KeepTune,

Big DATA• AI 기반 음악 스트리밍 서비스 마케팅 대시보드

SKN25-2nd-1Team 🎧

김나연, 박범수, 양예승, 이근혁, 최현우

CONTENTS

문제 정의 및 프로젝트 목표

기업 및 비즈니스 배경

광고 + 유료 구독 기반 수익 구조
구독 갱신 여부 → 매출에 직접적 영향

💡 **구독 유지율**
기업 수익과 직결되는 핵심 지표

프로젝트 문제 정의

Churn Prediction (이탈 예측)
사용자가 30일 이내에 재구독 X → 이탈

✓ 프로젝트 목표
이탈 가능성이 높은 고객 조기 식별
마케팅 타겟팅 자동화 기반 구축

본 프로젝트 접근 방식

- 머신러닝 기반 확률 예측 모델 구축
- SHAP을 통한 이탈 요인 해석
- 마케팅 활용 대시보드 설계
- 타겟 세그먼트 도출

타겟 이용자

- 구독 기반 디지털 콘텐츠 플랫폼의 CRM/Retention 마케팅 담당자
- 데이터 기반 퍼포먼스 마케팅 전략 수립 담당자

목표 : Big DATA• AI 기반 음악 스트리밍 서비스 마케팅 대시보드

사용 데이터 소개

KKBOX kaggle Data

아시아 최대 음악 스트리밍 플랫폼, 3,000만 곡 이상 보유

01

train.csv

- **사용자 ID**
- 이탈 여부(정답값)

02

members.csv

- **사용자 ID**
- 성별
- 연령대
- 가입 채널
- 거주 도시

03

transactions.csv

- **사용자 ID**
- 결제 금액
- 결제 방식
- 자동 갱신 여부
- 구독 기간
- 해지 여부

04

user_logs.csv

- **사용자 ID**
- 총 청취 시간
- 곡 재생 횟수
- 완청 비율
- 청취 아티스트 수

전처리 & 파생변수 생성

데이터 통합

분산된 4개 테이블을 사용자 ID 단위의 단일 데이터셋으로 통합

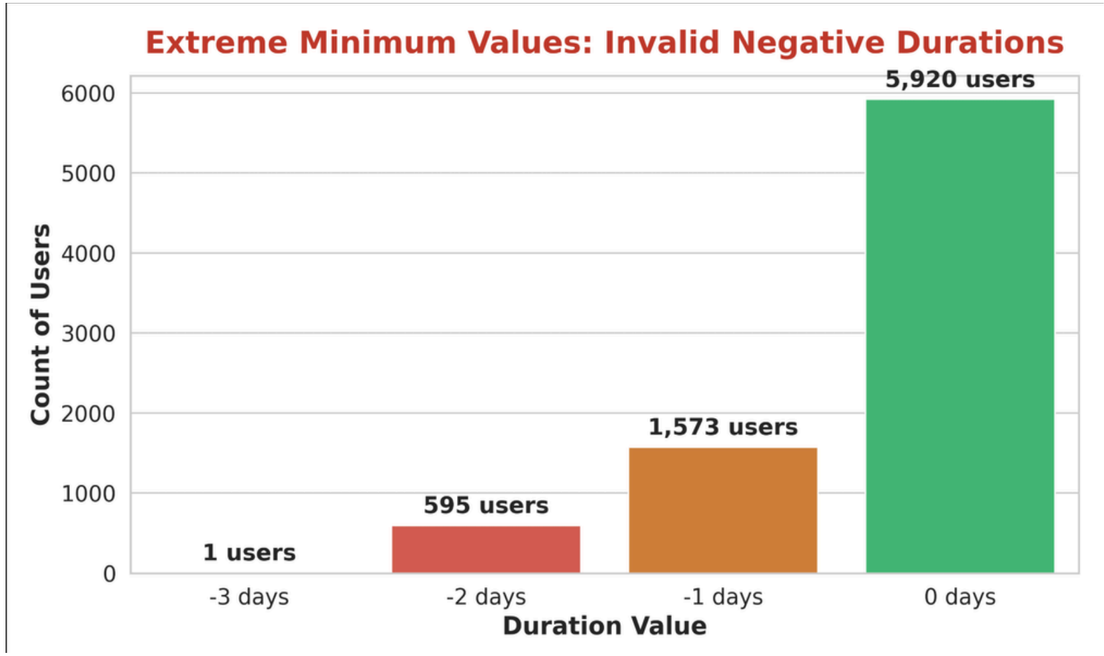
사용자별 수치형 특성 집계

- 결제 패턴 (transactions): 총 결제액, 평균 구독 기간, 자동 갱신 및 해지 비율 등 (1인 1행 요약)
- 행동 로그 (user_logs): 총 청취 시간, 유니크 아티스트 수, 구간별 재생 횟수 등 집계

사용자 ID 기준 데이터 병합

사용자 ID와 이탈 여부 있는 train.csv 기준으로 **사용자 ID 기준 3개 테이블을 순차적으로 결합**
결과: 모든 feature가 한 줄에 담긴 사용자 중심 통합 데이터셋 구축

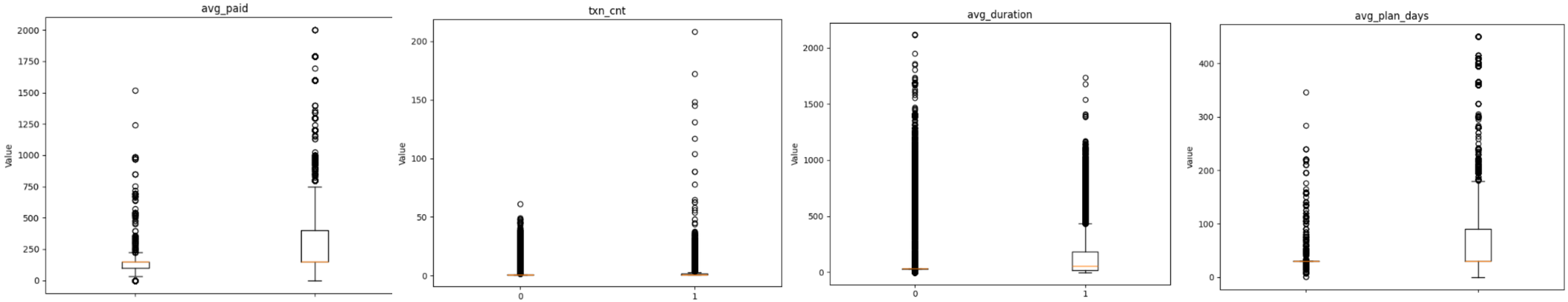
결측치 처리 예시- 최대 구독 유지 기간



- 분석 대상: max_duration 최하위 수치
- 발생 문제: 논리적으로 불가능한 음수 데이터
- 오류 건수: 총 2,169건 (약 0.2% 비중)
- 추가 확인: 기간이 0일(활동 없음)로 집계된 유저 5,920명

→ 음수데이터는 시스템 에러값으로 간주하고, 전처리 과정 진행

이상치 처리 예시- boxplot



- 이상치 보존
  - 고객 통계(결제액, 이용 시간 등)의 극단값은 단순 오류가 아닌 '헤비 유저' 또는 '특이 행동 패턴'을 지닌 실제 고객의 특성
  - 일괄 삭제 시 핵심 정보 손실 우려

- 데이터 삭제 지양: 고과금/장기 이용 유저의 이탈 패턴 등 중요한 정보 손실 방지

전처리 & 파생변수 생성

변수 범주화 & 결측치 처리

데이터의 특성에 따른 보간 전략 및 형변환

나이(bd) 데이터 구간으로 나누어 범주형 변수 생성

- 이상치 처리: 0~100세 범위를 벗어난 비정상 데이터 변환
- 구간화: **10대~70대 연령대별로 카테고리 변수** 생성

→ 이상치를 unknown으로 처리하여 모델 안정성 향상

로그 데이터 결측

- **no_log_flag** 파생 변수 생성 → 로그 특성 없는 데이터에 no_log_flag 값 1 할당
- 결측치는 0으로 채움

→ 결측치 보간을 0으로 채운 것에 대한 의미 보강

범주형 변수 최적화

- 대상: 거주 도시, 성별, 가입 경로
- 결측치를 단순 삭제 X, unknown 카테고리로 보존

→ 정보 손실 최소화

수치형 결측

- 방법: **다중 대치법 기반 반복 추정** 활용
- 중앙값으로 초기화 → 다른 변수들과의 관계 반복 계산 → 최적값 추정
- 단순 평균 대입보다 통계적 왜곡을 줄이고자 함

전처리 & 파생변수 생성

최종 파생변수 구조 • RFM 기반

FREQUENCY

txn_cnt	총 거래 횟수
auto_renew_rate	자동 갱신 비율
cancel_rate	해지 비율
total_cancel	총 해지 횟수
avg_duration	평균 구독 유지 기간
max_duration	최대 구독 유지 기간

auto_cancel_inter	자동 결제 및 취소 상호작용

age_group	연령대
city	거주 도시
gender	성별
registered_via	가입 경로

RECENCY

no_log_flag	로그 미존재 여부
total_secs_sum	총 청취 시간
total_secs_mean	평균 청취 시간
num_unq_mean	평균 유니크 곡 수
num_100,985,75,50,25_sum	구간별 재생 횟수

MONETARY

total_paid	총 결제 금액
avg_paid	평균 결제 금액

최종 데이터

총 사용자 수: **970,960명**
총 변수 수: **24개**
Target 변수: **이탈 여부**

최종 모델 성능 지표

머신러닝

XGBoost 👍

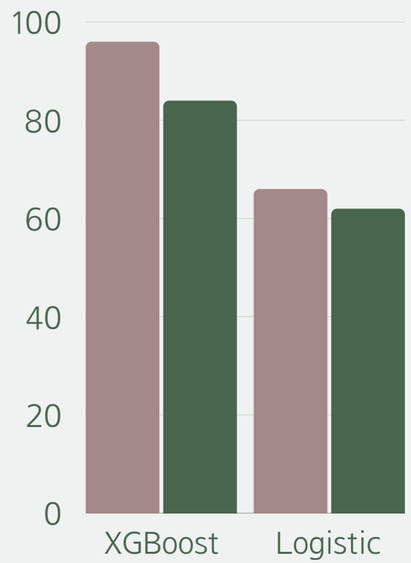
Recall : 0.959

Precision : 0.8434

Logistic

Recall : 0.6638

Precision : 0.6239



딥러닝

Resnet 👍

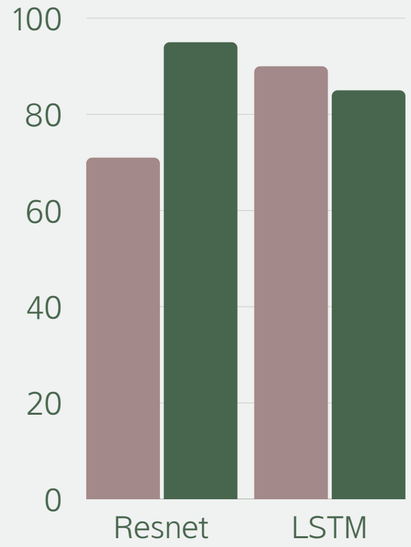
Recall : 0.7089

Precision : 0.9586

LSTM

Recall : 0.8920

Precision : 0.8528



최종 모델 성능 지표

모델	구분	Precision	Recall	F1-score	TP	FP	FN	TN	Accuracy
XGBoost (0.6)	유지 (0)	0.99	0.98	0.99	-	3,336명	-	173,390명	**0.98**
	이탈 (1)	0.83	**0.95**	**0.88**	**16,509명**	-	**957명**	-	
ResNet (0.8)	유지 (0)	0.97	**1**	0.98	-	678명	-	**176,048명**	0.97
	이탈 (1)	**0.95**	0.7	0.81	12,310명	-	5,156명	-	

성능지표에 따른 용도

XGBoost - Recall 👍

실제 이탈자 87,330명 중 83,778명 맞춤
→ 놓치는 이탈자 거의 없음

이탈 아닌데 이탈로 잘못 예측 : 15,557명

소극적 마케팅 범위 설정에
활용하기 좋은 모델

ex) 소규모 쿠폰, 맞춤형 추천

Resnet(finetuned) - Precision 👍

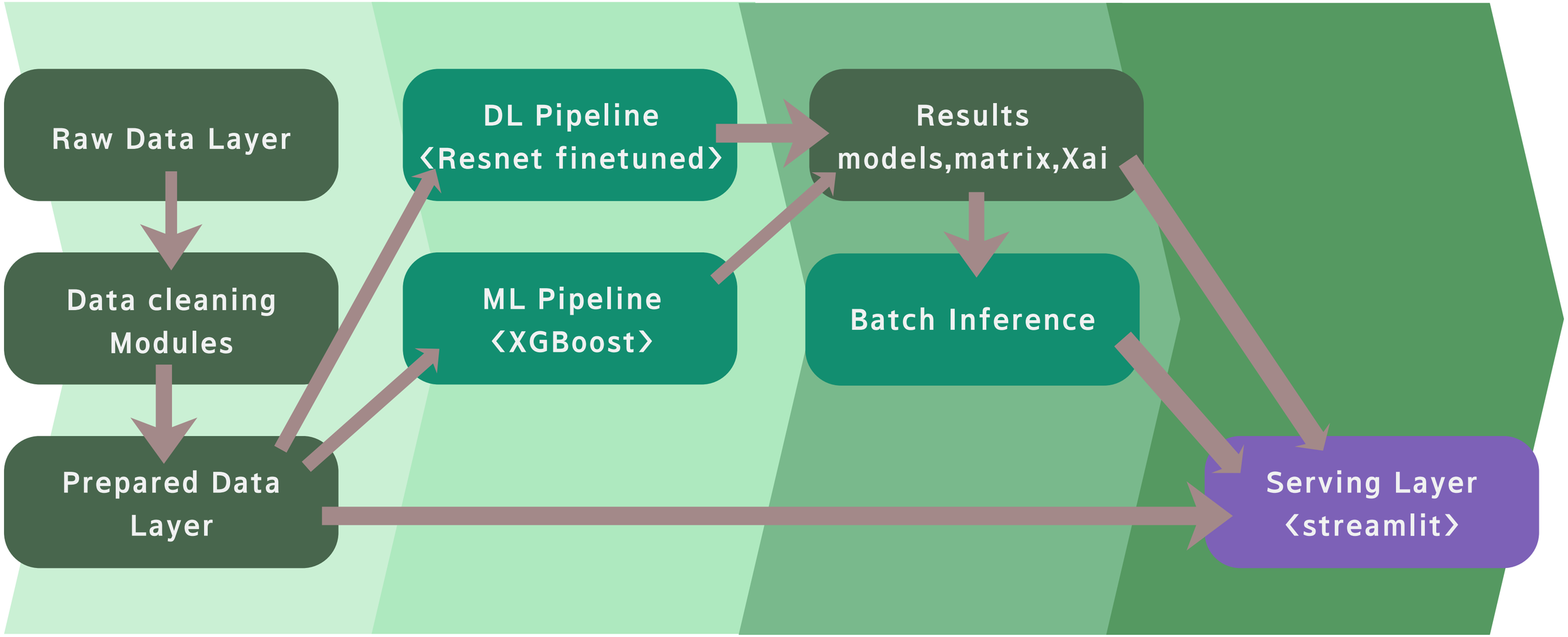
이탈 예측 64,582명 중 61,909명 맞춤
→ 이탈이라 판단 시 거의 맞음

놓치는 이탈자 : 25,421명

공격적 마케팅 범위 설정에
활용하기 좋은 모델

ex) 고액 쿠폰, 행사 등

시스템 아키텍처


Raw Data Layer
Data cleaning
Modules
Prepared Data
Layer
DL Pipeline
〈Resnet finetuned〉
ML Pipeline
〈XGBoost〉
Results
models,matrix,Xai
Batch Inference
Serving Layer
〈streamlit〉

대시보드 기능

📊 유저 행동 인사이트

- 프로젝트 핵심 지표
- 분석 아키텍처 및 하이브리드 엔진 소개
- 기대 효과 및 비즈니스 임팩트

🔮 이탈 위험도 시뮬레이터

- 하이브리드 AI 모델 기반 기업 맞춤형 전략 진단
- 각 변수를 조절해보면서 이탈 예측이 가능
- 모델별 분석 가중치를 보여줌으로써 해석력 부여

💡 비즈니스 전략

- 고객 유지 가이드
- 비즈니스 상황별 타겟팅 전략 시뮬레이션
- 초고위험군 집중 마케팅 전략 제공

대시보드 기능

📊 유저 행동 인사이트

이탈 핵심 요인 탐색

- PDP기반 각 변수 탐색
- 변수의 수치에 따른 예측 영향력 해석 제공

핵심변수 영향력

- SHAP 평균 영향력 Top8
- 특정 변수에 따른 예측 방향
- 개별 고객 별 SHAP 시각화

데이터 시각화

- 카테고리 변수별 이탈률 시각화 제공
- 수치형 변수 구간별 이탈률 시각 제공
- 최소 표본 수, 정렬 기준, 등 변형 가능

대시보드 기능

🔮 이탈 위험도 시뮬레이터 & 💡 비즈니스 전략

AI 하이브리드 진단

- 시뮬레이션 기반 진단 리포트 제공
- AI 전략 실행 계획 제공

고객 유지 가이드 제공

- 수치 기반 마케팅 우선순위 설정
- 비즈니스 상황별 타겟팅 전략 시뮬레이션

Thank you

감사합니다