

# 실패한 10대 투표 앱의 화려한 부활

## : 분류모델 기반 이탈 감지 파이프라인 및 대시보드 구축

---

쿼리커리(4팀) | 이수연 신종훈 이동하 조성찬  
카레를 좋아하는 데이터 분석가들  
QueryCurry

# 목차

01

프롤로그

02

분류 모델을 통한  
정착 실패와  
이탈 요인 분석

03

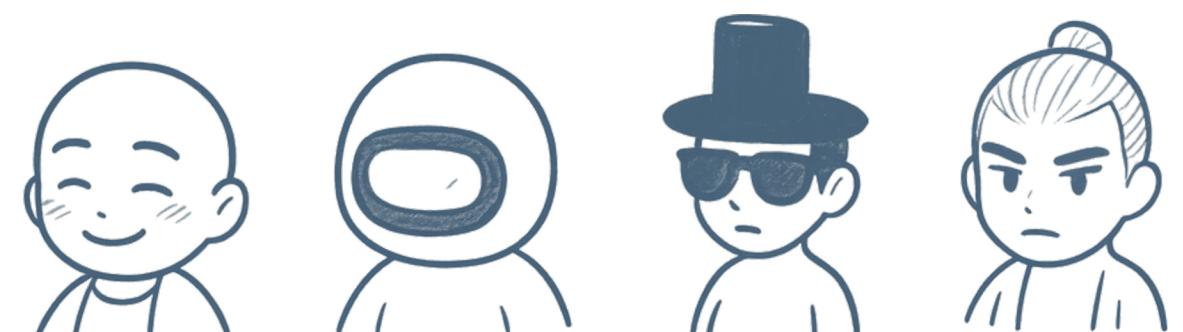
데이터 파이프라인  
재설계

04

새롭게 설계하는  
VOTEE

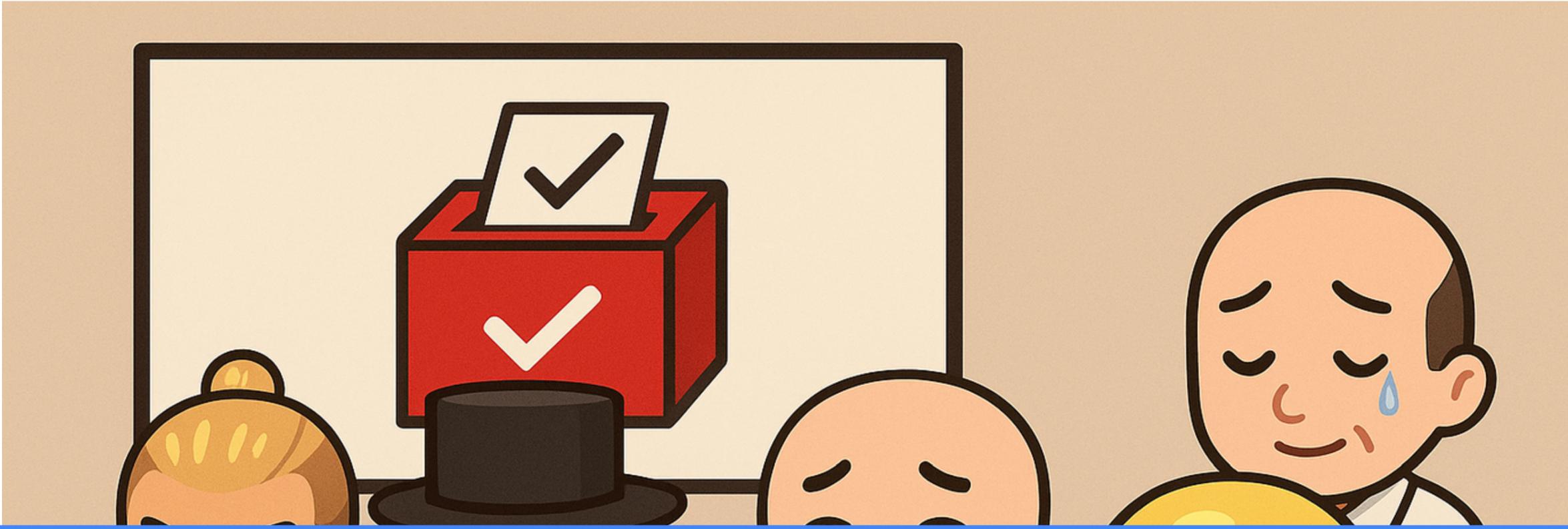
05

결론 및 요약

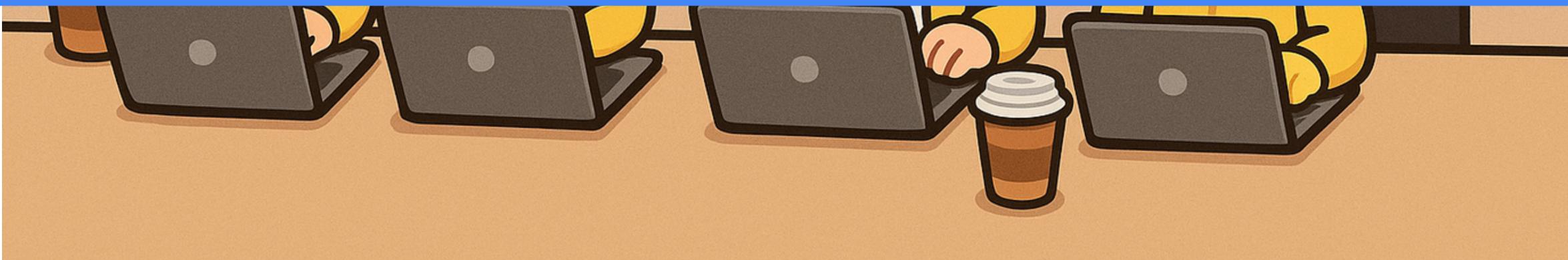


# Part 1. 프롤로그: 쿼리커리에게 주어진 미션





이제부터 팀 쿠리커리의 도전이 시작됩니다!



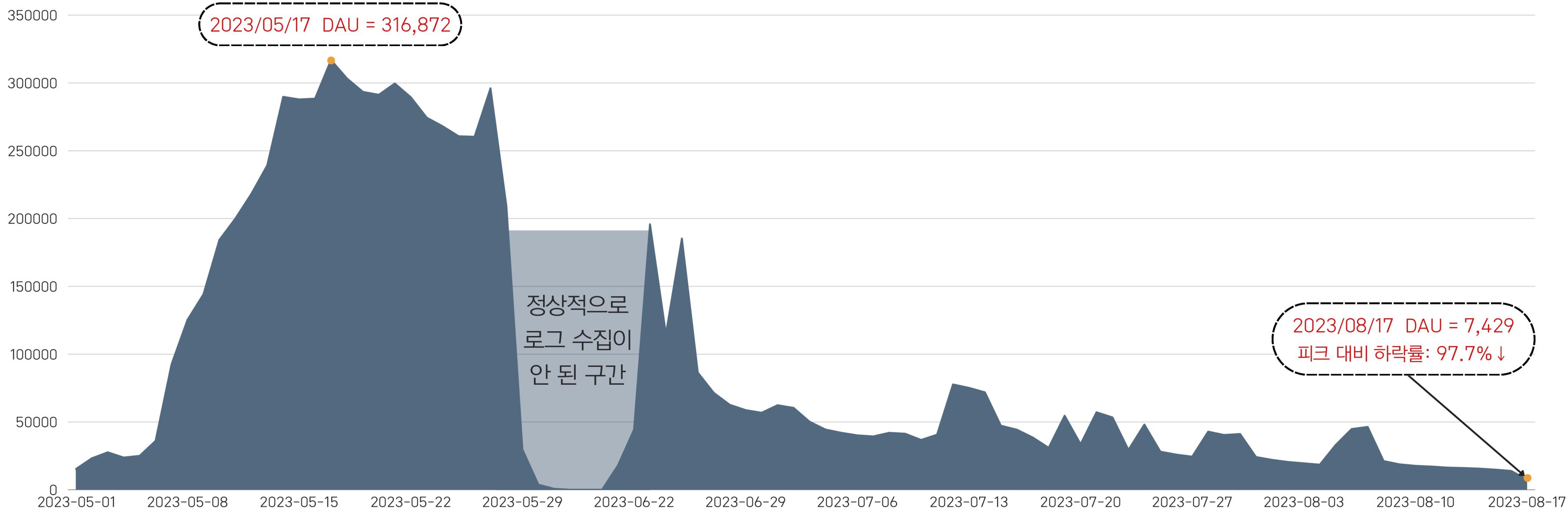
“실패한 서비스지만... 저는 이 서비스에 여전히 마련이 남네요... 😂

여러분에게 미션을 하나 주겠습니다. 이 앱이 남긴 데이터를 분석해 실패의 원인을 찾아주세요  
그리고 그 분석들을 바탕으로, 새로운 서비스를 기획해봅시다!!”

# 01. 프롤로그: 주어진 미션 (1)

## Part 1. 프롤로그

폭발적 시작 → 단 3개월 만에 유저 약 98% 하락



# 01. 프롤로그: 주어진 미션 (2)

Part 1. 프롤로그

과거 서비스  
실패 원인 규명

유저 행동 데이터에서  
인사이트 추출



새로운 서비스  
**VOTEE** 기획

- Vote + Teen → 투표하는 10대

## Part 2. 분류 모델을 통한 정착 실패와 이탈 요인 분석



# 01. 분석 관점 : 두 시점에 집중

## Part 2. 분류 모델을 통한 정착 실패와 이탈 요인 분석

유저 생애주기의 두 순간에 집중

- 단순히 과거의 실패 원인을 확인하는 것만으로는 재도약이 불가능
- 사후적 분석 → 실패 원인의 설명에 그침
- 필요한 것은 사전적 개입 → 실패 징후를 조기에 감지하고 대응하는 체계

정착 실패

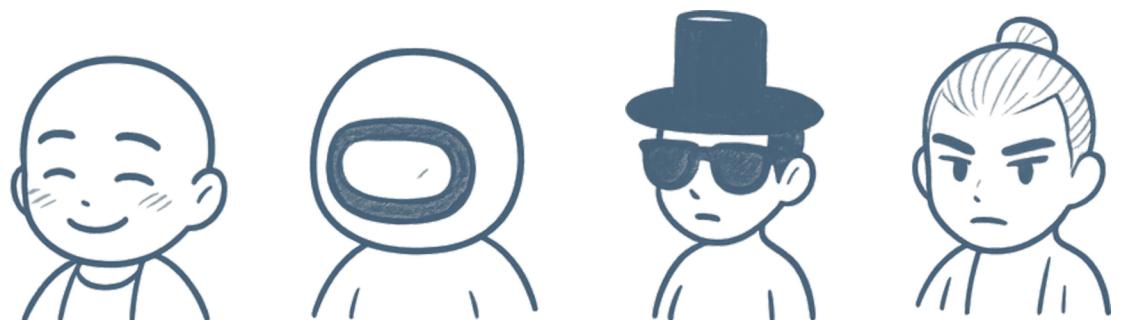
골든타임 24시간

이탈

이탈 예측의 바이탈 사인

EDA - 모델링 - 추가 EDA - 개선안

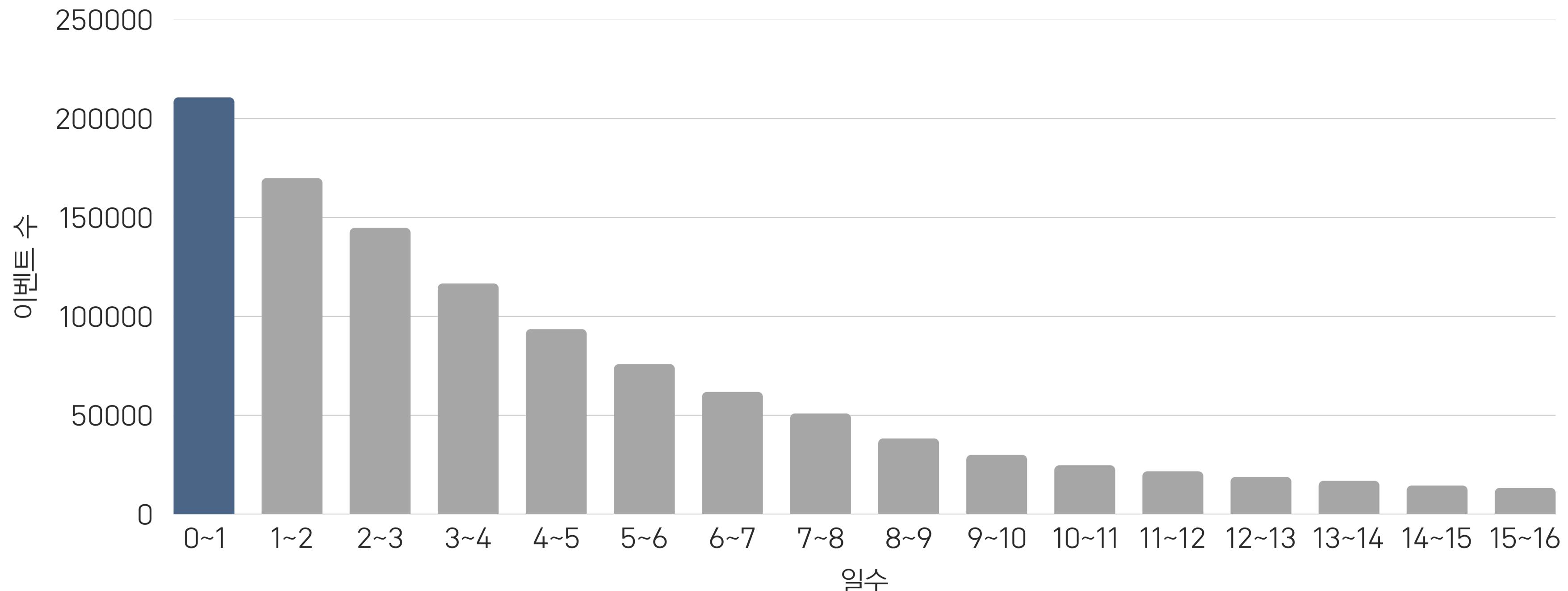
## 2-1. 정착 실패 예측 모델



## 02. 정착 실패: 골든타임 24시간

### Part 2. 분류 모델을 통한 정착 실패와 이탈 요인 분석

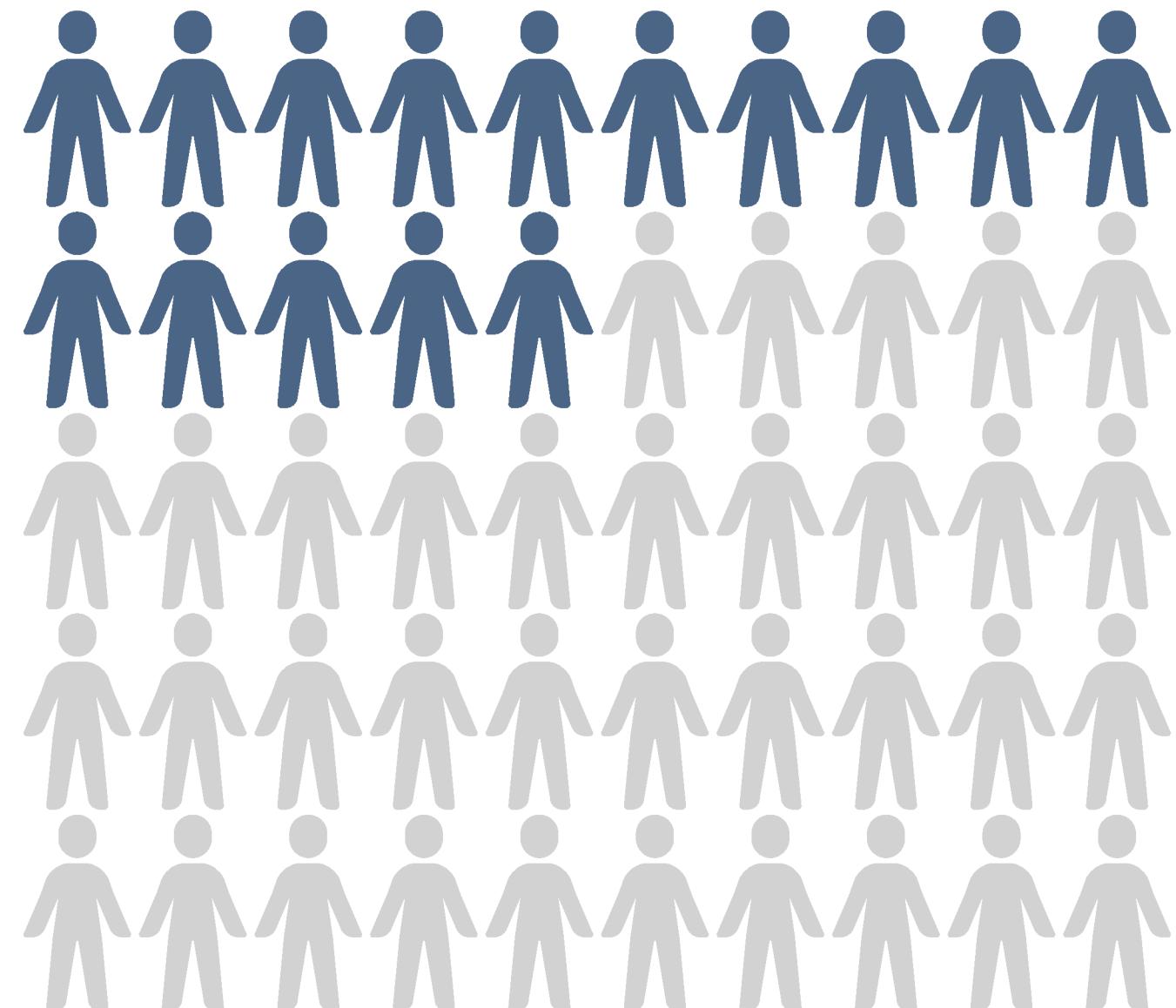
가입 후 **1일** 이내의 활동량이 가장 많았음



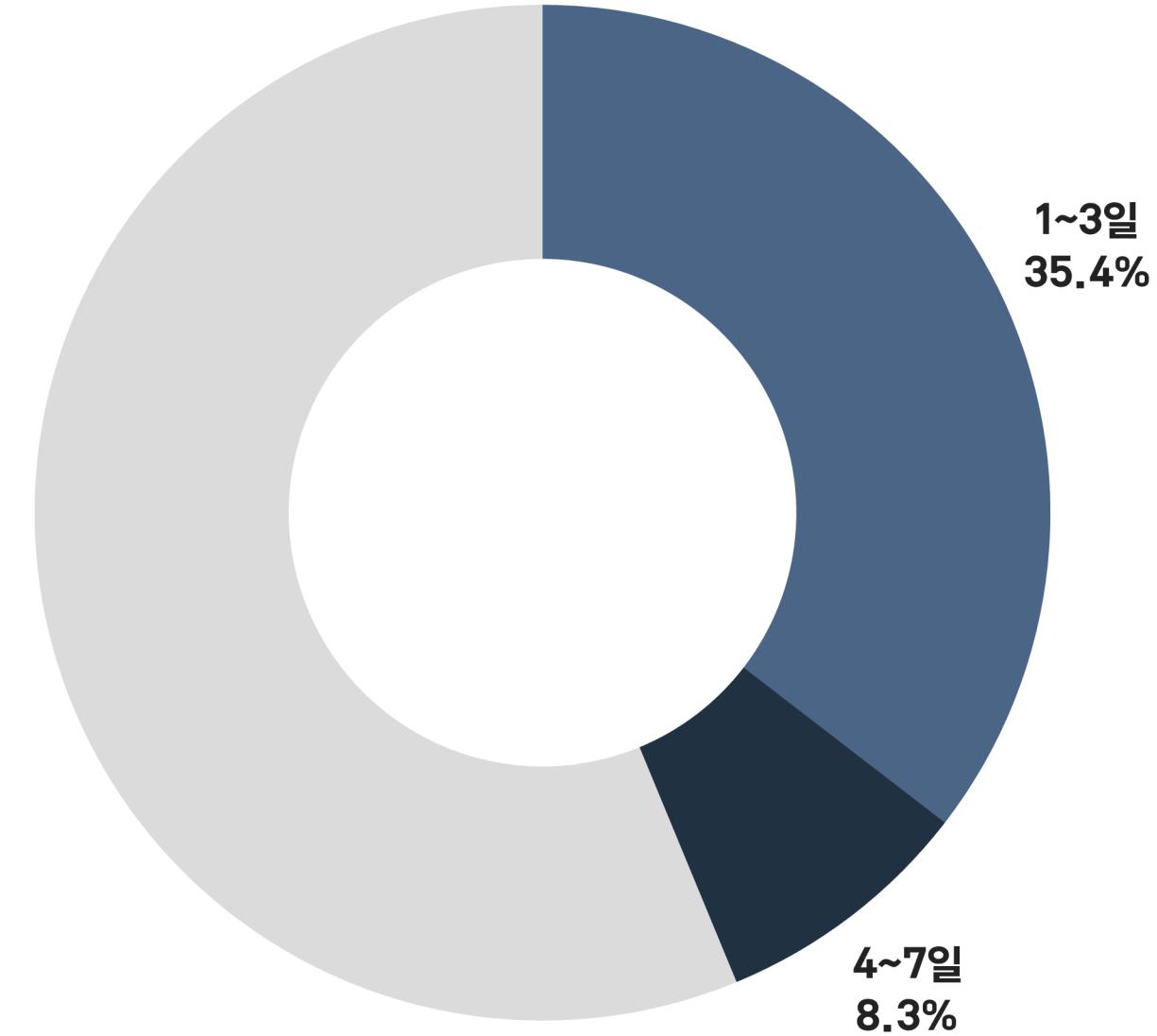
## 02. 정착 실패: 잘못된 첫 만남

### Part 2. 분류 모델을 통한 정착 실패와 이탈 요인 분석

전체 유저의 **약 70%**는  
질문/투표라는 핵심 기능을 단 한번도 경험하지 않았음



대부분의 유저가 가입 후 **첫 주 이내 44%** 사라졌고,  
그 중 35%는 3일 안에 이탈함



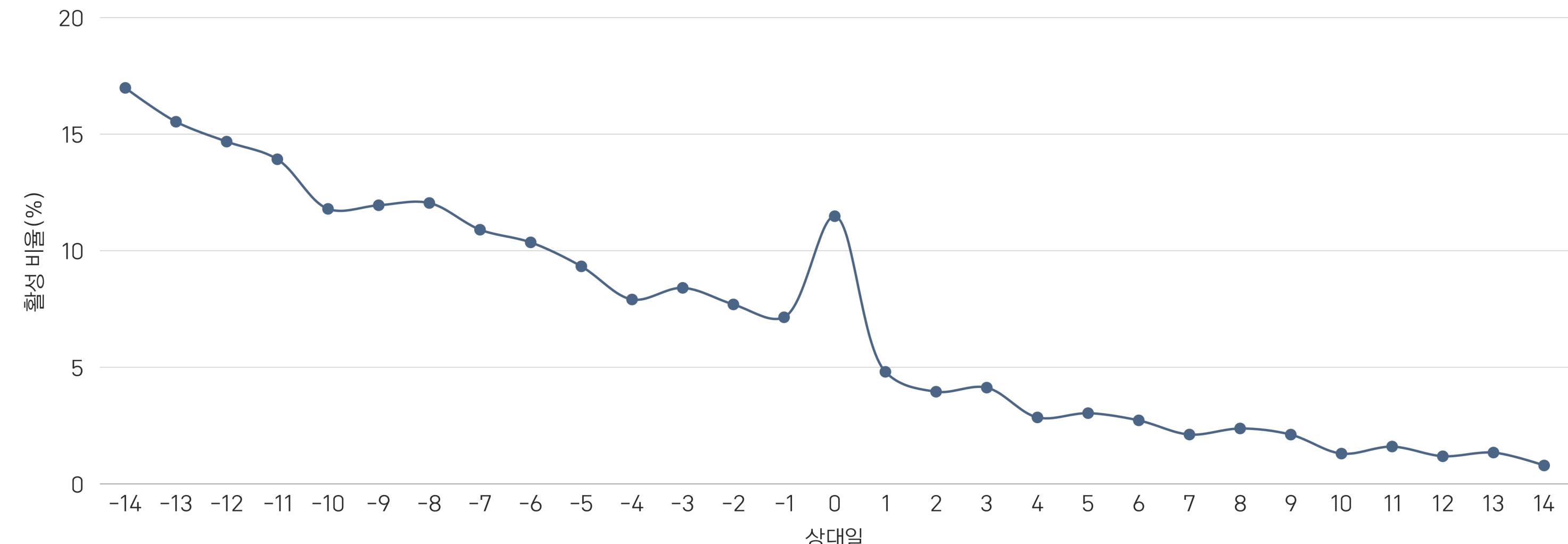
## 02. 정착 실패: 친구 네트워크 분석

### Part 2. 분류 모델을 통한 정착 실패와 이탈 요인 분석

중심 학생이 이탈하면 일반 학생들의 출석과 투표도 급감하며 반 전체의 활동이 무너짐

- **중심 학생 같은 반 활성 비율 ↓**, 투표·포인트 활동 중단 ( $p<0.001$ )
- 평균 15일 후 일반 학생 이탈 본격화
- 친구 요청은 받기 > 보내기 → 인플루언서적 특징

중심 학생 이탈 전후 같은 반 활성 비율

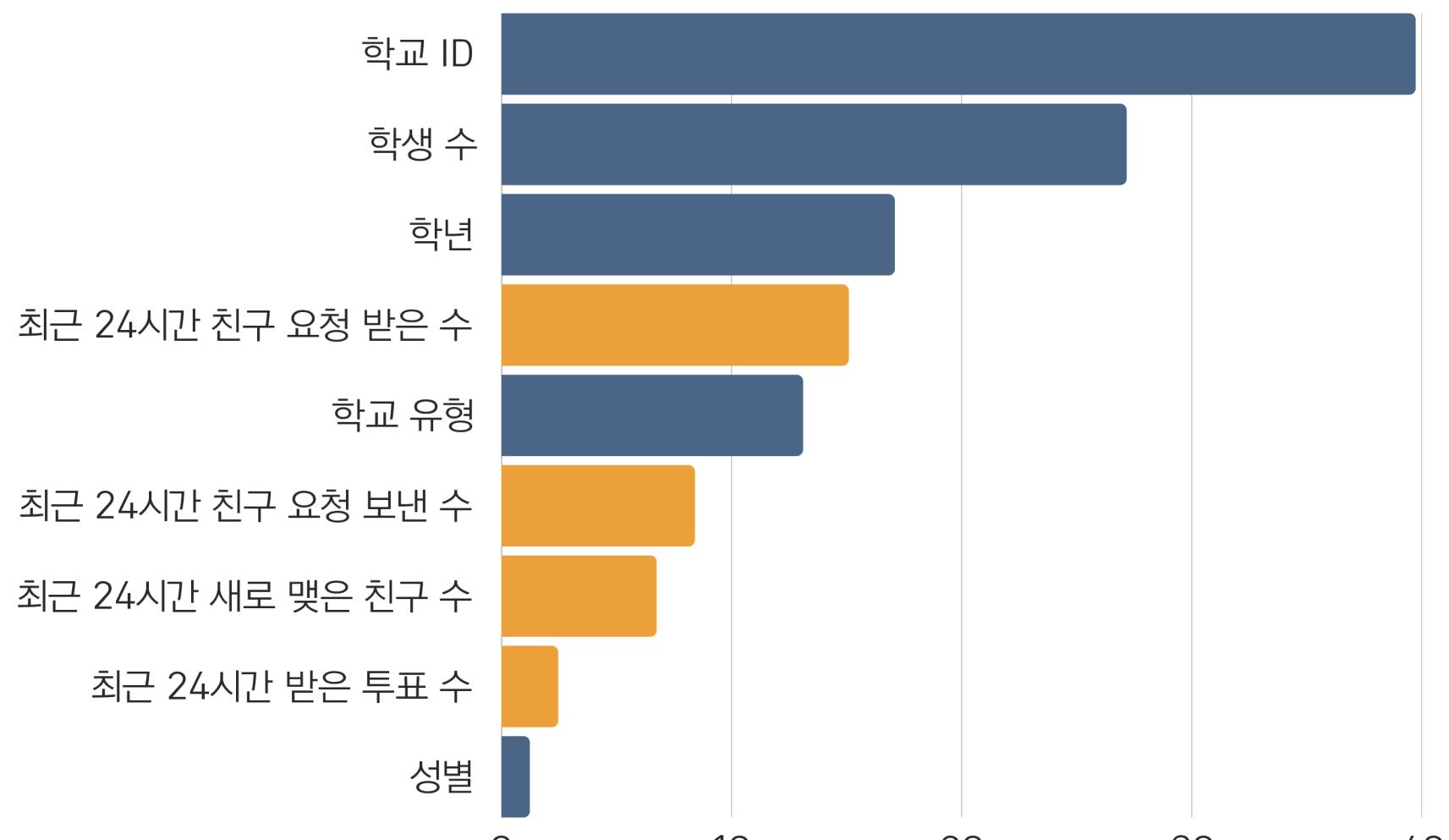


## 02. 정착 실패: 잘못된 첫 만남

### Part 2. 분류 모델을 통한 정착 실패와 이탈 요인 분석

가입 첫 24시간의 로그 데이터 → 2주 뒤 정착/이탈 예측

LGBM 분류 정착 예측 모델 속성 중요도



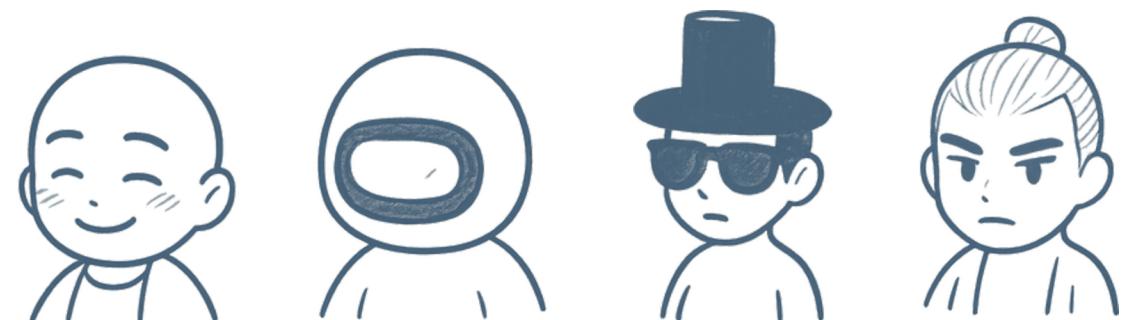
2주차 시점: 정착 vs 이탈 비율 (5:5)

- 정적 정보(성별·학교·학년) + 행동 데이터(투표·친구·포인트·결제)
- 알고리즘: LightGBM (Feature Importance + SHAP 해석)

학교·학년 같은 **태생적 요인 (배경요인)**이 가장 큰 영향  
24시간 내 '받은 친구 요청 수'가 정착을 결정하는 핵심 요인 중 하나  
환경은 바꿀 수 없어도 **초기 상호작용 설계**를 통해 결과를 바꿀 수 있음

Accuracy (0.73) **Recall (0.88)** F1 (0.81) ROC AUC (0.77)

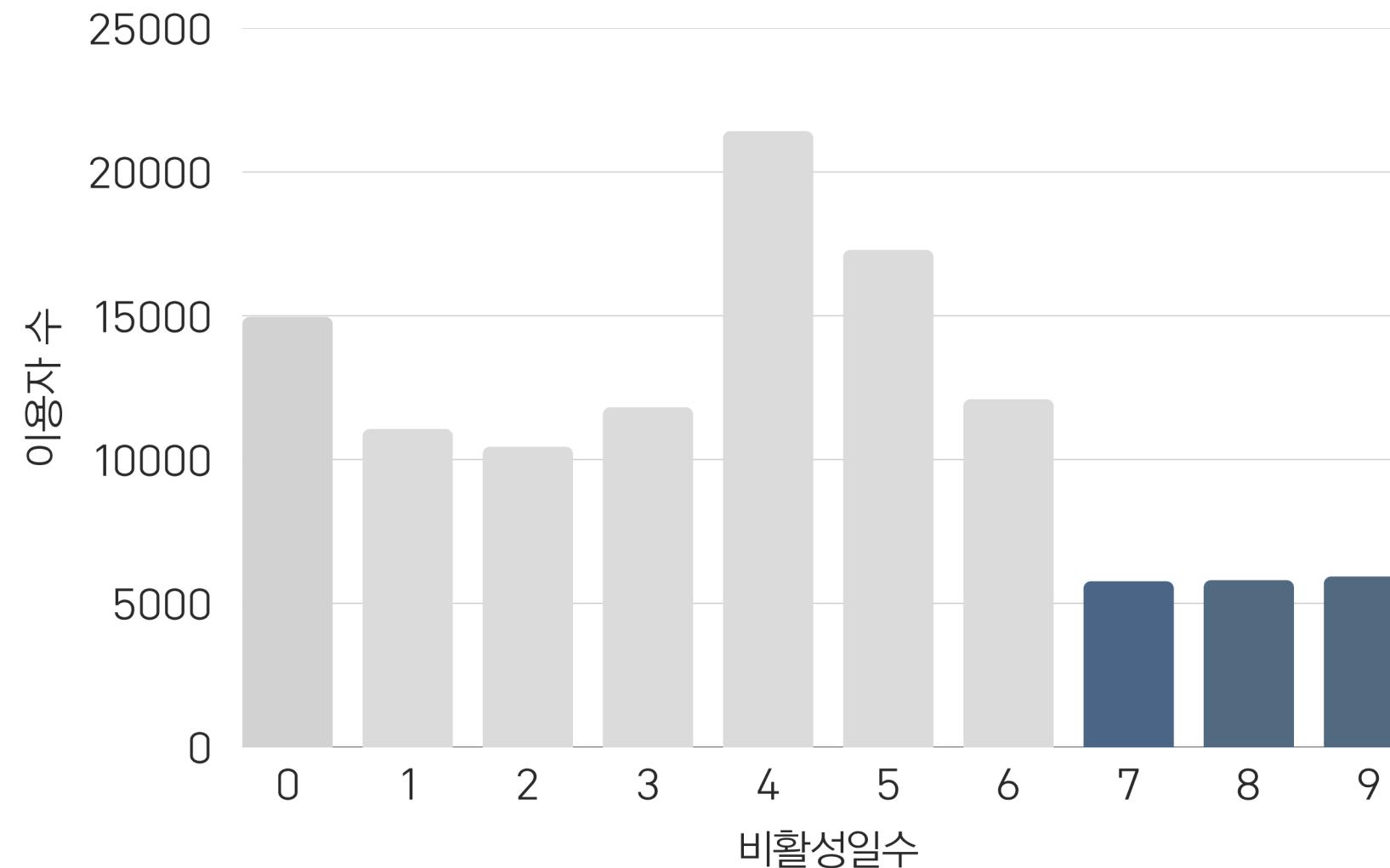
## 2-2. 이탈 예측 모델



# 03. 이탈 예측: 분석 준비

## Part 2. 분류 모델을 통한 정착 실패와 이탈 요인 분석

유저 이탈 패턴 - 분석 개요



마지막 사용 이후 7일이 넘어가는 경우, 복귀 유저의 수가 일정 수준 미만으로 감소

**이탈** : 마지막 접속 이후 **7일 이상** 사용을 하지 않은 경우

### 분석 개요

- **실제 환경**에서의 이탈 예측 필요성 확인
- 특정 시점에서 **활성 유저**의 향후 **이탈 여부 예측**
- 분석의 기준일과 기준일 이전의 **사용 로그** 활용
- 기준일 이후 **7일 이상 미접속 시 '이탈 유저'**로 분류
- 실제 환경과 동일한 조건에서 분류 모델 성능 및 주요 이탈 요인 파악

### 분석 대상

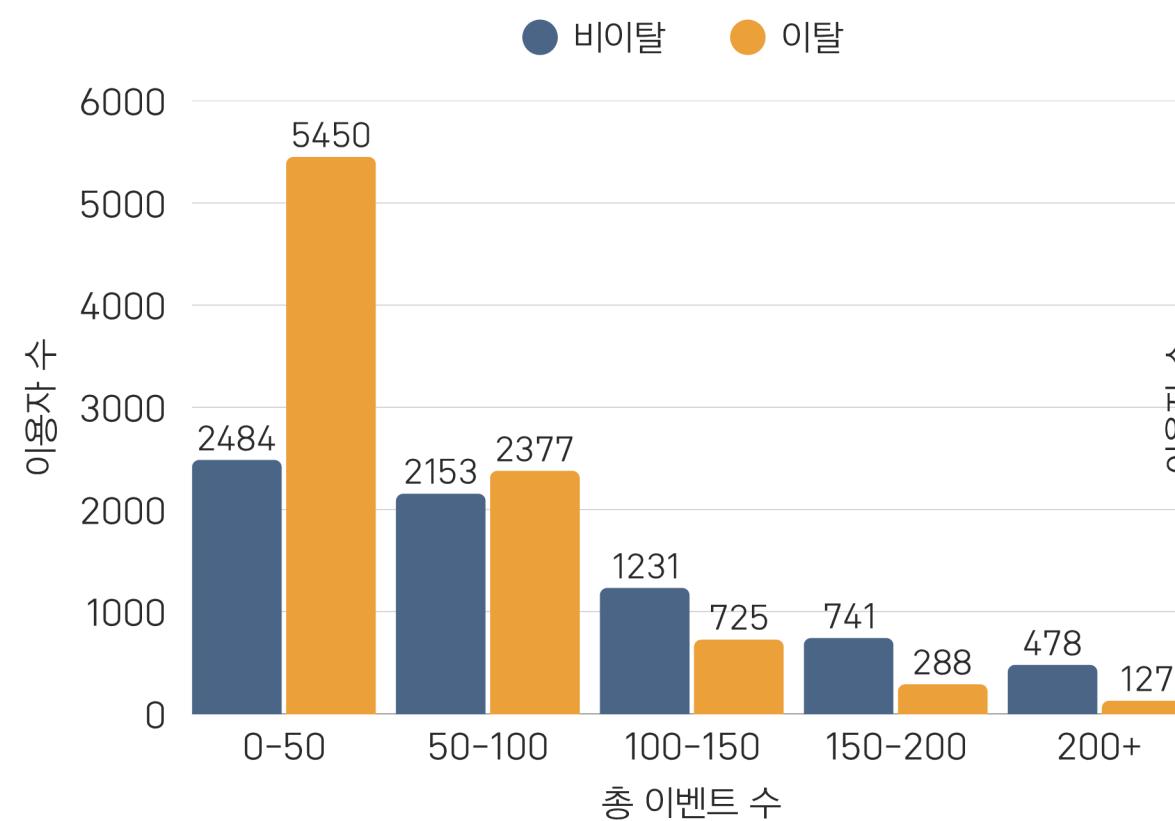
- Hackle 데이터 기준 '**2023-07-29**' 활성 유저
- 총 이벤트 수 5개, 평균 세션 길이 1초 이상
- 최소 **3일** 간 앱을 이용한 유저

# 03. 이탈 예측: 유저 이탈 패턴 (1)

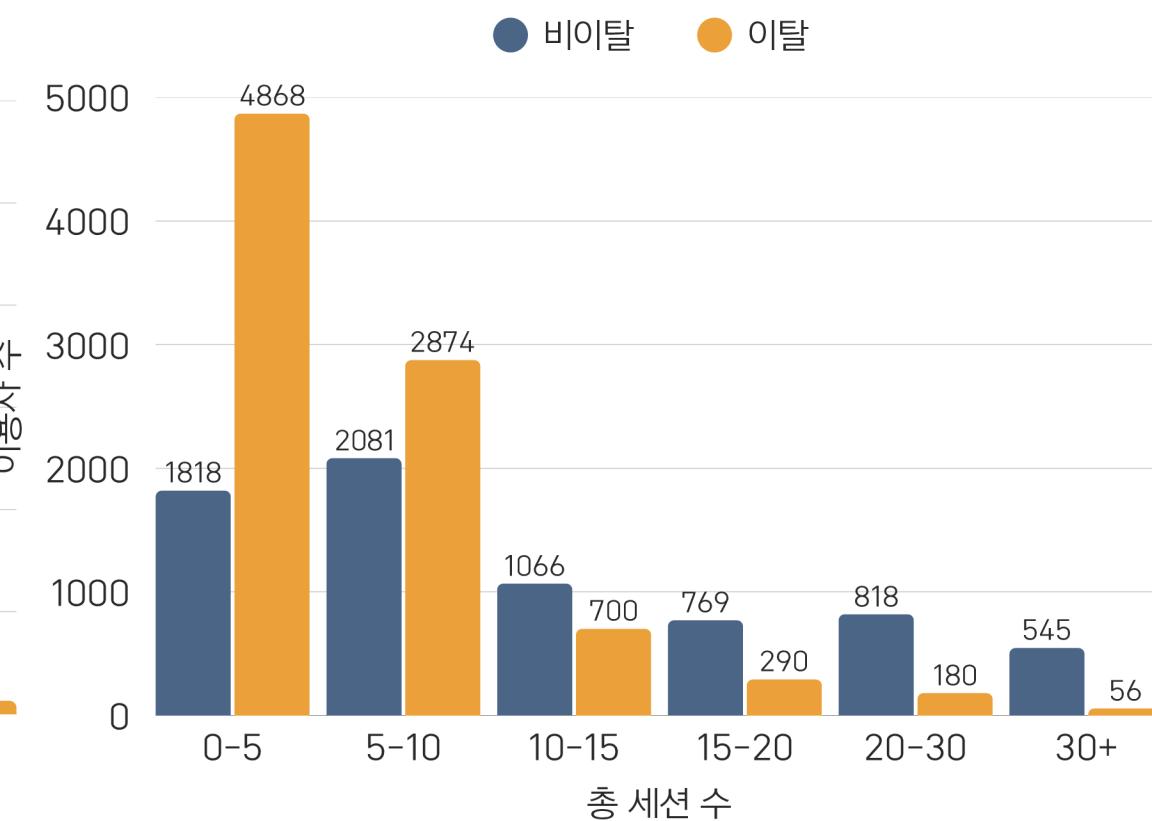
## Part 2. 분류 모델을 통한 정착 실패와 이탈 요인 분석

### 유저 이탈 패턴 - 기본 행동 지표

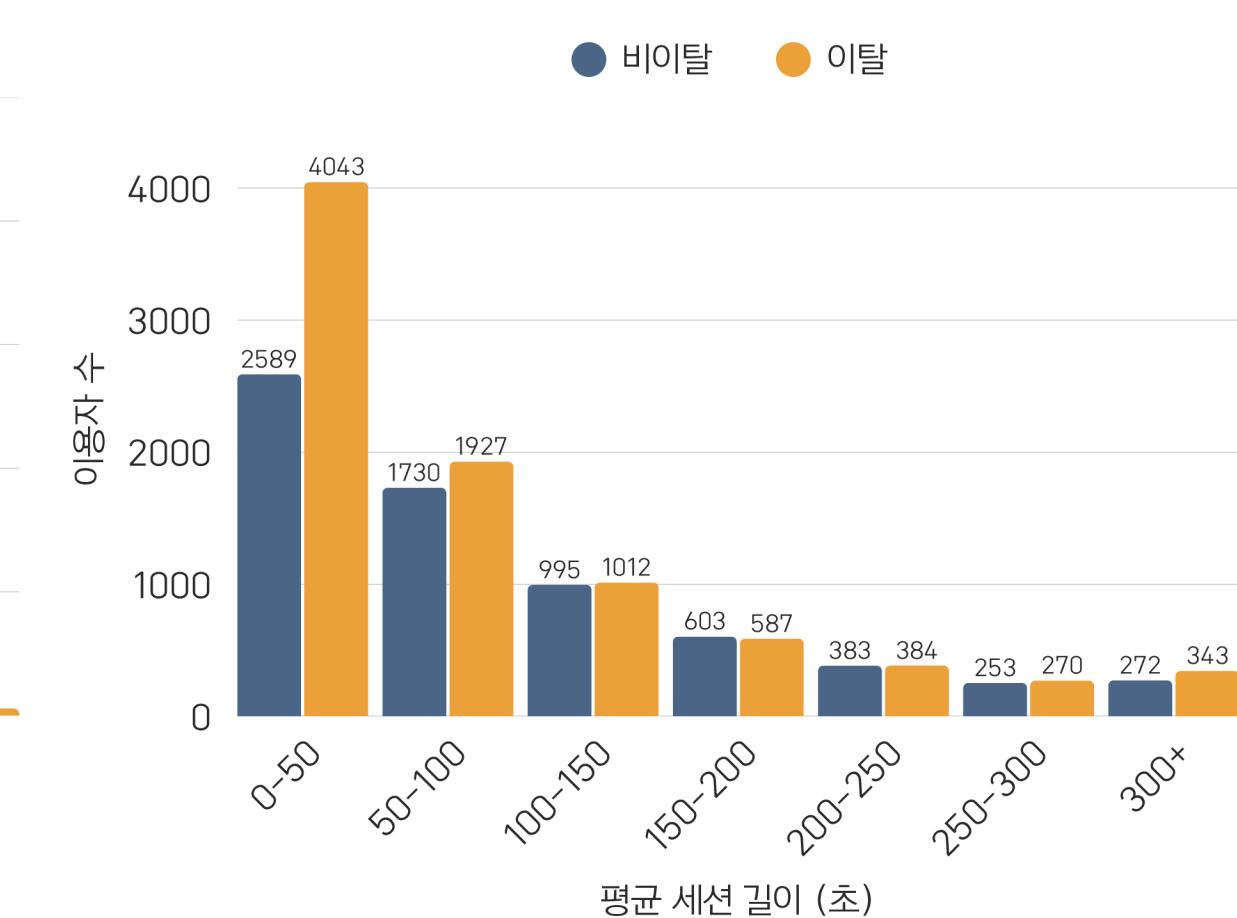
이탈 여부에 따른 총 이벤트 수



이탈 여부에 따른 총 세션 수



이탈 여부에 따른 평균 세션 길이



이탈여부	평균	중앙값	표준편차
비이탈	87.36	72	60.43
이탈	53.91	41	43.05

이탈여부	평균	중앙값	표준편차
비이탈	11.30	9	7.33
이탈	6.73	5	4.53

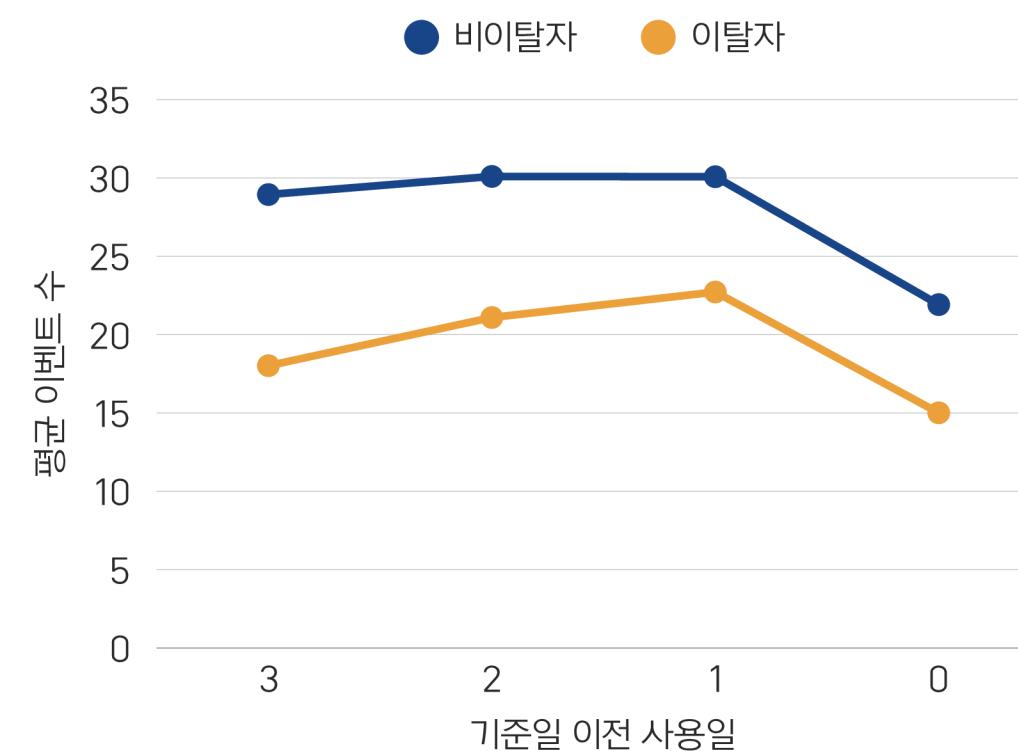
이탈여부	평균	중앙값	표준편차
비이탈	112.35	85	93.52
이탈	87.37	56	84.93

# 03. 이탈 예측: 유저 이탈 패턴 (2)

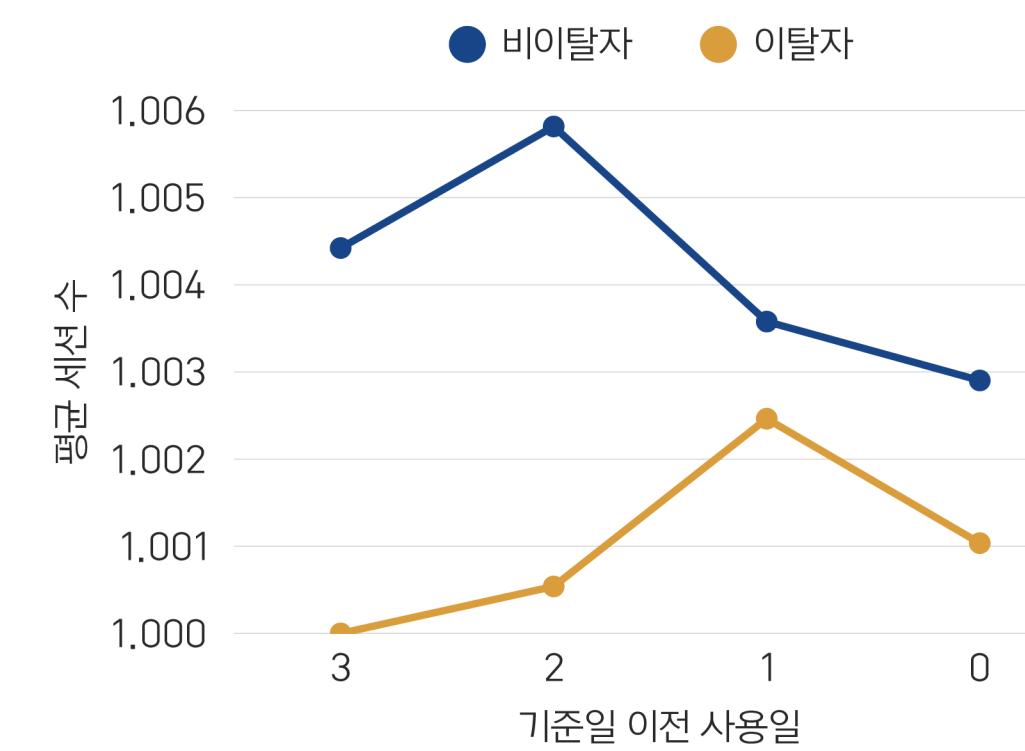
## Part 2. 분류 모델을 통한 정착 실패와 이탈 요인 분석

### 유저 이탈 패턴 - 기본 행동 지표 변화율

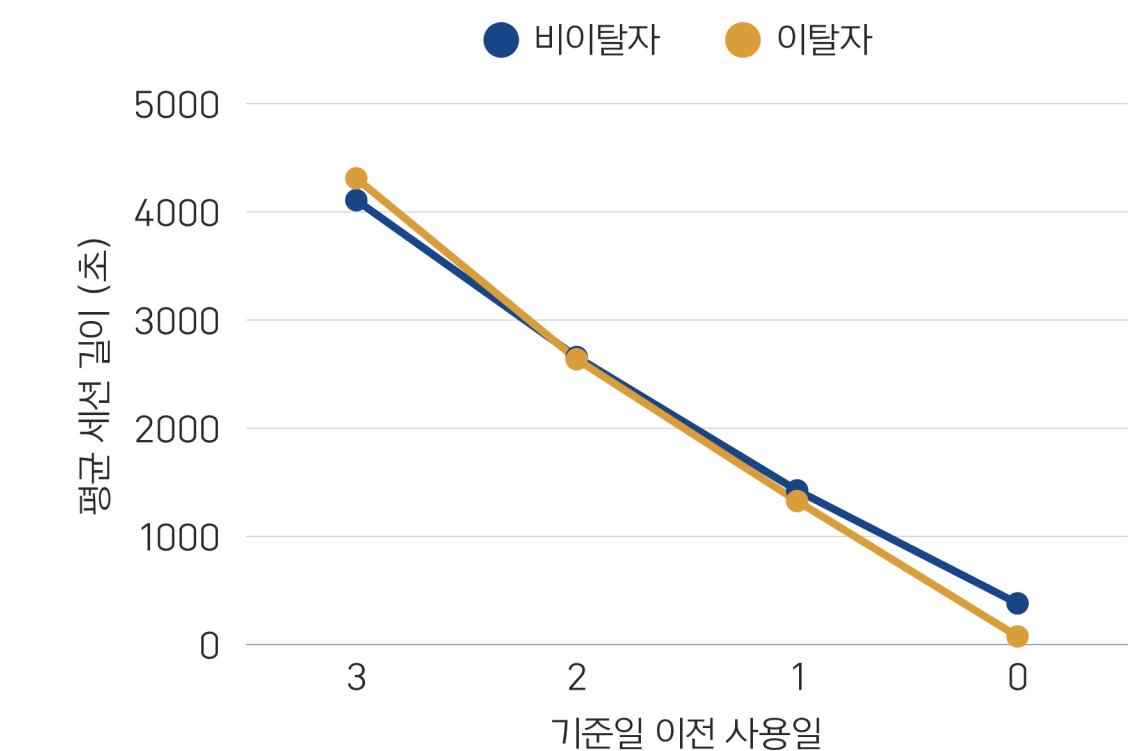
이탈 여부에 따른 이벤트 사용량 변화



이탈 여부에 따른 세션 사용량 변화



이탈 여부에 따른 평균 세션 길이 변화



이탈여부	기준일 - 1일	기준일	이벤트 변화율
비이탈	30.08	22	-24.46%
이탈	22.75	15	-31.51%

이탈여부	기준일 - 1일	기준일	세션 변화율
비이탈	1.00358	1.00290	-0.06%
이탈	1.00246	1.00104	-0.14%

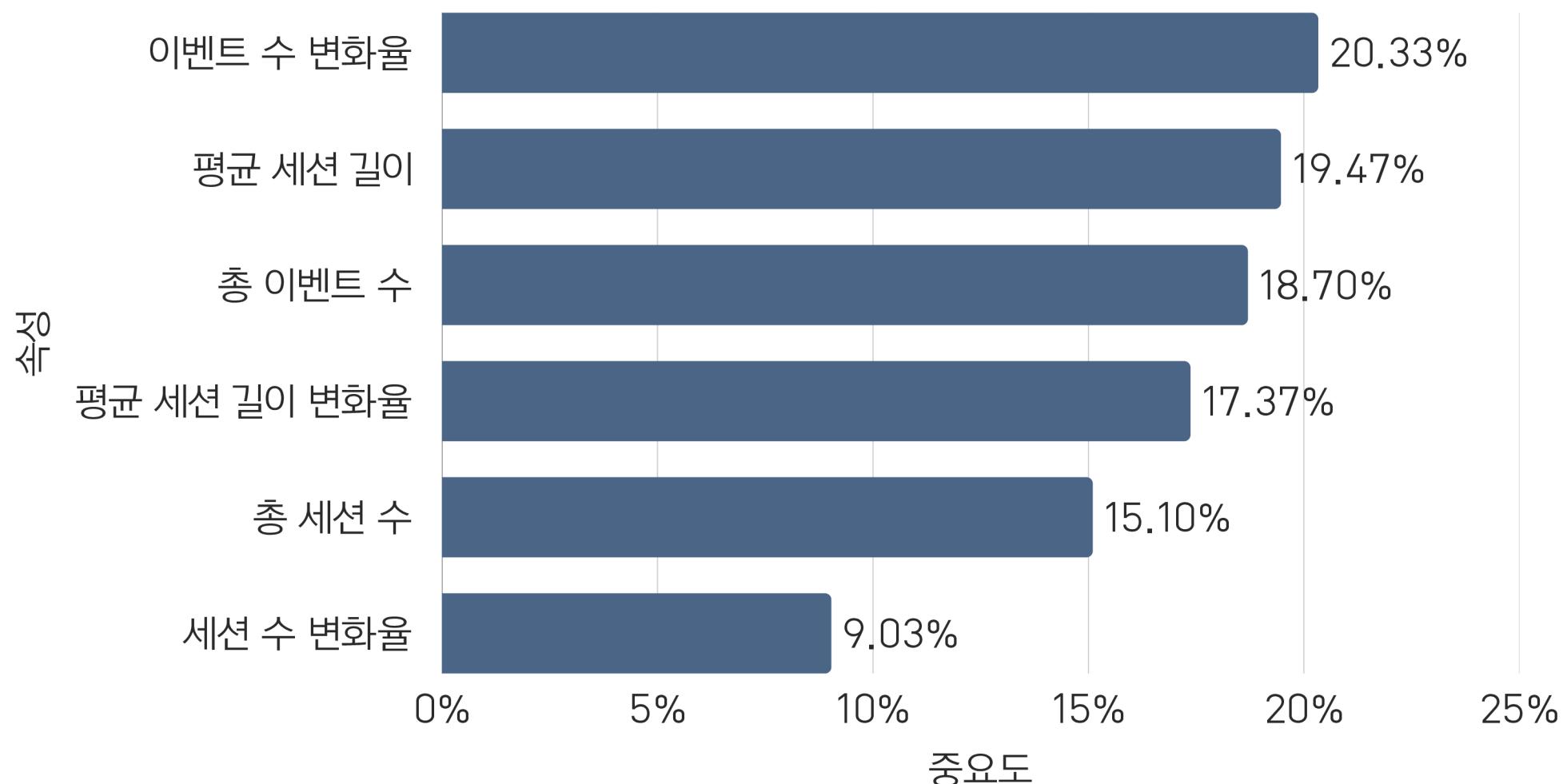
이탈여부	기준일 - 1일	기준일	세션 길이 변화율
비이탈	1424.55	380.81	-70.27%
이탈	1326.71	75.57	-94.31%

# 03. 이탈 예측: 분류 모델링

## Part 2. 분류 모델을 통한 정착 실패와 이탈 요인 분석

이탈 예측 분류 모델 - Light GBM Classifier 채택

LGBM 이탈 예측 분류 모델 속성 중요도



평가 지표	정확도 (Accuracy)	재현율 (Recall)	F1 Score	ROC AUC
Light GBM	0.6920	0.8189	0.7279	0.7448

약 81.9%의 높은 재현율을 가진 이 모델은 잠재 이탈 유저를 더 효과적으로 감지

총 이벤트 수, 평균 세션길이, 총 세션 수와 그 속성들의 변화율을  
**지속적으로 모니터링**하고 **감소 패턴을 감지할 체계** 필요

## Part 3. 데이터 파이프라인 재설계



### 데이터 파이프라인 재설계의 이유

#### 부정확한 세션 데이터

- 유저의 핵심 체류 시간을 제대로 측정할 수 없었음

#### 일관성 없는 이벤트 로그

- 날짜별로 이벤트 이름과 값이 달라 장기적인 행동 추적이 불가능했음

#### 데이터 유실 및 누락

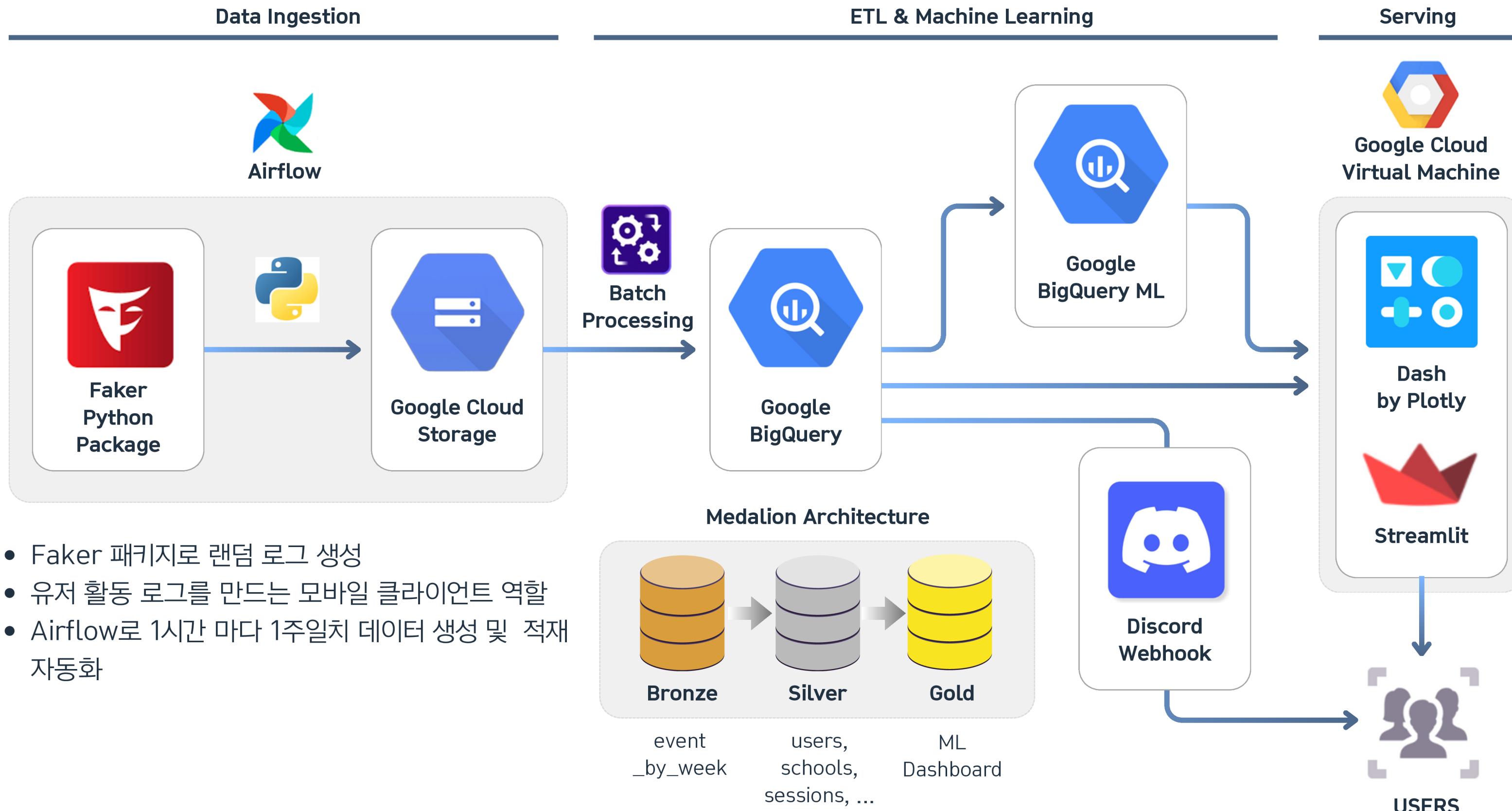
- 비정상적인 로그 수집으로 인해 분석에 필요한 데이터가 소실되었음

#### 통일되지 않은 데이터 구조

- 갑작스러운 컬럼명 변경으로 분석 코드가 깨지고 유지보수가 어려웠음

# 01. 데이터 파이프라인 개요

## Part 3. 데이터 파이프라인 재설계



- Faker 패키지로 랜덤 로그 생성
- 유저 활동 로그를 만드는 모바일 클라이언트 역할
- Airflow로 1시간마다 1주일치 데이터 생성 및 적재 자동화

event\_by\_week  
users, schools, sessions, ...  
ML Dashboard

# 02. Data Ingestion: 랜덤 로그 생성

## Part 3. 데이터 파이프라인 재설계

랜덤 로그를 생성하여 파이프라인 처리 및 대시보드 반영 여부 검증

1. 시뮬레이션 환경 및 초기 유저 생성



2. 일별 유저 활동 및 이탈 시뮬레이션



3. 핵심 기능(질문/투표) 사용 및 결제/포인트 기록

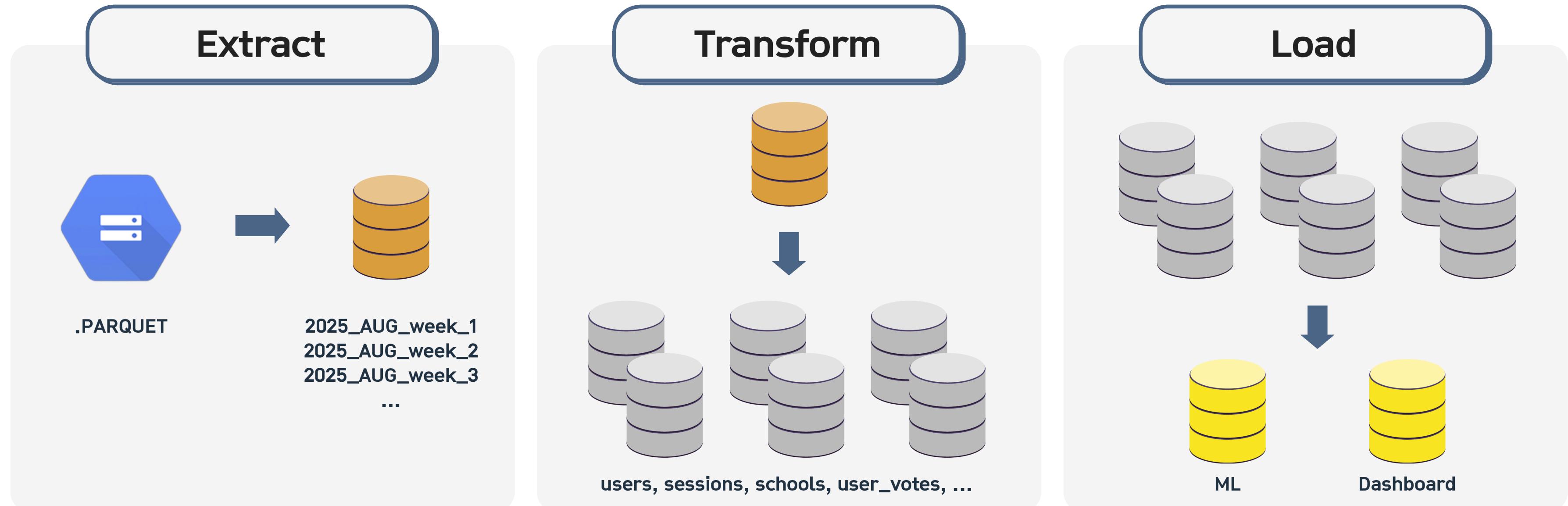


4. Airflow 파이프라인 구축 및 GCS 업로드

user_properties	device_properties
{"status": "active", "grade": 1, "class_num": 1, "has_push_permission": false, "school_name": "가락중학교", "points_balance": 1000}	{"device_id": "a47de979-d5a1-45c2-bfd0-11761a475132", "acquisition_channel": "Referral", "os_type": "iOS"}
event_properties	
{"votee_id": "VOTEABLE_7749", "content_text": "가고 싶은 여행지는?", "option_text": ["바다", "산", "도시", "시골"]}	

# 03. ETL: Medallion Architecture

## Part 3. 데이터 파이프라인 재설계



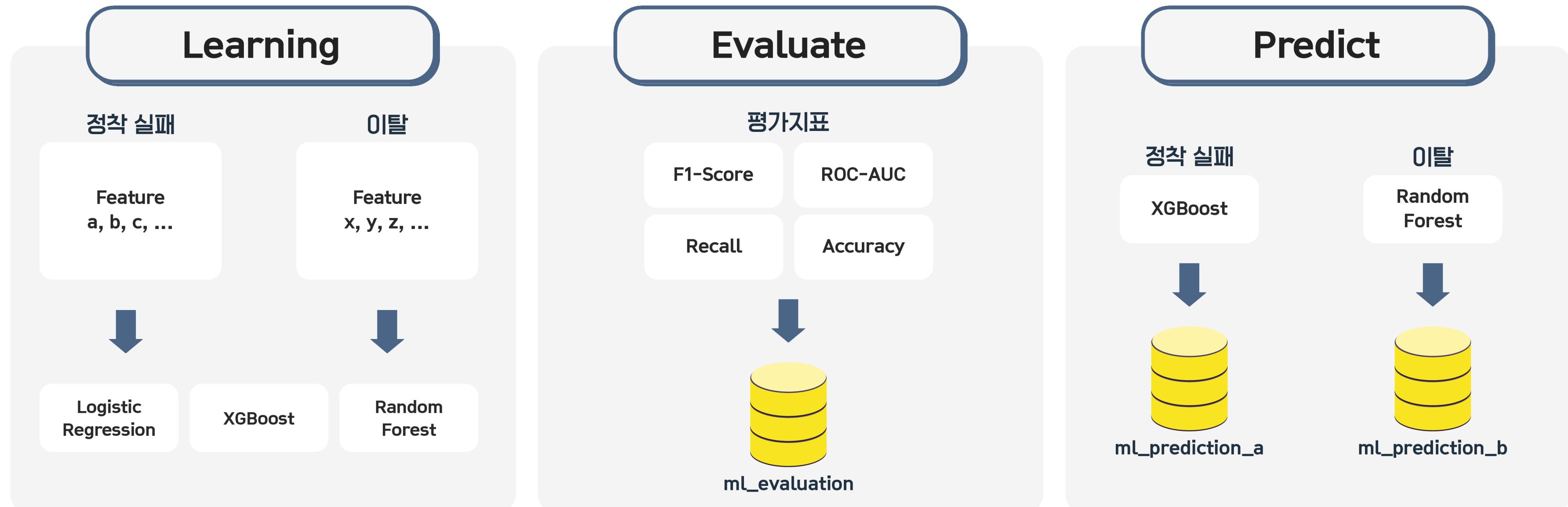
- GCS에 Faker에서 생성된 데이터가 PARQUET 파일로 저장되면 BigQuery에서 인식
- 원시 데이터를 Bronze table에 데이터 기간별로 파티셔닝하여 적재

- Bronze table의 데이터를 각 silver table의 역할에 맞게 자료형을 변환하고 분류하여 적재

- silver table에 적재된 데이터를 ML 학습용과 Dashboard 시각화 용으로 집계하여 Gold table에 적재

# 03. Machine Learning

## Part 3. 데이터 파이프라인 재설계

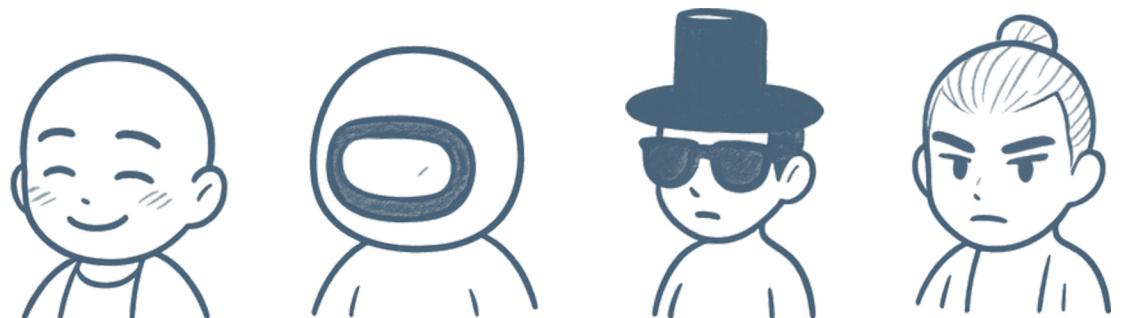


- 데이터 기간 기준으로 주 1회 분류 모델 학습
- 앞선 정착 실패, 이탈 분류 분석에서 도출한 중요  
도 높은 변수 사용
- 가능한 모든 모델 학습

- ML.EVALUATE 쿼리로 모델 평가 지표 도출
- 평가 지표를 Gold table로 저장하여  
Dashboard 시각화 용으로 사용

- 가장 지표가 좋은 모델로 새로운 데이터에 대해  
분류 진행
- 분류 결과를 Gold table로 저장하여  
Dashboard 정착 실패/ 이탈 감지에 사용

## Part 4. 새롭게 설계하는 익명 투표 앱 ‘VOTEE’



# 01. 대시보드: 신규 유저 정착 과정

Part 4. 새롭게 설계하는 익명 투표 앱 'VOTEE'

## 신규 유저 24시간 행동 대시보드

시작일

2025/08/01

종료일

2025/08/19

기간

Daily

차트 유형

Area

## 전체 통계

신규 가입자 수

12,855

↓ -97 (-13.36%)



친구 추가 수

2,143

↓ -18 (-14.52%)



투표 횟수

10,712

↓ -79 (-13.12%)



## 선택한 기간 데이터

신규 가입자 수

6,812

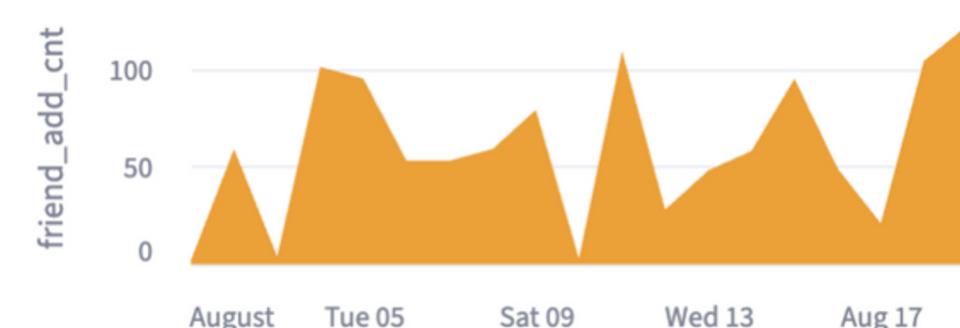
↑ +191 (+32.10%)



친구 추가 수

1,142

↑ +18 (+17.31%)



투표 횟수

5,670

↑ +173 (+35.23%)

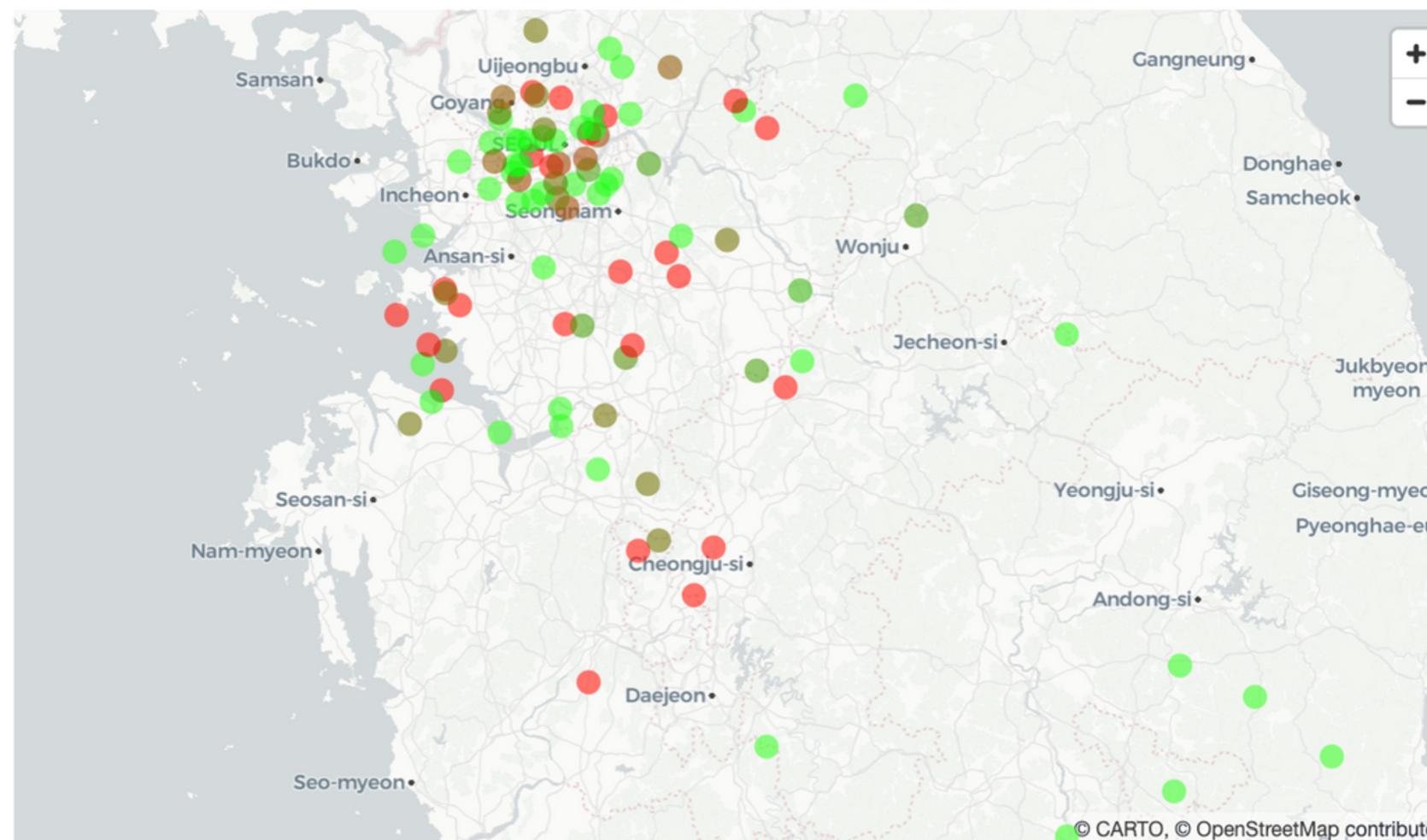


# 01. 대시보드: 신규 유저 정착 과정

## Part 4. 새롭게 설계하는 익명 투표 앱 ‘VOTEE’

## 정착 실패 모니터링 대시보드

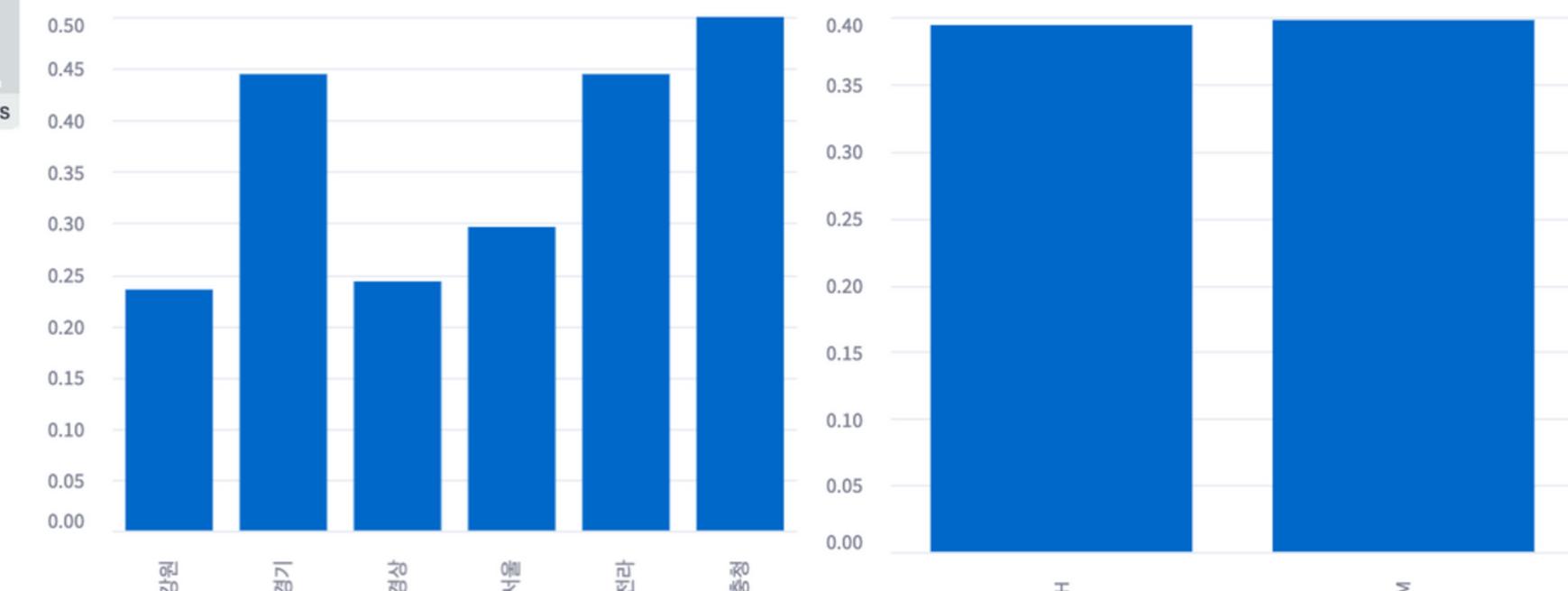
### 학교별 위험 유저 현황



## 일별 이탈 위험 비율



### 지역 및 학교 유형별 이탈 위험 비율



# 02. 대시보드: 이탈 감지

## Part 4. 새롭게 설계하는 익명 투표 앱 'VOTEE'



## 모니터링 기간 설정

시간 프리셋  
● 최근 1시간 ○ 최근 6시간  
○ 최근 24시간 ○ 최근 7일

커스텀 범위

시작 2025. 08. 01. 오전 12:00 종료 2025. 08. 26. 오전 11:20

리프레시(초)

새로고침 리셋

리프레시 간격 (초)

10 40 70 100 130 160 190 220 250 280 300

## 핵심 지표 현황

실시간 머신러닝 기반 이탈 위험  
**13,017(98.2%)**  
고위험 사용자 수 (0.7 이상)

현재 동시 세션 수  
**37(+1)**

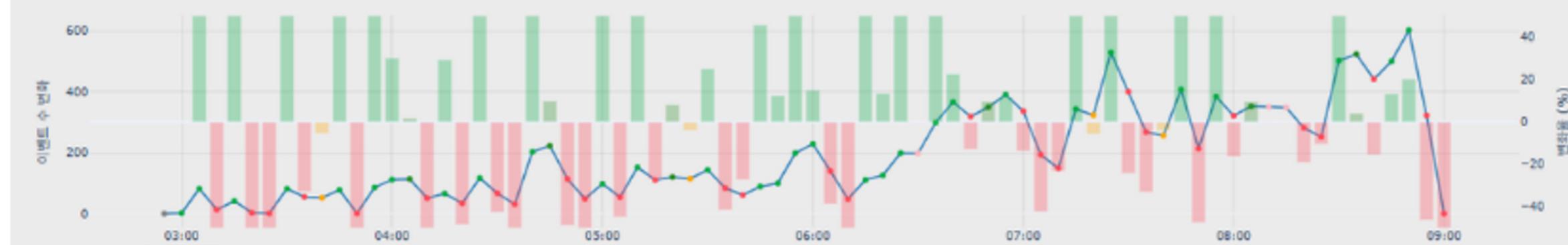
최근 이벤트/분  
**179**

최근 평균 세션 시간(분)  
**5.4**

## 패턴 및 추세 분석

### 이벤트 수 변화 시각화

이벤트 수 변화



# 03. Discord 위험 알림 시스템

## Part 4. 새롭게 설계하는 익명 투표 앱 'VOTEE'



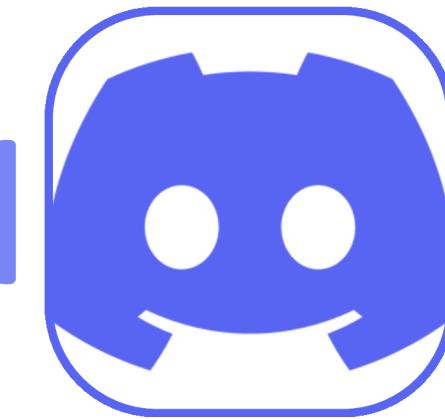
BigQuery

① 데이터 수집 및 분석



Airflow

② 위험 분석 로직 실행



Discord

③ 위험 감지 시 알림

모니터링 봇 앱 25. 8. 19. 오전 9:18

쿼리커리 데이터 모니터링 봇

⚠️ 데이터 검증 실패:

검증 구간: 2025-08-01  
데이터에서 16개의 이슈가 발견되었습니다.

📑 Parquet 스키마 불일치

- Table: events(parquet)
- Details: 추가:[device\_id] / 제거:-

📑 JSON 키 불일치(event\_properties)

- Table: events(parquet)
- Details: 추가:- / 제거:[class\_no, content\_text, grade, new\_status, option\_text, option\_user\_ids, page\_name, prev\_status, school\_id, target\_user\_id, user\_id, vote]

📑 JSON 키 불일치(user\_properties)

- Table: events(parquet)
- Details: 추가:- / 제거:[class\_num, grade, has\_push\_permission, school\_name, status]

📑 JSON 키 불일치(device\_properties)

데이터 품질 모니터링

모니터링 봇 앱 25. 8. 19. 오후 5:58

⚠️ 전체 이탈 위험 감지

사용자 수: 14988

- 👉 평균세션/유저: 1.026
- 🎯 평균이벤트/유저: 25.435
- ⌚ 평균세션길이(초): 308.6

⚠️ 주요 위험 신호

- ⚡ 이벤트/유저 하락 추세 (-5.2%/day)

Airflow Churn Watchdog • 25. 8. 19. 오후 5:58

⚠️ 전체 이탈 위험 감지

사용자 수: 14988

- 👉 평균세션/유저: 1.026
- 🎯 평균이벤트/유저: 25.435
- ⌚ 평균세션길이(초): 308.6

⚠️ 주요 위험 신호

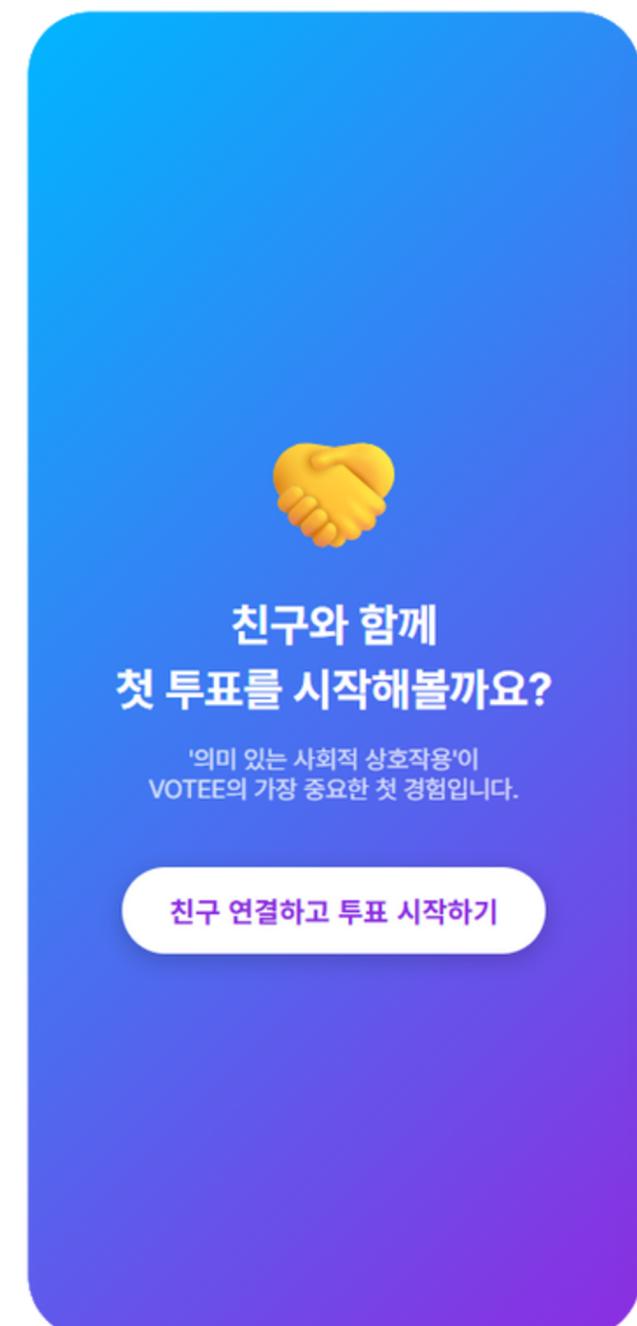
- ⚡ 이벤트/유저 하락 추세 (-5.2%/day)

이탈 위험 모니터링

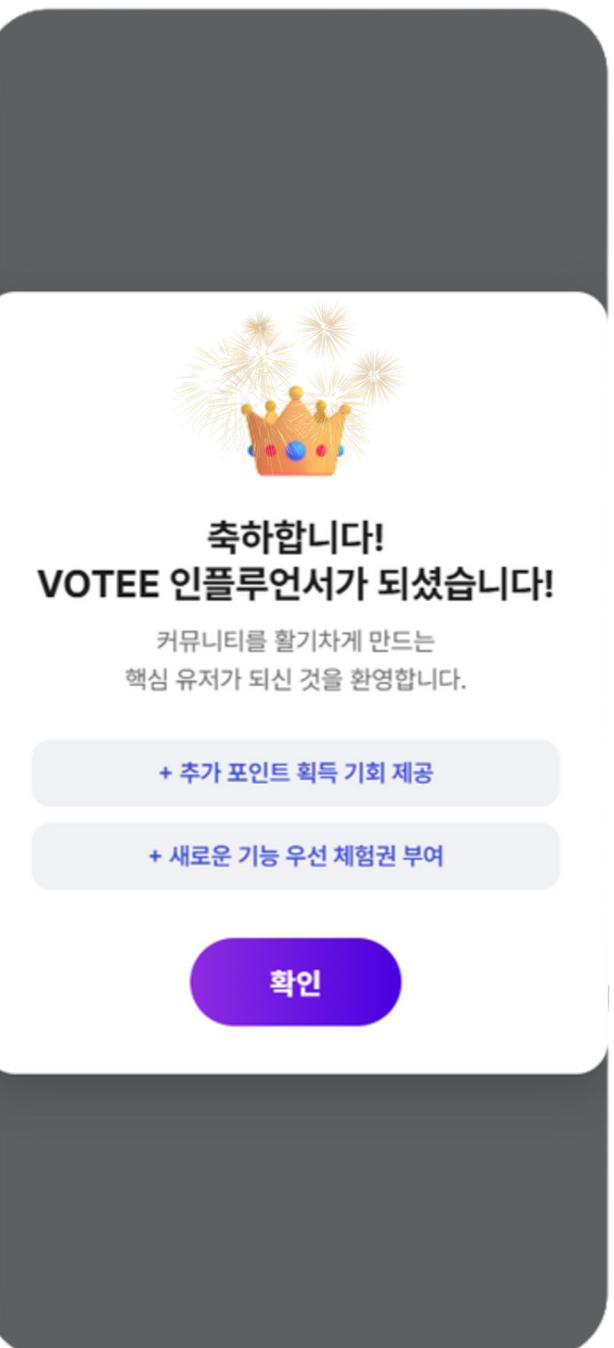
# 04. VOTEE 개선안: 운영 전략

## Part 4. 새롭게 설계하는 익명 투표 앱 'VOTEE'

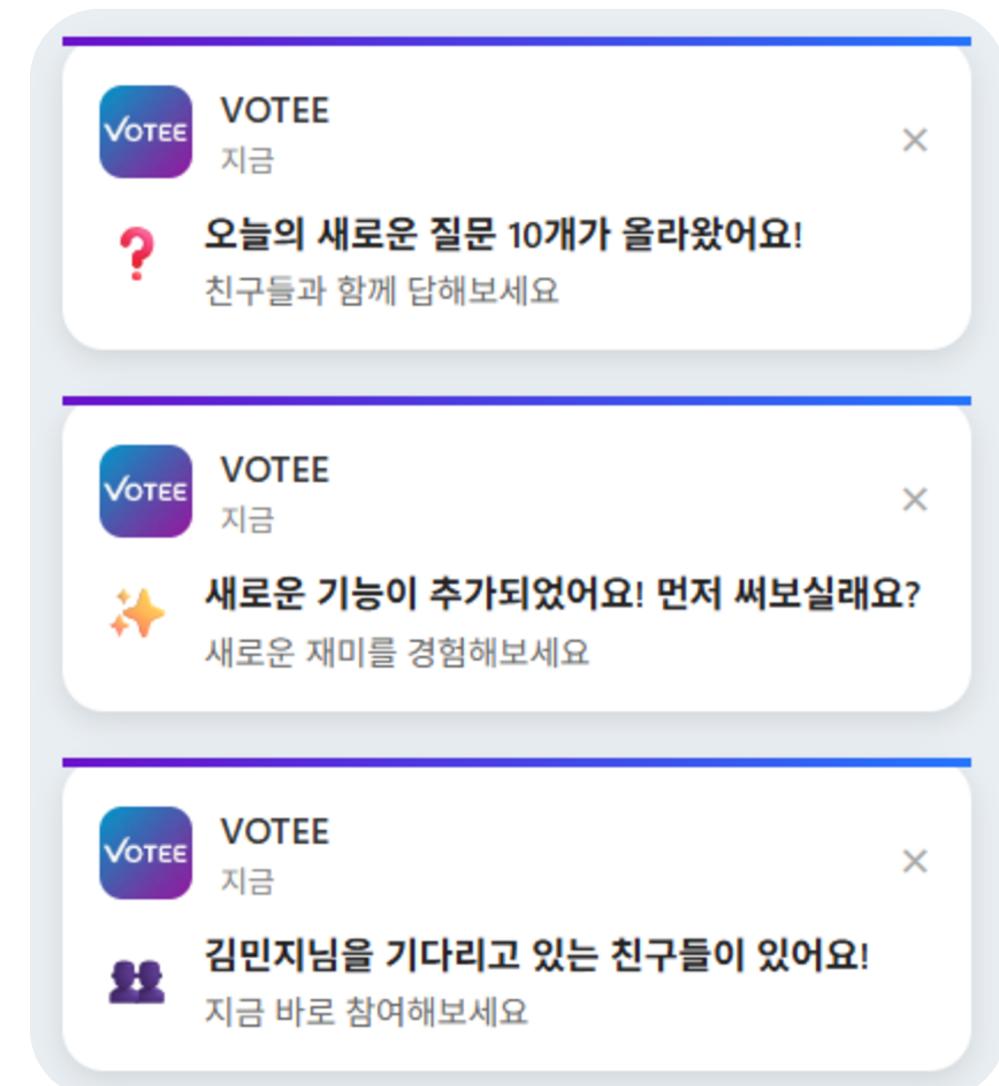
**정착 유도** : 첫 세션에서 첫 투표 경험



**중심 학생** : 기준 달성 시 뱃지 & 보상



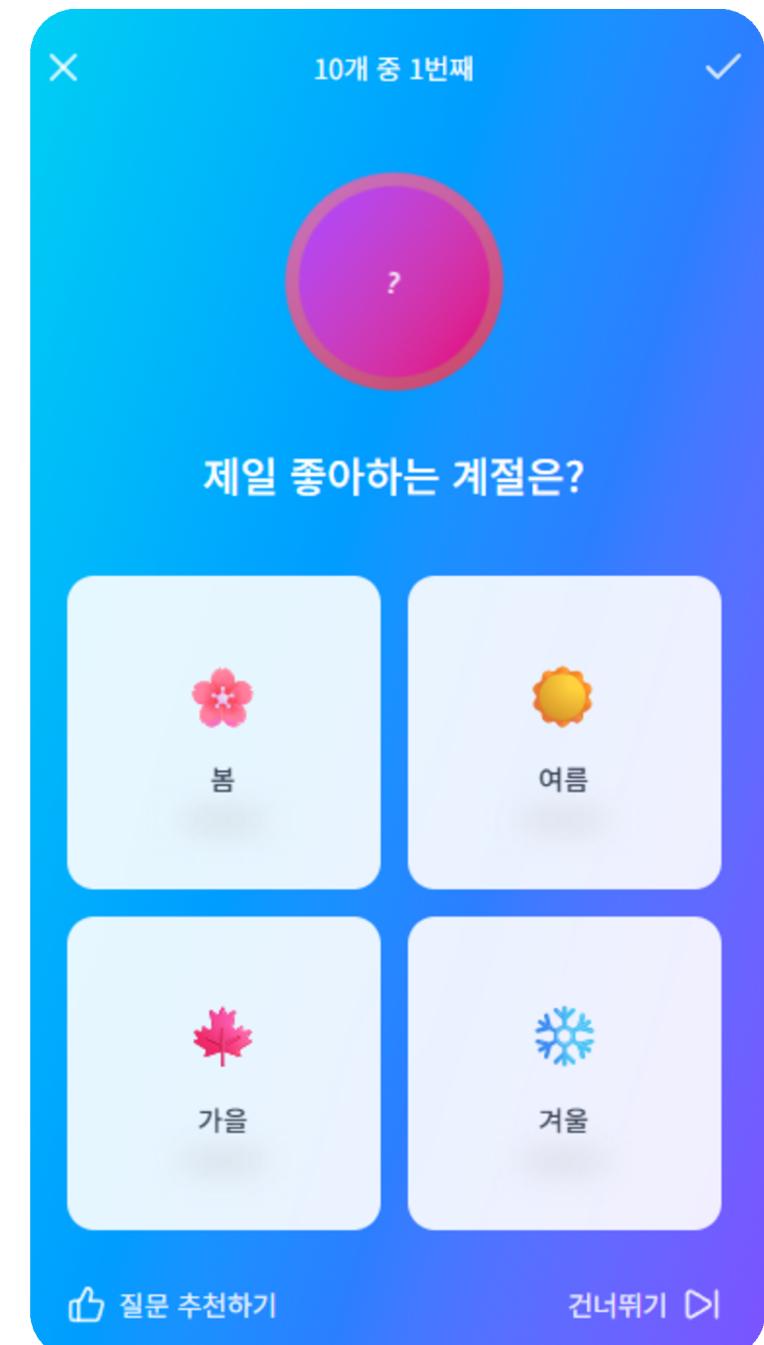
**이탈 방지** : 맞춤형 푸시 알림



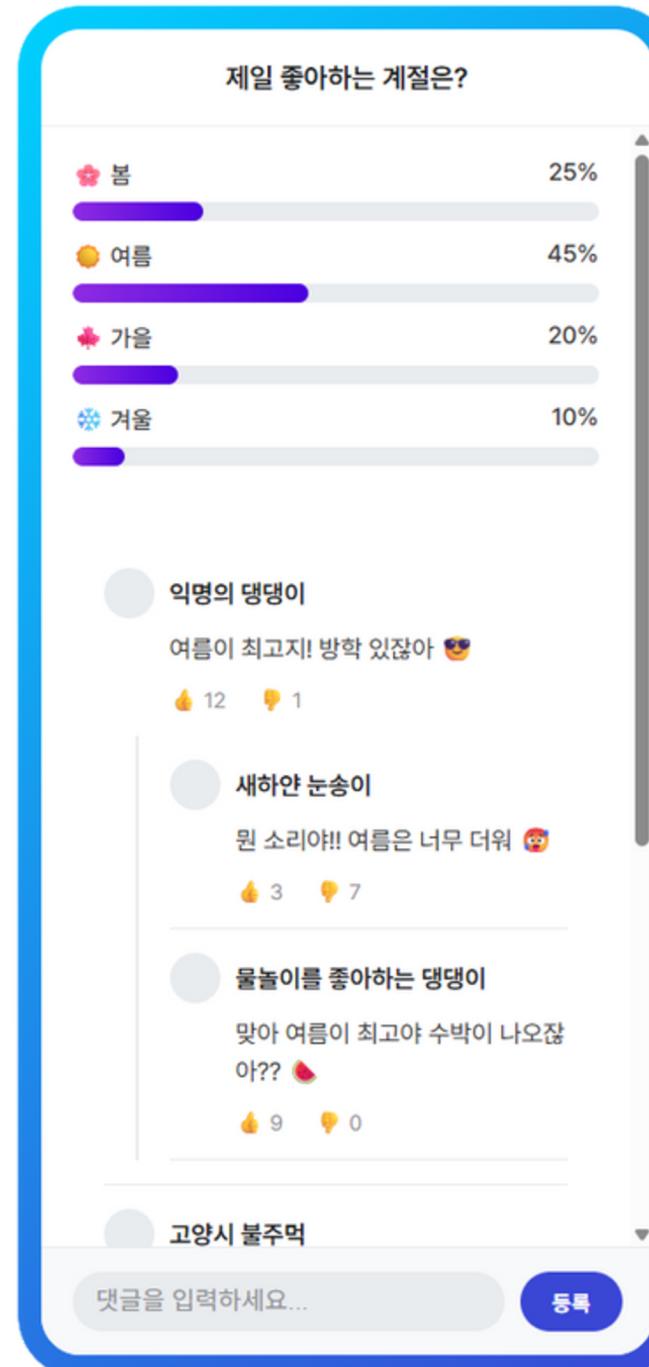
# 05. VOTEE 개선안: 신규 기능

Part 4. 새롭게 설계하는 익명 투표 앱 'VOTEE'

## 질문 서비스 개편



## 커뮤니티 기능 추가



## Part 5. 결론 및 요약



# 결론

기존 서비스의 문제점을 보완하여 VOTEE 출시

데이터 수집·적재

파이프라인 개선

실시간 대시보드·알림

이탈 분석 & 대응

맞춤형 전략

신규 기능 기획

## VOTEE 출시!

가장 솔직한 10대들의 놀이터, 지금 바로 참여하세요.



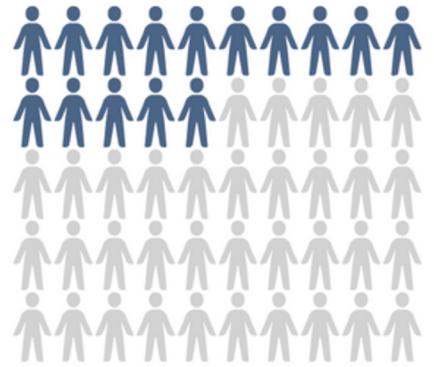
들어주셔서 감사합니다  
모두 고생 많으셨습니다!



# 요약

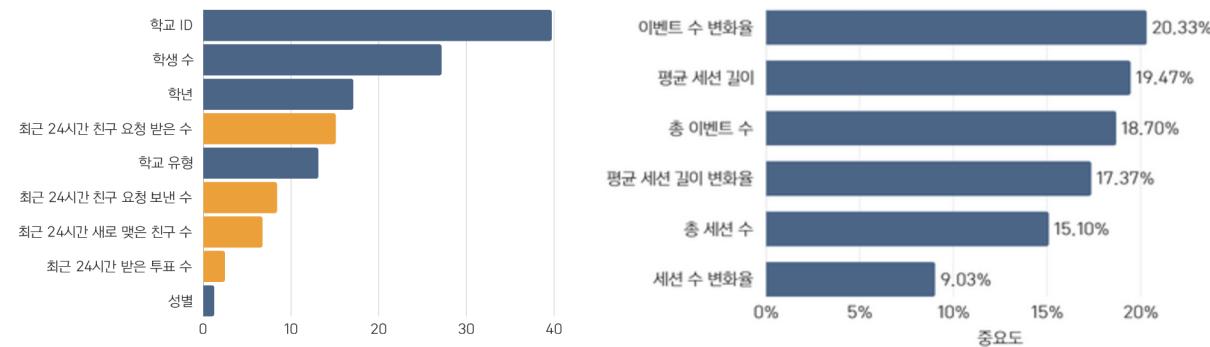
## 문제 제기 및 분석

활성 유저 수가 급격하게 줄었고, 로그도 정상적으로 수집이 안됐음



학교·학년 같은 **태생적 요인 (배경요인)**이 가장 큰 영향 24시간 내 “받은 친구 요청 수”가 진존을 결정하는 핵심 요인 중 하나 항경은 바꿀 수 없어도 **초기 상호작용 설계**를 통해 결과를 바꿀 수 있음

LGBM 분류 정차 예측 모델 속성 중요도



## 파이프라인, 대시보드 설계

부정확한 세션 데이터

- 유저의 핵심 체류 시간을 제대로 측정할 수 없었음

일관성 없는 이벤트 로그

- 날짜별로 이벤트 이름과 값이 달라 장기적인 행동 추적이 불가능했음

데이터 유실 및 누락

- 비정상적인 로그 수집으로 인해 분석에 필요한 데이터가 소실되었음

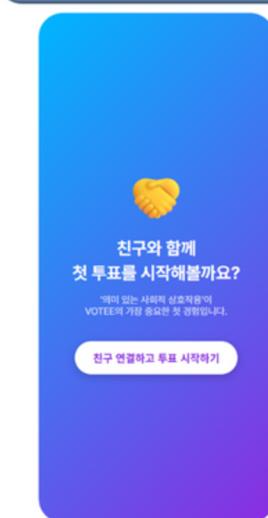
통일되지 않은 데이터 구조

- 갑작스러운 컬럼명 변경으로 분석 코드가 깨지고 유지보수가 어려웠음

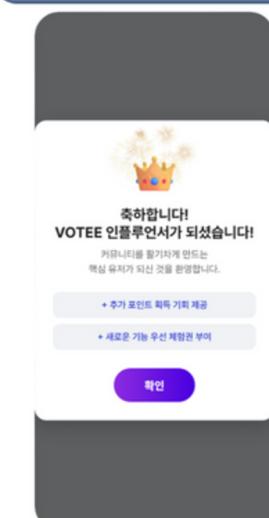


- Faker 패키지로 랜덤 로그 생성
- 유저 활동 로그를 만드는 모바일 클라이언트 역할
- Airflow로 1시간마다 1주일치 데이터 생성 및 적재 자동화

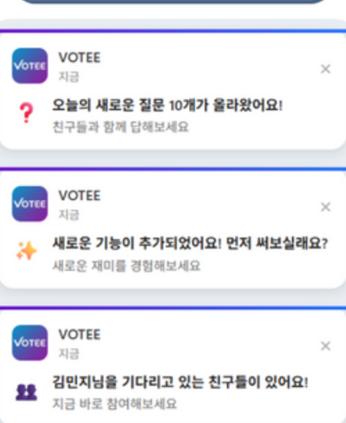
정학 유도 : 첫 세션에서 첫 투표 경험



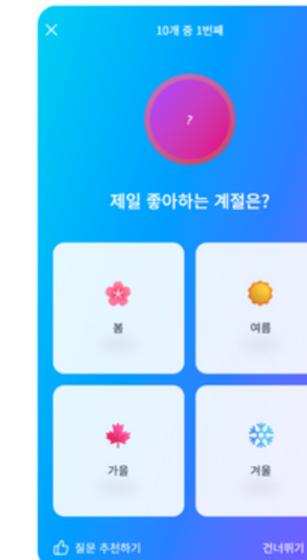
중심 학생 : 기준 달성을 뱃지 & 보상



이탈 방지 : 맞춤형 푸시 알림



질문 서비스 개편



커뮤니티 기능 추가

