PORTFOLIO · 2026

학부 연구 및 프로젝트 기록.

현실의 데이터는 결측과 노이즈로 가득하지만, 그 안에는 여전히 가치 있는 신호가 있다고 믿습니다.
학부에서는 그런 신호를 안정적으로 학습해내는 모델과 시스템을 만드는 연구를 해왔습니다.

NAME

이윤서 (Yoonseo Lee)

AFFILIATION

가톨릭대학교 컴퓨터정보공학부
Applied AI Lab (지도교수 장재연)

EMAIL

leeyoonseo@catholic.ac.kr

CONTENTS

학회 발표 · 데이터톤 · 산학협력 · 특허

포트폴리오 목차

결측치에 강건한 Gradient Boosting 모델.

결측치를 보간하지 않고, 학습 과정에서 각 피처의 예측 기여도(S_1)와 학습 유용성(S_2)을 동적으로 평가해 feature sampling 확률에 반영하는 boosting 기반 모델을 제안하고 실험했습니다.

TYPE	학회 발표 · 공동 1저자 (이윤서*, 유준영*, 장재연†)
DATASETS	Ion (35×351) · Hotel (21×1,194) · Adult (13×326)
BASELINES	XGBoost · CatBoost · LightGBM
CASES	**Case 1.** 모든 열에 균등한 결측률 적용 **Case 2.** 일부 열에만 결측
METRIC	Accuracy · F1-score · ROC-AUC

방법 요약

- 각 boosting round 후 별도의 **scoring round**를 두어, 피처 f에 대해
  S_1 = `loss(X without f) - loss(X)` (예측 기여도)
  S_2 = `loss(X trained w/o f) - loss(X with f)` (학습 유용성)
- 두 score를 결합해 다음 round의 **feature sampling 확률**을 갱신.
- 결측이 있어도 학습에 유용한 피처는 살아남고, 결측이 적어도 기여가 낮으면 가중치를 낮춥니다.

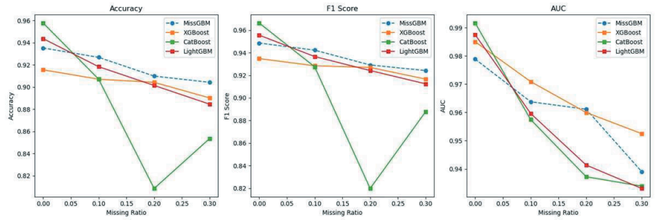
구현 · 실험

- GBDT 코드 베이스에 **S_1 · S_2 scoring 모듈**과 sampling 갱신 로직 추가.
- Case 1 / Case 2 실험 셋업, 데이터셋별 하이퍼파라미터 조정.
- 실험 결과 정리, 표·그래프 작성 및 발표 자료 구성.

Case 1 — 결측률 변화에 따른 성능 추이.

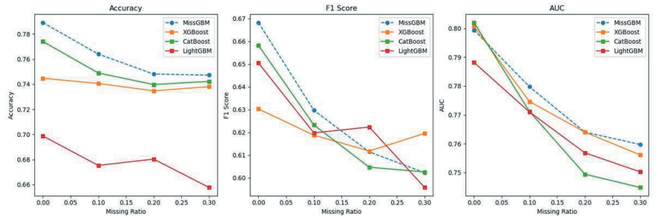
세 개의 벤치마크 데이터셋(Ion · Adult · Hotel)에서 결측률을 0.0에서 0.3까지 늘려가며 XGBoost · CatBoost · LightGBM과 비교했습니다.

Ion 35 feat · 351 samples — ACCURACY · F1 · AUC



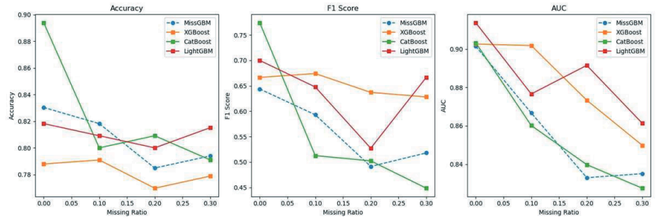
Ion. 결측률이 늘어도 MissGBM의 Accuracy / F1이 안정적으로 유지됨. CatBoost는 0.2 구간에서 큰 폭으로 하락.

Adult 13 feat · 326 samples — ACCURACY · F1 · AUC



Adult. 절대 성능은 모델 간 격차가 작지만, MissGBM이 전 구간에서 Accuracy 상위. F1·AUC는 노이즈가 큼.

Hotel 21 feat · 1,194 samples — ACCURACY · F1 · AUC



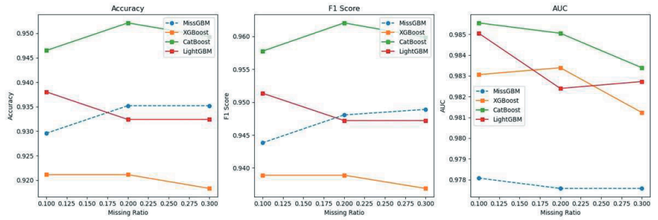
Hotel. 결측률 0.2 부근에서 MissGBM이 F1 기준 가장 낮게 측정되어, 데이터 특성에 따라 trade-off가 있음을 확인.

Case 2 — 일부 열에만 결측이 있을 때.

실제 산업 데이터에 더 가까운 결측 패턴입니다. 모든 열이 아닌 일부 열에만 결측을 주입한 후 같은 비교를 진행했습니다.

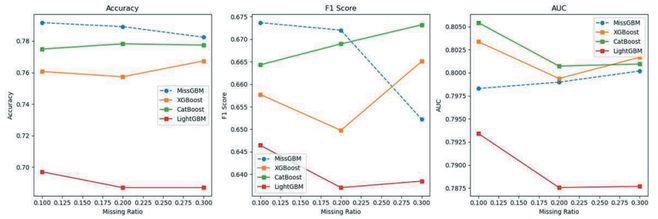
Ion

ACCURACY · F1 · AUC



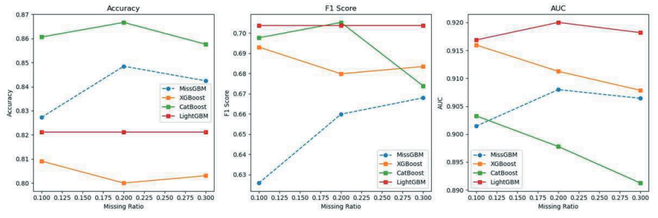
Adult

ACCURACY · F1 · AUC



Hotel

ACCURACY · F1 · AUC



관찰한 점. Case 2는 Case 1보다 결측의 분포가 비균질해 모델 간 차이가 줄어듭니다. 그래도 **Adult에서는 MissGBM이 Accuracy 기준 모든 구간 1위**를 기록했고, Ion · Hotel에서는 CatBoost와 비슷한 수준이거나 약간 뒤처졌습니다.

한계. 결측률이 0.1 → 0.3으로 변할 때의 성능 변화 폭이 작아서, 결측률에 대한 강건성을 충분히 강조하기엔 표본이 부족합니다. 추가 실험이 필요한 부분입니다.

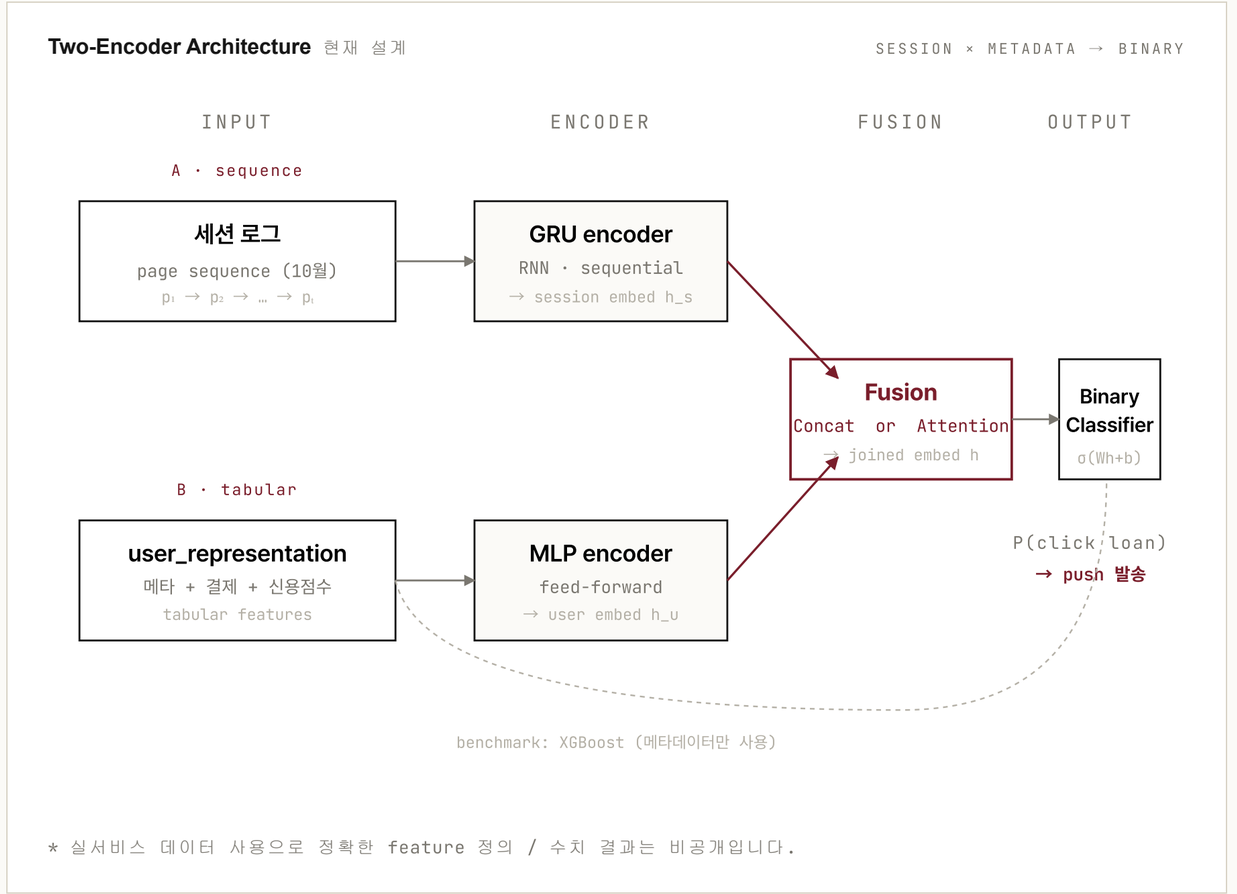
세션 시계열 + 유저 메타데이터 기반 전환 예측.

푸시 알림 채널에서 수익을 최적화하기 위해, **전환 가능성이 높은 유저를 선별하는** 작업입니다. 형식적으로는 추천이지만 본질은 “해당 유저가 11월에 대출 관련 페이지를 클릭할 것인가”의 이진 분류에 가깝습니다.

PARTNER	네이버파이낸셜 (실서비스 데이터)
TASK	11월 대출 관련 페이지 클릭 여부 — 이진 분류
DATA	10–11월 유저 페이지 이동 세션 로그 user_representation (메타 + 결제 · 신용점수 등 금융 특성)
BENCHMARK	XGBoost · 머신러닝 (메타데이터 only)
STATUS	두 인코더 결합 방식 실험 진행 중

맡은 부분

- 세션 인코더 (GRU 기반) 구현 — 페이지 이동 시퀀스 학습.
- 유저 메타데이터 MLP 인코더 구현.
- 두 인코더 출력을 합치는 두 가지 방식 (**Concat / Attention**) 비교 실험.
- 세션 로그 정제 — 세션당 평균 81회 / 최대 7,155회의 페이지 이동을 정제하는 기준 설계.



* 실서비스 데이터 사용으로 정확한 feature 정의 / 수치 결과는 비공개입니다.

Concat vs Attention — 결합 방식 비교.

세션 임베딩 h_s와 유저 임베딩 h_u를 합칠 때 단순 concat과 attention 두 방식을 실험했습니다. 전체 평균 성능은 비슷하지만, **특수 시나리오에서 attention이 일관되게 우수**했습니다.

Fusion 방식

H_S + H_U → H

A · CONCAT

`h = [h_s ‖ h_u]`

두 임베딩을 그대로 이어 붙임. 파라미터가 적고 학습 안정적이지만, 두 modality 간 상호작용을 명시적으로 모델링하지 않음.

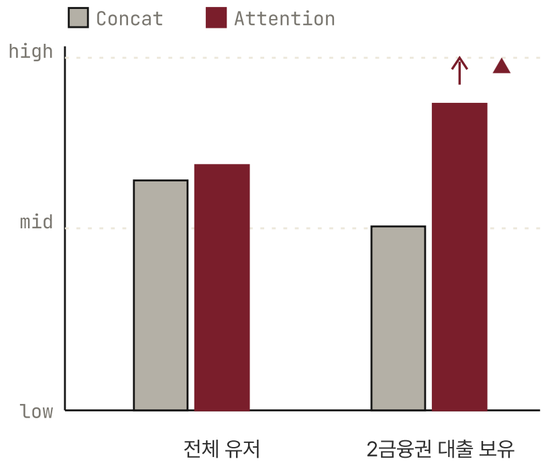
B · ATTENTION

`α = softmax(Q·K) ;  h = α·V`

메타데이터 특성 중 어느 차원이 세션 패턴과 강하게 상호작용하는지 학습. 파라미터 증가, 그러나 표현력 ↑.

Key finding

SCENARIO × FUSION



전체 유저 분포에서 두 방식은 비슷한 성능. 그러나 **2금융권 대출 보유 유저**처럼 메타데이터 분포가 특수한 부분집합에서는 attention이 concat보다 일관되게 높은 정확도를 보였습니다.

* 정확한 수치는 비공개. 그래프는 상대 비교용 도식.

사전 Quick Model

중간보고 이전

현재 설계 이전에 “다음 페이지 예측” 타겟으로 GRU4REC · SR-GNN을 우선 돌려보고, 학습 시간과 메모리·정확도 trade-off를 확인했습니다.

Model	Runtime	Hit@20	MRR@20
GRU4REC	138 m	0.9882	0.7487
SR-GNN	396 m	0.9881	0.7386

참고한 모델

- **GRU4REC** ICLR 2016 · RNN/GRU 세션 기반 추천.
- **SASRec** ICDM 2018 · Self-attention 시퀀셜 추천.
- **SR-GNN** AAAI 2019 · 세션 그래프 + GNN.
- **TGN** ICML 2020 W · 동적 그래프 추천.

전환 경로 분석 — 세션을 서비스 단위로.

고객이 대출 페이지로 전환하기까지 **어디를 거쳐 어떻게 도달하는지**를 분석했습니다. 페이지 단위는 너무 잘게 쪼개져 있어, 세션마다 대표 페이지를 뽑고 그것이 속한 서비스로 묶었을 때 유의미한 패턴이 드러났습니다.

분석 과정

PAGE → REPRESENTATIVE → SERVICE

계층구조상 이동 가능한 페이지가 정해져 있음

세션을 대표하는 핵심 페이지 추출

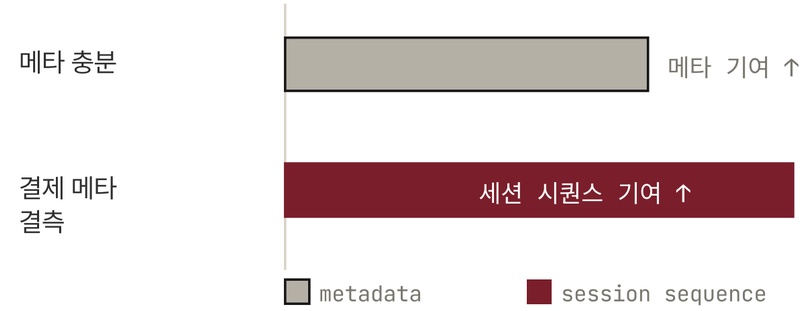
서비스로 묶었을 때 유의미한 패턴 확인

* 세션당 평균 81회 / 최대 7,155회의 페이지 이동 — 그대로 학습하기엔 노이즈가 커, 대표 페이지 · 서비스 단위 추상화가 필요했습니다.

KEY · 세션이 빛나는 지점

COLD-START

유저를 여러 조건으로 나눠 시나리오를 시도한 결과, **유저 메타정보가 비어 있는 cold-start 상황**, 특히 **결제 관련 메타데이터(결제금액 등)가 없을 때** 세션 시퀀스가 그 공백을 메우는 강한 신호가 되었습니다.



— 메타가 충분한 유저는 기존 tabular 모델로도 충분.

— **결제 메타가 비는 유저일수록 세션 시퀀스의 상대적 기여가 커짐** — 두 인코더를 합치는 설계의 근거.

이종 데이터 결합 GNN 콘텐츠 추천.

Steam(게임)과 IMDb(영화)처럼 분류 체계가 다른 플랫폼의 콘텐츠를 하나의 이분 그래프로 연결하고, 그래프 임베딩으로 도메인을 넘나드는 추천을 구현했습니다.

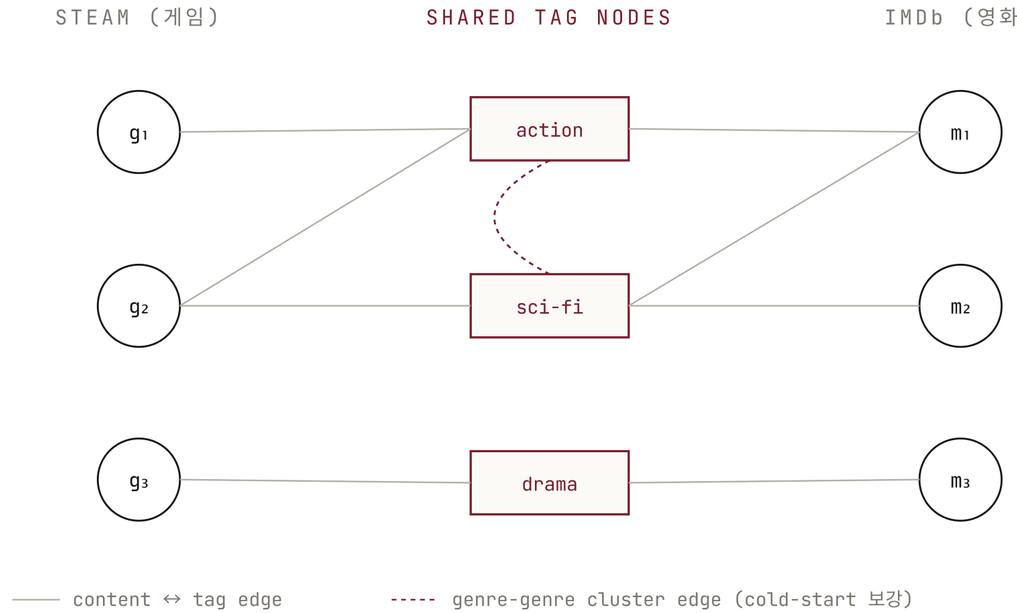
TYPE	교내 프로젝트 (Academic) · 서비스 A.O.D 운영 중
ROLE	AI Modeling & Data Engineering
STACK	Python · PyTorch · NetworkX · Selenium
PROBLEM	같은 장르라도 플랫폼마다 명칭(Tag/Genre)이 달라 통합 추천이 어려움

접근

- Steam · IMDb 메타데이터를 크롤링해 **이분 그래프(Bipartite Graph)**로 구조화.
- 태그를 억지로 매핑하지 않고 **고유 노드로 유지**, 연관성을 그래프가 학습하도록 설계.
- 플랫폼별 추천 리스트를 크롤링해 초기 그래프 밀도(density)를 보강.

Bipartite Graph Steam × Tag × IMDb

콘텐츠 통합 맵



핵심은 Cold-start — LLM으로 의미 기반 엣지 생성.

신규 콘텐츠는 상호작용 기록이 없어 추천이 어렵습니다(cold-start). 이 프로젝트의 핵심은 그래프를 만들 때 **의미가 비슷한 아이템 노드끼리 엣지를 추가**해 정보 전파 경로를 미리 열어 둔 점입니다.

KEY · Cold-start 완화 전략

LLM-ASSISTED EDGE

아이템 노드를 만들 때 "어떤 콘텐츠끼리 의미가 비슷한가"를 판단해 엣지를 추가하는데, 단순 태그 문자열 매칭으로는 `action` ↔ 액션 ↔ `FPS` 같은 표현 차이를 잡지 못합니다. 이 의미 판단에 **LLM을 부분적으로 도입**했습니다.

표현이 달라도 의미가 가까우면 → 엣지 연결 → cold-start 노드도 전파 경로 확보

- **역할:** LLM은 아이템 간 의미 유사도만 판단 — 추천 자체는 그래프 모델이 수행.
- 여기에 플랫폼 자체 추천 리스트도 크롤링해 초기 그래프 밀도(density)를 함께 보강.

임베딩 학습 — 3가지 비교

- **Random Walk** — 경로 탐색으로 구조적 유사성 학습, 소비 패턴 군집화.
- **LightGCN** — 비선형 제거, 이웃 정보 전파에 집중한 고차 상호작용 학습.
- **GraphSAGE** — 신규 노드도 재학습 없이 임베딩 생성 (inductive).

최종 매칭 & 맡은 부분

- 임베딩 벡터 간 **코사인 유사도**로 cross-domain 콘텐츠 매칭.
- 크롤링(Selenium) · 그래프 구성(NetworkX) · LLM 의미 엣지 · 세 모델 비교 담당.
- 결과물은 A.O.D(all of dopamine) 서비스로 운영 중.

음성 상담 기반 환자-간병인 매칭.

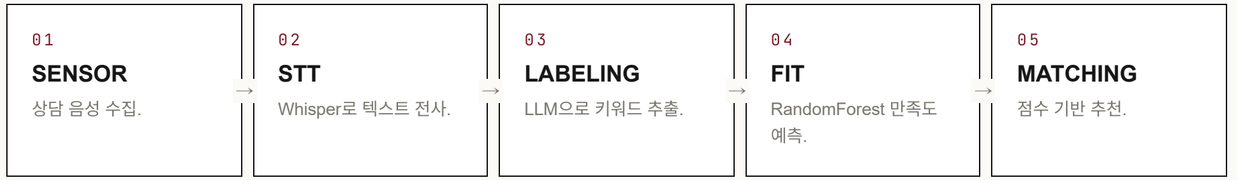
환자·보호자 상담 음성에서 키워드를 추출하고 간병인 메타데이터와 결합해, 매칭 모델 학습용 데이터셋을 자동 생성하는 방법을 설계했습니다. 해당 방법으로 특허를 출원했습니다.

PATENT	**출원번호 제10-2024-0160868호** “돌봄 서비스 제공을 위한 기계학습 모델의 학습 데이터 생성 방법 및 이를 수행하는 전자 장치”
INVENTORS	장재연, **이윤서**, 유준영, 김민식, 이홍주
STACK	OpenAI Whisper · LLM keyword extraction · RandomForest
FUNDING	한국연구재단(NRF) · 코드블라썸 산학협력

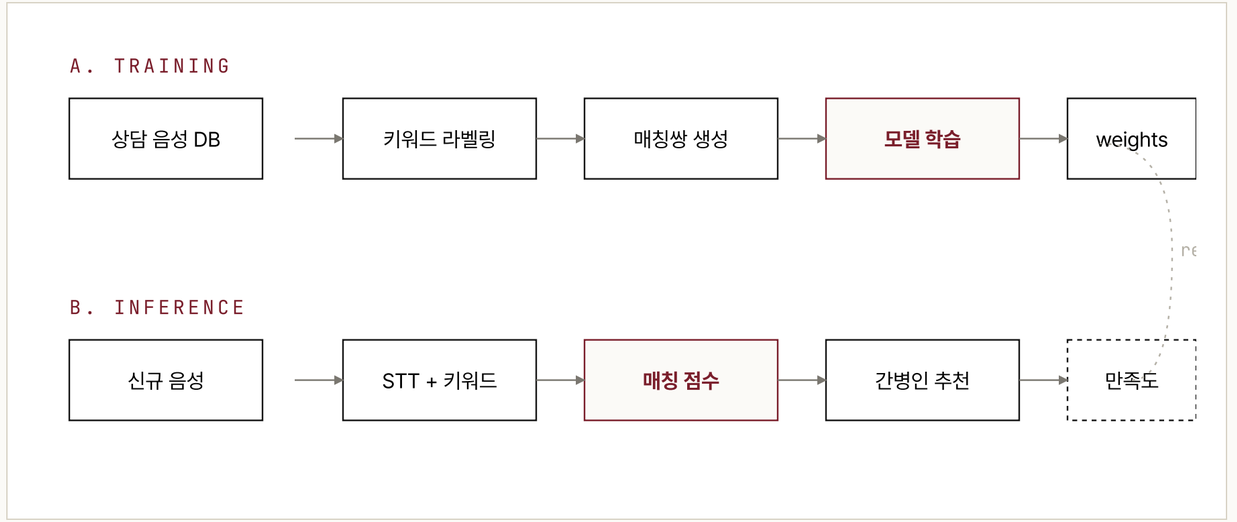
맡은 부분

- 5개 모듈 인터페이스 설계 (SENSOR · STT · LABELING · FIT · MATCHING) — I/O 명세 작성.
- 학습 파이프라인과 예측 파이프라인 분리. 데이터가 누적될 때마다 모델이 갱신되는 구조 설계.
- 오디오 분할 기반 STT 정확도 개선 실험 (다음 슬라이드).
- 특허 명세서 작성 참여 — 발명자 등재.

PIPELINE 5 MODULES



학습 / 예측 구조



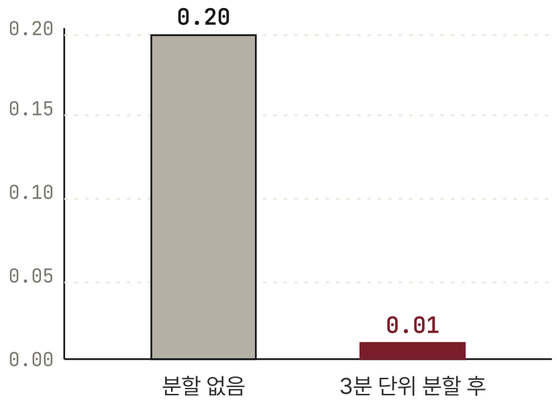
환자 데이터 생성 — 기술 디테일.

특허 명세서에서 학술적으로 의미 있는 세 가지: 음성 분할 기반 STT 정확도 개선 실험, 환자-간병인 결합 데이터 스키마 설계, 학습용 데이터 셋을 위한 전처리 표준화입니다.

STT 정확도 개선 실험

CER · CHARACTER ERROR RATE

긴 통화 녹음을 통째로 Whisper에 넣으면 의료·간병 전문 용어에서 오차가 누적됩니다. 오디오를 **3분 단위 세그먼트**로 분할한 뒤 변환하도록 설계해 CER을 측정했습니다.



RESULT 평균 CER을 **0.20** → **0.01**로 약 20배 감소 — 명세서 도 6(b)에 기재.

환자-간병인 결합 데이터 스키마

SCHEMA DESIGN

상담 음성에서 추출한 환자 키워드와 외부에서 받은 간병인 메타데이터를 결합해 단일 학습 튜플을 만듭니다.

PATIENT (from voice)

병명 · 환자 나이 · 의식 상태 · 식사 보조 · 화장실 도움 · 피딩 · 석션 · 욕창 · 치매 · 선호 간병인 성별 / 국적 … 등 키워드 약 30종

CAREGIVER (external)

경력(개월) · 자격증 7종 (요양보호사 · 간호조무사 · 호스피스 · 노인심리상담사 …) 선호도 8종 / 기피도 7종 … 등 약 38 특성

JOINED (target = 매니저 만족도)

간병번호 · 병원ID · 매니저 만족도 (target) · 환자 + 간병인 결합 약 **60+ 특성**. 만족도 라벨이 누적되면 학습 파이프라인이 자동으로 모델을 갱신.

전처리 표준화

PREPROCESSING

이질적인 두 데이터를 학습 가능한 단일 행렬로 만들기 위해 다음 절차를 표준화했습니다.

STEP 01 **Binary encoding** — 의식 상태 · 감염성 질환 여부 등 이진 특성을 0/1로.

STEP 02 **One-Hot encoding** — 다중 카테고리 특성(자격증·선호도)을 수치형으로.

STEP 03 **결측치 처리** — 결측률 ≥ 5%이면 해당 특성/행 제거, 그 외엔 평균·중앙값 대체.

STEP 04 **Scaling** — 단위가 다른 특성을 표준화/정규화.

STEP 05 **Serialization** — UTF-8-SIG / CP949 인코딩으로 CSV 변환, 학습 파이프라인 입력.

* 특허 명세서 청구항 7 · 8 항에 기재된 전처리 절차를 코드로 구현.

ICU 환자의 AKI 발생 조기 예측.

MIMIC-IV의 ICU 코호트에서 시간 윈도우별로 AKI 발생 여부를 예측한 데이터톤 작업입니다. 5개 트리 모델을 F1-score 가중치로 앙상블했습니다.

TEAM	**NARAK** — 박준범 (서울대) · 김시훈 (서울대) · **이윤서 (가톨릭대)** · 손재혁 (서울대)
TASK	ICU 환자의 AKI 발생 여부 이진 분류
DATA	MIMIC-IV ICU 코호트 · 24h 단위 Time Window
MODELS	XGBoost · CatBoost · AdaBoost · LightGBM · RandomForest → F1-score 가중 ensemble

맡은 부분

- 피처 추출 및 결측 비율 계산 · 결측치 처리 로직 작성.
- 24시간 단위 Time Window 데이터셋 빌더 구현.
- 5개 모델 학습/추론 파이프라인 및 F1 가중 앙상블 결합.

PIPELINE

01 Filter
조건 기반 코호트 필터링 후 병합.

→

02 Missing
결측 비율 계산 · 결측치 처리.

→

03 Merge
피처 병합 및 정리.

→

04 Time Window
24h 단위로 분할.

→

05 Ensemble
F1 가중 앙상블.

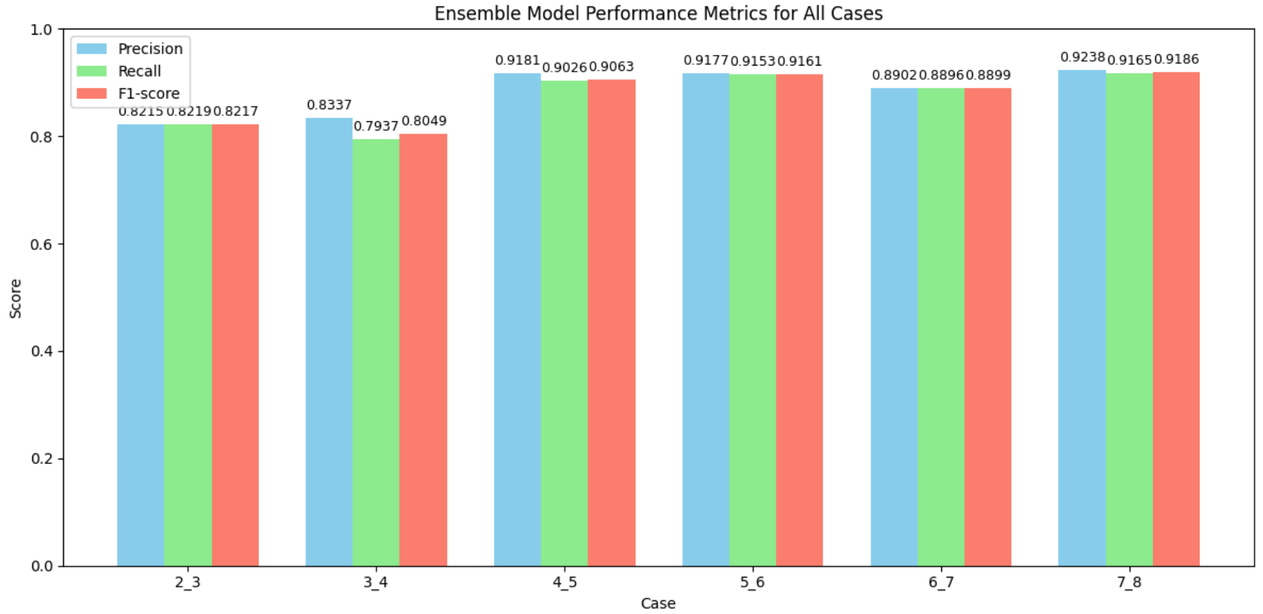
EXTRACTED FEATURES 20 ITEMS

Age · Gender · Race · Weight · BMI · Admission Time · Length of Stay · Baseline Creatinine · eGFR · Urine Output · Vasopressor Amount · # of Diagnoses · # of Procedures · Drug Severity · Drug Mortality · Nephrotoxic · Renal Replacement Therapy · Invasive Mechanical Ventilation · Lab — Hemoglobin · Potassium · … · Vitals — HR · Resp · …

시점 윈도우별 앙상블 성능.

AKI 발생 시점 윈도우를 2→3일, 3→4일, …, 7→8일까지 6가지로 나눠 평가했습니다. 후반 윈도우일수록 lab 데이터가 안정화되어 F1이 단조 증가합니다.

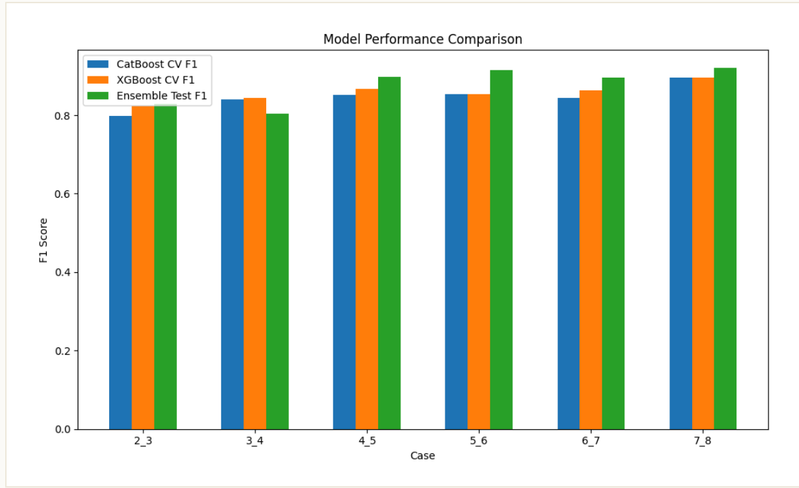
Ensemble Model Performance Metrics for All Cases PRECISION · RECALL · F1-SCORE



HIGHLIGHTS

- **7_8 (입원 7일 → AKI 8일):** F1 **0.9186** · Precision 0.9238 · Recall 0.9165
- **5_6:** F1 **0.9161** · Precision 0.9177 · Recall 0.9153
- **2_3 (가장 이른 시점):** F1 0.8217 — 절대값은 낮지만 임상적으로는 가장 가치 있는 구간

모델별 비교 CV F1 VS ENSEMBLE TEST F1

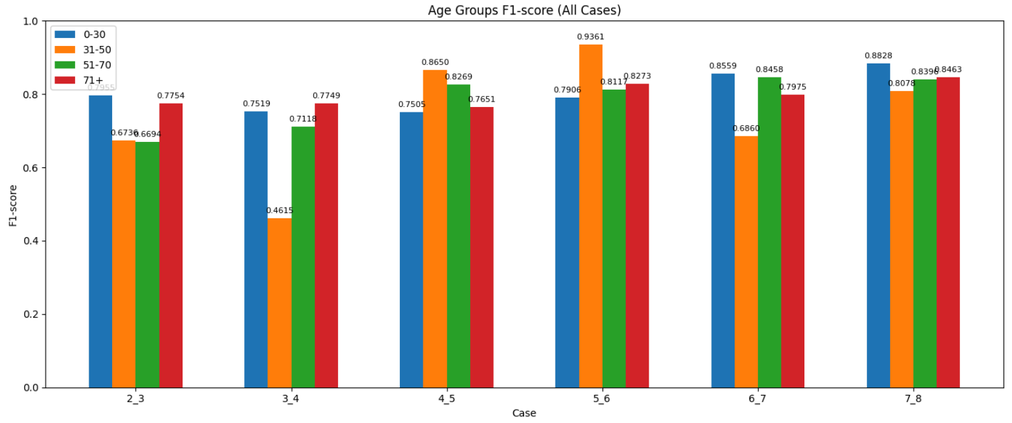


대부분의 윈도우에서 단일 모델 CV보다 앙상블 Test F1이 더 높음을 확인.

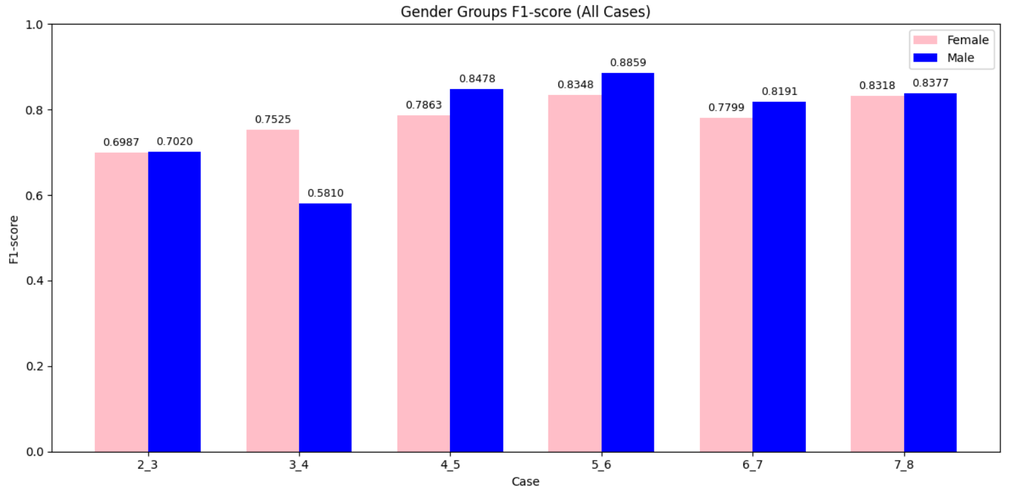
공정성 평가 — 그룹별 F1 분해.

평균 성능만 보고 끝내지 않기 위해, 연령(4개) · 성별(2개) 그룹별로 F1을 분해해 보았습니다. 후반 윈도우로 갈수록 그룹 간 격차가 줄어듭니다.

Age Groups F1-score 0-30 · 31-50 · 51-70 · 71+ ALL CASES



Gender Groups F1-score Female · Male ALL CASES



읽은 점

- **3_4 구간 · 31-50 그룹**에서 F1 0.4615로 큰 dip. 표본 수가 작은 그룹의 분산 영향으로 해석했습니다.
- **5_6 구간 · 31-50 그룹**은 반대로 0.9361로 최고치. 그룹 단위 분산이 크다는 의미입니다.
- **성별 격차**는 후반 윈도우로 갈수록 좁아져 7_8 구간에서 Female 0.8318 / Male 0.8377.

감사합니다.

연구 인턴 참여는 언제든 가능합니다.
대학원에서 본격적으로 연구를 이어가고 싶은 마음이 크고,
한번 이야기 나눠볼 기회가 있다면 진심으로 감사하겠습니다.

NAME	이윤서 · Yoonseo Lee
EMAIL	leeyoonseo@catholic.ac.kr
PORTFOLIO	leeyoonseo.com
PHONE	+82 10-8604-5870
LAB	Applied AI Lab, 가톨릭대학교 · 지도교수 장재연