LIKE 어플리케이션 활성화 및 유저 리텐션 개선 방안

행동 특성에 따른 유저 세분화 분석 보고서

3팀

김주형
강보민
구지수
양윤석
최하예

목차

1. 개요

1.1 LIKE 앱 소개

LIKE는 해외 서비스인 'GAS'를 벤치마킹한 것이다. GAS는 청소년을 대상으로 한 익명 칭찬 기반 SNS로, 퀴즈 형식의 질문을 통해 친구를 선택하면 상대에게 익명으로 긍정적 메시지가 전달된다. 투표 기반이기 때문에 익명성이 유지되며, 비교적 악플의 위험이 적다는 점이 GAS의 특징이자 장점이다.

GAS를 벤치마킹하여 LIKE를 런칭한 이유는 현재 한국 청소년의 SNS 이용 시간이 꾸준히 증가하고 있으며, 동시에 자존감 회복에 대한 수요가 계속되고 있기 때문이다. 이러한 예상은 맞아떨어졌다. 청소년들의 자연스러운 바이럴을 통해 빠르게 확산되어, 출시 2개월 만에 청소년 이용자 수가 70만 명을 돌파하는 성과를 기록하였다.

1.2 분석 배경

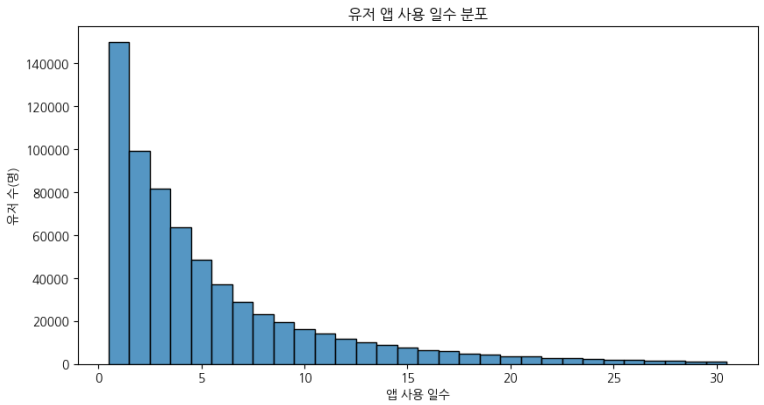
LIKE 어플리케이션은 초반의 급격한 사용자 유입과 달리, 일정 시점 이후의 리텐션 지표에서는 하락세를 보이고 있다. 이는 서비스 구조상 반복적인 질문 패턴과 제한적인 피드백 방식으로 인해 사용자 경험의 신선도가 빠르게 소진되기 때문이다. 특히 퀴즈 기반의 익명 투표는 단기적으로 흥미를 유도하지만, 장기적으로는 상호작용의 깊이가 부족하여 사용자 만족도로 이어지지 못하는 한계를 지닌다.

결과적으로 LIKE 어플은 콘텐츠 반복, 상호작용 단절, 커뮤니티 기능 부재 등 복합적인 요인으로 인해 리텐션 유지에 어려움을 겪고 있으며, 이는 서비스의 지속 가능성과 향후 성장 가능성에 있어 핵심적인 문제로 작용하고 있다.

1.3 문제 정의

LIKE 어플리케이션의 핵심 과제는 리텐션 확보에 있다. 리텐션은 단순한 재방문 여부를 넘어서, 서비스의 지속 가능성과 유저 기반의 안정성, 더 나아가 바이럴 확산을 위한 필수 지표이기 때문이다. 특히 10대를 중심으로 한 소셜 앱은 바이럴 유입이 빠른 만큼 이탈도

빠르게 발생하며 사용자의 재방문을 유도하지 못할 경우 앱 전체 생명 주기가 짧아질 위험이 크다. 또한 **유저 리텐션이 확보되어야만 입소문과 자연 확산, 즉 바이럴 루프의 선순환이 형성될 수 있으며, 이는 소셜 앱이 지속적인 성장을 이루는 데 필수적인 조건**이다. 따라서 이번 프로젝트는 LIKE 어플리케이션의 구조적 문제를 진단하고, 리텐션을 중심으로 한 실질적 활성화 방안을 도출하는 것을 목표로 한다.



cohort_month	volume	1day	2days	3days	4days	5days	6days	7days	8days	9days	10days
23-05	635505	38.83	29.68	23.84	20.17	17.69	16.36	15.46	14.14	12.77	11.9
23-06	16737	38.83	28.42	23.24	19.54	16.47	14.26	13.16	12.05	11.13	9.86
23-07	1849	28.61	20.71	17.25	15.79	12.76	11.95	10.38	10.01	9.41	8.33
23-08	524	9.92	5.92	3.05	2.48	3.24	2.1	2.67	2.29	2.48	2.29
23-09	604	17.22	10.1	7.62	5.79	6.79	4.47	4.8	3.81	3.15	2.81
23-10	409	10.02	5.13	3.67	2.93	3.67	3.42	3.18	1.96	2.69	1.22
23-11	731	9.03	3.69	1.78	1.23	0.41	1.23	0.82	0.68	0.68	0.55
24-01	231	8.66	2.6	1.73	1.3	2.6	1.3	1.73	1.3	0.43	1.3
24-02	455	7.47	3.08	1.1	0.66	1.1	1.1	1.1	0.44	0.88	0.44
24-03	255	6.27	5.49	1.57	1.96	1.18	0.39	1.18	0.39	0	0
24-04	446	6.95	3.81	3.81	2.91	2.24	3.36	1.57	0.67	0.45	0.45
24-05	196	3.57	1.53	0.51	1.02	0.51	0.51	0.51	0	0.51	0.51
24-06	49	4.08	2.04	0	0	0	0	0	0	0	0

1.4 분석 목표

- 사용자 행동 데이터를 기반으로 유저를 세분화하고 각 군집의 특성을 도출하고자 한다.
- 지속성, 활동 빈도 및 상호작용 의지, 구매 포인트 양을 중심으로 클러스터링을 수행하고자 한다.
- 군집별 행동 특성 비교를 통해 이탈 위험이 높은 집단과 충성도 높은 집단을 식별하고자 한다.
- 분석 결과를 바탕으로 각 집단에 적합한 맞춤형 유지 전략을 제안하고자 한다.
- 최종적으로는 리텐션 지표를 향상시키고 LIKE의 지속가능한 성장 기반을 확보하고자 한다.

2. 데이터 설명 및 전처리

2.1 데이터 설명

데이터 출처

- 서비스 운영 데이터(출석, 차단 기록, 친구 요청 등)

- 데이터 수집 기간: 2023년 4월 ~ 2024년 4월 (테이블 별 상이)

2.2 데이터 전처리

데이터 필터링

테이블 간 수집 기간이 상이하며 가장 빠른 수집기간을 가진 테이블은 가입 시점 컬럼이 포함된 accounts_user(유저 테이블)이었다. 대부분의 데이터가 2023년 4월 말 부터 수집되었음을 고려하였을 때, 이전에 가입한 유저들의 정확한 앱 사용 기록을 집계하기 힘들다고 판단하였으며, 2023년 5월 이전의 가입자가 19,092명으로 전체 유저 중 약 2.9%만을 차지하기 때문에 **가입 시점이 2023년 5월 이후인 유저를 대상으로 분석**을 진행하였다.

중복값 처리

포인트 기록, 구매 기록 테이블 등 사용자의 행동이 시점에 따라 기록되는 테이블을 대상으로 시점을 포함하여 전체적으로 동일한 데이터를 중복값으로 판단하여 제거하였다.

결측치 처리

전체 657,993명의 유저 중 회원 가입에 필수적인 요소인 성별이 결측치였던 2명의 유저를 제거하였다.

3. PTF 유저 세분화 분석

3.1.PTF란?

3.1.1개요

본 분석은 전통적인 고객 가치 분석 방법인 RFM(Recency, Frequency, Monetary)모델을 기반으로 하여 현재 서비스의 특성과 사용자 행동 패턴에 맞도록 핵심 지표를 재구성한 PTF 모델을 통해 유저 세분화를 시도하였다.

RFM 분석은 고객의 Recency(최신 구매일), Frequency(구매 빈도), Monetary(구매 금액)를 기반으로 고객 가치를 평가하는 기법이다. 특히 구조화된 지표를 바탕으로 고객을 정량적으로 분류할 수 있어, 개인화 마케팅과 리텐션 전략 설계에 강점을 지닌다.

그러나 본 분석에서는 분석 대상으로 삼은 서비스는 단순 구매가 아닌 앱 내에서의 사용성과 상호작용 빈도, 그리고 포인트 기반 전환 행동이 핵심이기 때문에, 기존 RFM 요소를 그대로 사용하는 데에는 한계가 있다. 이에 따라 본 분석에서는 다음과 같이 RFM 구성 요소를 서비스 특성에 맞춰 재정의하였다.

- Recency -> Total Using Days (T) : 사용자의 총 앱 접속일 수를 통해 '지속성' 평가
- Frequency -> Friend Request Count (F) : 친구 요청 횟수를 통해 활동 빈도와 상호작용 의지를 측정
- Monetary -> Point Sum (P) : 구매 포인트 양을 통해 직접적인 구매 전환 기여도 반영

이러한 PTF 지표는 사용자 행동을 보다 정밀하게 설명할 수 있는 프레임워크로 판단되며, 이를 기반으로 유저를 클러스터링하여 향후 유지율 향상과 전환율 유도를 위한 전략적 근거로 활용하고자 한다.

3.1.2 분석 지표 정의

본 분석에서 사용한 PTF 점수는 Total Using Days, Friend Request Count, Point Sum의 세 가지 지표를 기반으로 산출되며, 각각은 다음과 같은 의미와 해석 기준을 갖는다.

(1) Point Sum (P) – 구매 포인트 수

- 사용자가 앱 내에서 구매한 포인트 총량을 의미하며 이는 서비스의 핵심 기능인 투표의 '힌트 보기' UI에 사용된다.
- 해당 값이 높을수록 열의가 넘치는 유저라고 파악할 수 있으며 헤비유저일 확률이 높다. 전통적인 RFM의 Monetary 요소를 대체한다.

(2) Total Using Days (T) – 앱 사용일수

- 사용자가 앱에 접속한 총일수를 의미하며, 서비스와의 지속적인 관계 유지 여부를 나타낸다.
- 해당 값이 높을수록 사용자와의 장기적 접점이 형성되어 있을 가능성이 높고, 향후 전환 가능성 또한 상대적으로 높게 해석될 수 있다.

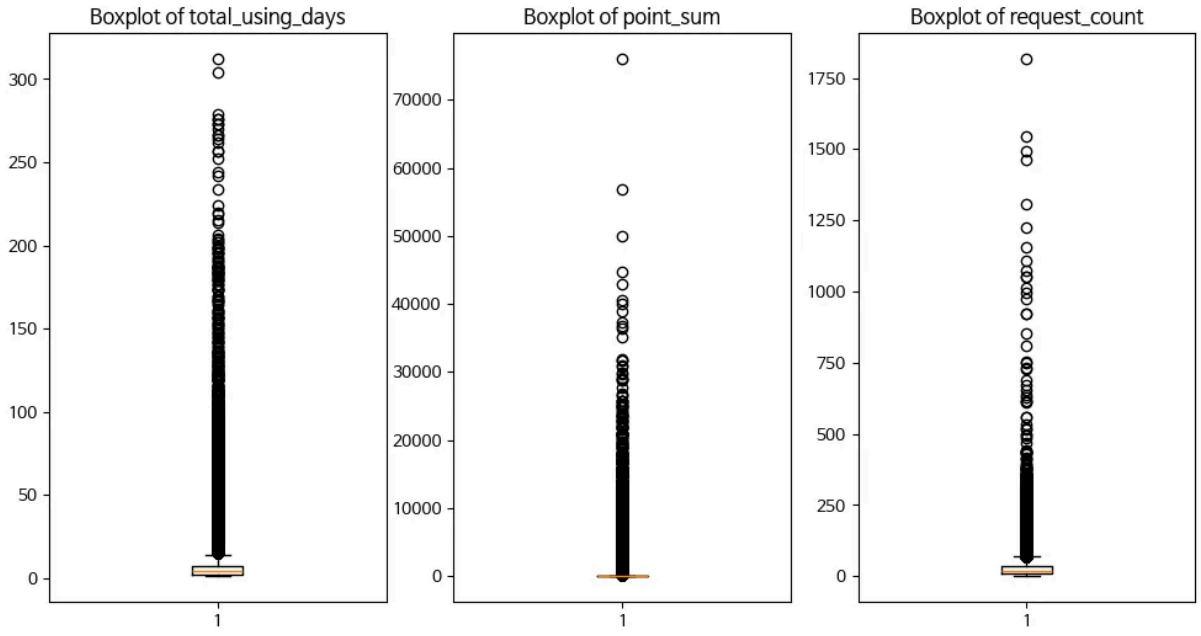
(3) Friend Request Count (F) – 친구 요청 수

- 앱 내에서 사용자가 보낸 친구 요청 횟수로, 사용자의 상호작용 의도와 커뮤니티 기능 활용도를 측정하는 지표이다.
- 빈번한 친구 요청은 단순 사용을 넘어, 타 유저와의 네트워크 형성 및 활동 확산**의 주체**가 될 가능성을 나타낸다.

3.2 1차 클러스터링 : K-means & GMM

3.2.1 데이터 분포 및 이상치 확인

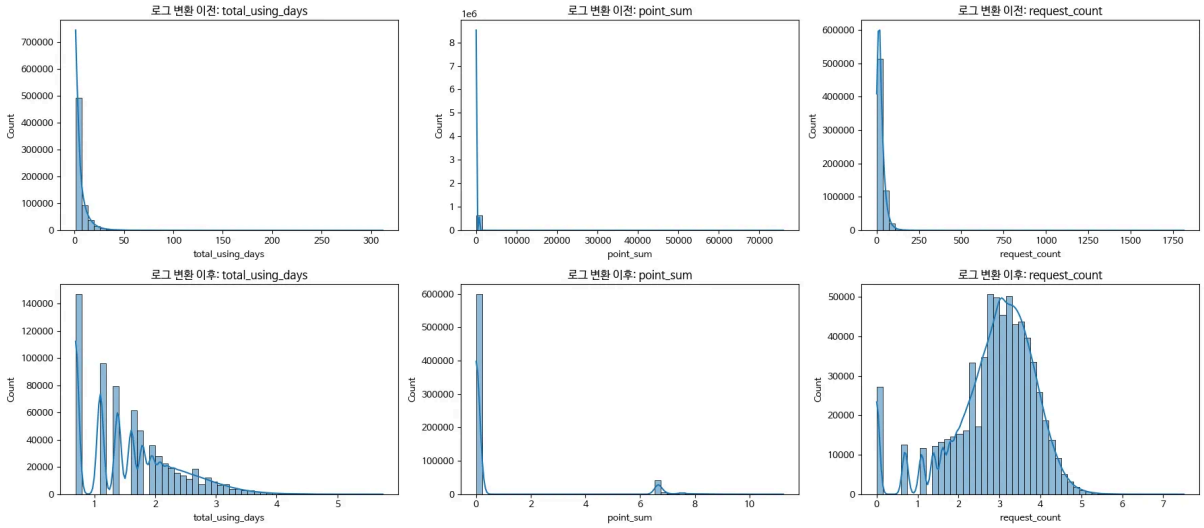
세부 분석에 앞서 사용된 세 지표(Total Using Days, Friend Request Count, Point Sum)의 분포를 확인한 결과, 전반적으로 비대칭적 분포와 이상치가 존재하는 것으로 나타났다. 특히 Point Sum의 경우 대부분의 값이 0에 집중되어 있고 일부 사용자에게서만 높은 값을 보였다. 그러나 본 분석의 목적은 극단값 자체를 특성으로 반영하여 유저를 세분화하는 것이므로, 이상치 제거는 수행하지 않고 스케일링 기반 정규화 방식을 적용하였다.



3.2.2 로그 변환 및 스케일링

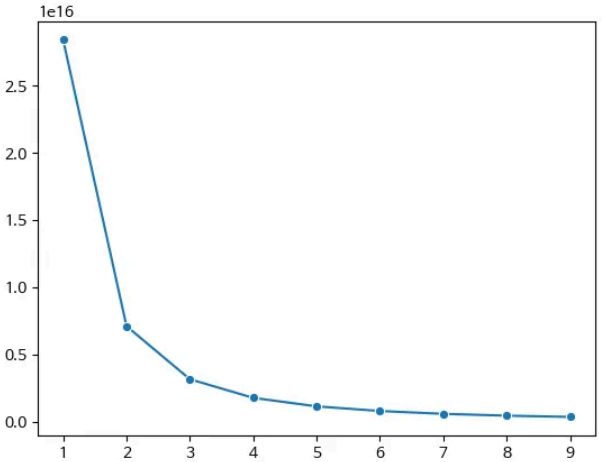
데이터 분포의 왜도를 줄이기 위해 로그 변환을 적용한 결과, 세 지표 모두 왜도 값이 크게 감소하였다. 특히 Total Using Days와 Friend Request Count는 정규 분포에 가까운 형태로 변환되었고, Point Sum 역시 왜도가 완화되었다.

왜도 로그 변환 전후		
변수	원본 왜도	로그 변환 후
Total using days	5.896748	0.645622
Point_sum	22.337908	2.936333
Friend request count	6.1641	-0.955475

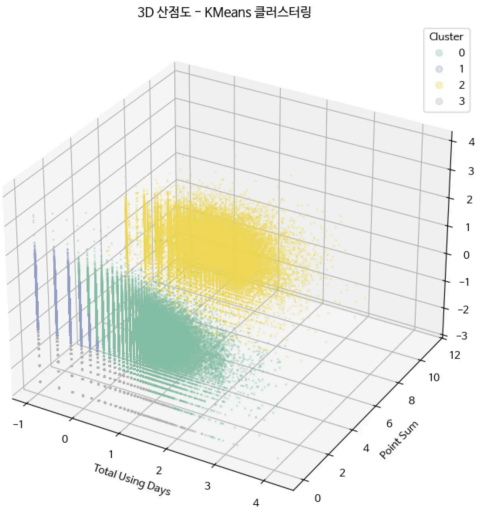


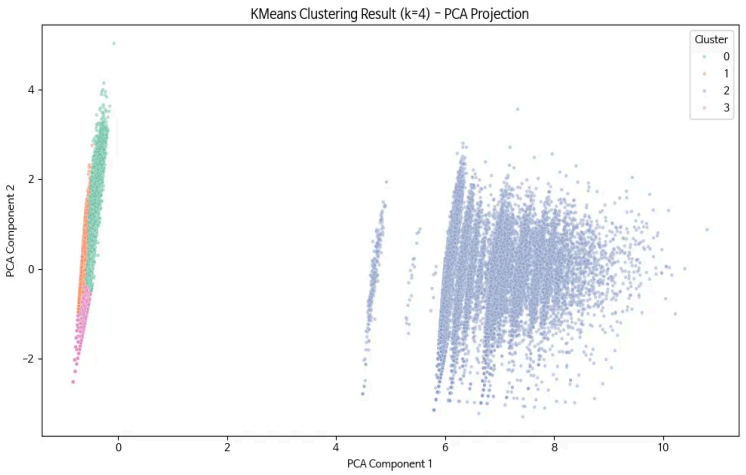
이상치와 관련해서는 로버스트 스케일링을 적용하였다. 이는 중앙값과 IQR을 기준으로 데이터를 변환함으로써 이상치에 덜 민감한 정규화 방식이다. 특히 Point Sum 과 같은 변수는 대다수의 값이 0에 몰려 있고 일부 사용자만 매우 높은 값을 보이는 심각한 비대칭 분포를 보이기 때문에, 평균 기반 스케일링(StandardScaler)보다 로버스트 스케일링이 더 적합하다 판단하였다.

3.2.3 K-Means 클러스터링 수행



위의 그래프로 미루어볼 수 있듯이 Elbow Method를 통해 최적의 클러스터 수(k)를 판단한 결과, k = 4일 때 군집 내 분산이 급격히 감소하는 경향이 나타났으며, 이를 기준으로 KMeans 클러스터링을 수행하였다.





3차원 시각화 2차원 시각화

클러스터 간 분포 차이를 정량적으로 평가하기 위해 실루엣 점수[1]를 샘플 수 기준으로 반복 계산한 결과, 평균적으로 약 0.42 수준의 점수가 도출되었다. 이는 군집 간 적절한 분리도와 응집도를 갖춘 군집화 결과임을 나타낸다.

---

1. 군집의 공간적 분리 여부 확인

- Cluster 2가 오른쪽에 뚜렷하게 분리되어 있음 → 명확한 특성을 가짐.
- Cluster 0, 1, 3은 왼쪽에 서로 겹치며 밀집되어 있음 → 서로 비슷하거나, 클러스터 경계가 명확하지 않음.

2. 클러스터 밀도/크기

- Cluster 2가 데이터의 다수를 차지하는 것처럼 보임

[1] 본 보고서에 정의되는 실루엣 점수의 방식은 데이터의 크기가 약 65만개로 전체 실루엣 계수 파악이 어렵다. 이에 랜덤 샘플링을 통해서 실루엣 계수 파악을 정의하고자 한다.
(샘플 수 : 실루엣 점수) 10,000개 : 0.42466, 20,000개 : 0.42314, 30,000개 : 0.42481, 40,000개 : 0.42470, 50,000개 : 0.42256

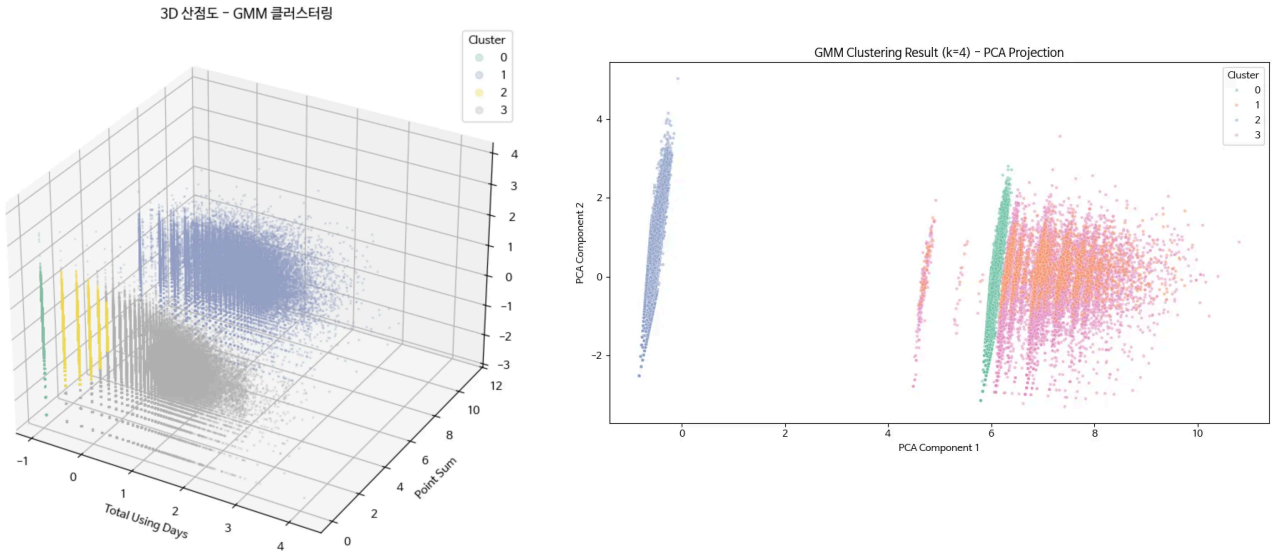
- Cluster 3은 점 크기나 밀도로 보아 소수의 특이한 그룹일 수 있음 (예: 극단적 사용자)

군집별 분포를 시각화한 결과, 한 클러스터는 명확히 분리된 특성을 보인 반면 나머지 세 클러스터는 서로 유사한 위치에 분포하며 부분적인 **중첩**을 보였다. 이로 인해 K-Means만으로 완벽한 분리는 어려울 수 있으며, 필요시 GMM과 같은 확률 기반 클러스터링 방식 적용을 고려할 수 있다.

3.2.4 GMM 클러스터링 수행

앞서 KMeans 분석에서는 거리 기반의 명확한 중심 분할이 가능했으나, 일부 클러스터 간 거리가 가까워 구분이 애매한 경우가 존재하였다. 이에 따라 보다 유연한 군집화가 가능한 GMM을 도입하여 동일한 변수(총 사용일수, 포인트 사용량, 친구 요청 수)에 대해 클러스터링을 수행하였다.

GMM은 데이터가 여러 개의 정규분포의 조합으로 구성되어 있다고 가정하고, 각 클러스터를 정규분포 형태로 표현한다. 이 방식은 KMeans에 비해 타원형, 비대칭, 중첩된 분포에도 더 유연하게 대응할 수 있다는 장점이 있다.



(좌) 3차원 시각화 결과 (우) 2차원 시각화 결과

분석 결과, GMM 역시 4개의 클러스터로 데이터를 나누었으며 다음과 같은 분포를 확인할 수 있었다.

1. 군집의 공간적 분리 여부 확인
   - **Cluster 2는** 가장 왼쪽에 뚜렷하게 분포되어 있음.
   - Total Using Days, Friend Request Count, Point Sum 모두 매우 낮은 경향
   - 비활동 사용자 집단으로 해석 가능

2. 클러스터 밀도/크기
   - **Cluster 0, 1, 3**: 오른쪽 영역에 분산된 상태로 존재함.
   - 상대적으로 활동량이나 포인트 사용량이 높은 유저군이 포함됨
   - Cluster 3은 좁고 밀도가 높으며, Point Sum값이 가장 높음 → 활성 사용자 집단

GMM의 실루엣 점수의 평균은 약 0.288~0.29 수준[2]으로, KMeans보다는 다소 낮았지만, 중첩된 사용자 집단에 대한 분류는 보다 자연스럽게 수행되었다는 점에서 의의가 있다.

3.2.5 KMeans와 GMM 결과 요약 및 결론

KMeans와 GMM은 사용자 활동 데이터를 기반으로 각각 서로 다른 접근 방식으로 군집화를 수행하였다.
KMeans는 거리 기반 중심점 클러스터링으로 뚜렷한 분리를 보여주었으며, 실루엣 계수가 상대적으로 높아 구조적 안정성을 확인할 수 있었다. 반면 GMM은 타원형 분포를 가정하여 더 유연한 경계와 분포를 형성하며, 특히 중첩된 사용자 그룹을 보다 현실적으로 표현하는 데 유리했다.

두 모델 모두 비활동 사용자(낮은 Total Using Days, Friend Request Count, Point Sum)와 고활동 사용자(모든 지표가 높은 집단)를 명확하게 구분하였으며, 이로써 사용자 행동 양극화가 존재함을 정량적으로 확인할 수 있었다.

결론적으로, **두 모델은 이후 진행되는 PTF 점수 산정과 클러스터링 전략 수립에 있어 데이터 전처리와 지표 선택의 방향성을 명확히 제시해준 기초 실험으로서 중요한 의미**를 지닌다.

[2] (샘플 수 : 실루엣 점수) 10,000개 : **0.29001**, 20,000개 : **0.28752**, 30,000개 : **0.28900**, 40,000개 : **0.29002**, 50,000개 : **0.28334**

3.2.6 Feature Importance

1차 클러스터링 분석 이후, 각 사용자 집단을 구분하는 데 있어 어떤 변수가 가장 결정적인 역할을 했는지를 파악하기 위해 다중 분류 모델 기반의 변수 중요도 분석을 수행하였다. 이 분석은 이후 PTF 점수 산출 시 가중치 설정 또는 주요 지표 선정의 근거로 활용될 수 있다. 사용한 클러스터링 모델은 K-Means 기반 클러스터 레이블로 구조적 안정성이 GMM보다 상대적으로 높기에 해당 모델을 채택하였다.

예측 대상: K-Means 기반 클러스터 레이블

사용된 입력 변수: Total Using Days, Friend Request Count, Point Sum

적용한 모델:

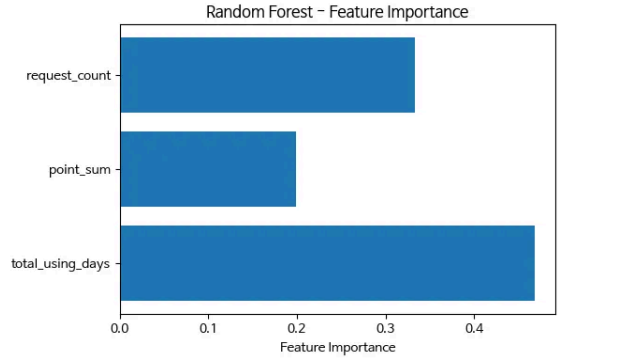
- Random Forest Classifier
- Decision Tree Classifier
- XGBoost Classifier
- LightGBM Classifier

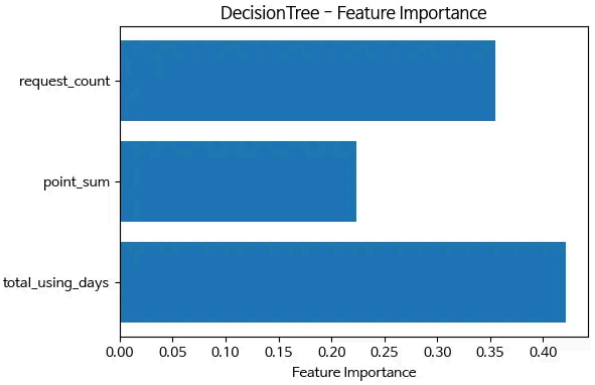
모든 모델은 학습/검증 분할 및 교차검증을 통해 학습되었으며, 변수별 중요도는 모델별 방식에 따라 계산되었다.[3]

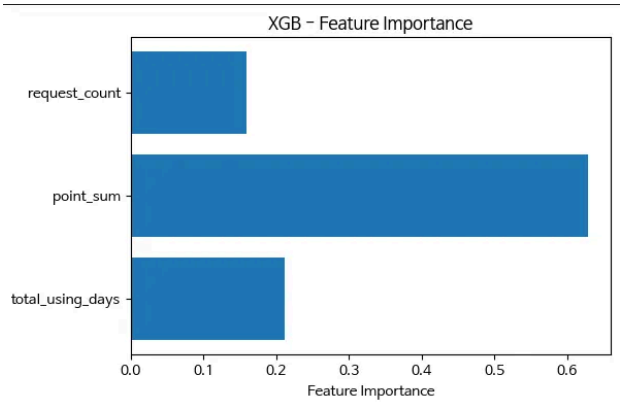
3.2.7 변수 중요도 결과

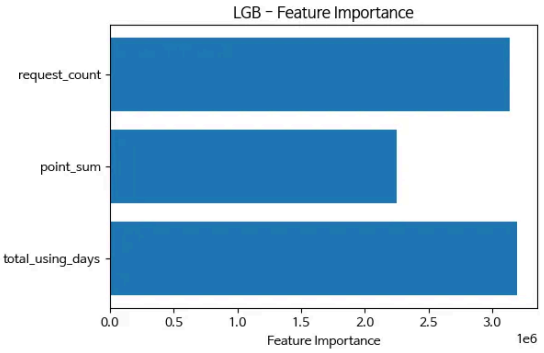
모델 전반에서 가장 높은 중요도를 보인 변수는 Total Using Days 였다. 이 지표는 사용자가 서비스에 얼마나 오래 머물렀는지를 반영하는 직관적인 지표로, 클러스터 간 가장 뚜렷한 분리 기준이 되었다. 결국 "이 사람은 얼마나 이 서비스를 이용했는가?"라는 기초적이면서 강력한 차별화 포인트라 판단된다.

[3] RandomForest와 DecisionTree는 Gini 기준 불순도 감소량, XGBoost와 LightGBM은 gain 기반 중요도 사용









- Point Sum은 XGBoost에서 가장 높은 중요도를 보였으며, 이는 해당 모델이 포인트 사용 여부에 민감하게 반응했음을 의미한다.
- Friend Request Count는 세 모델에서 상대적으로 덜 중요한 변수로 나타났지만, 특정 군집 구분에서는 보조 역할을 수행했다.

3.2.8 인사이트 및 결론

Total Using Days는 모든 모델에서 공통적으로 가장 중요한 변수로 도출되었으며, 이는 사용자의 앱 체류 기간이 전반적인 활동성과 충성도를 반영하는 핵심 지표임을 시사한다. 또한 Point Sum은 비활동 사용자와 고활동 사용자를 분리하는 데 효과적인 지표로 확인되며, 사용자의 금전적 참여도 측정 지표로 유효하다. Friend Request Count는 상대적으로 낮은 중요도를 보였지만, 소셜 기능 활용에 초점을 맞춘 유저 그룹 구분에는 여전히 의미 있는 변수로 판단된다.

3.3 2차 클러스터링 : 변동계수 가중치

3.3.1 점수 계산 방식

기존 RFM 기반 사용자 점수화에서 가장 큰 고민 중 하나는 지표별 가중치를 어떻게 설정할 것인가이다. 동일 가중치를 부여하거나 전문가 직관으로 정한 가중치는 신뢰성과 객관성에 한계가 있다.

이에 따라 PTF 점수는 각 변수의 패턴 일관성과 군집 구분 가능성을 함께 반영할 수 있는 변동계수 기반 가중치 방식을 채택하였다. 변동계수는 데이터의 흩어짐을 나타내는 지표이며, 변동 계수가 작을수록 데이터의 흩어짐이 작다는 것을 의미한다. 해당 열의 w_n (n = P, T, F)이 크다는 것을 n열에 대하여 군집화 결과가 다른 열에 비해 더 작은 변동 계수를 갖고 있어서, 데이터의 흩어짐이 작다는 것을 의미한다. 단순한 등급화보다 데이터 기반의 정밀한 사용자 구분이 가능하다는 점에서 전략적 장점을 갖는다.

데이터 전처리 과정에서는 로그 변환을 통해 각 변수의 왜도를 완화한 뒤, Total Using Days와 Friend Request Count는 StandardScaler를, Point Sum은 RobustScaler를 적용하여 이상치의 영향을 최소화하였다. 이와 같은 스케일링 전략은 변수의 분포 특성과 현실적 맥락(예: 대부분의 사용자가 포인트를 사용하지 않음)을 고려한 결과이며, 사용자 행동의 상대적 차이를 과장 없이 반영할 수 있다는 점에서 적절한 선택이다.

그 후, 각 변수에 대한 K-Means 클러스터링을 통해 군집 내 변동계수를 산출하고, 루트 보정과 로그 역수 보정을 통해 정규화된 가중치를 계산하였다. 이렇게 하면 극단적으로 작은 CV가 과도한 가중치를 가져가는 것을 방지하고, 보다 균형 잡힌 가중치 분포를 유도할 수 있다. 결과적으로, 세 변수의 변동 계수는 다음과 같이 설정되었다.

- W_P (Point Sum): 0.25
- W_T (Total Using Days): 0.43
- W_F (Friend Request Count): 0.31

이는 사용자의 전체 체류 일수가 가장 높은 설명력을 가지며, 친구 요청, 포인트 사용 순으로 군집화를 설명하고 있음을 의미한다.

3.3.2 점수 산출 및 클러스터링 적용

최종적으로 산출된 가중치를 바탕으로 PTF 점수를 다음과 같은 방식으로 계산하였다.

$$PTF\_SCORE = W\_T \cdot T + W\_F \cdot F + W\_P \cdot P$$

해당 점수는 사용자별 활동성, 참여도, 전환력을 하나의 정량화된 지표로 통합한 값이며, 클러스터링의 핵심 기준으로 사용되었다. 이후 PTF 점수를 기반으로 다시 KMeans 클러스터링을 적용하였고, Elbow Method 분석을 통해 k=4가 최적의 군집 수로 결정되었다.

이 과정을 통해 생성된 4개의 사용자 군집은 단순히 점수의 크기뿐 아니라 행동 변화 횟수, 참여의 지속성, 구매 활동의 복합적 특성을 반영하고 있다.

해당 과정을 통해 다음과 같이 PTF 점수가 반영된 데이터 프레임을 생성하고 클러스터별 통계를 요약한 값이다.

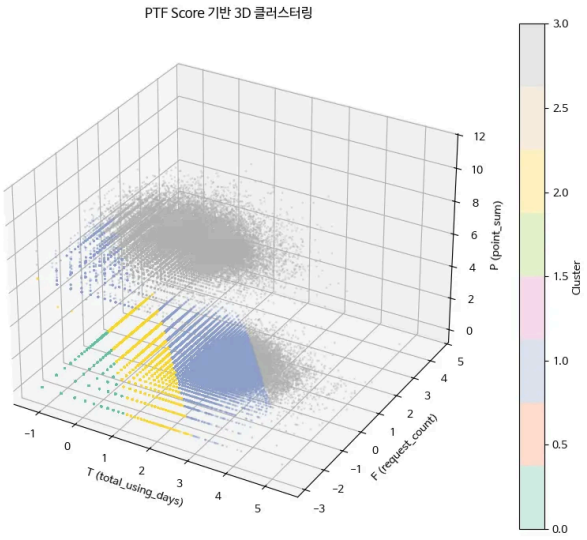
PTF_cluster	PTF_SCORE	P	T	F	Count
0	-0.8012	0.0000	-1.0423	-1.1177	160,859
1	0.6988	0.2349	-1.0421	0.6020	168,785
2	-0.0310	0.0003	-0.2129	0.1938	272,015
3	2.1614	6.4497	0.8820	0.4520	56,322

위 통계 요약값은 각 클러스터의 평균적인 PTF 점수 및 구성 요소(Point Sum, Total Using Days, Friend Request Count)의 경향을 수치적으로 보여준다. 그러나 수치만으로는 각 클러스터 간의 공간적 분리, 밀집도, 상호 유사성 등을 충분히 파악하기 어렵기 때문에, 다음 단계에서는 클러스터 분포를 3차원 시각화 및 PTF 점수 기반 밀도 분포 그래프를 통해 시각적으로 확인하고자 한다.

3.3.3 최종 클러스터링 결과 및 분포 해석

앞선 절차를 통해 산출된 PTF 점수를 기반으로, 최적의 클러스터 개수인 k=4로 K-Means 클러스터링을 수행하였다. 클러스터링 결과 각 군집은 사용자의 활동 특성에 따라 의미 있는 형태로 구분되었으며, 이를 통해 다양한 사용자 집단의 행동 패턴을 식별할 수 있었다.

특히, 클러스터링의 유효성을 보다 직관적으로 파악하기 위해 **3차원 시각화와 PTF 점수 기반 분포도 시각화를 함께 활용**하였다. 이를 통해 수치로 파악하기 어려운 클러스터 간의 거리, 밀도, 중첩 정도를 시각적으로 확인할 수 있었다.



먼저, 오른쪽의 3D 산점도는 총 앱 사용일수(T), 친구 요청 수(F), 포인트 합계(P)를 각 축으로 설정하고, 클러스터별 색상으로 구분하여 시각화하였다. 결과적으로 다음과 같은 인사이트를 확인할 수 있었다.

---

1. 🟢Cluster 0 (비활성 유저 그룹)

- 전체적인 PTF 점수가 낮은 편에 속한다.
- 거의 사용 이력이 없는 이탈 직전 유저 or 이탈 유저라고 봐도 무방하다

2. 🔵 Cluster 1 (안정적 유저 그룹)

- 모든 특성값(T, F, P)이 준수하게 보유된 집단이다.
- 주기적 활동 유저로서 안정적인 사용층이 확보된 집단으로 해석 가능하다.

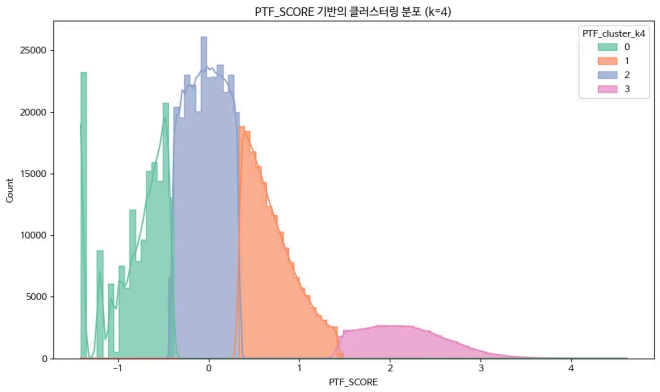
3. 🟡Cluster 2 (저활동 유저 그룹)

- 세 지표 모두 0에 가장 가까운 값을 보이며, 그래프 하단에 몰려 있는 것이 특징이다.
- 포인트 사용량이 거의 없고, 사용일도 짧은 비활성 사용자 또는 이탈 위험 유저로 해석된다. 가장 많은 인원으로 분류된 집단이다.

4. ⚪Cluster 3 (고활동 유저 그룹)

- 그래프의 상단 및 우측에 위치하며, 모든 지표에서 높은 값을 보이는 고활동 사용자 집단이다.
- 활용 빈도와 충성도가 모두 높은 핵심 유저층으로 분류 가능하다.

다음은 PTF 점수 기반 분포도 시각화를 살펴보았다.



1. Cluster 3(고활동 유저 그룹)의 점수 분포는 오른쪽 꼬리 방향으로 길게 형성되어 있으며, 점수의 분산도 큰 편이다. 이는 고활동 사용자가 다양한 수준으로 분포하고 있음을 시사한다.

2. Cluster 2(저활동 유저 그룹)는 매우 좁은 범위에서 점수가 밀집되어 있으며, 전체 분포의 왼쪽에 위치해 있다. 이는 해당 군집이 낮은 활동성을 지닌 균한 사용자로 구성되어 있음을 의미한다.

3. Cluster 0(비활성 유저 그룹)과 Cluster 1(안정적 유저 그룹)은 점수 분포가 다소 중첩되어 있으나, 평균값과 분포 폭에서 차이를 보이며 일부 분리된 특성이 확인되었다.

3.4 클러스터 가설 검정

앞서 PTF 점수 기반으로 수행한 클러스터링으로 사용자 집단을 구분하였다. 그러나 이러한 분류가 단순한 수학적 나눔에 그치지 않고 실제 사용자 간 특성이 통계적으로 의미 있는 차이를 보이는지 여부를 확인하는 것은 세분화 전략의 타당성을 확보하는 데 중요한 절차이다.

3.4.1 실루엣 점수 계산

이에 따라 본 절에서는 클러스터링의 품질을 정량적으로 평가하기 위해 실루엣 계수(Silhouette Score)를 활용하여 분석을 진행하였다. 실루엣 계수는 각 데이터 포인트가 해당 클러스터에 얼마나 잘 속해 있는지를 나타내는 지표로, 군집 간 분리도와 군집 내 응집도를 종합적으로 반영하는 평가 방법이다.

실루엣 점수는 PTF_cluster 컬럼을 기준으로 샘플 개수에 따라 다르게 평가되었고 실루엣 점수 평균은 약 0.31[4]로 양호한 분리 수준을 나타낸다. 이는 전체 데이터가 4개의 사용자 군집으로 나뉘는 구조가 비교적 분명하며, 실질적인 행동 기반 차이를 반영한 결과임을 시사한다. 또한 클러스터링은 통계적으로도 분리도가 존재함을 확인하였다. 이는 단순 군집화가 아닌, 실제 사용자 간의 행동 패턴 차이를 반영한 세분화 결과로 해석할 수 있다.

3.4.2 ANOVA 및 사후 검정

클러스터링을 통해 사용자들을 네 개의 집단으로 세분화하였지만, 이러한 구분이 실제로 의미 있는 행동 차이를 반영하는지 여부는 추가적인 통계 검정을 통해 확인할 필요가 있다. 특히 본 분석에서는 사용자들의 앱 사용 패턴(총 사용일수, 친구 요청 수, 구매 포인트)에 대해, 클러스터 간 평균 차이가 존재하는지를 확인하고자 하였다.

이를 위해 다음과 같은 형태의 가설을 설정하였다.

- 귀무가설 : 클러스터 간 각 지표의 평균에는 유의미한 차이가 없다.
- 대립가설 : 클러스터 간 적어도 하나의 집단 사이에는 평균 차이가 존재한다.

먼저,각 변수에 대해 클러스터 컬럼을 독립변수로 하여 일원분산분석(One-way ANOVA)을 실시하였다. ANOVA는 세 개 이상의 집단 간 평균 차이를 검정하는 대표적인 방법으로, 전체적인 평균 차이가 있는지를 먼저 판별한다.

만약 ANOVA 결과에서 통계적으로 유의미한 차이가 확인되면, 이후 Tukey's HSD(Honestly Significant Difference) 사후 검정을 통해 어떤 클러스터 쌍이 유의미하게 다른지를 세부적으로 분석하였다. 이는 ANOVA가 전체 차이만 확인하는 반면, Tukey는 개별 집단 간 차이까지 정량적으로 보여주는 장점이 있다.

[4] (샘플 수 : 실루엣 점수) 10,000개 : 0.29001, 20,000개 : 0.28752, 30,000개 : 0.28900, 40,000개 : 0.29002, 50,000개 : 0.28334

각 변수에 대해 ANOVA 및 Tukey's HSD를 실시한 결과는 다음과 같다.

Total Using Days

	group1	group2	meandiff	p-adj	lower	upper	reject
0	0	1	10.8391	0.0	10.7804	10.8977	True
1	0	2	2.3929	0.0	2.3399	2.4458	True
2	0	3	11.9189	0.0	11.8365	12.0014	True
3	1	2	-8.4462	0.0	-8.4983	-8.3940	True
4	1	3	1.0799	0.0	0.9979	1.1618	True
5	2	3	9.5260	0.0	9.4481	9.6040	True

- ANOVA **결과**: p-value < 0.001
  → 클러스터 간 평균 방문일 수에 유의미한 차이가 있음

- Tukey HSD **해석 요약**:
  - Cluster 0 vs 2: 유의미한 차이 없음
  - Cluster 1, 3은 다른 대부분의 군집과 유의한 차이를 보임
  - 특히 Cluster 3은 가장 높은 방문일 수 평균을 기록함

Friend Request Count

	group1	group2	meandiff	p-adj	lower	upper	reject
0	0	1	32.1651	0.0	31.9787	32.3516	True
1	0	2	17.5350	0.0	17.3667	17.7033	True
2	0	3	28.0850	0.0	27.8230	28.3469	True
3	1	2	-14.6301	0.0	-14.7959	-14.4643	True
4	1	3	-4.0801	0.0	-4.3405	-3.8198	True
5	2	3	10.5500	0.0	10.3023	10.7977	True

- ANOVA **결과**: p-value < 0.001
  → 친구 요청 수 역시 클러스터에 따라 통계적으로 유의미한 차이가 있음

- Tukey HSD **해석 요약**:
    - 모든 클러스터 간 조합에서 유의미한 차이 존재
    - 요청 수가 클수록 높은 활동성을 의미하며, Cluster 3에서 집중적으로 나타남

Point Sum

	group1	group2	meandiff	p-adj	lower	upper	reject
0	0	1	29.9873	0.0	25.3465	34.6280	True
1	0	2	0.0265	1.0	-4.1626	4.2156	False
2	0	3	1233.9915	0.0	1227.4710	1240.5119	True
3	1	2	-29.9607	0.0	-34.0875	-25.8339	True
4	1	3	1204.0042	0.0	1197.5236	1210.4848	True
5	2	3	1233.9649	0.0	1227.7997	1240.1302	True

- ANOVA **결과**: p-value < 0.001
  → 포인트 사용량의 클러스터 간 차이도 통계적으로 유효

- Tukey HSD **해석 요약**:
    - Cluster 0 vs 2는 유의미한 차이가 없음
      기타 조합에서는 명확한 차이가 존재
    - 일부 클러스터는 포인트 사용이 거의 없거나 매우 낮은 수준에 머무름

본 분석을 통해, 클러스터링 결과가 단순 점수 기반 구분이 아닌 행동 패턴의 실제 차이를 반영한 구조임이 통계적으로도 입증되었다. 특히 다음과 같은 인사이트를 도출할 수 있다:

- Cluster 3(고활동 유저 그룹)은 모든 변수에서 평균이 가장 높고, 모든 다른 클러스터와 유의미한 차이를 보이며 가장 활동적이고 충성도 높은 사용자군으로 분류된다.

- Cluster 2(저활동 유저 그룹)는 포인트와 방문일 모두 낮은 값을 기록하며, 다른 클러스터와의 차이도 뚜렷하여 이탈 위험이 높은 비활성 사용자군임을 시사한다.

- 일부 클러스터(예: 0과 2)는 변수에 따라 평균이 유사하여 특정 지표에서는 구분이 어렵고, 보다 세밀한 기준 또는 추가 지표의 도입이 필요함을 나타낸다.

3.4.3 예측 모델 지표 평가

클러스터링은 비지도 학습 기반의 데이터 세분화 기법으로, 데이터에 내재된 구조를 지도 학습 없이 파악할 수 있다는 점에서 유용하다. 하지만 클러스터링 결과가 얼마나 잘 분리되었는지, 즉 실제로 유저 집단 간의 구분이 명확한지를 객관적으로 검증하는 데에는 한계가 존재한다.

이에 본 분석에서는 클러스터링을 통해 도출된 유저별 클러스터 레이블을 타깃 변수로 설정하고, 다양한 지도 학습 기반 분류 모델을 학습시켜 보았다. 이는 클러스터링 결과가 얼마나 예측 가능한지를 판단함으로써, 해당 클러스터 구조가 실제로 의미 있는 분리였는지를 확인하고자 하는 목적이다.

분류 모델의 정확도, 정밀도, 재현율, F1 점수 등 주요 성능 지표가 높게 나타날수록, 클러스터 간 경계가 뚜렷하고 재현 가능한 패턴이라는 해석이 가능하다. 결과적으로 본 절차는 비지도 학습 결과의 품질을 정량적으로 평가하고, 향후 신규 유저 분류 자동화 가능성을 판단하는 근거로 활용될 수 있다.

분석 대상 데이터는 앞서 산출된 PTF 점수 기반 클러스터링 결과를 종속변수로 설정하고, 총 사용일수, 친구 요청 수, 포인트 합계의 세 가지 지표를 독립변수로 사용하였다. 총 5개의 머신러닝 모델(Decision Tree, Random Forest, LightGBM, XGBoost, AdaBoost)을 선정하여 교차검증 기반 성능 비교를 진행하였다.

각 모델의 성능 평가지표 결과는 다음과 같다.[5]

Model	f1_weighted	balanced_accuracy	recall_weighted
Decision Tree	1.0000	1.0000	1.0000

[5] f1_weighted: 클래스 불균형을 고려한 f1-score

balanced_accuracy: 클래스 간 균형을 고려한 정확도

recall_weighted: 전체 리콜을 가중 평균한 지표

Random Forest	1.0000	1.0000	1.0000
LightGBM	0.9999	0.9999	0.9999
XGBoost	0.9997	0.9996	0.9997
AdaBoost	0.1980	0.5000	0.3421

전반적으로 매우 높은 성능을 보였으며, 특히 Decision Tree, Random Forest, LightGBM 모델은 거의 완벽한 수준의 분류 성능을 달성하였다. 이는 클러스터링 과정에서 생성된 종속변수수가 세 변수(P, T, F)의 조합으로 매우 명확하게 구분되었음을 의미한다. 즉, 각 클러스터 간의 경계가 선명하며, 예측 가능성이 매우 높은 구조임을 시사한다.

반면, AdaBoost의 성능은 모든 지표에서 현저히 낮게 나타났으며, 이는 기본 약한 분류기(weak learner)의 단순성이나 적합하지 않은 파라미터 설정에서 기인할 가능성이 있다. 특히 Random Forest와 XGBoost, LightGBM과 같은 앙상블 모델은 불균형한 분포에서도 강한 예측 성능을 보이는 것이 특징이다.

4. 맞춤형 제안 및 시사점

4.1 유지율 상승을 위한 맞춤 전략 제안

PTF 기반 클러스터링 결과를 바탕으로, 각 유저 집단은 서비스에 대한 참여 방식과 충성도에서 뚜렷한 차이를 보였다. 결국 갑 집단마다 향후 어플 사용에 대한 의지 차이가 발생할 수 밖에 없다. 그렇기에 클러스터별로 유지율을 높이거나 유지하기 위한 구체적인 전략을 제안하고, 각 전략이 실제 사용자 행동에 미치는 영향을 중심으로 설계 방향을 제시하고자 한다.

1. Cluster 0 : 비활성 유저 그룹

비활성 유저 그룹은 전체 유저 중 약 16만 명 규모를 차지하며, 모든 지표가 −1 이상으로 극단적으로 낮은 수준을 보인다. 이는 해당 집단이 앱을 거의 사용하지 않고 곧바로 이탈한 사용자들로 구성되어 있음을 시사한다. 이러한 유저는 이미 서비스와의 접점이 사라졌거나

관심도가 매우 낮기 때문에, 마케팅 전략보다는 서비스 진입 전 경험(온보딩 UX)개선이나 단기 복귀 유도 정도로 접근하는 것이 효율적이다.

2. Cluster 1 : 안정적 유저 그룹

안정적 유저 그룹은 비활성 유저 그룹과 약 8천 명 차이로 비슷한 규모로 구성되어 있다. 이들은 이들은 (F)지표에서 0.6 수준으로 비교적 높은 수치를 기록하고 있고, (P) 또한 0.23으로 어느 정도의 포인트 구매가 이뤄지고 있다. 반면 (T)는 -1.04로 여전히 낮은 수치를 보이며, 이는 이들이 짧은 기간 동안에만 비교적 앱을 사용하는 '휘발성 활동자'임을 시사한다.

따라서 이 집단은 지속적인 재방문을 유도하기 위해선 이들이 경험하는 사용 맥락을 파악해 습관화 요소를 강화하거나, 점진적으로 VIP 그룹으로 전환할 수 있도록 유도하는 전략이 효과적이다.

3. Cluster 2 : 저활동 유저 그룹

저활동 유저 그룹은 약 27만 명으로 전체 유저 중 가장 많은 비중을 차지하며, 실질적으로 더 주의가 필요한 집단임을 알 수 있다. 모든 지표에서 평균 이하의 낮은 수치를 보임에도 불구하고 PTF_SCORE는 -0.03으로 오히려 중간 수준에 위치한다. 이는 점수 계산 구조의 특성상 상대적으로 균형 잡힌 '무활동' 상태가 특징이다. 실제로 해당 집단은 '눈에 띄지 않는 대규모 비활동 유저 집단'으로 방치 시 유지율 하락과 이탈로 직결될 가능성이 매우 높다. 이에 따라 저활동 유저 그룹은 어플 접속 후, 첫 Action의 기회를 확대하는 전략을 구성해야 한다.

4. Cluster 3 : 고활동 유저 그룹

고활동 유저 그룹은 모든 지표에서 두드러진 수치를 기록하며, 명확한 고활동자 집단으로 분류된다. 세 지표 모두 상대적으로 가장 높은 수준을 보이며, 장기적 접속속 + 지속적 관계 형성성 + 적극적 구매 활동이 모두 병행된 사용자임을 나타낸다. 비록 규모는 5만 6천 명(8.5%)으로 적지만, 전체 매출 또는 활성이용률에서 핵심적인 기여를 할 가능성이 높은 VIP 유저층이다.

이 집단은 별도의 리텐션 유도 없이도 자체적으로 높은 유지율을 보일 수 있으나, 장기적 유지와 확장을 위해서는 심화 경험과 서비스 중심 홍보로 연결시켜야 한다. 특히 오피니언 리더로서 자발적으로 앱을 추천하거나 콘텐츠 생산자로 유도될 수 있도록 동기를 부여하면, 앱의 핵심 가치의 외부 전이도 기대할 수 있다.

4.2 향후 분석 및 개선 방향

본 프로젝트는 로그 변환, 스케일링, 클러스터링, 가설 검정, 예측 모델링 등 다양한 정량 분석 기법을 통해 사용자 세분화 및 맞춤 전략 수립에 도달하였다. 그러나 분석 과정에서 드러난 한계점과 데이터 구조상 보완이 필요한 요소들을 기반으로 향후 분석 방향을 다음과 같이 제시하고자 한다.

4.2.1. 주요 변수를 보강한 다층적 PTF SCORE 재정의

본 프로젝트는 PTF 모델을 단순화하여 구매 포인트 수(P), 앱 사용일 수(T), 친구 요청 수(F)의 3개의 변수로 구성하였다. 하지만 이 모델은 앱 내 다양한 기능[6]을 포괄하지 못해 사용자의 다층적 행동을 반영하는데 한계가 있다. 향후에는 이를 반영한 확장형 PTF+ 모델을 설계하여 사용자 유형을 보다 정밀하게 구분하는 것을 기대한다.

4.2.2 다양한 모델 비교 및 앙상블 가능성 모색

현재는 주로 K-Means, GMM, RandomForest등 비교적 모델을 제한적으로 사용하여 분석이 진행되었다. 하지만 사용자 행동 데이터는 비선형성과 상호작용이 복잡하므로, 여러 모델을 혼합하거나 더 정교한 비지도 모델을 적용할 수 있다. 클러스터링에는 DBSCAN, Spectral Clustering 등을 적용하거나, 예측 모델에는 Stacking Ensemble, Deep Neural Network 등의 확장을 고려할 수 있다.

5. 결론 및 제언

본 프로젝트는 RFM 분석 기법을 변형한 PTF 모델을 활용하여 사용자 행태를 기반으로 한 유저 세분화 분석을 수행하였다. 총 3개의 핵심 지표(총 사용일수, 친구 요청 수, 구매 포인트 수)를 통해 개별 사용자의 서비스 참여도와 충성도를 계량화하였고, 이후 가중치 계산과

[6] 투표, 차단 및 신고, 채팅, 학교 및 학급 정보, 질문 행동 등

K-Means 클러스터링을 통해 4개의 뚜렷한 사용자 집단을 도출하였다. 이 과정에서 단순한 수치 나열이 아닌, ANOVA, Tukey's HSD와 같은 통계적 검증을 통해 군집 간 실질적인 평균 차이를 확인하고, 실루엣 점수 및 분류 예측 모델을 통해 클러스터의 분류 가능성과 내부 응집력을 확인하는 과정을 병행하였다.

특히, 클러스터링 결과를 단순 분류에 그치지 않고 실루엣 계수를 활용한 군집 응집도 평가, 분류 예측 모델을 통한 f1_weighted 및 클러스터 재현성 측정 등은 단순 군집화에 그쳤던 기존 분석의 한계를 보완하였다. 이는 결과의 신뢰도를 높이는 역할을 수행하였다. 이러한 과정은 클러스터링 결과를 실제 마케팅 및 운영 전략에 반영할 수 있는 실용적 근거로 확장시켰다는 점에서 의미가 있다.

그 결과, 전체 유저 중 상당수가 이탈 가능성이 높거나 저활동 상태로 남아 있음을 확인하였고, **이러한 '위험 이탈군'에 대한 전략적 접근이 서비스 성장의 주요 관건임을 인식하게 되었다.** 특히, 활동이 거의 없는 저활동 유저 그룹의 규모는 전체 유저의 40%에 달해 서비스의 핵심 성장 모멘텀이 해당 집단에 있다는 판단을 가능하게 했다. 반면, 고활동 유저 그룹과 같은 고활동 VIP 유저는 수는 적지만 모든 지표에서 매우 우수한 값을 보였다. **이들을 통한 장기 유저 유지 및 바이럴 유입 효과에 집중할 필요가 있다는 점도 파악**되었다.

핵심은 저활동 유저 그룹을 중심으로 한 리텐션 전략이 서비스 성장의 핵심 과제가 되어야 한다는 점이다. 단순히 VIP 집단을 강화하거나 신규 유저만을 확보하려는 접근보다, 낮은 관심도& 높은 수를 보유한 집단을 되살리는 것이 ROI 관점에서도 훨씬 효율적인 전략이 될 수 있다. 해당 집단에게는 첫 행동 유도, 초기 보상 리뉴얼 등 접근 문턱을 낮춘 실용적 전략이 필요하며, 단기성과보다는 장기적 습관 형성 관점에서 단계적 개입 설계가 요구된다.

궁극적으로 이번 PTF 기반 유저 세분화 분석은 서비스 내 사용자의 행동을 구조화하고, 그에 따른 전략적 기획과 마케팅 실행을 가능하게 하는 출발점이자 근거 자료를 제공한다. 본 분석이 향후 서비스 운영과 유저 리텐션 전략 수립의 핵심 가이드라인으로 작용하길 기대한다.

서브프로젝트:
운영 대시보드 및 군집 별 대시보드

1. 제작 배경

LIKE 어플리케이션은 서비스 초기 빠르게 유저를 유입시켰으나, 지속적인 리텐션 유지에 어려움을 겪고 있다. **특히 유저 행동 데이터의 정합성이 낮아 로그 데이터 사용 및 분석에 한계가 있는 등 안정적인 데이터 수집 기반이 부족한 상황이다.** 또한 SNS 서비스의 특성상 지표의 실시간 변화에 빠르게 대응해야 하며, 운영팀이 즉각적인 판단과 조치를 내릴 수 있도록 돕는 도구가 필요하다. 이에 따라 **유저 행동을 직관적으로 파악할 수 있는 대시보드를 별도로 제안하고자 한다.** 대시보드는 실무 운영자의 빠른 의사결정과 유저 리텐션 개선 전략 수립에 기여할 수 있도록 구성되었다.

또한 메인 프로젝트에서 진행된 클러스터링 분석 외의 별도의 클러스터링 분석을 진행하여 이에 따른 결과를 대시보드에 추가하였다. 유저들과 앱의 상호작용을 트래킹하기 위해 구성된만큼, 클러스터링 분석은 서비스 내 활동성, 사회적 관계 형성, 네거티브 피드백(차단, 신고 등)을 지표로 삼아 진행되었다. 대표적인 군집들을 선정하여 추적하며, 중요도가 높은 군집의 유지율 변화를 실시간으로 모니터링하기 위함이다. 또한 유지율 개선에 있어 우선적으로 타게팅되어야 할 사용자 군집을 판단하고, 이를 효과적으로 추적하여 관리할 수 있도록 한다.

2. 파이프라인 설계 및 구축

2.1 개요

본 데이터 파이프라인은 GCP의 가상 서버를 활용하여 Airflow 환경에서 운영되며, 일 단위로 자동 업데이트되도록 구성되었다. 이를 통해 데이터의 최신성을 유지하고 지속적인 분석 및 대시보드 반영이 가능하도록 하였다.

2.2 데이터 파이프라인 구성

1. **데이터 수집 및 적재**
    - 원본 데이터는 GCS bucket에 주기적으로 자동 적재됨.
    - 다양한 소스로부터 수집된 데이터가 표준화된 형식으로 저장됨.
2. **데이터 전처리 및 지표 계산**
    - 수집된 데이터에 대해 대시보드에 포함될 지표 별 전처리 과정.
    - 앱 활성화, 유저 리텐션, 수익에 대한 핵심 지표(DAU, 투표 수, 구매 포인트 수 등)를 계산.
    - K-means를 활용하여 클러스터링을 수행하고, 유저 그룹을 세분화함.
3. **데이터 마트 구축 및 저장**
    - 전처리된 데이터와 계산된 지표를 바탕으로 데이터 마트를 구축함.
    - 데이터 마트는 분석 및 시각화를 용이하게 하기 위해 최적화된 스키마로 설계됨.
    - 생성된 데이터 마트를 새로운 GCS bucket에 적재하여 후속 분석에 활용 가능하도록 함.
4. **BigQuery 테이블 적재 및 Looker Studio 연동**
    - 데이터 마트를 BigQuery 테이블로 변환 및 적재함.
    - Looker Studio와 연동하여 대시보드에서 실시간 데이터 분석 및 시각화를 지원함.

3. 데이터 전처리 및 지표 계산

3.1 DAU, MAU, WAU

사용된 데이터는 accounts_attendance(출석 테이블), accounts_blockrecord(차단 기록 테이블), accounts_friendrequest(친구 요청 테이블), accounts_paymenthistory(결제 내역 테이블), accounts_pointhistory(포인트 내역 테이블), accounts_timelinereport(유저 신고 기록 테이블), accounts_user(유저 테이블), accounts_userquestionrecord(질문 테이블), polls_questionreport(질문에 대한 신고 목록 테이블)이다. LIKE 어플리케이션에는 명확한 '접속 기록' 테이블이 존재하지 않기 때문에, 유저의 앱 내 활동을 대표할 수 있는 운영 데이터를 통합하여 유저 접속 및 활동을 간접적으로 추정하였다.

각 테이블의 날짜 데이터를 일관된 datetime 형식으로 변환하고, 중복값 및 결측값을 제거하였다. 특히 accounts_user 테이블의 경우 성별 정보가 없는 유저는 분석의 정확성을 위해 제외하였다. 이후 user_id와 created_at을 기준으로 필요한 컬럼만 추출하여 하나의 통합 테이블로 결합하였다.

통합 테이블에서 created_at 기준으로 일자, 주차, 월 정보를 생성하여 각각 일간, 주간, 월간 단위로 고유 user_id 수를 집계함으로써 지표를 계산하였다.

3.2 리텐션

사용된 데이터는 위의 9개 테이블에 더해, 사전 클러스터링된 결과를 포함한 cluster_df.csv를 포함하였다. 각 테이블은 유저의 출석, 친구 요청, 결제, 포인트 활동, 신고, 생성 등 다양한 앱 내 활동을 반영하는 트랜잭션 데이터를 담고 있다. 다양한 활동을 포괄함으로써 클러스터별 유저들의 이탈 시점과 행동 흐름을 정밀하게 추적하고자 했다.

전처리 과정에서는 모든 테이블의 날짜 컬럼을 datetime 형식으로 통일하고, 중복값과 결측값을 제거하였다. accounts_attendance 테이블은 리스트 형태의 날짜 데이터를 개별 날짜로 분해하였으며, 각 테이블은 user_id와 created_at 컬럼을 기준으로 정리되었다.

통합 테이블에서 유저별 첫 활동일 대비 경과일을 계산하고, 이를 0일, 3일, 7일, 14일, 30일, 90일, 180일 구간으로 구분하였다. 이후 클러스터별로 구간별 고유 유저 수를 집계하고, 첫날 대비 유지 비율을 계산하여 월별 가입자 수를 3days, 7days, 14days, 30days, 90days, 180days 지표로 구성된 리텐션 테이블을 산출하였다.

3.3 주별/월별 판매량

사용된 원본 데이터는 accounts_paymenthistory.csv이며, 유저 ID, 결제 상품 ID, 결제 시각 등의 정보를 포함한다. LIKE의 수익 구조가 heart 상품 구매에 기반하고 있기 때문에, 판매 시점에 따른 heart 총량을 집계함으로써 유료화 흐름을 추적하고자 하였다.

먼저 created_at 컬럼을 datetime 형식으로 변환하였고, user_id, productId, created_at 중 결측치가 있는 행은 제거하였다. 이후 created_at을 기준으로 주간, 월간 단위의 컬럼을 추가하였다. productId에서 heart 수치를 정규표현식으로 추출하여 heart_count 컬럼을 생성한 뒤, 주간 및 월간 총 결제 수량을 산출하였다.

3.4 투표 수

사용된 데이터는 accounts_userquestionrecord.csv이며, 각 행은 유저 ID와 투표 시각을 포함한다. LIKE 어플리케이션의 핵심 기능이 퀴즈 형식의 투표를 기반으로 하기 때문에, 투표 건수와 참여 유저 수는 유저의 실제 사용 및 반응도를 반영하는 가장 직접적인 지표로 활용된다.

전처리 과정에서는 created_at을 datetime 형식으로 통일한 뒤, 특정 시점 이후의 데이터만 필터링하고, 유효한 투표 데이터만 남도록 결측치를 제거하였다. 그 다음, 일자 단위로 고유 유저 수와 전체 투표 건수를 각각 집계하여 voter_count, vote_count 지표를 생성하였다.

3.5 클러스터링 & PPF

accounts_user, accounts_friendrequest, accounts_pointhistory, accounts_paymenthistory, accounts_userquestionrecord, accounts_blockrecord, accounts_timelinereport, accounts_attendance, polls_questionreport 등 유저 활동 로그 기반의 9개 테이블을 사용하였으며, 클러스터 ID가 포함된 cluster_df.csv를 병합하여 활용하였다. 유저의 활동성, 사회적 관계 형성 정도, 부정적 피드백 경험(신고, 차단 등)을 종합적으로 반영하기 위해 total_votes_count, purchase_count, friend_count, pending_votes_count, friend_request_count, block_count, report_count, pending_chat 변수를 선정하였다.

전체 데이터는 user_id와 created_at을 기준으로 통합하였고, 결측값과 중복값을 제거하였다. created_at은 일관된 datetime 포맷으로 통일하였으며, 왜도가 높은 변수는 로그 변환 및 Box-Cox 변환을 통해 정규화하였다. 이후 PCA를 활용하여 차원을 2개로 축소하고, K-Means 클러스터링을 통해 유저 그룹을 분류하였다. 최적 군집 수는 Elbow Method와 실루엣 계수를 기반으로 결정하였으며, Random Forest를 통해 변수 중요도를 계산한 결과, pending_votes_count, purchase_count, friend_count가 핵심 변수로 도출되었다.

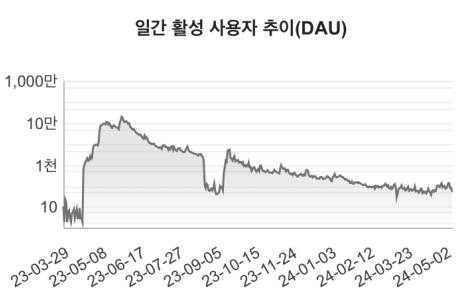
전처리를 거친 데이터는 분석 목적에 따라 각각 클러스터링과 PPF로 분리하여 사용되었다. 클러스터링은 유저의 리텐션 및 활동 패턴을 바탕으로 군집을 분류하고 각 그룹별 특성과 리텐션율을 파악하기 위함이다. 반면 PPF 분석은 플랫폼 기여도를 측정할 수 있는 핵심 지표를 바탕으로 개별 유저 단위의 점수를 산출하여 유저의 가치 및 우선순위 판단에 활용하기 위한 목적이었다. 이를 통해 유저를 집단 수준과 개인 수준에서 입체적으로 분석하였다.

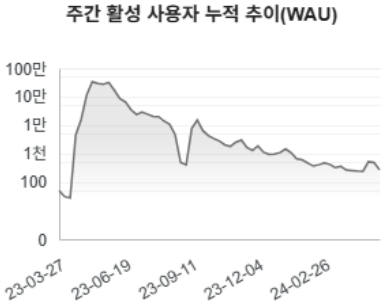
4. 대시보드 항목

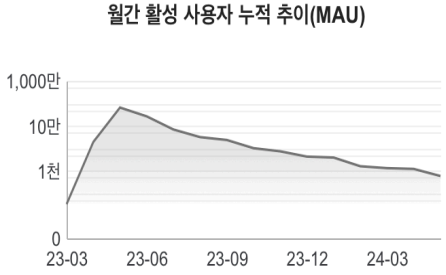
4.1 운영 대시보드

4.1.1 DAU / WAU / MAU 지표 트래킹

Daily, Weekly, Monthly를 트래킹하며 유저 활동 추이를 볼 수 있게 한다. 활성 유저수를 집계하여 어느 시점에 활성유저들이 급감하고 급증하는지 파악하는 데에 용이하다. 접속자 수가 몰려 서버가 다운되는 위험을 예측하는 방향으로도 사용이 가능하다.







4.1.2 가입 월별 리텐션 차트

서비스 운영 및 사용자 유지 전략을 최적화하기 위해, 가입 월별 리텐션 차트를 실시간 대시보드에 반영할 필요가 있다. 이를 통해 다음과 같은 이점을 얻을 수 있다:

월별 코호트 리텐션 차트

구분 ▲	가입자 수	3days	7days	14days	30days	90days	180days
2023-03	32	0	0	0	40.6	53.1	21.9
2023-04	19,060	50.5	34.7	30.6	31.8	31.8	6.0
2023-05	635,505	56.0	40.0	36.0	40.5	20.0	5.4
2023-06	16,737	50.0	31.3	24.2	20.8	14.7	6.7
2023-07	1,849	37.2	23.4	18.1	16.9	20.0	6.8
2023-08	524	12.8	6.1	6.5	6.5	11.3	4.0
2023-09	604	22.4	11.3	10.3	8.3	4.8	1.7

1 - 15 / 15

- **신규 유저 정착률 모니터링**: 특정 월에 가입한 유저들의 유지율 변화를 추적할 수 있음.
- **마케팅 및 운영 전략 최적화**: 특정 프로모션, 이벤트, 광고 캠페인이 리텐션에 미치는 영향을 실시간으로 확인하고, 성공적인 전략을 반복 적용 가능.
- **기능 업데이트 및 서비스 변화의 영향 분석**: 신규 기능 도입 또는 UX 개편이 리텐션에 미치는 영향을 파악하고, 사용자 만족도를 높이기 위한 개선점을 도출 가능.

- **문제 발생 시 신속한 대응**: 특정 기간의 리텐션 하락을 감지하여 원인을 분석하고, 즉각적인 조치를 취할 수 있음.
- **장기적인 성장 전략 수립**: 리텐션 데이터를 바탕으로 유저 유지율을 높이기 위한 지속적인 개선 방향을 설정할 수 있음.

4.1.3 주별 및 월별 판매량 변화

유료 결제를 진행 한 유저 수 및 판매 금액을 주 단위, 월 단위로 분석한다. 실질적 매출을 쉽게 파악할 수 있다. 어떤 제품(heart 종류)이 많이 판매되었는지, 판매량이 느는 시기는 언제인지, 더 나아가 데이터 정합성이 보완된다면 캠페인/이벤트의 실행 시점과 결제율 간의 연관성을 파악할 수 있을 것으로 예상된다.

월간 수익 현황

	구분	월간 수입액
1.	2023-05	66,182,897
2.	2023-06	5,971,828
3.	2023-07	1,598,704
4.	2023-08	1,028,353
5	2023-09	382,077

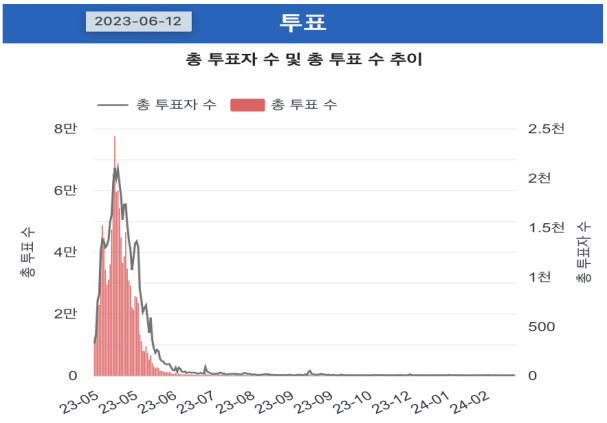
1 - 13 / 13

주간 수익 현황

	구분	주간 수입액
1.	2023-05-08	10,931,563
2.	2023-05-15	29,633,917
3.	2023-05-22	20,936,839
4.	2023-05-29	7,727,291
5.	2023-06-05	1,961,695
6	2023-06-12	569,444

1 - 53 / 53

4.1.4 투표 수 트래킹



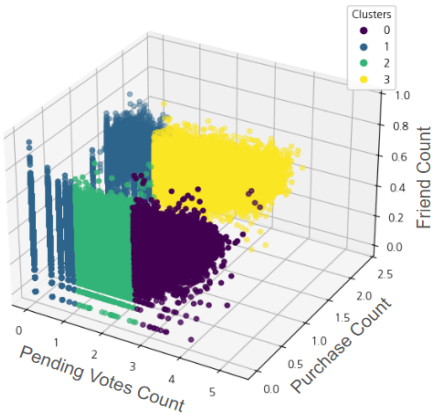
일별 투표 수, 질문별 응답 수 등을 시각화하여, 유저의 실제 반응성과 참여 밀도를 측정할 수 있다. 투표는 LIKE 어플리케이션의 핵심 작동 구조로, 이 지표의 변화는 리텐션 및 참여도에 직접적인 영향을 미친다. 따라서 투표수를 지속적으로 모니터링함으로써, 유저 활동 저하 시점을 조기에 파악하고, 플랫폼 구조 개선의 필요성을 진단할 수 있다.

4.2 군집 별 대시보드

4.2.1 클러스터링

유저의 리텐션과 행동 패턴을 기반으로 전체 유저를 4개 군집으로 분류하였다. 클러스터링 결과는 아래와 같은 목적을 달성하는 데 사용된다:

- 유저 구조의 전체적 이해: 전체 유저를 코어 유저, 활발한 유저, 일반 유저, 휴면 유저로 분류함으로써 서비스 이용자군의 분포를 직관적으로 파악할 수 있음.
- 특이 사용자 탐지: 평균과 다른 리텐션, 활동 패턴을 가진 군집을 식별하여 마케팅/서비스 전략상 주의가 필요한 타깃을 선별할 수 있음.
- 군집별 유저 규모 확인: 각 클러스터의 유저 수를 함께 표시하여, 비즈니스상 의미 있는 유저군에 대한 리소스 집중 여부를 판단 가능하게 함.



클러스터	특징	해석
Cluster 0 (보라색)	Pending Votes Count: 주로 2~5 사이 Purchase Count: 낮음 (0~1.5) Friend Count: 낮음 (0~0.6)	친구 수와 구매 횟수도 낮고 대기 중인 투표 수가 많은 비활동적 사용자 그룹
Cluster 1 (파란색)	Pending Votes Count: 0~3 사이 Purchase Count: 0~1.5 사이 Friend Count: 낮거나 중간 수준 (0~0.6)	친구 수는 적당하며, 대기 중인 투표 수가 적고 구매 횟수도 낮은 일반적인 사용자 그룹
Cluster 2 (초록색)	Pending Votes Count: 0~3 사이 Purchase Count: 0~1.5 사이 Friend Count: 중간 이상 (0.5~1.0 이상)	친구 수가 많고, 대기 중인 투표 수도 적절한 수준인 소셜 활동이 활발한 그룹
Cluster 3 (노란색)	Pending Votes Count: 3~5 사이 Purchase Count: 높음 (1.5~2.5) Friend Count: 높음 (0.6~1.0 이상)	친구 수도 많고 구매 횟수도 높은, 가장 활발한 VIP 그룹

4.2.2 PPF

PPF는 Pending Votes, Purchase Count, Friend Count의 세 가지 핵심 지표를 가중 평균하여 산출한 활동성 지표로, 아래와 같은 목적에서 활용된다.

- 유저 기여도 정량화: 각 유저가 플랫폼 내에서 생성한 투표 수, 구매 이력, 친구 관계 형성 등을 반영하여 개별 유저의 플랫폼 기여도를 수치화함.

- 핵심 유저 탐색: 전체 유저 중에서 기여도가 높은 상위 유저군을 추출하고, 이들의 활동 패턴을 분석하여 리텐션/활성화 전략에 활용 가능.

- 클러스터 내 상위 유저 식별: 동일한 군집 내에서도 PPF Score가 높은 유저를 식별함으로써, 보다 정밀한 타깃팅이 가능해짐.
- 리텐션 지표와의 비교 분석: PPF Score가 높은 유저와 실제 재방문율 간의 관계를 파악하여, 유저 가치와 잔존율 간의 상관성을 분석할 수 있음.

5. 활용 기대효과

운영 대시보드는 데이터 정합성이 낮은 상황에서도 핵심 지표를 직관적으로 확인할 수 있도록 돕는다. 반복적인 수작업 보고서 작성 없이 운영자가 유저 현황을 실시간으로 파악할 수 있으며, 앱 사용 지표의 변화에 빠르게 대응할 수 있도록 한다. 또한, 실무 운영에 필요한 인사이트 도출을 자동화함으로써, 장기적인 유저 리텐션 향상 전략 수립에 기반이 되는 도구로 기능할 수 있다.

한편, 본 대시보드는 정제된 지표를 중심으로 구성되었지만, 유저 행동 데이터 전반에 걸쳐 여전히 정합성이 낮고 누락이 발생하는 문제가 존재한다. 따라서 단순 시각화와 트래킹을 넘어서, 근본적으로 로그 수집의 일관성 확보와 데이터 저장 구조의 정비가 병행되어야 하며, 이는 향후 분석 신뢰도를 높이고 실질적인 인사이트 도출로 이어지기 위한 핵심 기반이 된다.