[프로그래머스_데브코스_최종 프로젝트]

Product 매출 증대를 위한

고객 로그 분석 및 해결방안 제안서

퍼널·세그먼트 기반 사용자 행동 분석을 통한 전환율 개선 전략 및 개인화 추천 시스템 제안

[6기] LOG - in
조원: 김예원, 박수민, 성은영, 염용기
기간: 2025.04.28 ~ 2025.05.13

목차

1. 프로젝트 개요

1.1 프로젝트 배경

디지털 환경에서의 소비자 행동은 점점 더 복잡해지고 있으며, 수많은 선택지 속에서 사용자의 이탈은 점점 더 빠르게 일어난다. 이커머스 산업 전반에서 다양한 채널을 통해 많은 사용자를 확보하고 있음에도 불구하고 유입이 실제 구매로 이어지지 않고 빠르게 이탈하는 현상이 지속되면서 단순 유입 증가 전략의 한계가 분명히 드러나고 있다.

많은 고객이 유입되고 있음에도 실제 구매까지 이어지는 전환율이 현저히 낮다는 점에서 어려움을 겪고 있다. 사용자의 구매 전환율을 높이기 위해 어떤 전략이 필요할까?

Product의 매출 증대를 위해 현재 수집되고 있는 고객의 웹사이트 행동 로그 데이터를 기반으로 현황을 대시보드로 시각화한다. 세션 단위로 정리된 데이터를 변수별로 분석하여 사용자의 전환 흐름을 각 단계별로 정의하고 분석을 통해 각 단계에서의 고객 이탈 지점 및 원인을 파악하고 이를 기반으로 매출 증가 전략을 도출하는 것을 목표로 한다.

1.2 분석 목표

1. 퍼널 분석을 통한 이탈 구간 식별
고객의 행동 흐름을 퍼널 형태로 시각화하고, 각 단계에서 이탈률이 높은 구간을 식별하여 개선의 우선순위를 설정한다.

2. 이탈 방지를 위한 전략 도출
퍼널 내 핵심 이탈 구간을 중심으로 사용자를 두 그룹을 나누고 분류모델을 활용하여 두 그룹에 각 그룹에 영향을 미치는 주요 속성을 분석한다.

3. 고객 세그먼트 기반 타겟 전략 수립
전환 이력을 기반으로 고객을 다양한 속성 기준으로 세분화하고, 각 세그먼트 별로 이탈 가설을 수립하고 검증한다.

4. 추천 시스템을 통한 제품 소개 확대
고객 행동 데이터를 기반으로 개인화된 추천 시스템을 적용하여, 관련성 높은 제품을 효과적으로 제안하여 구매 전환율을 높인다.

1.3 기대 효과

1. 매출 증대
이탈률 감소 및 제품 추천 최적화를 통해 고객 1인당 평균 구매 금액을 증가시킨다.

2. 고객 행동 기반의 데이터 중심 의사결정 강화
퍼널 분석 및 세그먼트 전략을 통해 마케팅과 제품 운영에 있어 보다 정량적인 인사이트를 제공한다.

3. 지속 가능한 고객 여정 관리 체계 구축
고객 이탈 원인을 지속적으로 모니터링하고 개선할 수 있는 분석 프레임워크를 정립 한다.

1.4 프로젝트 팀 구성 및 역할

역활	이름	담당
멘토	이가람	프로젝트 질의 응답 및 피드백
팀원	김예원	EDA, 데이터 추출 및 전처리, 퍼널 정의 및 분석, 분류 모델 생성 및 분석
팀원	박수민	EDA, 장바구니 추천 시스템 학습 및 모델 고도화
팀원	성은영	EDA, 시각화 및 세션 별 대시보드 생성
팀원	염용기	EDA, 퍼널 정의 및 분석, 구매 기반 추천 시스템 학습

2. 데이터 수집 및 준비

2.1 데이터 소개

데이터는 **Google Analytics Merchandise Store**의 실제 방문 데이터를 기반으로 생성된 샘플 데이터로 사용자의 행동 로그와 사용자 및 세션에 대한 속성 정보가 포함되어 있다.

- **출처**: Google BigQuery 공개 데이터셋 - Google Analytics Sample Dataset
- **총 데이터**: 25,343,052행
- **전체 컬럼**: 338개
- **수집 기간**: 2016년 8월 1일 ~ 2017년 8월 1일

주요 컬럼 설명

1. **기본 정보**

변수명	설명	값 예시
fullVisitorId	방문자 ID (쿠키 기반 고유 식별자)	1234567890123456789
visitId	세션 방문 ID (timestamp 기반)	1505670000
visitNumber	해당 사용자의 방문 횟수 (방문자 ID 기준)	4
date	방문 날짜 (YYYYMMDD 형식)	20160801

2. **세션 요약 통계**: totals 필드

변수명	설명	값 예시
hits	세션 중 발생한 전체 히트 수	27
pageviews	세션 중 발생한 총 페이지뷰 수	14
timeOnSite	세션의 총 체류 시간	37,000 (ms 단위)
transactions	세션 중 발생한 거래 건수	1
bounces	진입 직후 이탈 여부	1(이탈), null(잔존)

3. **유입 경로**: trafficSource 필드

변수명	설명	값 예시
campaign	마케팅 이름	AW - Accessories (AW: Google Ads)
referralPath	유입 외부 웹페이지 URL 경로	/yt/about/ (유튜브 소개 페이지)
medium	트래픽 유입 마케팅 매체	affiliate (제휴 마케팅 링크)

4. **디바이스 정보**: device 필드

변수명	설명	값 예시
browser	사용자 이용 브라우저 이름	Chorme, Safari
deviceCategory	접속 디바이스 유형	desktop, mobile, tablet

5. **지리 정보**: geoNetwork 필드

변수명	설명	값 예시
continent	사용자 접속 위치의 대륙	Americans, Asia
country	사용자 접속 위치의 국가	United States, India
city	사용자 접속 위치의 도시	Mountain View, New York

6. **세션 내 행동 로그**: hits 필드

변수명	설명	값 or 컬럼 예시
hitNumber	세션 내 히트 순서	3
hour, minute	히트 발생 시간	13(HH), 07(MM)
page 필드	방문한 페이지 URL 및 HTML	page.pagePath
product 필드	상품 이름,수익 / 클릭, 노출 여부 등	product.isClick
promotion 필드	프로모션 고유 ID, 노출 페이지 위치 등	promotion.promoId
eCommerceAction 필드	전자 상거래 행동 유형	eCommerceAction.action_type
social 필드	유입 소셜 네트워크 이름 등	social.socialNetwork

2.2 데이터 전처리

1) 기본 전처리

전체 338개의 컬럼 중 기본 전처리 후 분석에서 사용 가능한 71개의 컬럼을 선별하였다.

- **결측 컬럼 제거**: 컬럼의 전체 값이 해당 필드에 아예 존재하지 않거나(null), 사용자 개인 식별 우려 또는 광고 설정 보호 등의 이유로 데이터 공개 시 고의적으로 마스킹 되어있는 경우(not available in demo dataset) 분석에서 제외하였다.

- **단일값 컬럼 제거**: 컬럼의 전체 값이 같은 값으로 정보를 포함하고 있지 않은 경우 분석에서 제외하였다. ex. 세션 방문 수(visits) 컬럼의 모든 값은 1 고정

- **동일값 컬럼 제거**: 다른 컬럼과 모든 값이 동일하여 중복된 정보를 담고 있어 분석에서 제외하였다. ex. 거래 총 수익과 현지 통화 기준 거래 총 수익은 모두 USD로 동일

2) 파생변수 생성

- **session**: fullVistorId와 visitId 두 컬럼을 조합하여 한 세션을 구분하는 단위 컬럼을 생성하였다. ex. 76529729049367813_1482366901

- **converted**: transactions(세션 중 발생한 거래 건수) 컬럼에서 값이 1 이상인 경우 1 나머지를 0으로 구매 여부를 나타내는 컬럼을 생성하였다.

3) 세션 단위 정보 요약 전처리

전체 데이터는 한 사용자 세션 안에서 발생하는 전체 사용자 행동(hits)을 기록한 형태로 세션 단위 필드들이 여러 행에 반복되어 나타나는 총 2,500백만 건의 대용량 데이터이다.

예시)

fullVistorId	visitId	hitNumber	pagePath
12345	111	1	/home
12345	111	2	/products
12345	111	3	/cart

전체 기간에 걸친 고객들의 데이터를 활용하기 위해 hits 단위의 데이터를 세션 단위 데이터로 정보를 요약하여 총 903,653건의 데이터를 추출하였다.

- **사용자 행동 경로 요약**: 히트 흐름에 따라 변하는 경로를 요약하여 Path,Flow 컬럼을 새로 생성하였다. ex. Home > products > products > Log In > Shopping Cart …

- **상품·마케팅 세부 정보 요약** : 여러 개의 세부 정보를 하나의 컬럼으로 요약하여 Summary 컬럼을 새로 생성하였다. ex. 1.프로모션 클릭 여부 | 2.프로모션 이름 | 3. 프로모션 위치…

- **중복 경로 요약**: 동일한 경로가 반복해서 나타난 경우 직관적인 이해와 데이터 압축을 위해 반복되는 값을 숫자로 표현하였다. ex. Home > products(2) > Log In > …

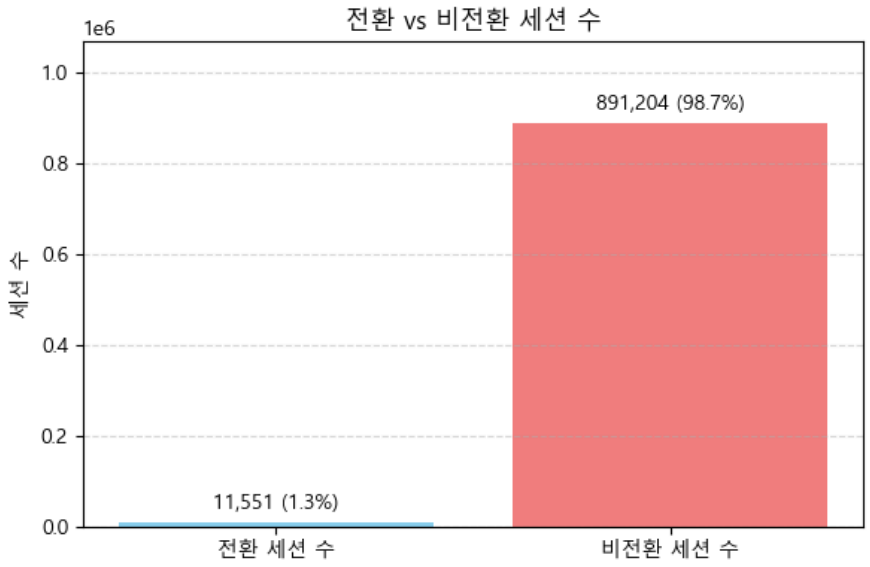
분석 사용 컬럼

전체 전처리 과정 이후 분석에서 사용되는 컬럼은 총 40개로 다음과 같다.

컬럼명	설명	컬럼명	설명
visitStartTime	세션 시작 시간	longitude	경도
dates	방문 날짜	latitude	위도
channelGrouping	트래픽 소스 그룹화 채널	city	도시
visitNumber	특정 방문자의 방문횟수	region	지역 (국가 내 시/도)
hits	세션 중 전체 히트 수	country	국가
pageviews	세션 중 페이지뷰 수	subContinent	하위 대륙
timeOnSite	세션의 총 체류 시간(초)	session	세션 단위
newVisits	신규 방문 여부	converted	전환 여부
transaction Revenue	총 구매 금액	promotionSummary Path	페이지 방문 경로
transactions	세션 중 거래 건수	hitTimeFlow	히트 시작 시간 경로
bounces	첫 화면 이탈 여부	hitTimeSpentPath	히트 체류 시간 경로
totalTransaction Revenue	전체 세션 기준 누적 거래 금액	typeFlow	상호작용 유형 흐름
campaign	마케팅 이름	pageTitlePath	페이지 타이틀 경로
referralPath	리퍼럴 URL	eventActionPath	주요 이벤트 흐름
medium	유입 매체 (어떻게)	eventLabelPath	이벤트 분석 흐름
browser	브라우저 이름	pagePathFlow	promotion 요약 흐름
deviceCategory	디바이스 유형	actionTypePath	전자상거래 행동 유형 흐름
isMobile	모바일 여부	maxActionType	전자상거래 행동 최대 깊이
operatingSystem	운영체제	socialNetwork	소셜 네트워크 이름
continent	대륙	contentGroupPath	콘텐츠 카테고리 흐름

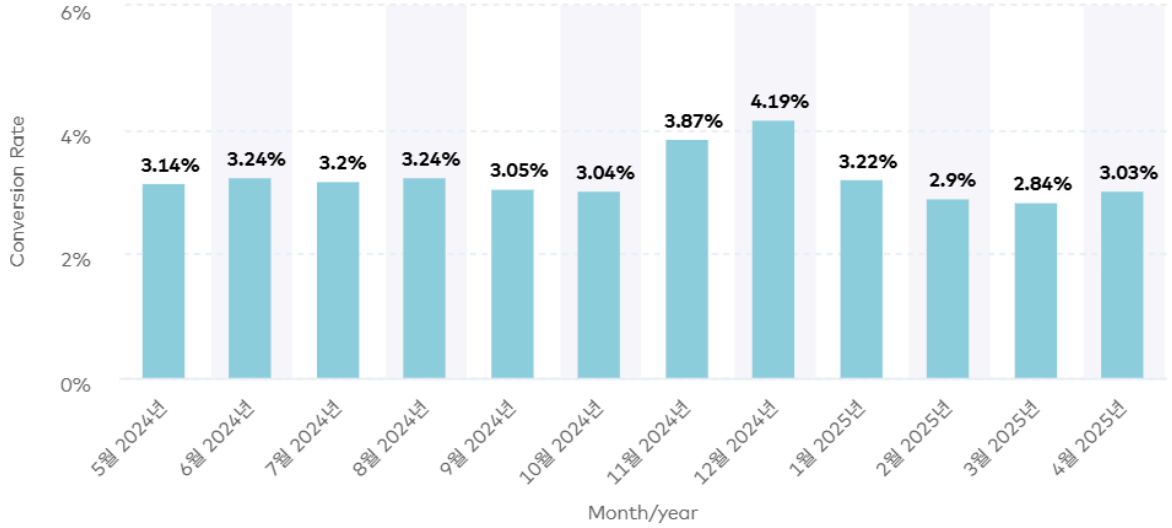
3. 탐색적 데이터 분석 (EDA)

3.1 전환율 분포 확인



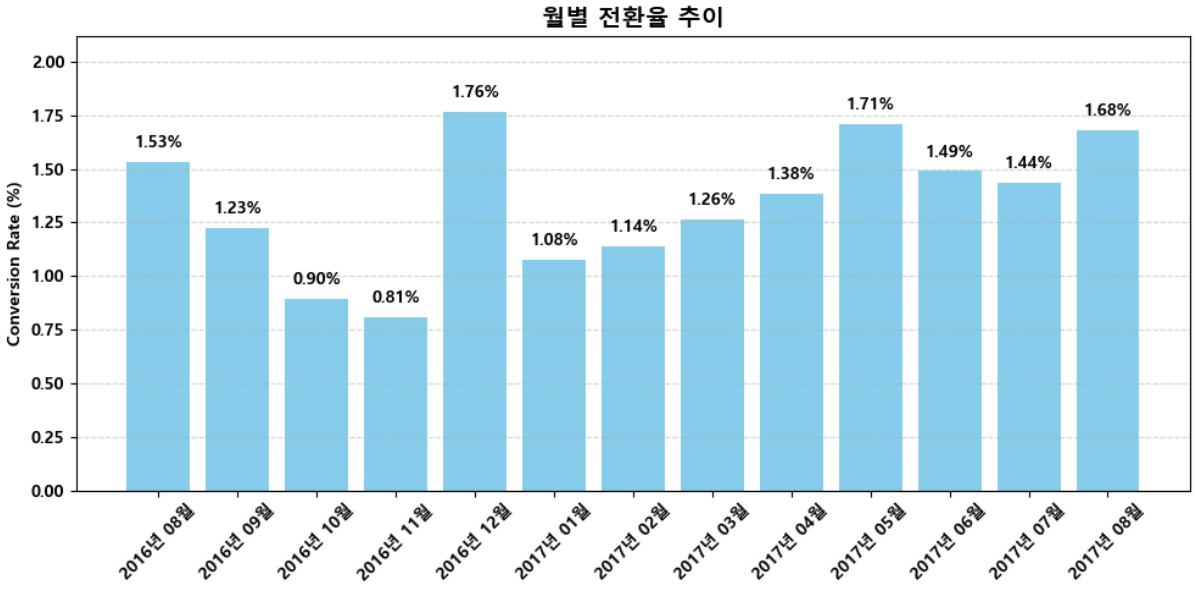
세션별 구매 건수를 기준으로 하나 이상의 구매가 발생한 세션은 전환 세션, 구매가 없는 세션은 비전환 세션으로 정의하였다.

전체 902,755건의 세션 중 구매가 발생한 세션 비율은 1.3%로 전 세계 전자상거래 평균 전환율인 3.17%[1]에 비해 굉장히 낮은 전환율을 보이고 있다.



글로벌 전자상거래 (패션, 액세서리 및 의류 산업) 구매 완료 기준 월별 평균 전환율

[1] "The average eCommerce conversion rate globally is 3.17%", DynamicYield, last modified 29 April, 2025,https://marketing.dynamicyield.com/benchmarks/conversion-rate/#:~:text=In%20April%2C%20the%20conversion%20rate,18



Google Analytics Merchandise Store(분석 데이터) 구매 건수 기준 월별 평균 전환율

글로벌 전자상거래 산업의 평균 전환율과 프로젝트 분석 대상 데이터의 전환율을 구매 완료/건수 기준 월별 평균 전환율로 비교하였다.

항목	글로벌 전자상거래 데이터	프로젝트 데이터
분석 기간	2024.05 ~ 2025.04	2016.08 ~ 2017.08
전환율 범위	약 2.84% ~ 4.19%	약 0.81% ~ 1.76%
피크 시점	2024.12(4.19%) - 연말	2016.12(1.76%) - 연말
저점 시점	2015.03(2.84%)	2016.11(0.81%)

두 데이터 모두 12월 연말에 전환율이 상승하는 공통된 시간적 패턴을 보이며 연말 쇼핑 시즌의 수요 증가와 같은 시기적 요인이 전환 행동에 영향을 미칠 가능성을 시사한다.

산업 평균 전환율은 전반적으로 3%대를 유지하며 안정적인 흐름을 보인 반면, 프로젝트 데이터의 전환율은 약 0.81% ~ 1.76% 수준으로 낮고 월별 변동 폭도 상대적으로 크게 나타났다.

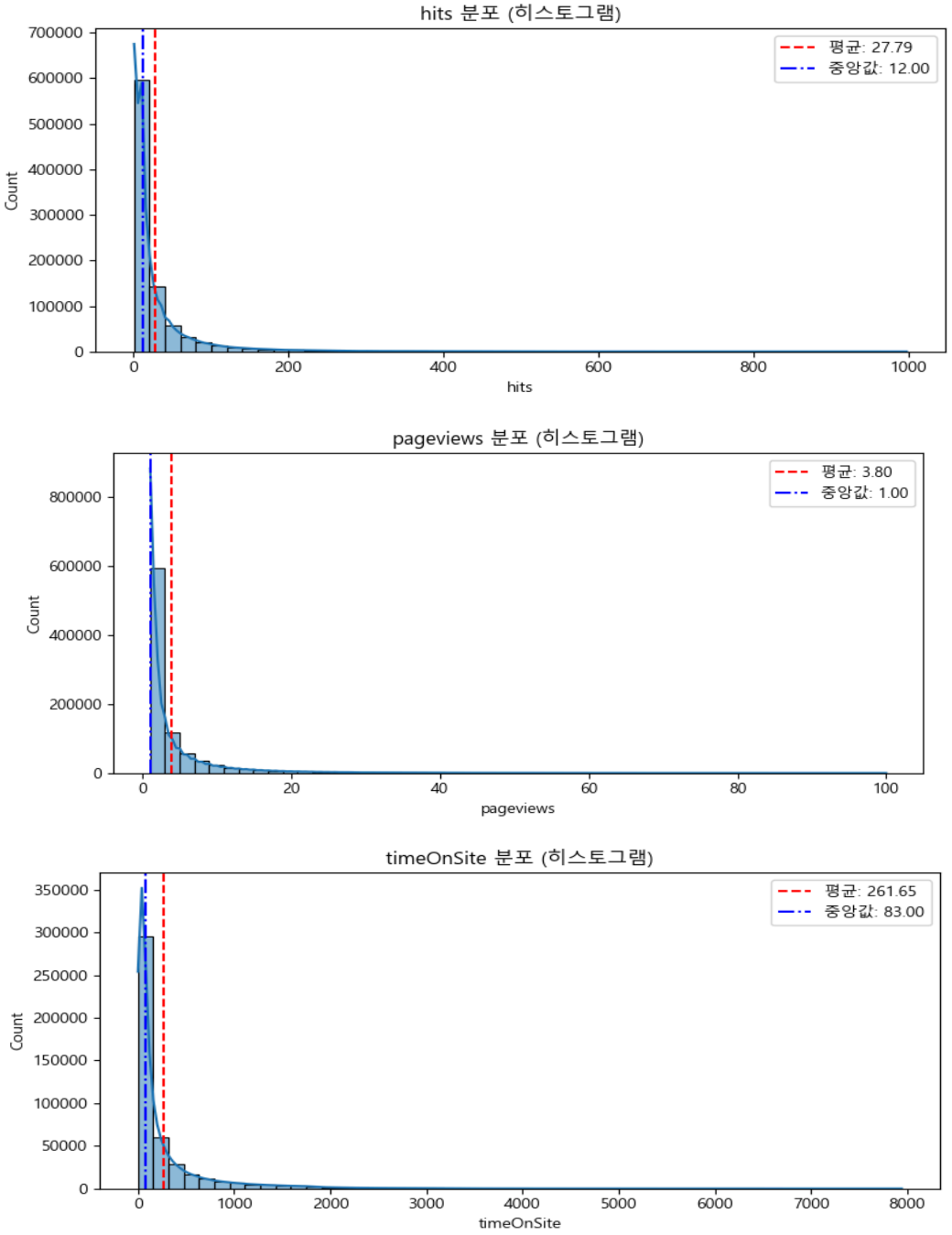
이는 전체적인 전환 효율이 낮고 사용자 행동이 일정하지 않다는 신호로 해석할 수 있으며 전환율을 향상시키기 위한 전략적 접근이 필요함을 의미한다.

따라서 본 분석에서는 전환율이 낮은 원인을 파악하고 이탈을 유발하는 주요 지점과 사용자 특성을 식별하여 개선점을 도출하는 것을 핵심 목표로 설정하였다.

3.2 주요 변수 분포 분석

사용자의 행동 로그 데이터를 세션 단위로 요약한 totals 필드의 통계 정보를 주요 변수로 선정하여 해당 컬럼들의 분포를 시각화를 통해 분석하였다.

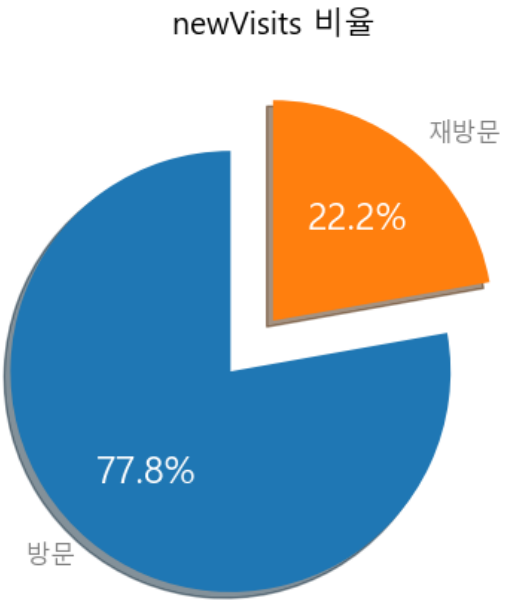
1) hits, pageviews, timeOnSite



낮은 전환율에 비해 세션 중 발생한 전체 히트 수와 페이지뷰 수, 세션의 총 체류 시간이 전반적으로 높게 나타나는 경향이 있을 수 있다.

일부 세션에서 상당히 높은 수준의 사용자 활동이 관찰되었고 이러한 세션에서 많은 페이지를 조회하고 오랜 시간 머무르며 다양한 상호 작용을 수행한 것으로 보이지만 실제 전환으로 이어지지 않는 사용자 행동 패턴이 존재할 가능성을 시사한다.

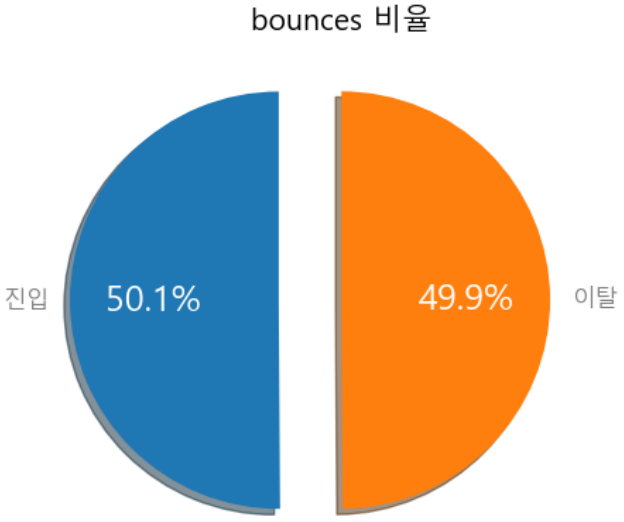
2) 신규 방문 여부



전체 사용자의 상당 비율이 사이트를 처음 방문한 사용자로 구성되어 있으며 이는 초기 유입은 활발하지만 재방문 비중은 상대적으로 낮다는 점을 보여준다.

신규 고객의 재방문을 유도하거나 기존 고객의 재방문을 유지하는데 어려움이 존재할 가능성이 있으며 사이트 경험이나 서비스 만족도가 충분히 다음 방문으로 이끌지 못하고 있음을 알 수 있다.

3) 진입 직후 이탈 여부



전체 사이트 사용자 중 약 49.9%가 사이트 진입 직후 첫 화면만 보고 이탈한 것으로 나타났다.

사이트의 첫 인상 또는 진입 페이지 품질에 문제가 있을 가능성이 존재하며 광고나 검색 유입 등에서 사용자 기대와 콘텐츠가 일치하지 않거나 실수로 클릭하는 등의 이유로 낮게 나타날 수 있다.

3.3 문제 정의 및 분석 방향 도출

EDA 결과 전체적인 전환율이 산업 평균에 비해 낮은 수준이며 사이트 내 활동량은 많지만 전환으로 이어지지 않는 사용자 행동 패턴이 존재할 가능성이 확인되었다. 또한 신규 방문자의 비중이 높고 진입 직후 이탈한 사용자가 약 49.9%로 사이트 이용 흐름 전반에서 이탈 가능성이 반복적으로 나타나고 있다.

전환율이 전반적으로 낮고 사용자 행동 데이터에서 **이탈 가능성이 반복적으로 관찰된다는 점**에서 단순히 전환 여부를 확인하는 수준을 넘어 사용자가 어떤 경로를 따라 사이트를 이용하며 **어느 단계에서 이탈이 발생하는지를 파악**하는 것이 중요하다.

본 분석에서는 사용자의 주요 행동 흐름을 퍼널 구조로 정의하고 각 단계 간 전환율을 정량적으로 분석함으로써 **이탈이 집중되는 구간**과 **전환으로 이어지지 못하는 원인**이 무엇인지에 대해 파악하고자 한다.

이후 병목 구간에서 이탈한 사용자와 전환에 성공한 사용자를 세그먼트별로 비교 분석하여 전환을 저해하는 주요 요인을 식별하고 **실질적인 개선 방향을 도출**하는 것을 분석의 핵심 목표로 설정하였다.

4. 사용자 행동 분석

4.1 퍼널 단계 정의 및 전환 흐름 분석

1) 분석 목표

고객 행동 데이터를 기반으로 사용자의 주요 행동 단계를 퍼널로 정의하여 각 단계별 전환율을 기반으로 핵심 이탈 지점을 식별하고 전환 저해 요인을 분석하여 전환율 개선 전략 수립을 목표로 한다.

- **전환 목표 및 KPI**: 구매 완료, 구매 전환율
- **사용자 경로**: 홈페이지 방문 및 탐색 → 제품 상세 페이지 조회 → 장바구니 → 결제 진행 → 구매 완료

2) 퍼널 설계

전자상거래 행동 유형을 나타내는 action_type과 세부 구매 진행 단계를 나타내는 step 컬럼들을 통해 세션별 최대 행동 깊이를 해당 세션의 최종 퍼널 단계로 정의하였다.

action_type-step	컬럼 설명	퍼널 단계	퍼널 단계 설명
0	진입 및 확인	1(유입점)	방문 및 카테고리 탐색
1	이벤트 클릭	2	상세 페이지 조회
2	이벤트 페이지 진입	2	
3	장바구니 물건 담기	3	장바구니
4	장바구니 물건 제거	3	
5 - 1	구매 목록 확인	4	결제 진행
5 - 2	결제창 호출	4	
5 - 3	구매 확정 확인 창	4	
6	결제 완료	5(목표 전환)	구매 완료

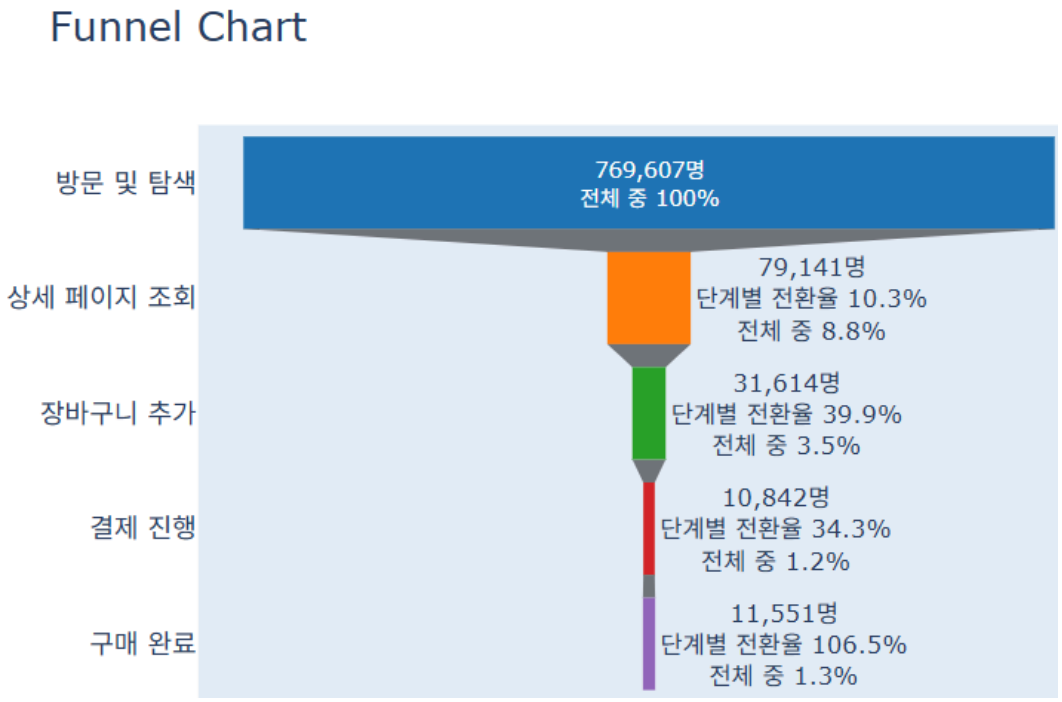
최종 퍼널 구조

[방문 및 카테고리 탐색] → [상세 페이지 조회] → [장바구니 추가] → [결제 진행] → [구매 완료]

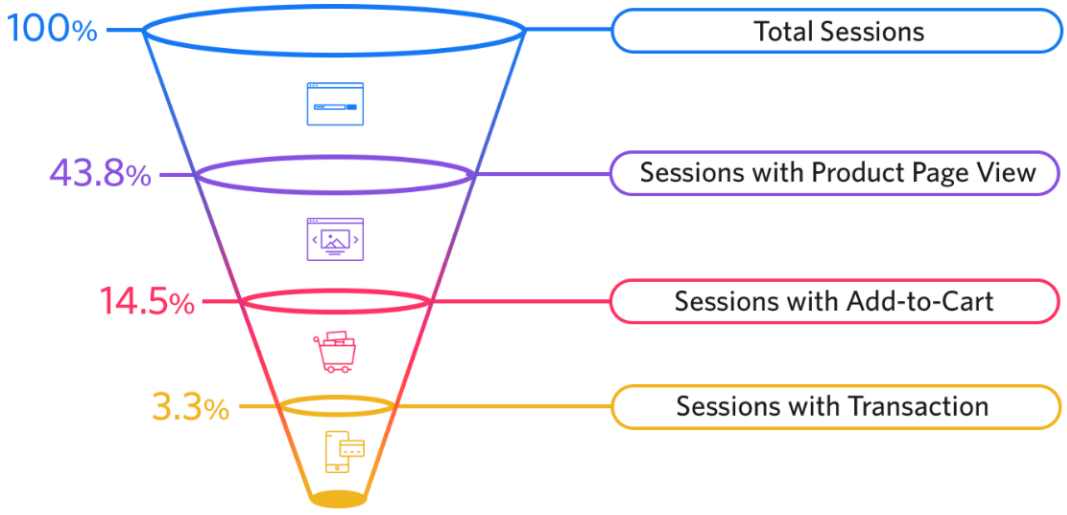
3) 지표 계산 및 결과 해석

각 단계별 세션 수를 집계하여 현재 단계에서 다음 단계로 진행하는 비율(단계 전환율)과 이전 단계 대비 이탈자 비율을 계산하여 주요 이탈 구간을 식별하였다.

퍼널 단계	사용자 수	단계별 전환율(퍼센트)	이탈률(퍼센트)
1: 방문 및 탐색	769,607	10.28%	89.72%
2: 상세 페이지 조회	79,141	39.95%	60.05%
3: 장바구니 추가	31,614	34.29%	65.71%
4: 결제 진행	10,842	106.54%	-6.54%
5: 구매 완료	11,551	(없음)	(없음)



방문 및 탐색 과정에서 상세 페이지 조회로 넘어가는 단계에서 가장 낮은 단계 전환율을 보이며 상품 조회 전환이 저조한 상황이다.



[2]

이커머스 업계 평균 전환율인 43.8%보다 크게 낮은 전환율로 현재 전환율인 10.28%는 일반적이라 보기 어렵고 해당 구간에 대해 전환을 저해하는 요인이 존재한다고 판단하여 개선이 시급한 핵심 이탈 구간으로 선정하였다.

---

[2] "E-commerce conversion rate benchmarks – 2025 update",SmartInsights,last modified 02 Jan,2025,https://www.smartinsights.com/ecommerce/ecommerce-analytics/ecommerce-conversion-rates/?utm_source=chatgpt.com

4.2 세그먼트 기반 가설 검증 분석

퍼널 분석 전환 결과 가장 큰 이탈률을 보이는 단계 기반으로 사용자 그룹을 분류하였다. 이탈 그룹과 잔존 그룹 간의 행동 패턴 차이를 비교하고 이탈을 유발하는 주요 원인을 파악하기 위해 탐색 활동 지표를 기반으로 가설을 다음과 같이 설정하였다.

가설: 이탈한 방문자는 페이지 조회 수, 히트 수, 체류 시간 등에서 일관되게 낮은 행동 패턴을 보일 것이다.

- **전환 목표 재정의**: 상세 페이지 조회
- **이탈 그룹**: 방문 후 상세 페이지 이전까지 탐색한 사용자 (퍼널 1단계)
- **잔존 그룹**: 상세 페이지 조회와 그 이후 단계를 넘어간 사용자 (퍼널 2~5단계)

그룹별 주요 지표 비교

가설에서 선택한 지표들이 실제로 두 그룹을 구분하는 핵심 변수로 작용하는지 검증하고 이탈을 유발하는 주요 행동 패턴을 도출하기 위해 그룹별 주요 지표를 비교하였다.

현재 데이터는 총 40개의 변수를 포함하고 있어 단순 통계 비교만으로는 차이를 효과적으로 파악하는데 한계가 있다.

- 모든 변수를 요약 통계하는 방식으로 주요 차이 요인을 파악하기엔 변수가 많으며
- 변수 간의 복합적인 상호작용을 반영하기 어렵다.

이에 따라 두 그룹을 대상으로 이진 분류 모델을 적용하여

- 다변량 분석을 통한 복합적인 주요 영향 요인을 탐색하고
- 그룹 분류 기여도를 정량적으로 평가하며
- SHAP값을 통해 변수별 영향 방향과 크기를 분석하였다.

모델링

1) 전처리

전체 40개의 컬럼을 전처리 하여 모델에서 사용할 140개의 피처를 선별하였다.

- **불필요한 컬럼 제거**: 타겟 레이블과 직접적으로 연결되어 예측할 수 있는 정보를 포함하고 있거나 타 컬럼의 정보를 포함하고 있는 변수는 분석에서 제외하였다.

- **그룹화**: 방문 날짜 컬럼(dates)을 월 단위로 그룹화하여 특정 시점에서 두 그룹 간의 행동 차이가 존재하는지 파악하고자 하였다.

- **인코딩**: 히트 단위별 전처리로 생성된 변수들은 사용자 행동의 자주 등장하는 흐름을 수치화하기 위해 Frequency Encoding을 적용하고 범주형 변수들은 가장 빈번하게 노출된 상위 카테고리를 One-hot Encoding을 통해 변환하였다.

2) 샘플링

maxActionType 컬럼을 활용해 세션별 최종 퍼널 단계가 1인 이탈 그룹과 나머지 잔존 그룹을 나누어 두 그룹의 비율을 기반으로 10만개의 데이터를 샘플링하였다.

maxActionType	퍼널 단계	퍼널 설명	비율
== 0	1	방문 후 탐색	0.8524
!= 0	2~5	상세 조회 ~ 최종 전환	0.1475

3) 기본 성능 모델 비교

두 사용자 그룹의 핵심적인 행동 패턴 차이를 분석하기 위해 모델 선택 기준을 설정하고 4개의 모델에 기본 성능을 비교하여 최종 사용 모델을 선택하였다.

1. 예측이 아닌 두 그룹 간의 분포 차이 탐색이 목표이므로 **변수 해석이 쉬워야함**
2. 신뢰할 수 있는 변수 해석을 위해 모델의 **기본적인 성능 필요**
3. 총 10만건의 사용자 세션 데이터로 **대용량 데이터를 처리 가능한 모델**
4. 두 그룹의 비율이 약 85:15로 **불균형하여 이에 대응할 수 있는 모델**

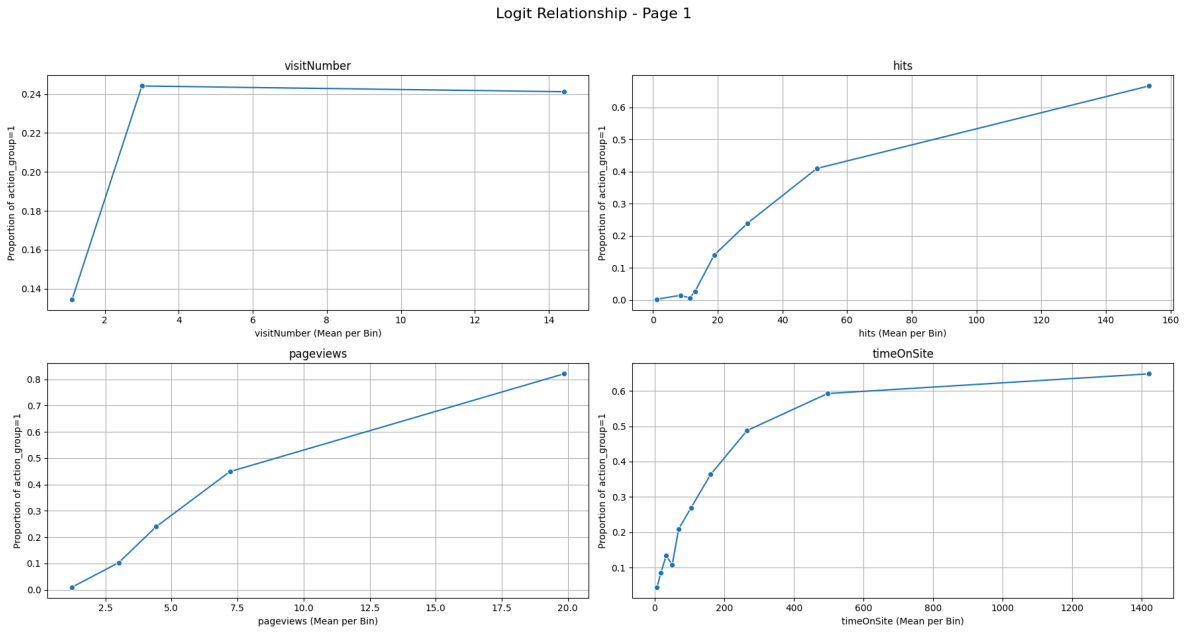
모델 성능 비교 결과)

모델	Accuracy	F1 Score	AUPRC
LightGBM	0.94400	0.80636	0.89859
CatBoost	0.94420	0.80751	0.89861
Random Forest	0.93465	0.75525	0.86034
Logistic Regression	0.92570	0.71224	0.83771

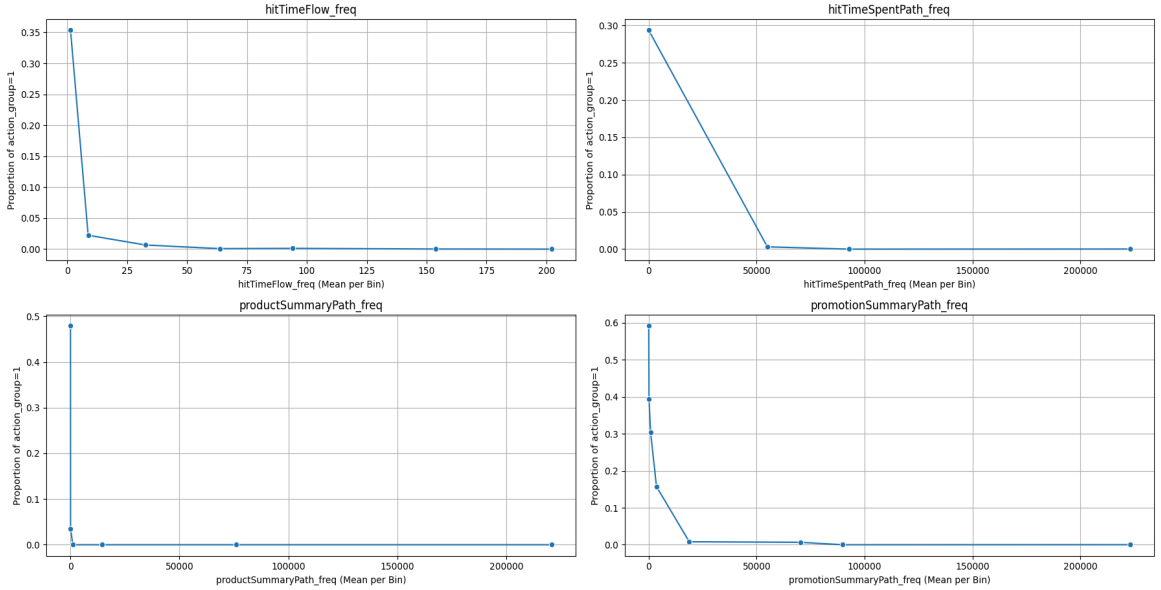
- **Logistic Regression**: 변수의 영향 방향을 계수로 명확히 확인할 수 있어 기본 모델로 선택
- **LightBGM**: 모델 성능이 높고 학습 속도가 빨라 최종 모델로 선택

4) Logistic Regression

a. 수치형 변수와 타겟의 선형 관계 확인 - Binning 후 시각화



visitNumber(세션 방문 횟수) hits(전체 히트 수)
pageviews(총 페이지 뷰 수) tImeOnSite(체류 시간)



hitTimeFlow(히트 시간 경로) hitTimeSpentPath(히트 체류시간 경로)
productSummaryPath(상품 요약 경로) promotionSummaryPath(캠페인 요약 경로)

hits와 pageviews를 제외한 나머지 변수들은 타겟과의 관계에서 비선형적인 형태를 보여 변수 해석의 용이성과 로지스틱 회귀 모델의 가정 충족을 위해 해당 변수들에 대해 로그 변환을 통해 선형성을 확보하였다.

b. 다중공선성 확인 - VIF 지표 (분산팽창 계수)

변수명	VIF 값	변수명	VIF 값
hitTimeFlow	14,8	pageviews	4.94
hitTimeSpentPath	12.7	hits	4.54
productSummaryPath	4.09	newVisits	2.33
promotionSummaryPath	5.59	visitNumber	2.21
timeOnSite	7.88		

- **hitTimeFlow & hitTimeSpentPath**
    - 하나의 컬럼으로 다른 컬럼의 데이터를 유추할 수 있으며 두 변수 모두 히트별 시간 흐름을 나타내고 있어 둘 중 변수 해석이 용이하고 분류 모델에 더 적합하다고 판단되는 hitTimeSpentPath를 선택하고 hitTimeFlow는 제거한다.
- **timeOnSite & promotionSummaryPath**
    - 다중공선성이 강하게 나타나지 않으며 분석 목표가 변수별 해석으로 두 컬럼 모두 비즈니스적 의미가 있다고 판단되어 분석에서 사용한다.
    - 사이트 체류 시간(timeOnSite)은 총 히트 수와 상관 관계를 가지지만 항상 비례하지 않으며 각 컬럼을 통해 해석 가능한 정보가 다르기에 분석에서 사용한다.

c. 표준화 - StandScaler

- 계수 해석을 위해 변수 간 단위 차이를 제거하고, 계수 크기 비교를 통한 상대적 영향력 해석이 가능하도록 StandScaler를 활용해 표준화를 진행하였다.

d. 모델 학습 및 성능 평가

- **F1 Score**: 0.7336
- **AUPRC**: 0.8506

로지스틱 회귀 모델을 해석용으로 사용하기에 충분히 신뢰할 수 있는 수준에 성능으로 추가적인 튜닝 과정 없이 해당 모델을 통해 변수의 영향 정도를 해석한다.

e. 주요 변수 해석

계수 기준 상위 5개 **양**의 영향 변수)

Feature	계수	오즈비	해석
social Network _Twitter	0.902	2.47	Twitter 유입 사용자는 다른 소셜 네트워크 대비 전환 가능성이 약 **2.47배** 더 높음.
country _Vietnam	0.635	1.89	베트남 사용자는 전환 확률이 **1.89배** 높음.
city_ Los Angeles	0.536	1.71	LA 거주자는 타 지역 대비 전환 확률이 **1.71배** 높음.
city_ San Francisco	0.486	1.63	샌프란시스코 사용자는 **1.63배** 높은 전환 가능성.
pageviews	0.475	1.61	페이지를 많이 본 사용자일수록 전환 가능성이 **1.61배** 높음.

- 특정 소셜 채널과 지역, 페이지뷰 수가 전환 가능성을 높이는 요인으로 나타났다.

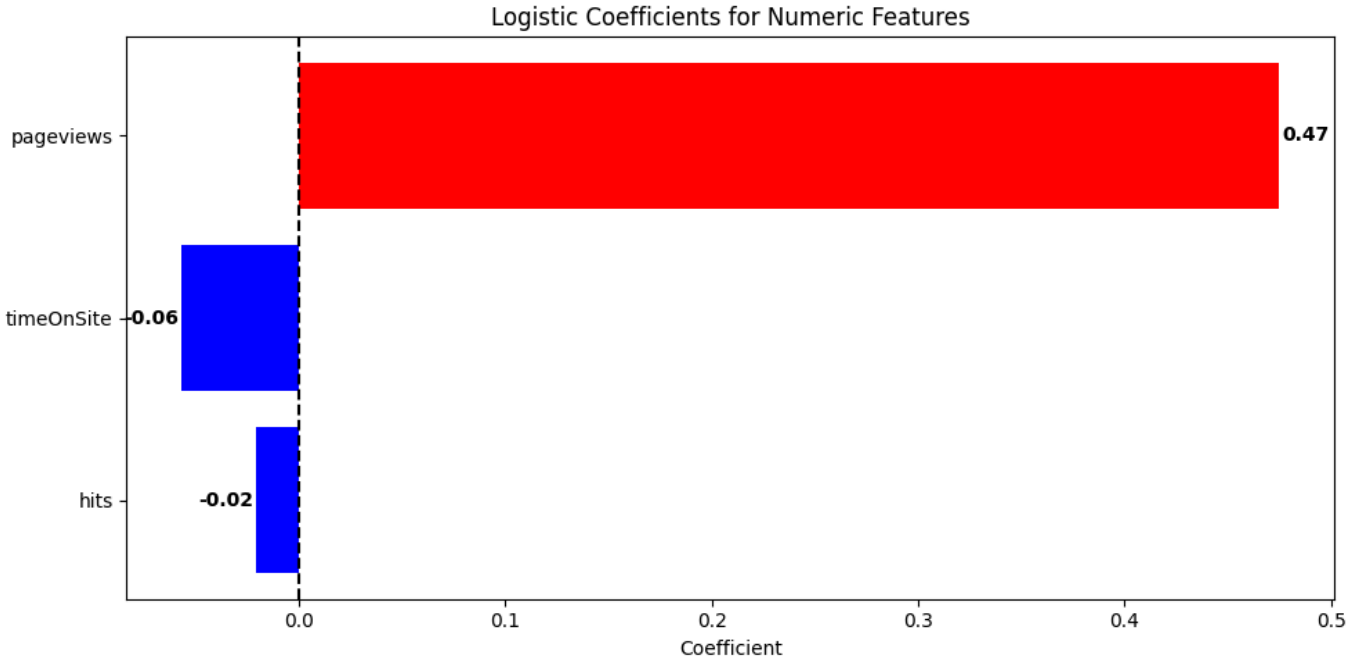
계수 기준 상위 5개 **음**의 영향 변수)

Feature	계수	오즈비	해석
year_month_ 201611	-1.62	0.197	2016년 11월 방문자는 2016년 8월 대비 이탈 확률이 약 **80.3% 낮음**
year_month_ 201703	-1.60	0.20	2017년 3월 방문자는 2016년 8월 대비 이탈 확률이 약 **79.9% 낮음**
year_month_ 201612	-1.55	0.22	2016년 12월 방문자는 2016년 8월 대비 이탈 확률이 약 **78.8% 낮음**
year_month_ 201702	-1.55	0.21	2017년 2월 방문자는 2016년 8월 대비 이탈 확률이 약 **78.8% 낮음**
year_month_ 201610	-1.55	0.21	2016년 10월 방문자는 2016년 8월 대비 이탈 확률이 약 **78.8% 낮음**

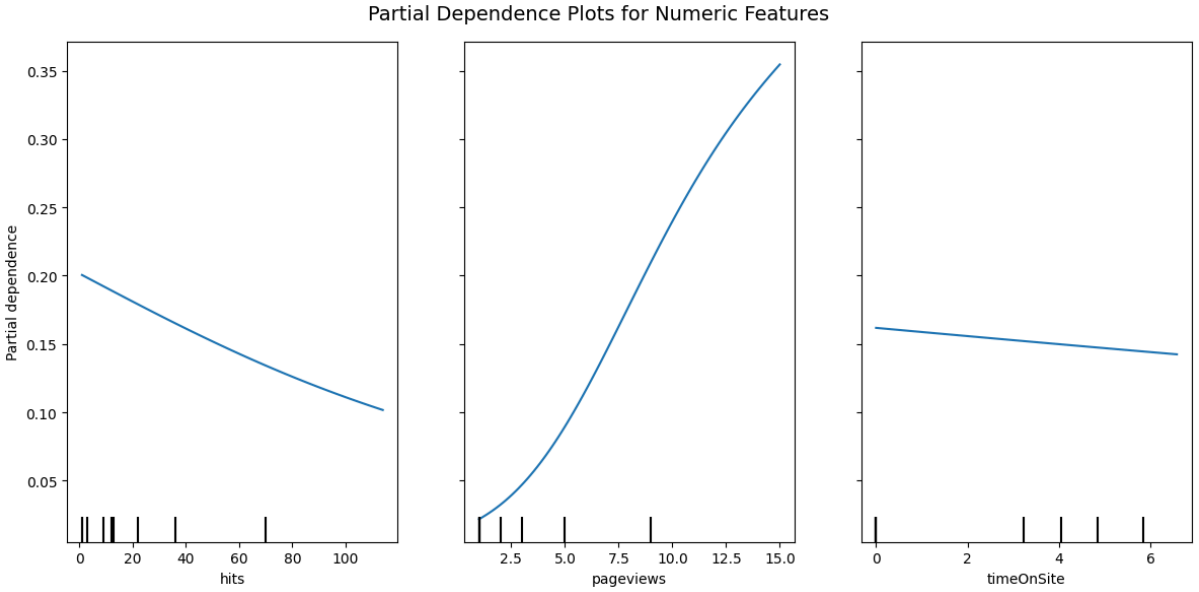
- 특정 기간에서 전체 평균보다 전환 가능성이 약 80% 낮게 나타났다.

f. 가설 검증

가설: 이탈한 방문자는 페이지 조회 수, 히트 수, 체류 시간 등에서 일관되게 낮은 행동 패턴을 보일 것이다.



페이지를 많이 본 사용자는 상세 페이지 조회 이후 단계로 전환 가능성이 높지만 사이트 체류 시간과 전체 히트 수는 전환 가능성이 약하게 감소하는 경향을 보였다.



각각의 피처들이 전환 확률에 미치는 영향을 Partial Dependence Plot(PDP)를 통해 시각화하여 변수별로 결과를 해석하였다.

- **hits**
  - 히트 수가 증가할수록 전환 확률은 점점 감소하는 것으로 보아 불필요한 활동이 많을수록 전환 가능성이 낮아질 수 있음을 시사한다.

- **pageviews**
  - 페이지뷰가 많아질수록 전환 확률이 빠르게 증가하며 특히 5페이지 이상부터 가파르게 상승한다. 해당 페이지 이상의 탐색은 의도 있는 탐색의 신호로 보여진다.

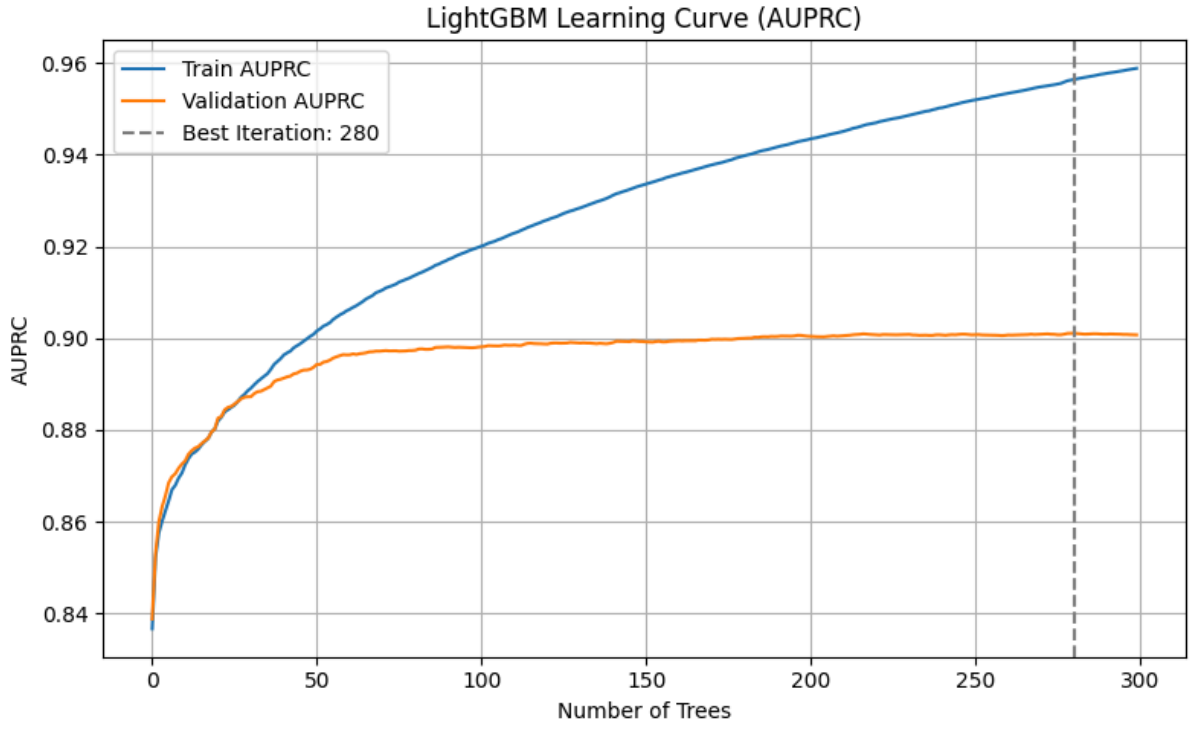
- **timeOnSite**
  - 체류 시간의 증가가 전환 확률에 미치는 영향은 거의 없으며 약간 감소하는 경향을 보인다.

5) LightBGM

a. 기본 모델 학습 및 과적합 여부 확인

i. 학습과 검증 데이터를 80:20으로 나누고 이탈 그룹과 잔존그룹은 불균형한 분포를 가져 성능지표로 F1 Score와 AUPRC를 사용하였다.

ii. 기본 모델 학습 성능 F1 Score: 0.8088 / AUPRC: 0.8980



iii. 학습 데이터와 검증데이터 간 격차가 크지만 Validation 성능이 일정 수준 이상에서 안정적으로 유지되므로 일반화 성능이 잘 유지되는 학습 패턴으로 판단하였다.

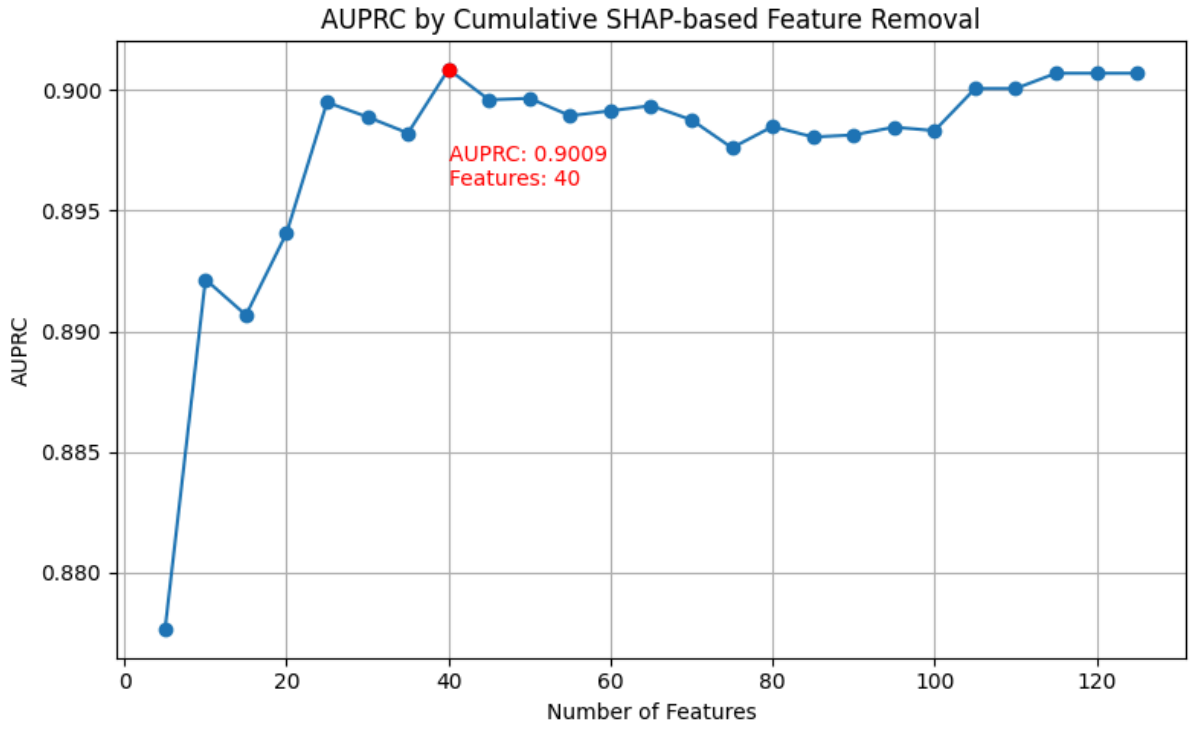
iv. 280번째 트리에서 Validation AUPRC가 최고점에 도달하며 이후 트리 수를 늘려도 증가하지 않거나 미세하게 감소하여 해당 트리 수에서 과적합을 방지하고 최적의 성능을 낸다.

b. SHAP 기반 최적 피처 조합

SHAP 기준 상위 5개 피처)

Feature	평균 절대 SHAP값 (mean_abs_shap)
productSummaryPath_freq	2.8032
pageviews	0.5567
hitTimeSpentPath_freq	0.5141
hits	0.2502
timeOnSite	0.2235

같은 피처도 학습 때마다 예측에 영향을 주는 정도가 달라 모든 예측에서의 평균 기여도를 확인하기 위해 SHAP 평균 절댓값을 통해 주요 변수를 선정하였다.



SHAP값이 낮아 상대적 중요도가 낮은 피처들을 제거하며 변하는 AUPRC를 시각화한 그래프로 피처가 40개일 때 AUPRC=0.9009로 가장 높은 성능을 보였다. SHAP 상위 40개의 피처를 모델 학습 변수로 선택한다.

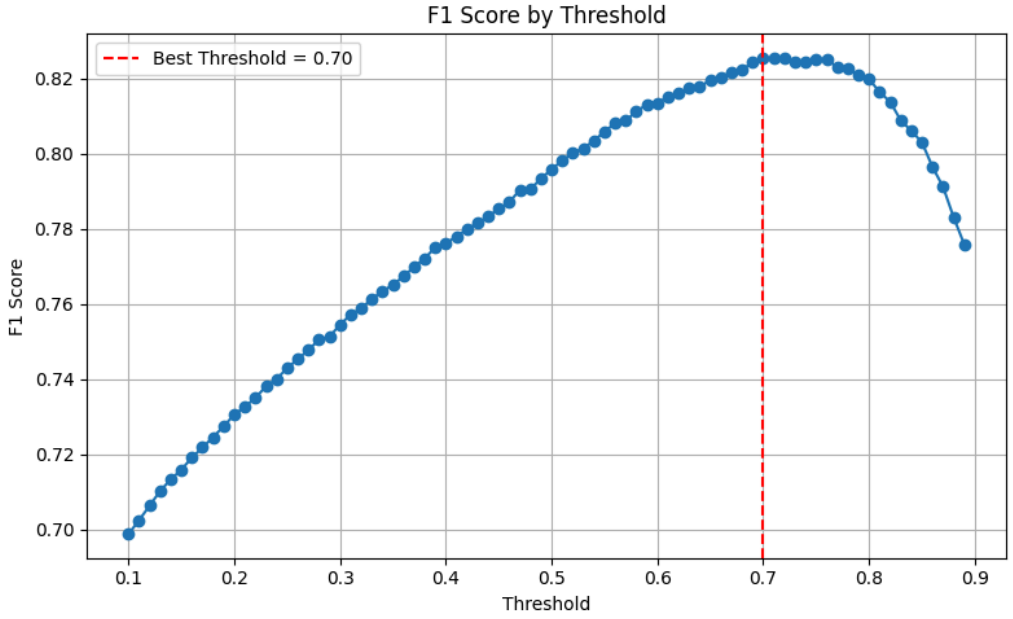
- **최적 피처 조합 모델 학습 성능** F1 Score: 0.79508 / AUPRC: 0.8980

c. 최적 하이퍼파라미터 및 Threshold

i. 분석 목표에 맞춰 적은 계산으로 빠르게 실용적인 성능을 낼 수 있는 RandomizedSearch를 사용해 최적 하이퍼파라미터를 탐색하였다.

하이퍼파라미터	값	하이퍼파라미터	값
subsample	0.8	max_depth	-1
num_leaves	63	learning_rate	0.05
min_child_samples	50	colsample_bytree	0.8

ii. 최적 피처 조합과 최적 하이퍼파라미터을 적용하여 F1 Score를 통해 최적 Threshold를 탐색하였다.

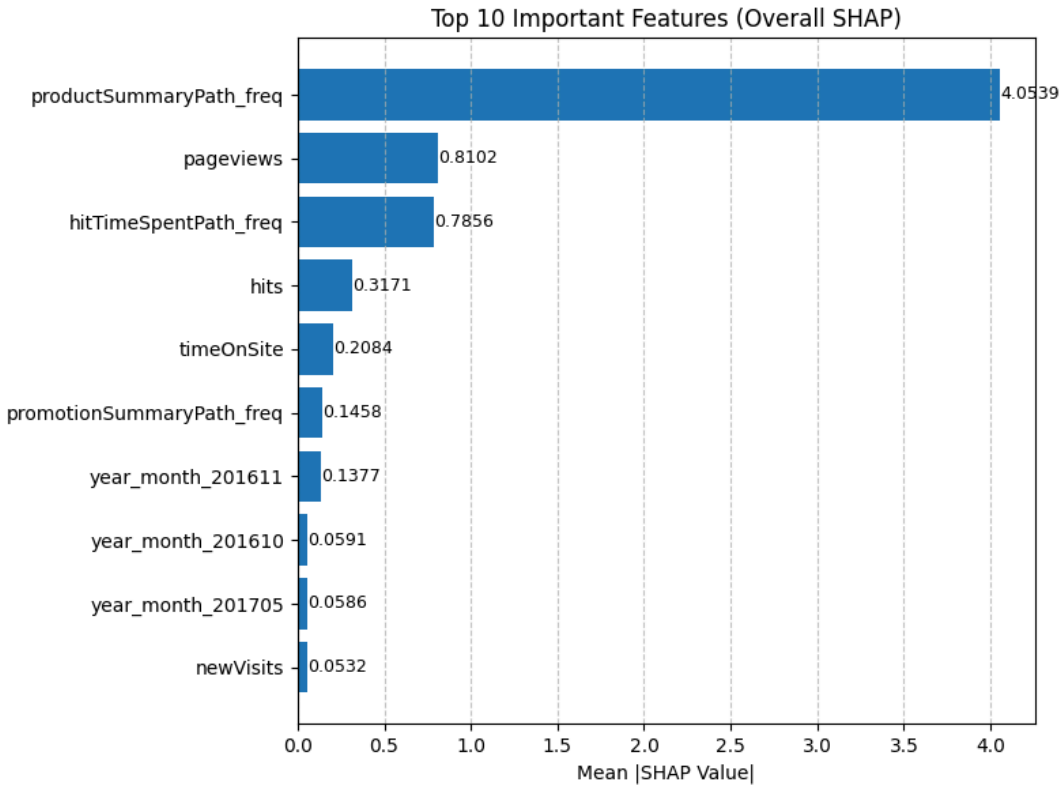


- **최종 모델 학습 성능** F1 Score: 0.82545 / AUPRC: 0.90137

d. SHAP 기반 주요 변수 해석

이탈 그룹과 잔존 그룹을 분류하는 LightBGM 최종 모델의 예측에 영향을 미친 상위 10개 피처를 SHAP 평균 절대값 기준으로 선정하여 막대그래프로 시각화하였다.

모델은 이탈 여부를 예측하는데 있어 페이지 탐색 수(pageviews), 히트 수(hits), 체류 시간(timeOnSite) 등이 주요한 영향을 미치는 변수로 나타났다. 이는 해당 변수들이 이탈 사용자와 잔존 사용자 간의 행동 패턴 차이를 구분하는 주요 요인임을 알 수 있다.



e. 가설 검증

가설: 이탈한 방문자는 페이지 조회 수, 히트 수, 체류 시간 등에서 일관되게 낮은 행동 패턴을 보일 것이다.

- **hits (전체 히트 수)**

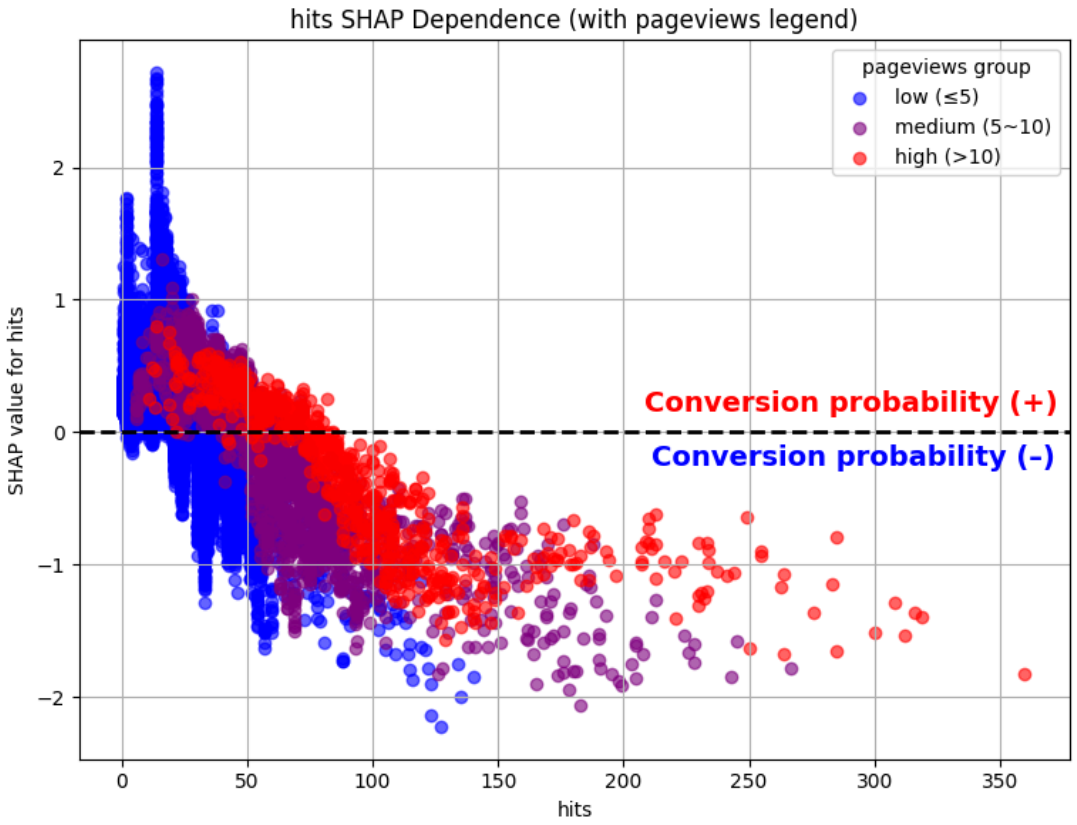
  로지스틱 회귀 분석 결과 hits 수가 많을수록 전환 확률이 감소하는 경향이 나타났다.

  [전체 hits 증가] → [페이지 탐색량 증가] → [상세 페이지 조회]

  일반적으로 다음과 같은 흐름을 통해 전환 확률이 높아질 것으로 기대되지만 실제로는 전환 없이 이탈하는 패턴이 확인되었다.

  전환 저해 원인으로 추정되는 사용자 행동)
  1. 원하는 정보를 찾지 못해 여러 페이지를 방황하는 경우
  2. 정보가 과도하게 많아 선택을 주저하는 경우
  3. 구매 의도 없이 단순 탐색만을 목적으로 방문한 경우

  hits 컬럼 단독으로는 사용자 탐색 행동의 질을 파악하는데 한계가 있어 pageviews(페이지 조회 수)와 함께 조합하여 사용자의 실제 탐색 깊이와 의도를 분석하고 탐색량은 많지만 전환율이 낮은 행동 패턴의 원인을 보다 명확하게 해석하고자 하였다.



그룹별 개별 해석)

1. **전반적으로** hit가 증가할수록 SHAP값이 음수로 감소하며 전환 확률을 낮추는 방향으로 기여하며 특히 hits 수가 50 이상일 때 더 뚜렷하게 나타났다. 활동이 활발한 사용자가 전환 그룹에 속하기 보다는 과도한 탐색이나 비효율적 흐름일 가능성이 존재한다.

2. **pageviews가 높게 나타나는 붉은색 점들**은 대부분 hits 수도 동일하게 높게 나타나며 상당한 수가 SHAP값이 음수이다. 사용자가 많은 페이지를 보고 활발히 행동하지만 전환에는 기여하지 못하는 정보 탐색형 방문자가 다수 존재한다는 사실을 알 수 있었다.

3. **SHAP 값이 1 이상에 분포하는 파란색 점들**은 전환 예측에 높은 기여도를 보이는 그룹으로 hits 수와 pageviews 모두 적게 나타났다.

   전환 기여 원인으로 추정되는 사용자 행동)
   a. 구매 의도가 명확하여 소수의 페이지만 탐색하고 전환하는 경우
   b. 재방문 등 이전 탐색 경험이 있어 무의미한 탐색 없이 행동하는 경우

- **timeOnSite (체류 시간)**

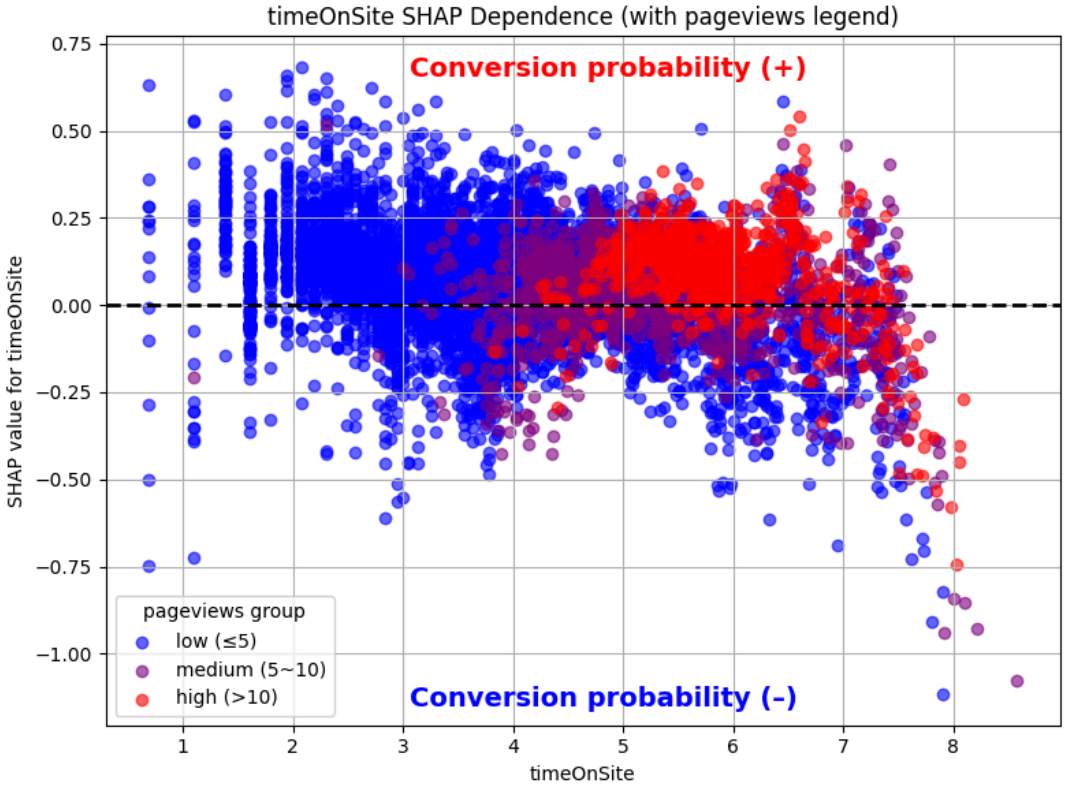
  로지스틱 회귀 분석 결과 timeOnSite는 전환 확률에 미치는 영향이 거의 없으며 약하게 감소하는 경향이 나타났다.

  체류 시간의 증가는 사이트 내 행동과 탐색의 증가를 예상하여 전환 확률을 높인다고 기대되지만 실제로는 전환에 유의미하게 기여하지 않는 것으로 확인되었다.

전환 저해 원인으로 추정되는 사용자 행동)

1. 실제 행동은 없지만 시간만 흐르는 경우 (ex. 탭 방치)
2. 정보 과잉이나 조회 유도 프로세스 부족 등으로 결정을 주저하는 경우

위와 같은 원인들로 사용자의 흥미를 유발하지 못하고 탐색 중 이탈한다고 예상하였다. timeOnSite 컬럼 단독으로는 사용자의 전환 의도를 명확하게 설명하기 어려워 pageviews(페이지 조회 수)와 함께 조합하여 단순히 머문 시간이 아닌 실제 탐색의 깊이와 적극성을 고려하여 전환 저해 요소를 파악하고자 하였다.



그룹별 개별 해석)

1. **전반적으로** timeOnSite가 증가할수록 SHAP값이 뚜렷하게 변화하지 않지만 6~9 사이에 높은 체류시간에서 전환 확률을 약하게 낮추는 방향으로 기여한다. 체류 시간이 긴 사용자가 전환 가능성이 높다기 보다는 정보 과잉 등으로 결정하지 못해 망설이거나 고민하여 체류 시간만 늘어났을 가능성이 존재한다.

2. **pageviews가 높게 나타나는 붉은색 점들**은 대부분 timeOnSite도 높게 나타나며 SHAP값이 전반적인 분포와 동일하게 -0.5에서 0.5 사이 범위의 해당하지만 pageviews가 적은 사용자들에 비해 낮은 SHAP값을 가진다.

3. 사용자가 많은 페이지를 탐색하더라도 원하는 정보를 빠르게 찾지 못하거나, 선택을 주저하며 오랜 시간 고민하는 상황으로 인해 높은 체류 시간과 페이지 뷰 수에도 전환 확률에 부정적인 영향을 준다고 예상하였다.

5. 전략 제안

5.1 전환율 개선 전략

퍼널 분석 결과, **전체 사용자 중 약 90%**가 인입 단계에서 제품 탐색 단계로 이어지지 않고 **이탈**하는 것으로 나타났다. 이는 **사용자가 사이트에 유입되었음에도** 불구하고 흥미를 느낄만한 제품을 발견하지 못했거나, **탐색할 동기를 부여받지 못했음**을 시사한다.

제품 탐색 구간을 보완하고, 제품 탐색 경험을 강화하기 위해 **장바구니 단계에 추천 시스템을 도입**하였다. 장바구니에 제품을 담은 고객에게 **관련 제품을 추가로 추천**함으로써, **구매 전환율을 높이고 전체 매출 증대에 기여**하는 것이 목적이다.

특히, 고객이 관심을 보인 제품과 유사하거나 함께 구매되는 제품을 제안함으로써 **제품 탐색 경험을 보완**하고, **추가 탐색을 유도**하여, **다음 방문 시 더욱 빠르게 제품 탐색**으로 이어지도록 설계하였다.

이러한 추천 시스템은 고객의 구매 여정 후반부에 제품 노출을 강화함으로써 **초기 탐색 의도를 강화**하고, 차후 **재방문 시 인입 단계에서의 이탈률을 구조적으로 낮추는 데 기여**할 수 있다.
궁극적으로는 **탐색 전환률을 높이고, 전체 퍼널 흐름의 효율성을 개선하는 효과**를 기대할 수 있다.

5.2 추천 시스템

추천 시스템은 사용자의 행동 데이터를 분석하여 **개인화된 콘텐츠나 상품을 제안**하는 기술이다. 일반적으로 사용자 식별 정보와 장기적인 행동 이력을 활용하지만, 사용자의 ID가 명확하지 않거나 비로그인 상태에서 발생하는 단기 행동 데이터에 주목할 필요가 있다.

이러한 상황에서 효과적인 방식이 바로 **세션 기반 추천 시스템**(Session-based Recommendation)이다. 세션 기반 추천은 사용자의 로그인 여부와 무관하게 하나의 세션 내에서 발생하는 클릭, 탐색, 장바구니 추가 등의 **순차적 행동(sequence) 을 분석**하여, 그 맥락(Context)에 맞는 **다음 행동**(또는 선택할 아이템)을 **예측**하는 방식이며 아래와 같은 장점을 가진다.

- **비식별 사용자도 추천 가능**: 사용자 ID 없이도 세션 흐름만으로 추천 가능
- **실시간 반응성**: 사용자의 현재 행동에 즉각적으로 반응하는 추천이 가능
- **순서 정보 반영**: 사용자의 행동 순서를 고려하여 맥락에 맞는 추천 제공
- **콜드 스타트 완화**: 신규 사용자나 신규 아이템에 대해서도 효과적으로 추천 가능

이러한 특성 때문에, 세션 기반 추천은 **사용자 행동 흐름이 빠르게 변화하는 환경**에서 매우 유용하게 적용되고 있으며 순차적 정보를 학습하는데 적합한 RNN 계열 모델(GRU4Rec, LSTM), Transformer 계열 모델 (SASRec)을 사용하였다

5.2.1 장바구니 추천 시스템

사용 모델의 장 단점 비교

사용모델	장점	단점
LSTM	hidden state, cell state를 활용하여 더 긴 시퀀스 학습에 강점 여러 단계의 순환을 통해 더 복잡한 시퀀스 학습 가능	더 많은 파라미터 사용으로 메모리 소모가 크고 복잡하여 훈련속도 느림
GRU4Rec	모델이 간단하고 사용 파라미터가 적어 메모리 사용량이 적고 학습이 빠름	간단한 모델 구조로 복잡한 상관관계를 놓칠 가능성 높음
SASRec	시퀀스 내 아이템 간 관계를 잘 포착하며 유연한 시퀀스 길이 처리 가능	많은 데이터와 긴 훈련 시간이 필요 및 과적합 위험도 존재

평가 지표 설명)

MBCC(Mean Binary Classification Confidence)– 정확도 기반 지표

평가 초점 : 정답에 대해 얼마나 강하게 추천했는가

단순 Hit/No-Hit이 아닌 연속적인 confidence 수준을 활용하는 정량적 확신 평가 지표

추천 모델이 특정 아이템에 대해 “정답”일 확률을 얼마나 확신하는지를 평균적으로 평가하는 지표

여러 모델이 같은 Hit Rate를 가질 때, 모델의 예측 강도를 비교하고 싶을 때 사용 가능

Hit Rate@10– 정확도 기반 지표

평가 초첨 : 정답을 포함했는가

추천 리스트 상위 10개 항목에 사용자가 실제로 상호작용(클릭, 구매 등)한 항목이 포함되어 있는지를 확인하는 지표

사용자가 관심 있어 할 아이템을 추천 상위 목록에 얼마나 잘 포함시켰는지를 평가

NDCG@10 (Normalized Discounted Cumulative Gain) – 순위 민감 정확도 지표

평가 초점 : 정답의 순위 위치

추천된 아이템이 실제 관심도와 얼마나 잘 일치하며, 높은 순위일수록 더 가중치를 두는 지표

정확도(Accuracy) 기준이지만, 순위 민감(Ranking-aware)하여 Hit Rate보다 정밀한 평가가 가능

모델 훈련 결과 및 비교

개발환경 : CPU, 8G RAM

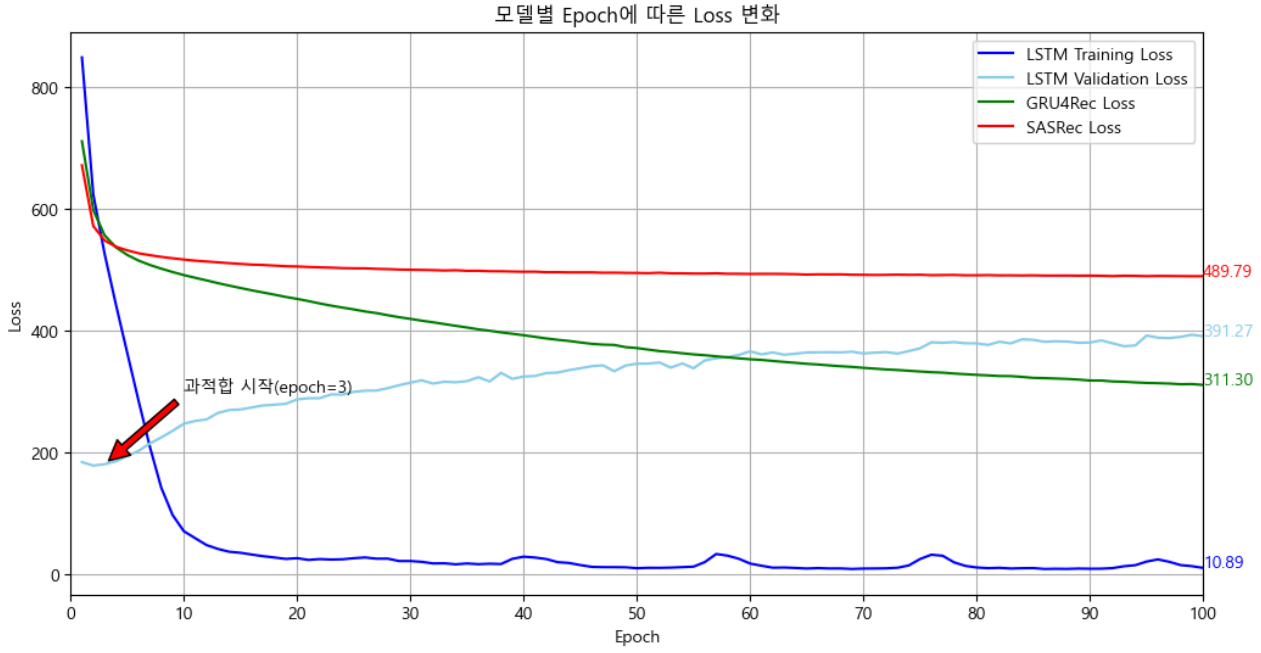
개발 모델	학습시간
LSTM	3m 29.8s
GRU4Rec	113m 13.3s
SASRec	1625m 23.7s

학습 속도 비교

모델의 학습 속도는 **구조적 복잡성**에 따라 달라진다. LSTM은 일반적으로 복잡한 게이트 구조로 인해 학습 시간이 오래 걸리지만, 이번 실험에서는 **시퀀스 길이를 최대 10개로 제한**하고, 작은 데이터셋과 배치 크기 32의 **미니배치 학습을 적용**하여 연산량을 줄였다. 또한, 단층 LSTM 구조로 파라미터 수를 최소화하여 학습 시간을 단축할 수 있었다. 이로 인해 **GRU4Rec이 가장 빠르게 수렴**하였고, **SASRec이 그 뒤를 이었다**.

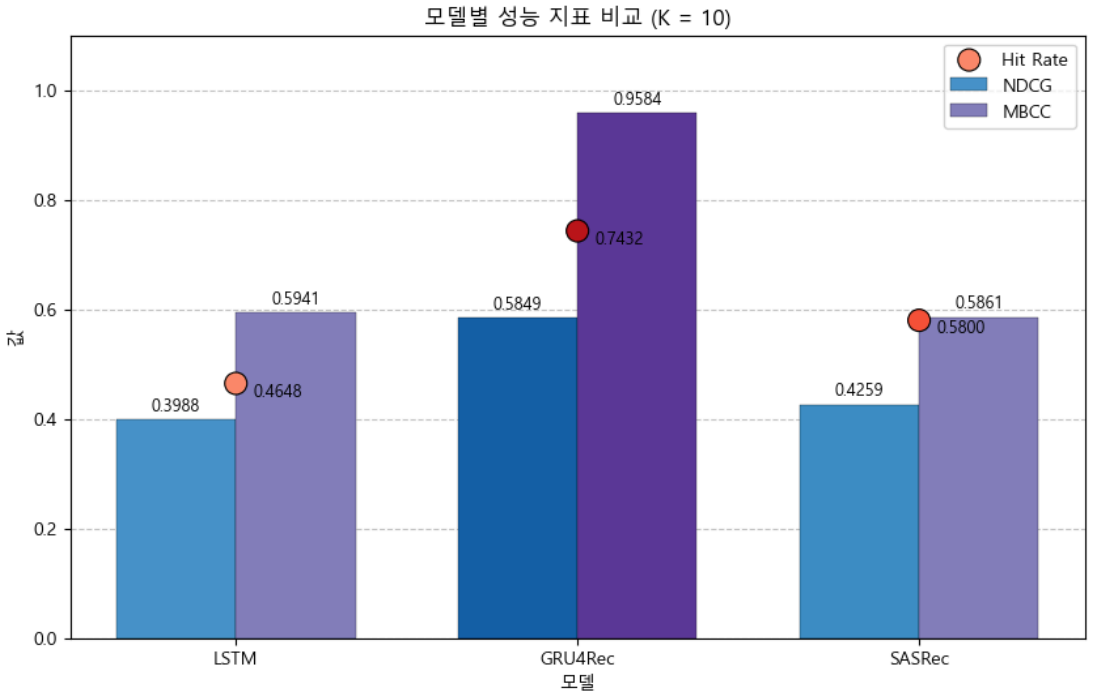
추천 시스템 구현을 위해 **시계열 기반 모델인 LSTM** (Long Short-Term Memory), **GRU4Rec** (Gated Recurrent Unit for Recommendation), **SASRec** (Self-Attentive Sequential Recommendation)을 **비교 평가**하였다. 각 모델은 순차적인 사용자 행동 데이터를 학습하는 데 차별화된 접근 방식을 가지고 있어, **다양한 상황에서의 추천 성능을 평가**하기 위해 선택되었다.

LSTM은 **장기 의존 관계를 학습**하는 데 강점을 지닌 모델이지만, 복잡한 구조로 인해 훈련 시간이 길어질 수 있으며, 과적합 위험이 상대적으로 크다. 반면, **GRU4Rec**은 **더 단순한 구조**를 갖추고 있어 학습 효율이 높고, **실제 서비스 환경에서의 적용 가능성이 높다**. SASRec은 Transformer 기반의 모델로, 장기 의존성 뿐만 아니라 아이템 간의 순서를 효과적으로 학습할 수 있는 구조를 가지고 있다.



해당 그림은 **모델별 Epoch에 따른 Loss 변화를 나타낸 그래프**이다. 그래프에서 LSTM 모델의 Training Loss는 Epoch이 증가함에 따라 지속적으로 감소하는 경향을 보였으며, Validation Loss는 Epoch 2에서 최저점을 기록한 후 Epoch 3부터 증가하기 시작하였다. 이는 Epoch 3부터 과적합이 시작되었음을 나타낸다. **GRU4Rec 모델의 Loss**는 전반적으로 **안정적인 감소세**를 보였고, **SASRec 모델의 Loss**는 **상대적으로 높은 값을 유지**하였다. 각 모델의 최종 손실값은 LSTM Training Loss가 10.89, LSTM Validation Loss가 489.79, GRU4Rec Loss가 391.27, SASRec Loss가 311.30으로 나타났다.

개발 모델	MBCC	Hit Rate	NDCG
LSTM	0.5941	0.4648	0.3988
GRU4Rec	0.9584	0.7432	0.5849
SASRec	0.5861	0.5800	0.4259



모델 성능 지표 비교 결과는 다음과 같다:

- **Hit Rate**: GRU4Rec (0.7432)이 가장 높은 정확도를 기록하였으며, SASRec (0.5800)이 그 뒤를 이었다. LSTM (0.4648)은 상대적으로 낮은 정확도를 보였다.
- **NDCG**: GRU4Rec (0.5849)이 가장 높은 순위 정밀도를 기록하였으며, LSTM (0.3988)은 중간, SASRec (0.4259)은 가장 낮은 값을 기록하였다.
- **MBCC**: GRU4Rec (0.9584)이 가장 다양한 아이템을 고르게 추천하였고, LSTM (0.5941)과 SASRec (0.5861)은 상대적으로 낮은 점수를 기록하였다.

5.3 추천 시스템 적용 방안

추천 시스템 구현을 위해 순차적 사용자 행동 데이터를 효과적으로 학습할 수 있는 시계열 기반 모델 두 가지, **LSTM(Long Short-Term Memory)** 과 **GRU4Rec(Gated Recurrent Unit for Recommendation), SASRec(Self-Attentive Sequential Recommendation)** 을 비교 평가하였다.

LSTM은 장기 의존 관계를 학습하는 데 강점을 지니지만, 복잡한 구조로 인해 훈련 속도와 과적합에 대한 이슈가 발생할 수 있다. 반면, GRU4Rec은 LSTM에 비해 상대적으로 구조가 단순하면서도 추천 정확도와 학습 효율성이 뛰어나 실제 서비스 환경에 적합하다고 판단했다.

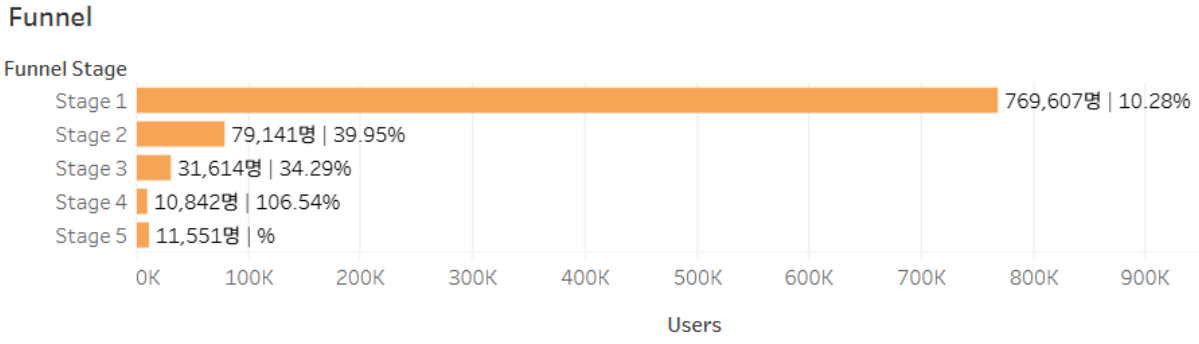
결론적으로, **GRU4Rec 모델**이 모든 성능 지표에서 가장 우수한 성능을 보여, 본 프로젝트의 장**바구니 단계 추천 시스템 구현에 최종적으로 채택**되었다. 해당 모델은 다양한 아이템을 고르게 추천하면서도 높은 정확도를 유지하여, **고객 경험 개선과 매출 증대에 기여**할 수 있을 것으로 기대된다.

6. 결론 및 한계

6.1 주요 분석 결과 요약

본 프로젝트는 고객 행동 데이터를 기반으로 사용자의 전환 흐름을 분석하고, 주요 이탈 지점을 식별함으로써 전환율 개선을 위한 전략적 인사이트를 제시하고자 하였다.

EDA 과정 중 사용자들의 사이트 내 탐색 활동량에 비해 낮은 전환율이 확인되어 퍼널 분석을 통해 전체 사용자 여정에서 이탈이 발생하는 핵심 구간을 파악하고 전환 저해 요인을 식별하기 위한 구체적인 가설을 도출하였다.



방문 및 탐색 단계에서 상세 페이지 조회로의 전환율이 가장 낮게 나타나 이 구간이 주요 병목 지점으로 확인되었으며, 해당 구간에 전환을 저해하는 요인이 존재한다고 판단하여 다음과 같은 가설을 제시하였다.

가설: 이탈한 방문자는 페이지 조회 수, 히트 수, 체류 시간 등에서 일관되게 낮은 행동 패턴을 보일 것이다.

가설에서 기대한 것과 달리 실제 분석 결과 이탈한 방문자 그룹은 페이지 조회 수, 히트 수, 체류 시간 등에서 높은 행동량을 갖는 것으로 나타났다.

전환하지 못한 사용자 그룹은 전환 그룹과 비교하여 사이트 내에서 많은 행동과 긴 체류시간의 행동 패턴을 보였으며 이러한 행동 패턴은 전환 확률에 부정적인 영향을 미치는 경향을 나타낸다.

사용자가 행동량이 많고 긴 체류 시간에도 불구하고

1. 원하는 정보를 찾지 못해 여러 페이지를 방황하거나
2. 정보 과부화, 조회 유도 프로세스 부족 등으로 결정을 내리기 어려운 상황이거나
3. 구매 의도 없이 정보 탐색만을 목적으로 방문하는 등과 같은 이유로 이탈했을 가능성이 크다.

반면 **전환에 긍정적으로 기여한 사용자 그룹**은 비교적 사이트 내에서 페이지 조회 수, 히트 수, 체류 시간과 같은 행동량이 낮은 경향을 보였다.

1. 명확한 구매 의도를 가지고 소수의 페이지만 보고 전환하거나
2. 재방문 등 이전 탐색 경험이 있어 무의미한 탐색 없이 전환에 성공했을 가능성이 높다.

전환 실패 그룹에는 사용자 선호 기반의 맞춤형 제품 추천 시스템을 통해 사용자가 원하는 정보를 쉽게 찾고 의미 있는 탐색을 유도하여 전환으로 이어질 수 있도록 탐색 흐름을 설계하는 전략이 필요하다.

전환 성공 그룹에는 전환 당시 조회한 제품군 기반 유사 상품을 추천하여 장바구니 내 업셀링 또는 크로스셀링 전략을 통해 추가 구매나 높은 객단가를 유도할 수 있으며 추천 상품 페이지 조회로 재전환과 재탐색 흐름을 만들어낼 수 있다.

결론적으로, 탐색 이상의 전환 단계를 가지는 사용자 비율을 높이기 위해서는

1. 방문자가 탐색 중 방황하거나 흐름이 단절되는 구간을 최소화하고
2. 의도가 분명해 전환에 성공한 사용자 그룹의 행동 패턴을 이탈 그룹 사용자 그룹에게 확산시키는 전략이 필요하다.

단순한 행동량의 증가보다는 사용자의 탐색 흐름의 질과 방향성, 구매 의도에 기반한 전략이 전환율 개선의 핵심임을 시사한다.

6.2 분석의 한계 및 향후 개선 방향

향후 개선 방향

고객 세그먼트 분석을 위해 진행한 분류 분석 결과, 이탈 여부에 영향을 주는 주요 변수들이 도출되었다. 향후에는 이 외에도 설명력이 높은 변수를 추가로 탐색하고 분석함으로써, 고객 이탈의 원인을 보다 세부적으로 규명할 계획이다.

또한, 장바구니 단계에 적용한 GRU4Rec 모델의 초기 하이퍼파라미터 튜닝 결과, **NDCG와 Hit Rate 지표에서 성능 향상**이 확인되었다. 이를 바탕으로 추가적인 모델 학습과 정교한 하이퍼파라미터 최적화를 통해 추천 시스템의 성능을 지속적으로 개선할 예정이다.

향후에는 장바구니 이후의 구매 이력을 기반으로, 세부 페이지 탐색 없이 이탈한 사용자의 행동 경로(페이지 탐험 흐름)를 분석하고, 해당 흐름에 맞춘 세션 기반 추천 시스템을 적용할 계획이다.

이를 통해 사용자가 원하는 정보를 더 효과적으로 노출시키고, 의미 있는 탐색을 유도하여 전환율을 높이는 방향으로 시스템을 고도화할 예정이다.

분석의 한계

본 분석에 사용된 데이터에는 프로모션 정보가 포함되어 있지 않아, 프로모션이 전환에 미치는 영향을 검증할 수 없었다. 이는 실제 구매 유도에 중요한 요인을 고려하지 못한 한계로 작용한다.

또한 유입경로 외에 방문 동기나 유입을 유도하는 외부 요인을 설명할 수 있는 변수들이 부족하여, 사용자의 방문 목적과 의도에 대한 정밀한 분석이 어려웠다. 이는 전환에 영향을 주는 다양한 요인들을 포괄적으로 파악하는 데 제약이 되었다.

특히 가장 큰 이탈이 발생한 구간은 상세 페이지에 진입조차 하지 않은 사용자 그룹이었는데, 이들의 로그 기록이 상대적으로 부족해 행동 분석의 정밀도가 낮을 수밖에 없었다.

Next Step

이러한 분석의 한계를 보완하기 위해, **프로모션 관련 데이터 확보**를 요청하고, **유입을 유도하는 추가 변수**에 대한 수집을 진행할 예정이다. 이를 통해 사용자 유입과 전환에 영향을 미치는 외부 요인을 보다 정밀하게 반영할 수 있도록 한다. 동시에, 기존 분석에서 도출된 주요 변수들에 대한 **지속적인 모니터링과 모델 성능의 최신화·최적화 작업**도 병행할 계획이다.