학습하는 조직 사례

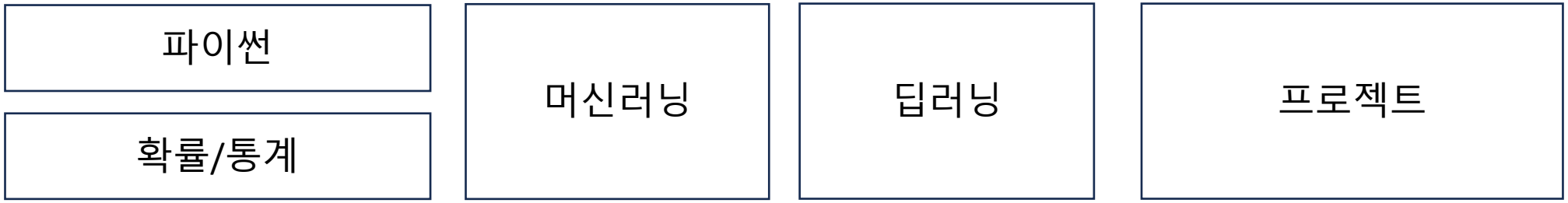
송태영 책임/CDO부문/AI빅데이터담당

Contents

- 연구 모임
- 연구 논문 공유
- Q&A

연구 모임

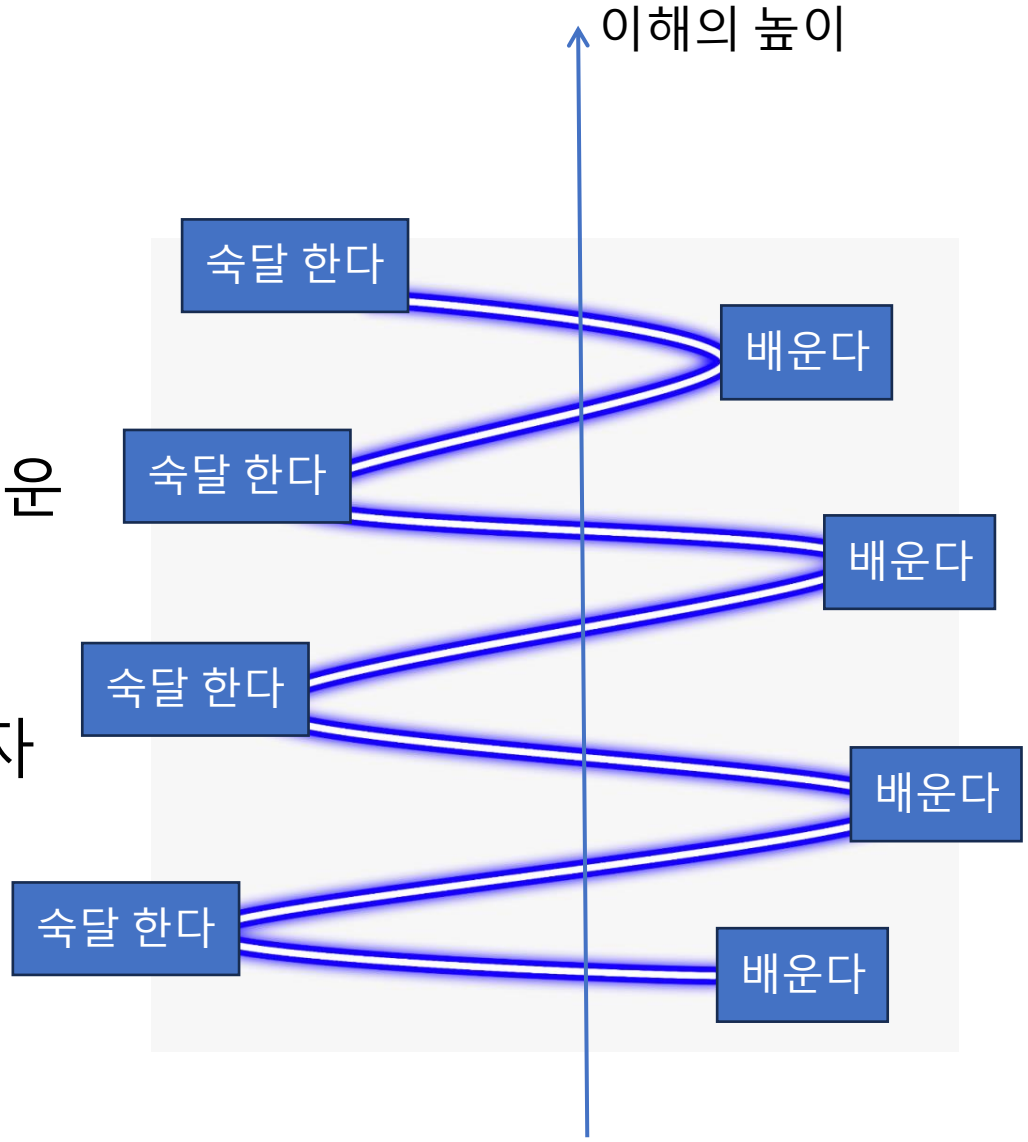
교육 과정을 거치면 생긴 고민



- 2달이란 기간에 (더 이상 젊지 않은 나의 뇌가) 소화 하기엔 너무 많은 내용을 배움
- 어떻게 하면 교육 과정의 경험들을 내 것으로 만들 수 있을까?

이해의 높이를 더 높게

- 배우는 것과 숙달하는 것은 나선형의 관계를 가지는 것이 아닐까?
- 새로운 것을 배운 후 숙달 과정이 있어야 새로운 배움에 대한 시야가 넓어 진다.
- 어떻게 숙달 시킬 것인가?
- 다양한 과제들을 해결을 통해 경험치를 늘리자

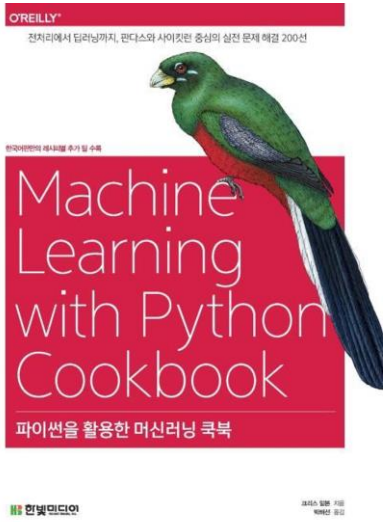


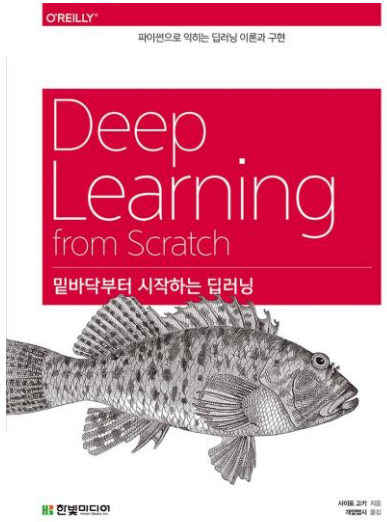
상반기 정리

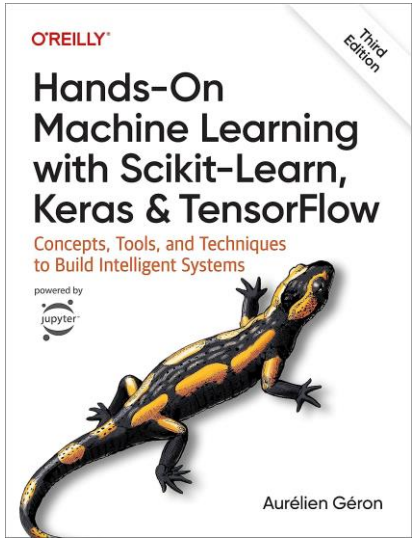
- 교육 과정중에 배웠던 머신러닝과 딥러닝의 이론들을 Kaggle 의 사례를 중심으로 복습
  - 교육에서 배웠던 대부분의 알고리즘을, Kaggle 의 사례(7개)를 통해 복습
  - LLM 등 참가자들의 자발적인 최신 기술 동향 공유
- Pytorch 를 기반으로 하는 딥러닝 개념 및 프로그래밍 숙달
  - Pytorch와 다양한 딥러닝 모델 (13개)

하반기 일정

- 머신러닝과 딥러닝의 이론에 대한 이해도 향상
- 최신 딥러닝 기술들에 대한 이해와 구현
  - NLP, 생성 모델

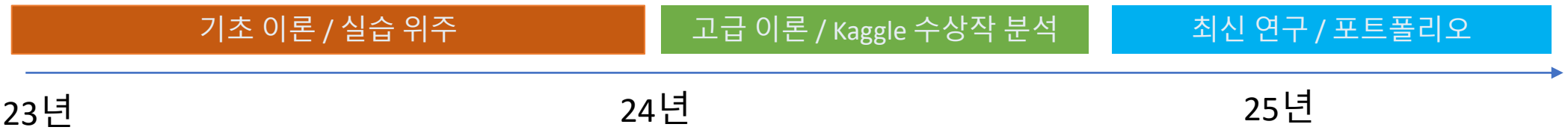






로드맵

- 23년
  - 교육 과정의 내용을 소화 하는데 집중
  - 머신러닝/딥러닝 기본 이론 및 응용 예제 중심 실습
- 24년 상반기
  - 고급 이론서 스터디 및 Kaggle 수상 팀 프로젝트 분석
- 24년 하반기 이후
  - 최신 연구 논문 분석 및 연구 주제 선택



스터디 광고

- 일시 : 매주 토요일 오전 9시 부터 2시간
- 장소 : Webex
- URL : https://github.com/restful3/ds4th_study
- 대상자 : 딥러닝과 머신러닝 기술에 관심이 있는 사람은 누구나
- 현재 참석자 수 : 7명
- 연락처 : 송태영 책임 (tyoung.song@lge.com)

연구 논문 공유

개요

- 제목 : 응급실 환자들의 악화를 예측하기 위한 머신러닝 기반 임상 의사 결정 지원 시스템의 개발
  - Development of a machine learning-based clinical decision support system to predict clinical deterioration in patients visiting the emergency department
  - URL : https://www.nature.com/articles/s41598-023-35617-3
  - 2023년 5월 논문 개재
- 배경 : 2022년 세브란스와 협업한 국책 과제의 결과
- 사용 데이터 :
  - 신촌 세브란스의 5년간의 303,345명의 4,787,121 건의 응급실 데이터
- 결과 :
  - AUROC > 0.9 (과거 6시간의 결과로 1시간 이내를 예측 하는 결과)

배경

- 세계적으로 AI/ML을 활용한 임상 의사 결정 지원 시스템(CDSS)이 개발 되고 있음
- 응급 의료 분야는 의사에 의한 신속한 임상 의사결정이 필요하기 때문에 ML 기반 CDSS 도입에 적합
- 기존의 연구들은 사용하는 데이터가 제한적, 다양한 응급실 환자의 상황에 적용하기 어려움이 있었음
- 본 연구에서는, 의사가 응급실에서 실제로 의사결정에 사용하는 임상 데이터를 활용하여, 실용적인 ML 기반 CDSS 개발 및 임상적 유용성 검증

데이터 수집

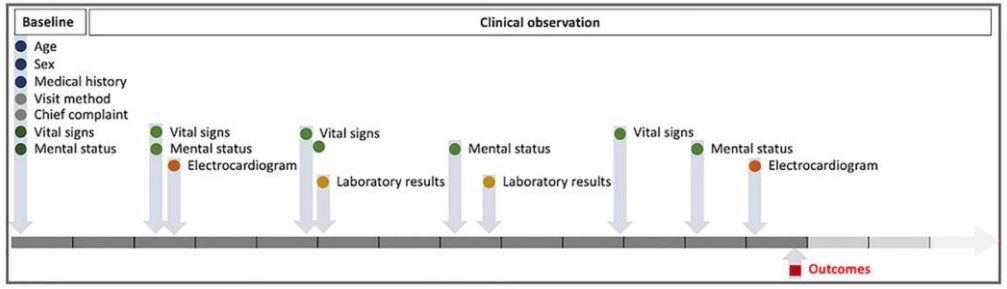
- 병원내 임상 연구 분석 시스템으로 부터, 환자에게 임의 식별 번호 부여, 익명으로 수집
- 독립변수 : 27개의 환자의 고유 (나이, 기존 병력 등) 특징 과 93개의 검사(피검사, 심전도 등) 결과 특징
- 종속변수 : 기도삽관, 중환자실 입원, 승압제 투여, 심정지

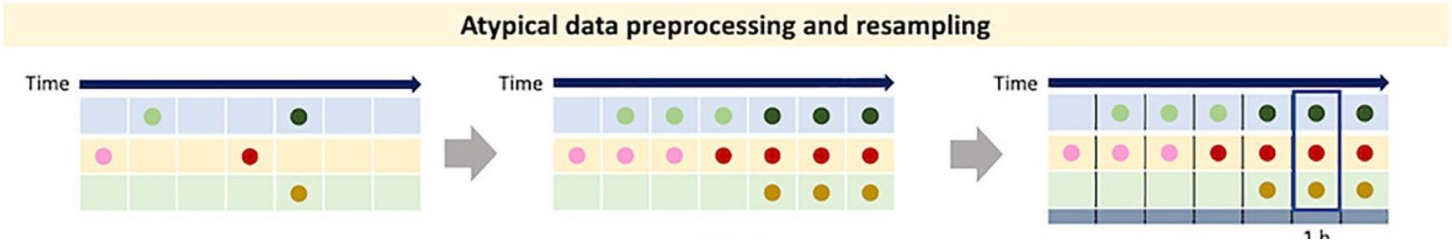
Features	Category	Number	Description
Fixed	Patient information	6	Sex, age, visit method(2), reason of visit(2)
	Past medical histories	7	Hypertension, diabetes mellitus, pulmonary tuberculosis, hepatitis viral carrier, allergies, medications, operations
	Chief complaints and duration	6	Chief complaints categorized into 15 groups; cardiovascular, pulmonary/respiratory, gastrointestinal organs, gastrointestinal bleeding, neurologic, genitourinary, obstetric/gynecological, musculoskeletal, ophthalmologic, ear/nose/throat, skin, orofacial, psychiatric, medical device related, other general issues. If multiple complaints exist in each patient, up to 3 categories (3) and its duration(3) were used
	Vital signs on arrival	8	Systolic blood pressure(2), diastolic blood pressure(2), respiratory rate, pulse rate, oxygen saturation, body temperature
Observation	Clinical observation	6	Systolic blood pressure, diastolic blood pressure, respiratory rate, pulse rate, oxygen saturation, body temperature
	Mental status monitoring	1	Change of mental status; alert, drowsy, stuporous, semi-comatose, comatose, confused and sedated
	Electrocardiogram diagnosis	23	Coronary(5), electrophysiologic(12), metabolic (5), unspecified(1)
	Electrocardiogram metrics	8	P axes, R axes, T axes, PR interval, QRS duration, QT, QTc, ventricular rate
	Laboratory test	55	Arterial blood gas analysis result(12), regular laboratory results(43)

Table 1. Summary of features.

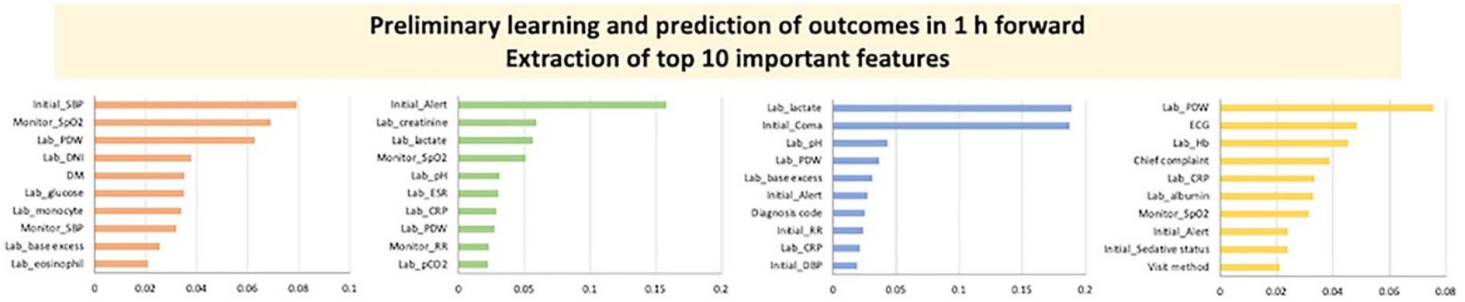
전처리

- 독립변수, 종속변수의 데이터를 환자별로 수집하여 시간에 따라 한시간 간격으로 테이블로 만듦
- 가장 가까운 값을 대입 하는 것으로 결측치 처리
  - 의사도 환자의 가장 최근에 측정된 신체 검사 및 생체 신호에 기초하여 즉각적인 응급 조치의 제공과 같은 의사 결정을 함
- 생체 신호 데이터의 이상치는 생리학적 범위를 적용
- 데이터셋의 불균형 특성 : 학습용 데이터는 언더샘플링 적용, 테스트용은 비적용 (실제 임상 환경과 유사하게 하기 위해)

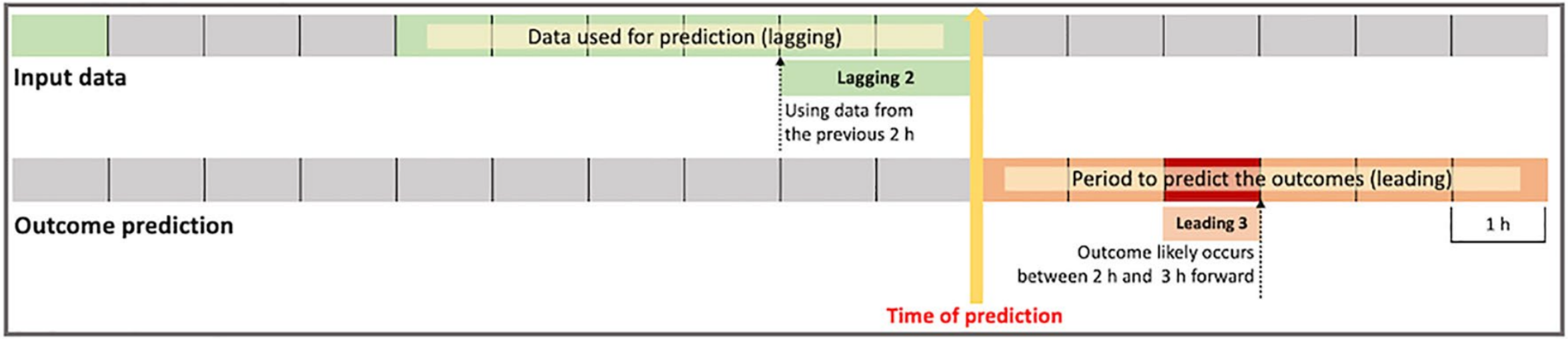




모델 개발



- Base line model : XGBoost으로 종속변수 별로 10개의 중요한 예측 변수 도출
- 학습에 사용할 중요한 예측 변수의 5가지 과거 값, 종속 변수의 5개 미래 예측, 총 25개의 예측 모델을 개발



결과

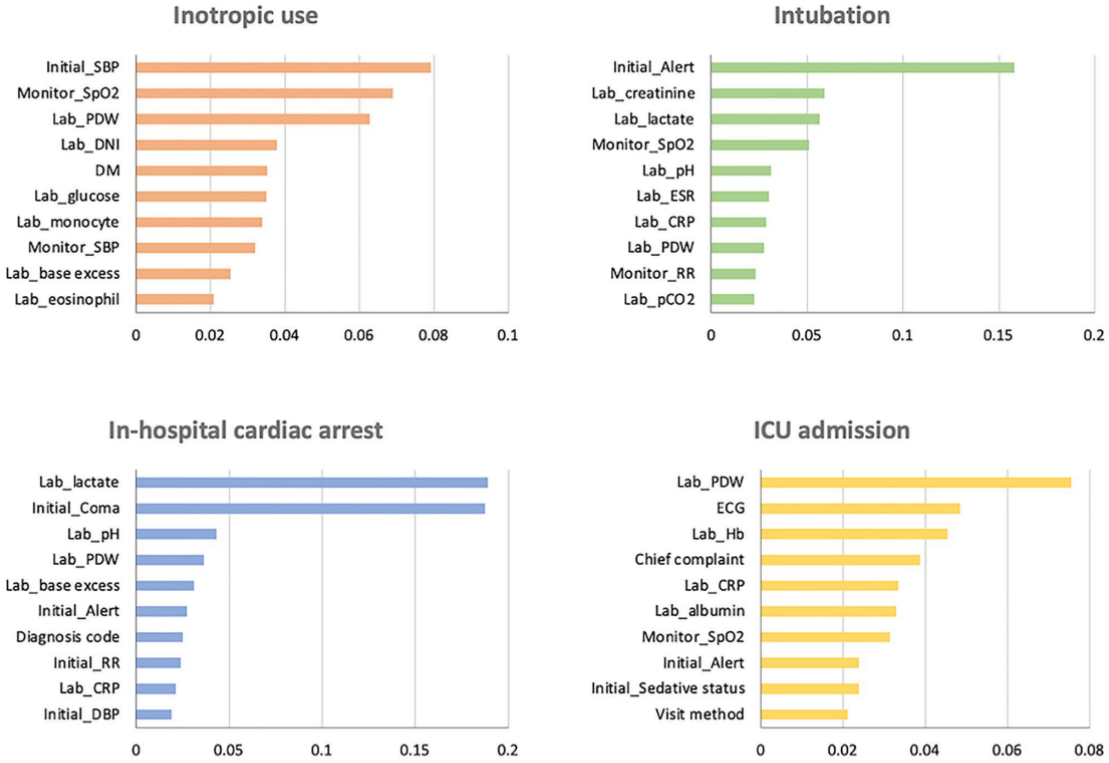


Figure 3. Top 10 important features in predicting each outcome. *ICU* intensive care unit, *SBP* systolic blood pressure, SpO_2 oxygen saturation, *PDW* platelet distribution width, *DNI* delta neutrophil index, *DM* diabetes mellitus, *ESR* erythrocyte sedimentation rate, *CRP* C-reactive protein, *RR* respiratory rate, *DBP* diastolic blood pressure, *ICU* intensive care unit, *ECG* electrocardiogram, *Hb* hemoglobin.

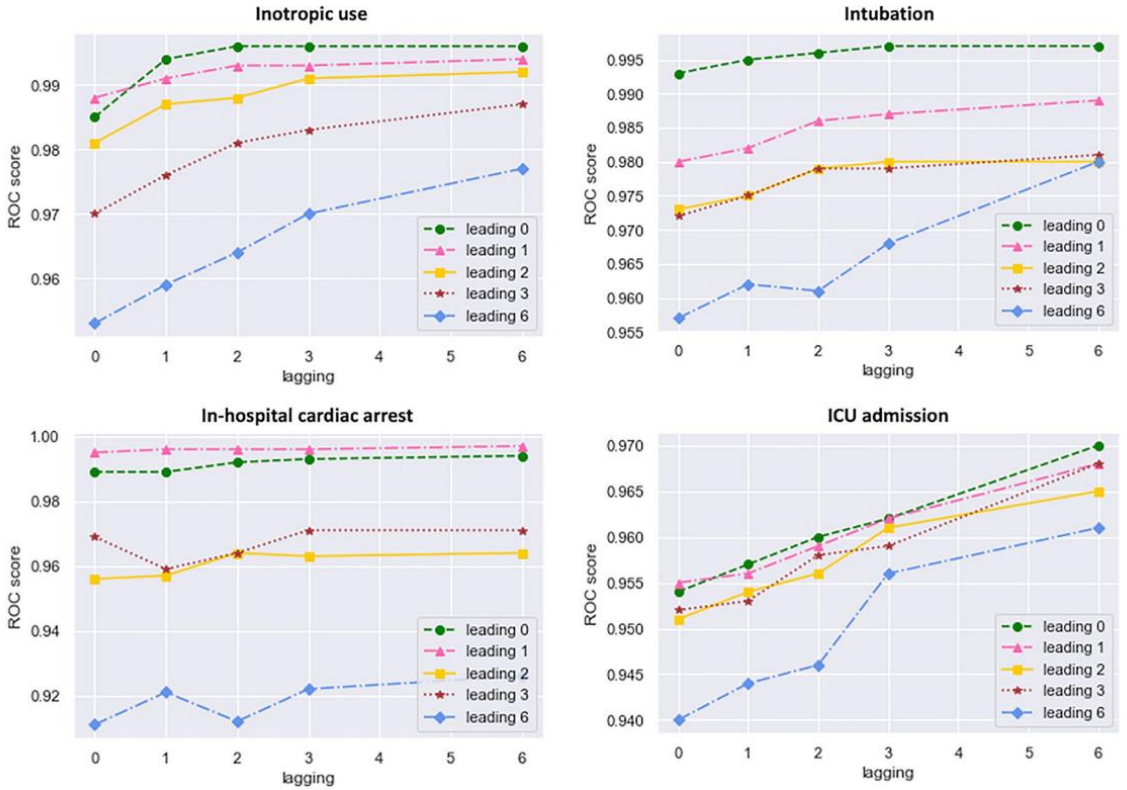


Figure 4. Area under receiver operating curve value for each outcome. *ROC* receiver operating curve, *ICU* intensive care unit.

AUROC는, 더 많은 과거의 정보를 학습 시키고, 예측 하는 미래가 가까울 수록 증가 한다. (심정지 제외)

Contributions

- 본 논문의 :
  - 이전 연구에서는 생명체의 바이탈 사인 같은 몇 가지 고정 변수를 예측 변수로 사용
  - 응급실에서 발생하는 모든 임상 소견, 검사 및 심전도와 같은 변수를 입력 변수로 활용
  - 모든 임상 변수의 업데이트를 최신 관찰로 간주 하여 판단
    - 응급실 의사들이 매 시간마다 환자를 재평가하고 해당 임상 소견을 기반으로 중요한 사건을 예측한다는 가정
    - 응급실 의사의 순차적 의사 결정 프레임워크를 모방하도록 설계
- LG전자의 :
  - 위와 같은 모델 및 데이터 처리의 설계를 제안 및 구현

느낀 점

- 1년간의 프로젝트 기간 중, 10개월을 데이터 전처리에 사용 해야 했음
  - 13개가 넘는 테이블과 200개 가까운 컬럼, 30기가 이상의 text 데이터
- 모델링, 테스트, 문서 정리, 결과 정리 에 2개월
  - 시간적 여유가 있었다면, 트렌스포머 등의 좀 더 성능이 좋은 모델을 검토 할 수 있었을 듯
- 커뮤니케이션이 매우 중요
  - 결정 사항을 꼼꼼하게 문서화 해서 여러 번 도움을 받았음
  - 주고 받은 카톡 까지 캡쳐 하여 문서화

Q&A