColour
- Season
- Year
- Usage
- productDisplayName

- device_id
- email
- first_name
- last_name
- Username
- Recency_Score
- Frequency_Score
- Monetary_Score
- Rank
- click_date
- Recency
- Frequency
- Monetary

선정이유

고객 구별을 위한 customer_id와 성별, 디바이스 정보, 주소, 최초 가입날짜 feature 사용

고객들이 구매했던 정보를 학습시키기 위해 session_id, booking_id와 프로모션 데이터, created_at, shipment_date_limit feature 사용

우수 고객일수록 많은 제품 주문과 매출을 올렸기 때문에 제품 정보와 구매 수량과 가격 feature 사용

다중공선성 문제를 일으키는 RFM 관련 칼럼과 고객의 이름, 이메일은 customer_id로 구분되고 feature importance 수치도 낮아 Drop

Part4 프로젝트 수행결과

5 비교 이탈 모형 설명

1 Logistic

사용 이유: 이진 분류 문제에서 가장 간단하고 해석하기 쉬운 모델로, 이탈 여부를 예측할 때 많이 활용

장점: 해석 가능하며, 기본적인 이탈 패턴을 파악하기에 적합

한계: 선형 경계를 가정하므로 비선형적인 패턴을 잡아내기 어려울 수 있음

2 Random Forest

사용 이유: 다양한 피처의 상호작용을 고려하여 복잡한 이탈 패턴을 예측가능

장점: 과적합을 줄이는 데 우수하며, 변수 중요도를 제공하여 피처 선택에 활용할 수 있음

한계: 해석력이 상대적으로 떨어지며, 특정 피처의 중요도에 치우칠 수 있음

3 XGBClassifier

사용 이유: 고차원 데이터와 불균형한 클래스를 다루는 데 효과적이며, 높은 예측 성능을 제공

장점: 과적합 방지와 예측 정확도 향상을 위해 그레이디언트 부스팅 알고리즘을 활용

한계: 하이퍼파라미터 튜닝이 필요

Logistic 모델

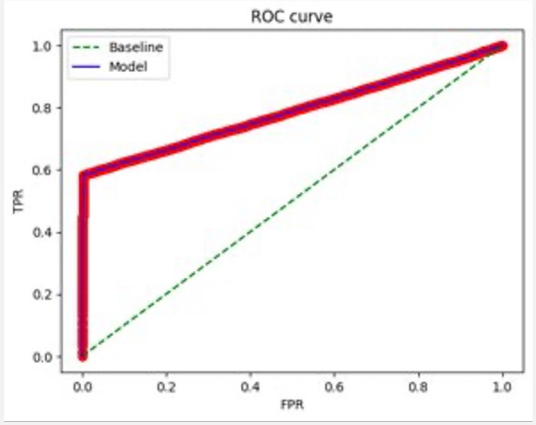
StandardScaler 미사용

검증 정확도 0.87124

	precision	recall	f1-score	support
0	0.97	0.89	0.93	228525
1	0.28	0.63	0.39	15953
accuracy			0.87	244478
macro avg	0.63	0.76	0.66	244478
weighted avg	0.93	0.87	0.89	244478

accuracy(정확도) is 0.87124
precision(정밀도) is 0.28184
recall(재현율) is 0.62866
f1점수(F1 score)is 0.38920

ROC-AUC: 0.78762



StandardScaler 사용

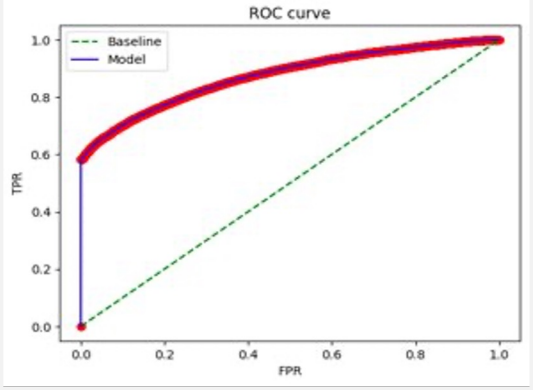
scale_pos_weight 미사용

검증 정확도 0.88599

	precision	recall	f1-score	support
0	0.98	0.90	0.94	228525
1	0.33	0.70	0.45	15953
accuracy			0.89	244478
macro avg	0.65	0.80	0.69	244478
weighted avg	0.93	0.89	0.90	244478

accuracy(정확도) is 0.88599
precision(정밀도) is 0.32653
recall(재현율) is 0.70319
f1점수(F1 score)is 0.44597

ROC-AUC: 0.87989



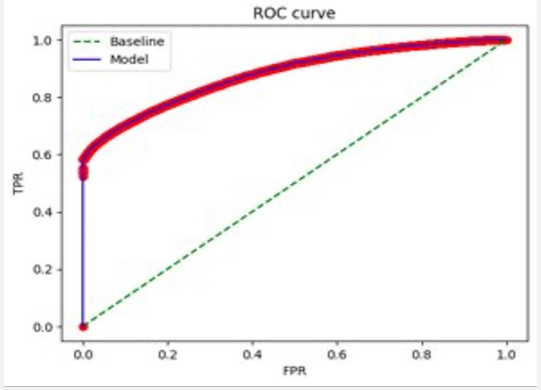
scale_pos_weight 사용

검증 정확도 0.88599

	precision	recall	f1-score	support
0	0.98	0.90	0.94	228525
1	0.33	0.70	0.45	15953
accuracy			0.89	244478
macro avg	0.65	0.80	0.69	244478
weighted avg	0.93	0.89	0.90	244478

accuracy(정확도) is 0.88599
precision(정밀도) is 0.32653
recall(재현율) is 0.70319
f1점수(F1 score)is 0.44597

ROC-AUC: 0.87989



Part4 프로젝트 수행결과

6 RandomForset 모델

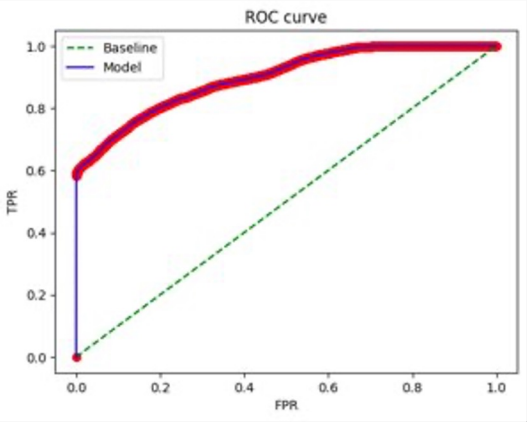
StandardScaler 미사용

검증 정확도 0.97271

	precision	recall	f1-score	support
0	0.97	1.00	0.99	228525
1	1.00	0.58	0.74	15953
accuracy			0.97	244478
macro avg	0.99	0.79	0.86	244478
weighted avg	0.97	0.97	0.97	244478

accuracy(정확도) is 0.97271
precision(정밀도) is 1.00000
recall(재현율) is 0.58177
f1점수(F1 score)is 0.73559

ROC-AUC: 0.89842



StandardScaler 사용

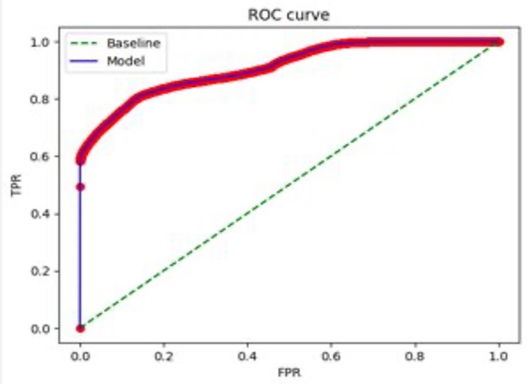
scale_pos_weight 미사용

검증 정확도 0.97263

	precision	recall	f1-score	support
0	0.97	1.00	0.99	228525
1	0.99	0.58	0.74	15953
accuracy			0.97	244478
macro avg	0.98	0.79	0.86	244478
weighted avg	0.97	0.97	0.97	244478

accuracy(정확도) is 0.97263
precision(정밀도) is 0.99360
recall(재현율) is 0.58434
f1점수(F1 score)is 0.73590

ROC-AUC: 0.90971



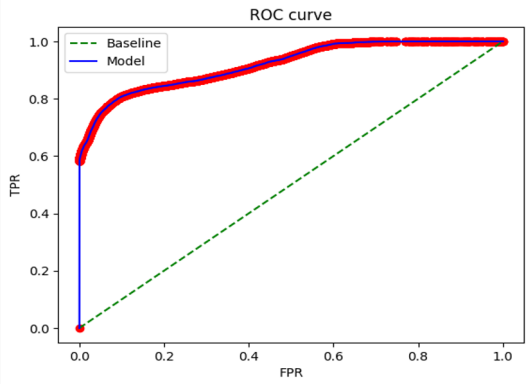
scale_pos_weight 사용

검증 정확도 0.96704

	precision	recall	f1-score	support
0	0.97	0.99	0.98	228525
1	0.82	0.63	0.71	15953
accuracy			0.97	244478
macro avg	0.90	0.81	0.85	244478
weighted avg	0.96	0.97	0.96	244478

accuracy(정확도) is 0.96704
precision(정밀도) is 0.82366
recall(재현율) is 0.62979
f1점수(F1 score)is 0.71379

ROC-AUC: 0.91983



6 XGBClassifier 모델

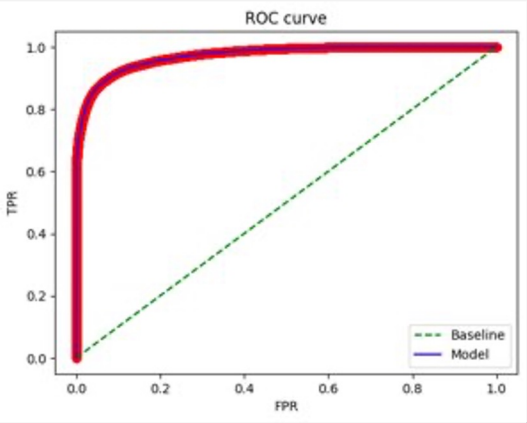
StandardScaler 미사용

검증 정확도 0.97513

	precision	recall	f1-score	support
0	0.97	1.00	0.99	228525
1	0.98	0.63	0.77	15953
accuracy			0.98	244478
macro avg	0.98	0.81	0.88	244478
weighted avg	0.98	0.98	0.97	244478

accuracy(정확도) is 0.97513
precision(정밀도) is 0.98468
recall(재현율) is 0.62860
f1점수(F1 score)is 0.76734

ROC-AUC: 0.97250



StandardScaler 사용

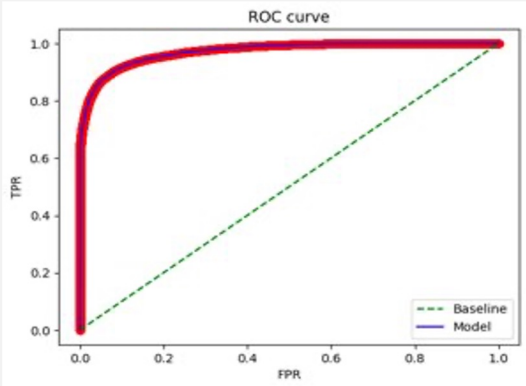
scale_pos_weight 미사용

검증 정확도 0.96004

	precision	recall	f1-score	support
0	0.99	0.97	0.98	228525
1	0.66	0.82	0.73	15953
accuracy			0.96	244478
macro avg	0.82	0.89	0.85	244478
weighted avg	0.97	0.96	0.96	244478

accuracy(정확도) is 0.96004
precision(정밀도) is 0.65530
recall(재현율) is 0.81784
f1점수(F1 score)is 0.72760

ROC-AUC: 0.97034



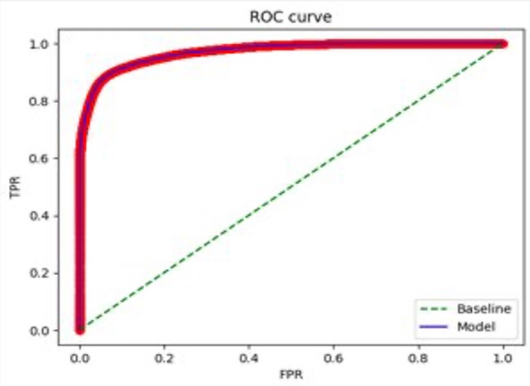
scale_pos_weight 사용

검증 정확도 0.96004

	precision	recall	f1-score	support
0	0.99	0.97	0.98	228525
1	0.66	0.82	0.73	15953
accuracy			0.96	244478
macro avg	0.82	0.89	0.85	244478
weighted avg	0.97	0.96	0.96	244478

accuracy(정확도) is 0.96004
precision(정밀도) is 0.65530
recall(재현율) is 0.81784
f1점수(F1 score)is 0.72760

ROC-AUC: 0.97034



7

Part4 프로젝트 수행결과

모델 비교 및 선정

MODEL	Accuracy (정확도)	Precision (정밀도)	Recall (재현율)	F1 score (F1 점수)
Logistic	0.88599	0.32653	0.70319	0.44597
RandomForest	0.96704	0.82366	0.62979	0.71379
XGBClassifier	0.96004	0.65530	0.81784	0.72760

데이터 전처리

범주형 변수 처리를 위해 차원을 더 늘리지 않는 OrdinalEncoder 사용
이상치 처리를 위해 StandardScaler 사용
데이터들을 취합한 후 churn의 비율이 0 : 1218797, 1 : 85083 로 불균형한 데이터이기 때문에 scale_pos_weight 사용하기로 결정, 실제 테스트 결과도 사용한 쪽이 더 정확

모델들을 검증 데이터로 비교해 봤을 때 F1 score가 가장 높은 XGBClassifier이 가장 좋기 때문에 선택

8

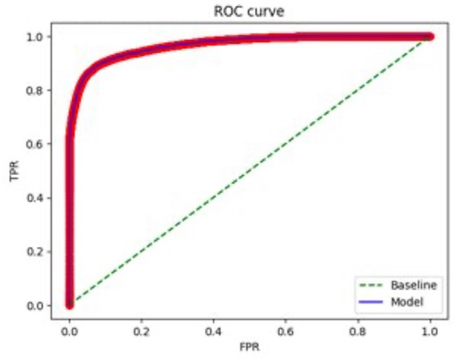
고객들의 이탈 예측 모델

MODEL	Accuracy (정확도)	Precision (정밀도)	Recall (재현율)	F1 score (F1 점수)
검증 데이터 수치	0.96007	0.65735	0.81076	0.72604
테스트 데이터 최종 결과	0.96002	0.65765	0.80781	0.72504

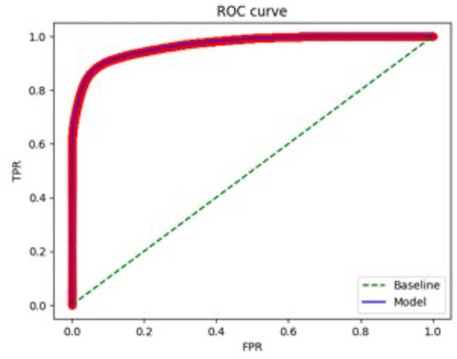
하이퍼 파라미터 튜닝

RandomizedSearchCV와 GridSearchCV 사용하여
max_depth, learning_rate, min_child_weight,
colsample_bytree 이 4가지 파라미터 들을 튜닝
best_parameter는 각각 7, 0.12, 4, 0.6

검증 데이터 사용



테스트 데이터 사용



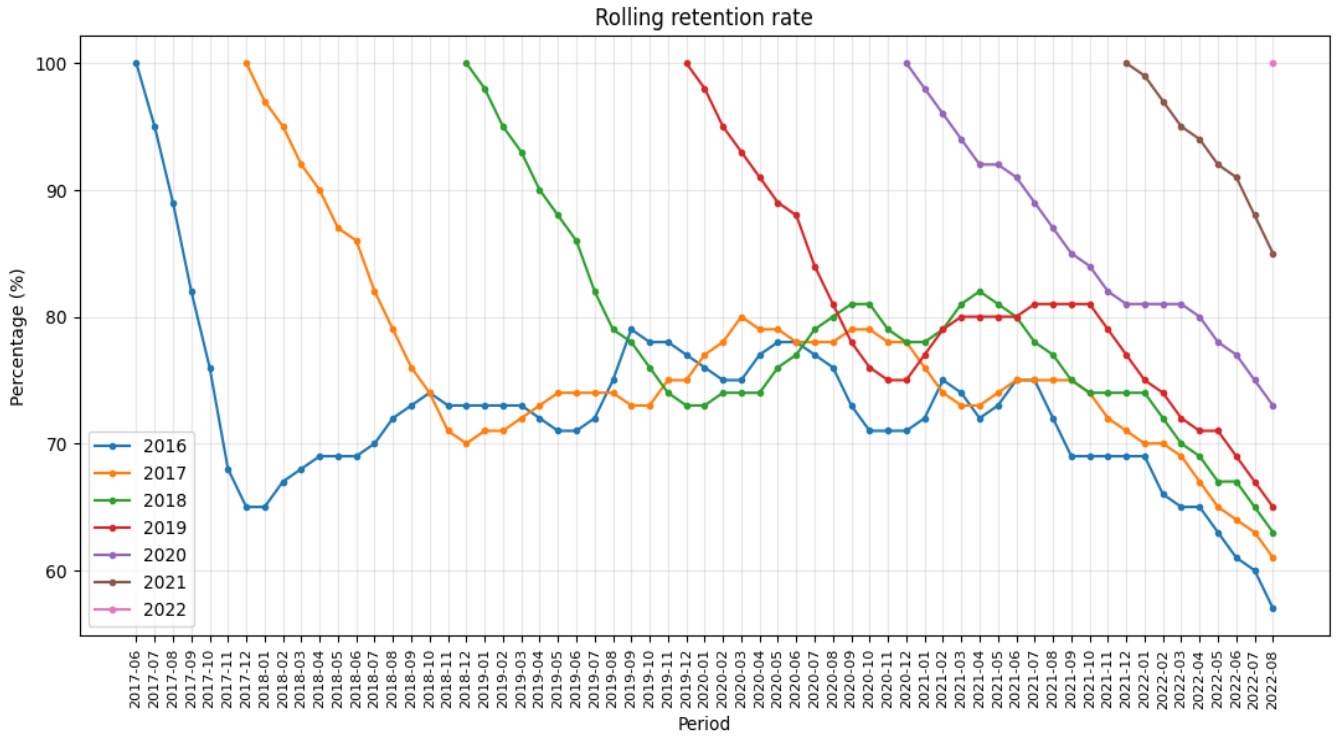
리텐션분석

1 리텐션

롤링 리텐션

리텐션 : 사용자들이 서비스를 얼마나 꾸준히 이용하고 있는지를 보여주는 지표.
롤링 리텐션 : 리텐션의 한 종류로 기준일을 포함하여 그 이후에 한 번이라도 재방문한 사용자의 비율을 나타냄.

구매 기록 데이터를 이용한 그래프.
구매일 포함, 1년 이내에 재구매가 없으면 이탈한 것으로 간주.
최근 약 1년 동안 구매 감소.



Part4 프로젝트 수행결과

2 고객 세분화 : 구매 기록에 따른 고객 분류

활동 고객
가입한 지 1년이 넘고 구매가 최근 1년 이내에 이루어진 고객

복귀 고객
가입한 지 1년이 넘고 6개월간 구매 이력이 없다가 최근 다시 구매가 이루어진 고객

신규 고객
최근 1년 이내에 가입과 첫 구매가 이루어진 고객

휴면 고객
가입한 지 1년이 넘고 최근 1년 동안 구매 이력이 없는 고객

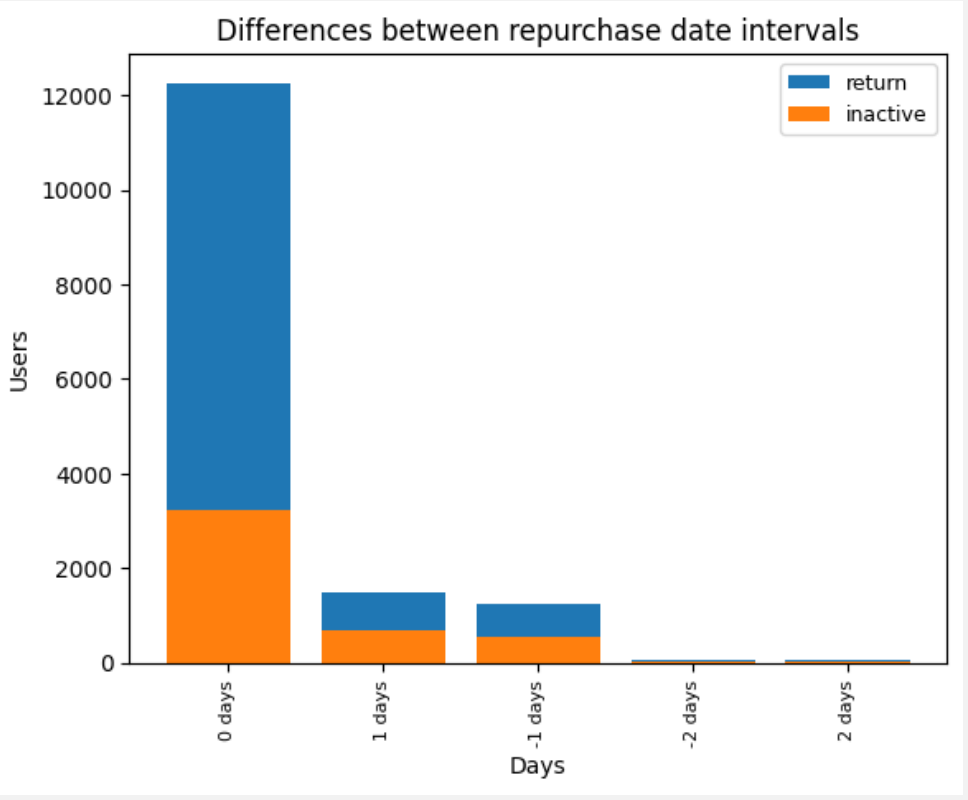
3 고객 분석 : 리텐션율 상승을 위한 고객 분석

구매 리텐션율을 올리기 위해 구매가 활발한 고객들보다 복귀, 휴면 고객의 구매를 늘리는 것이 중요. 따라서 복귀, 휴면 고객의 분석을 통해 리텐션율의 상승 방법을 구상.

복귀 고객과 휴면 고객의 재구매 날짜 간격 차이

복귀, 휴면 고객 중 세 번 이상 구매가 이루어진 고객들의 기록에서 구매간 간격의 차를 나타낸 그래프. 예를 들어, A의 재구매가 365일 뒤에 이루어졌다면 세 번째 구매는 약 363일에서 367일 후에 이루어짐.

재구매 날짜 간격의 차이를 줄인다면 휴면 고객과 복귀 고객을 활동 고객으로 전환할 수 있을 것.

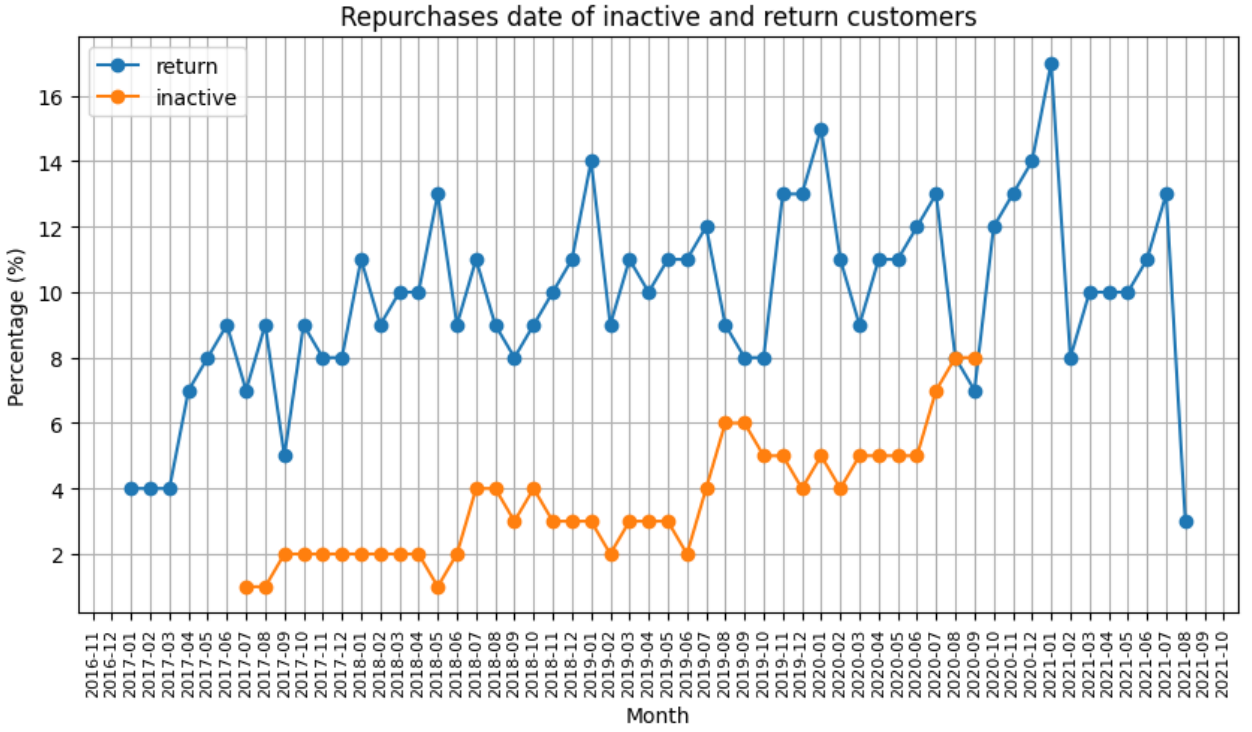


3 고객 분석 : 리텐션율 상승을 위한 고객 분석

복귀 고객과 휴면 고객의 재구매 시기

복귀 고객 : 대체로 1월과 7월에 재구매 수가 높고 그 후로 급락하는 모습. 약 6개월 간격으로 반복적인 패턴.

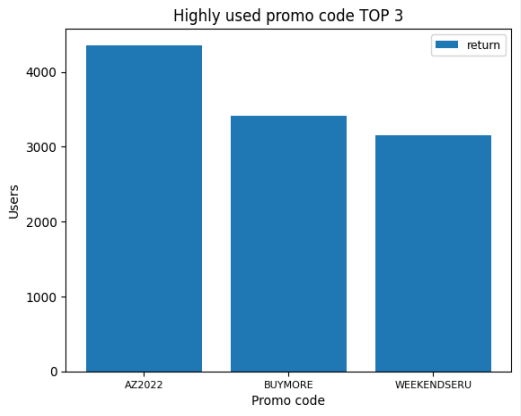
휴면 고객 : 미세하지만 대부분 여름에 재구매 수가 소폭 증가.

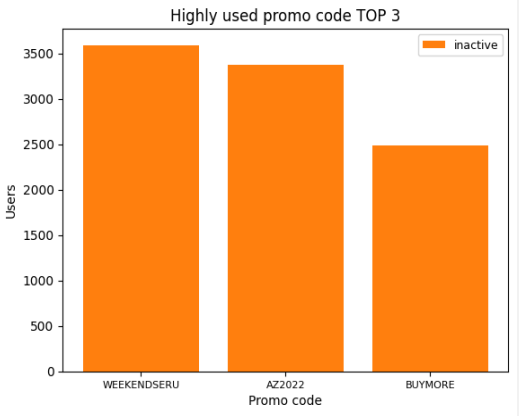


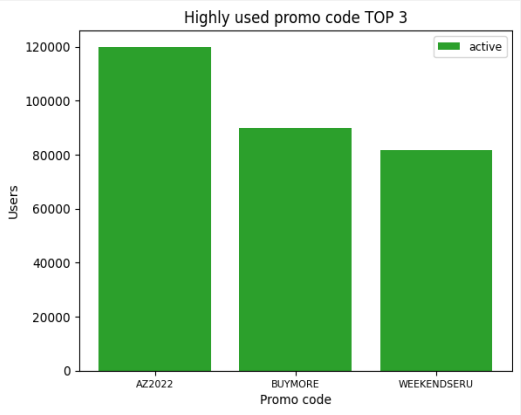
3 고객 분석 : 리텐션율 상승을 위한 고객 분석

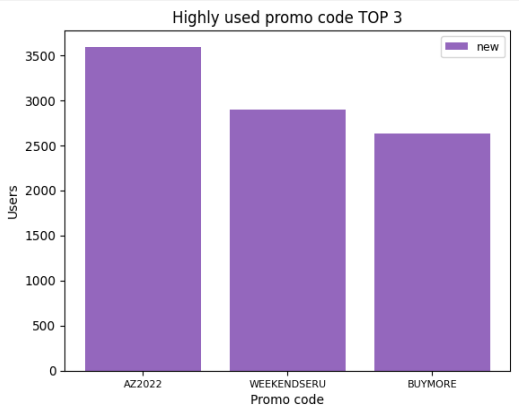
사용된 프로모션 코드

사용된 프로모션 중 상위 3개의 항목 확인.
가장 많이 사용된 코드는 휴면 고객은 WEEKENDSERU, 나머지 고객들은 AZ2022.







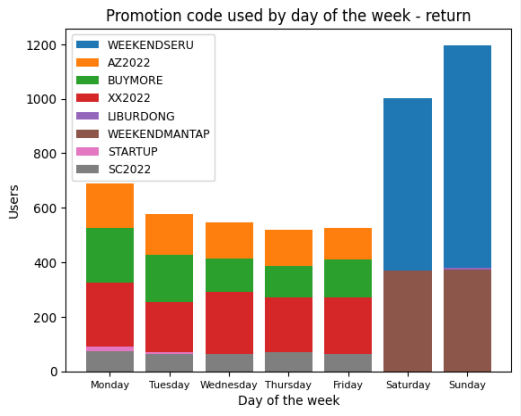


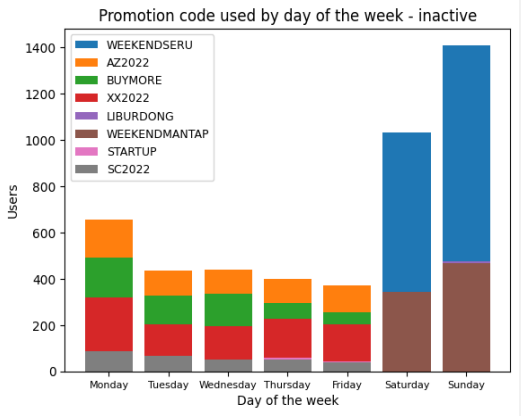
3 고객 분석 : 리텐션율 상승을 위한 고객 분석

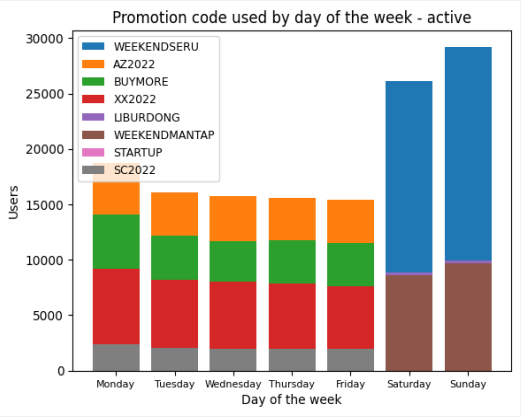
사용된 프로모션 코드

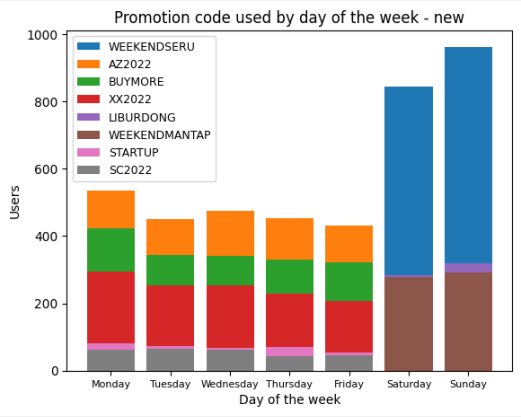
사용된 프로모션 중 상위 3개의 항목 확인.
가장 많이 사용된 코드는 휴면 고객은
WEEKENDSERU, 나머지 고객들은 AZ2022.

WEEKENDSERU는 주말에 사용되는 코드, AZ2022
코드와 BUYMORE 코드는 평일에 사용할 수 있는 코드.





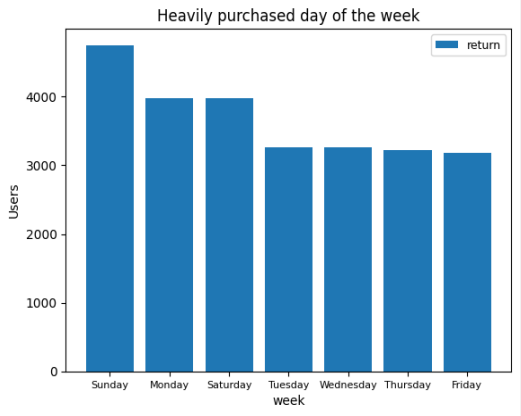


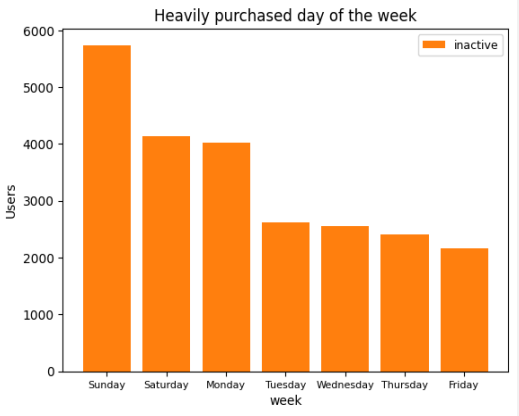


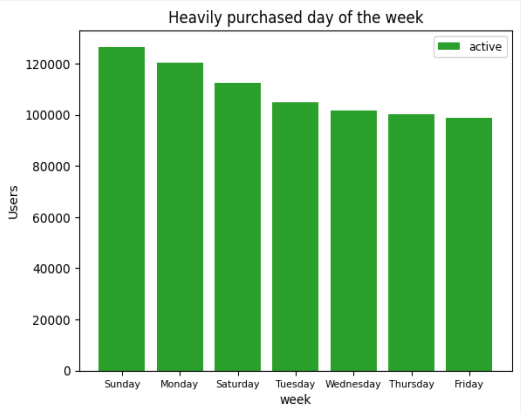
3 고객 분석 : 리텐션율 상승을 위한 고객 분석

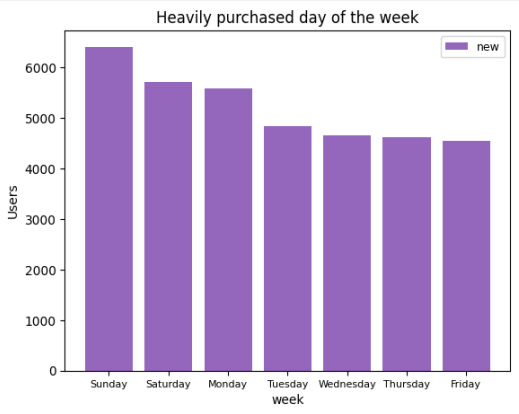
구매가 이루어지는 시간

휴면고객을 제외한 나머지 고객들은 평일과 주말의 구매 횟수 차이가 크지 않지만 휴면 고객은 월요일을 뺀 평일과 주말의 구매 횟수 차이가 큼.









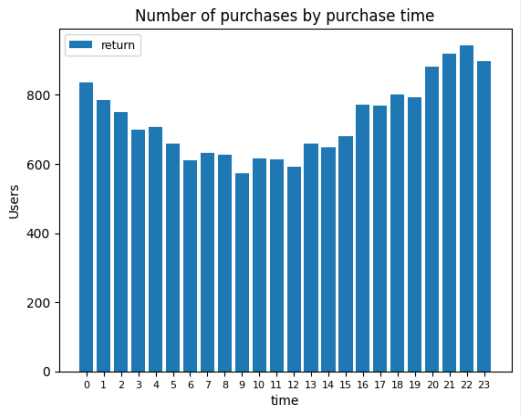
Part4 프로젝트 수행결과

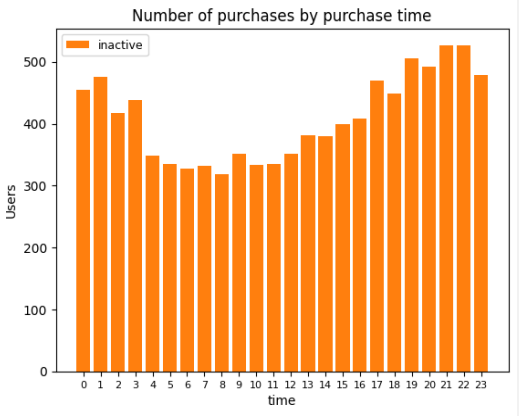
3 고객 분석 : 리텐션율 상승을 위한 고객 분석

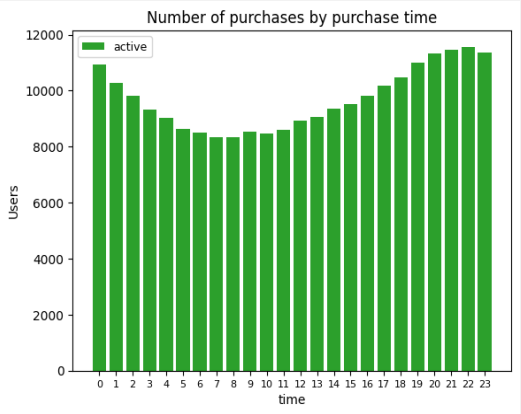
구매가 이루어지는 시간

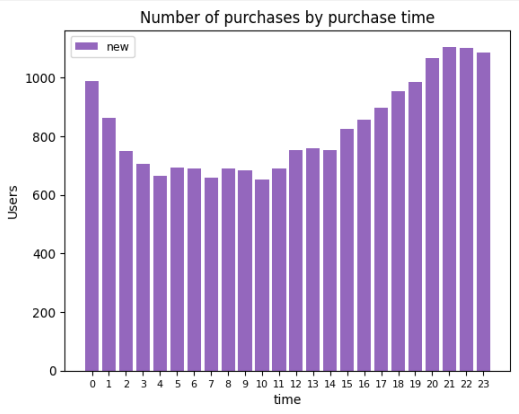
휴면고객을 제외한 나머지 고객들은 평일과 주말의 구매 횟수 차이가 크지 않지만 휴면 고객은 월요일을 뺀 평일과 주말의 구매 횟수 차이가 큼.

휴면 고객의 경우 오전 4시부터 오후 4시까지 13시간 동안 구매가 적고 오후 5시부터 오전 3시까지 11시간 동안 구매가 상대적으로 많음.
전체적으로 오전보다는 오후 시간대의 구매를 선호하고 대체적으로 오후 10시에 가장 활발한 구매가 이루어 짐.
7시에서 9시 사이의 구매 활동이 가장 적음.









Part4 프로젝트 수행결과

4 액션 아이템 제안

Item 01

알림 서비스 제공

비슷한 간격을 두고 재구매가 일어나므로 고객별 구매가 이루어진 날짜들을 계산하여 **다음 구매 예정일 이전(예: 30일 이전)에 알림 서비스를 제공**하여 간격을 줄여 나갈 수 있도록 함.

Item 02

기한제 포인트 도입

고객들의 재구매 간격을 줄이기 위해 구매 시 **구매 금액의 일정 부분을 포인트로 지급하고 포인트 사용 기한을 6개월로 지정**하여 더 많은 고객들의 재구매가 이루어지도록 유도.

Item 03

프로모션 개선

많은 구매가 이루어지는 것은 대부분 주말, 오전보다 오후에 구매가 활발.
상대적으로 구매가 적은 **평일 7시-9시 사이에 한시적으로 사용할 수 있는 프로모션 코드**를 도입하여 구매 유도. 구매 횟수가 좀 더 많은 시기인 7월에 진행.

Item 04

활동 고객 우대

기존 고객들의 이탈을 막기 위하여 1년 동안의 구매 횟수에 따라 **등급을 부여해 등급에 따른 혜택을 줌으로써 충성도를 높이고 지속적인 구매**가 이루어질 수 있도록 함.

Part 5

자체 평가 의견

Part5

1 프로젝트 회고

문제를 먼저 정의하고 분석을 진행한 것이 방향을 잡는데 도움이 되었다.
같은 데이터여도 배경지식과 도메인 지식을 알았을 때와 몰랐을 때 그래프에 대한 이해나 도출되는 인사이트가 달라지는 것을 느낄 수 있었다.
RFM 분석, 리텐션 분석 등의 여러 분석기법을 배우고 실습해볼 수 있어서 좋았다.
데이터로 고객들의 특징을 분석하여 이탈할 것으로 예측되는 고객을 분류하는 의사결정 과정을 경험할 수 있어서 좋았다.

경동연
팀장

프로젝트에 대한 이해도가 많이 낮아 개념적인 이해를 하는 것만으로도 많은 시간이 걸렸다. 그래도 팀원들이 자료들을 공유해주고 프로젝트의 전반적인 흐름과 큰 틀을 잡아주어서 버겁지만 따라갈 수 있었다.
업무의 방향을 잘 잡지 못해서 프로젝트에 많은 기여를 하지 못해 팀원들에게 미안하고 아쉽다.

이주연
팀원

처음 접해보는 다소 긴 기간의 프로젝트를 수행하며 주어진 시간을 효율적으로 관리하고 팀원들과 지속적으로 소통하는 방법을 익히는 경험을 할 수 있었다.
프로젝트의 경우 실제 업무에서 있을 법한 상황을 주제로 진행되어 후에 현장에서 분석가로 경험하게 될 상황을 체험하는 기분이었다.
다만 프로젝트의 완성도 측면에서는 아직 부족한 점이 많은 것 같아 프로젝트가 마무리된 이후에도 다시 정리하면서 채워 넣을 계획이다.

방은혜
팀원

긴 시간 발표를 들어주셔서 감사합니다.