Data Analytics TeamProject

환자정보를 이용한 사망위험도 분석

7조

2019204045 윤서환
2019204023 윤성호

Contents

1. 문제 상황 및 데이터 설명

- 문제 상황
- 활용 데이터 설명
- 전처리 및 EDA

2. 실험 내용

- 활용 알고리즘 및 지표
- bias-variance tradeoff
- Training & Optimization

3. 실험 결과 및 해석

- 실험 결과
- 결과 해석

4. 활용 방안

- Rule Extraction with Decision Tree
- 시나리오 분석

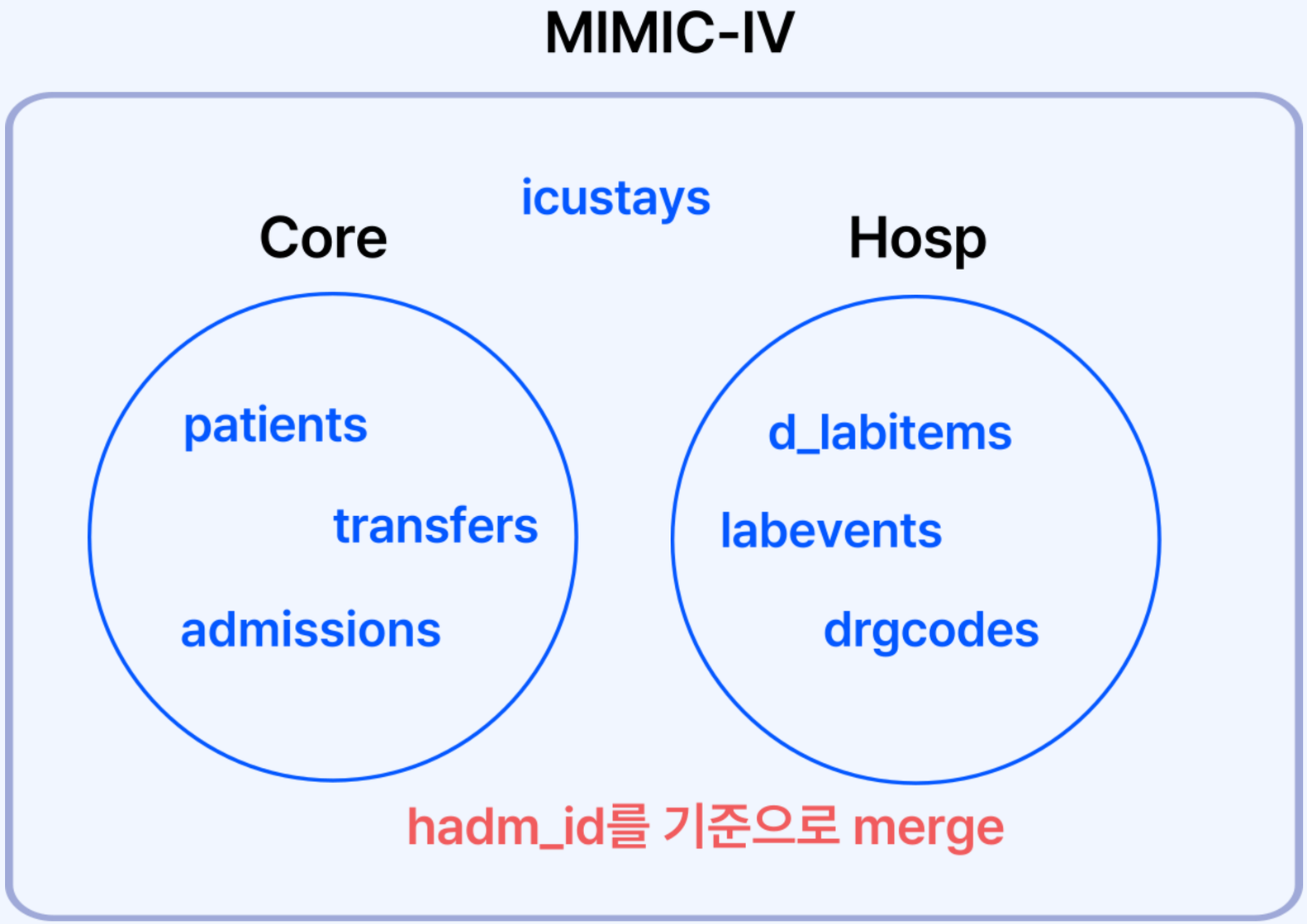
문제 상황

환자의 기본적인 입원 정보 + 의료 검사 결과

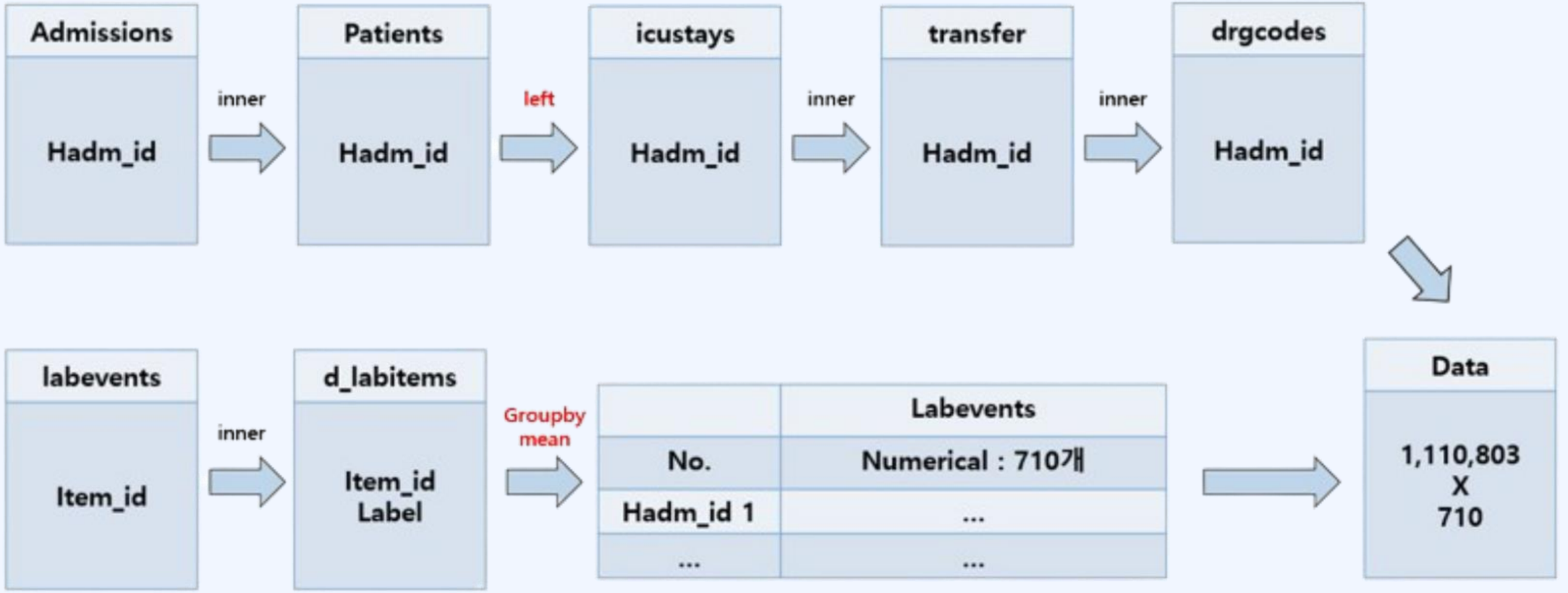
- 환자의 사망 예측 모델링
- 결과에 따른 케어 시나리오 생성

데이터 설명

활용 데이터


MIMIC-IV
icustays
Core
Hosp
patients
transfers
admissions
d_labitems
labevents
drgcodes
hadm_id를 기준으로 merge

데이터 전처리


Admissions
Hadm_id
inner
Patients
Hadm_id
left
icustays
Hadm_id
inner
transfer
Hadm_id
inner
drgcodes
Hadm_id
labevents
Item_id
inner
d_labitems
Item_id
Label
Groupby
mean
Labevents
No.
Numerical : 710개
Hadm_id 1
...
...
...
Data
1,110,803
X
710

데이터 전처리

- discharge_location, dod 등 leakage 위험이 있는 feature 삭제
- 모든 입/퇴원 시간을 퇴원시간-입원시간 값으로 변환, sqrt 적용 후 이산화
- object type features를 one-hot encoding
- 결측치 0 처리, 측정 오류 feature를 drop

활용 알고리즘 및 지표

DL Model	Tree-based Model
MLP	Decision Tree
TabNet	XGBoost

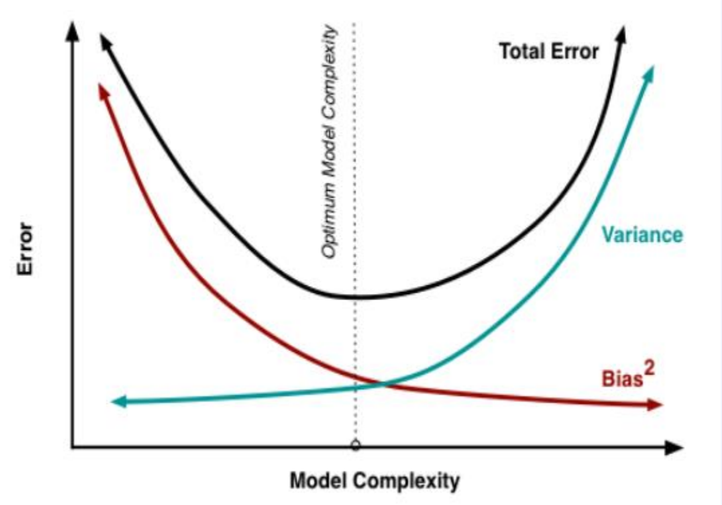
평가 지표는 f1-score, 2종 오류의 심각성을 고려하여 Recall 사용

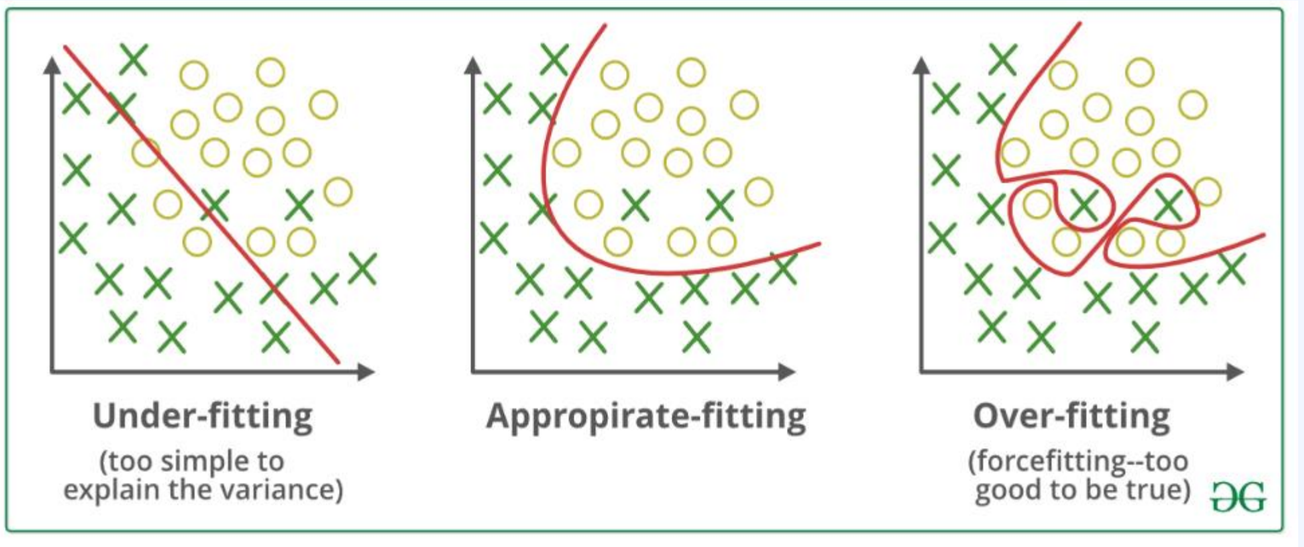
활용 알고리즘 및 지표

DL Model 복잡도가 높은 모델	Tree-based Model
MLP	Decision Tree
TabNet	XGBoost

bias-variance tradeoff를 고려하여 최적의 모델 탐색

Bias-Variance Tradeoff





모델 복잡도 낮음

DT, XGBoost

$Bias\uparrow$ $Variance\downarrow$

모델 복잡도 높음

MLP, TabNet

$Bias\downarrow$ $Variance\uparrow$

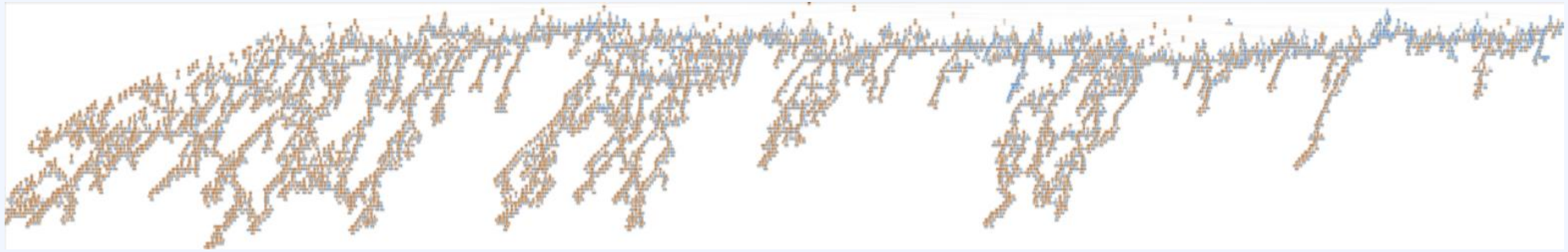
Training & Optimization

- Training
  - train, val, test 비율은 7:1:2
  - Seed, 실험 하드웨어 고정
  - feature importance 도출 후, importance 0인 feature drop
  - DL model에는 scaling 및 oversampling 진행

- Optimization
  - Tree-based model들은 grid search
  - DL model들은 Optuna로 tuning

실험 과정

Baseline - Decision Tree Classifier



```
Accuracy: 1.00
Recall: 0.99
Precision: 0.99
F1-Score: 0.99
Confusion Matrix:
 [[320087    184]
 [   190  12780]]
```

Pruning을 진행하지 않으면?

- Rule 해석의 어려움
- 일반화 성능 감소

⇨ Max Depth 조정 진행

실험 결과

다른 모델과 비교실험 결과

	Decision Tree	XGBoost	MLP	TabNet
f1-score	0.84	0.87	0.07	0.56
Recall	0.74	0.79	0.4	0.88
모델 복잡도 (파라미터 수 / 노드 수)	3351	9766	11697	28888
추론시간(초)	0.43	0.44	0.35	12.49

활용 방안

- **Decision Tree로 상위 Rule Extraction**

  ⇨ 도메인 지식이 있는 의료진에게 인사이트 도출

- **XGBoost로 시나리오 분석 진행**

  ⇨ 중요 변수 변화에 따른 환자 별 사망 위험도 변화 파악

Rule Extraction

```
if (drg_severity > 3.5) and (Lactate > 3.534) and (Bicarbonate <= 19.394) and (Lactate > 3.577) and (Base Excess <= -8.258) and (Prostate Specific Antigen <= 66.7) and (Bilirubin, Total <= 47.239) and (Absolute Neutrophil Count <= 66.033) and (Heparin, LMW <= 0.402) and (pO2 > 37.167) and (Calculated Bicarbonate, Whole Blood <= 19.5) and (Bilirubin, Total <= 23.546) then class: 1 (proba: 100.0%) | based on 1,708 samples
```

환자가 가장 많이 사망한 Rule

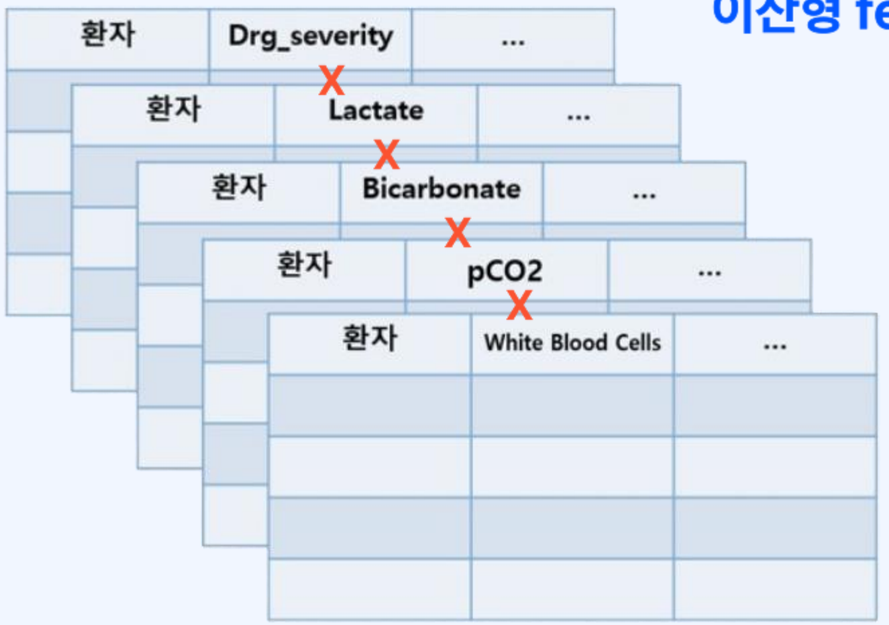
```
if (drg_severity <= 3.5) and (Lactate <= 4.572) and (Temperature <= 11.95) and (pCO2 <= 82.091) and (Sodium <= 148.36) and (los_1 > 0.5) and (Uric Acid <= 18.21) and (Lactate Dehydrogenase (LD) <= 1869.75) and (Alkaline Phosphatase <= 3177.375) and (Bilirubin, Total <= 31.251) and (Hyaline Casts <= 21.833) and (Blasts <= 27.833) and (drg_severity <= 2.5) and (Lactate Dehydrogenase, CSF <= 90.0) and (Creatine Kinase, MB Isoenzyme <= 298.333) and (INR(PT) <= 8.383) and (Gamma Glutamyltransferase <= 851.667) and (Free Calcium <= 1.474) and (Lactate <= 3.675) and (Amylase, Body Fluid <= 10610.0) then class: 0 (proba: 99.95%) | based on 301,803 samples
```

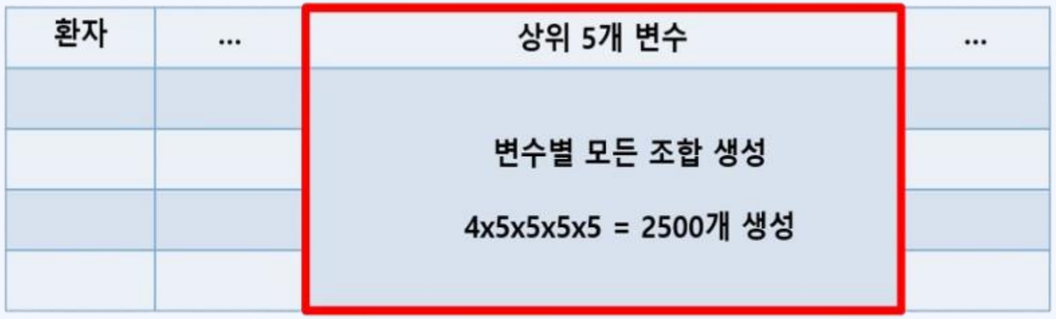
환자가 가장 많이 생존한 Rule

시나리오 분석

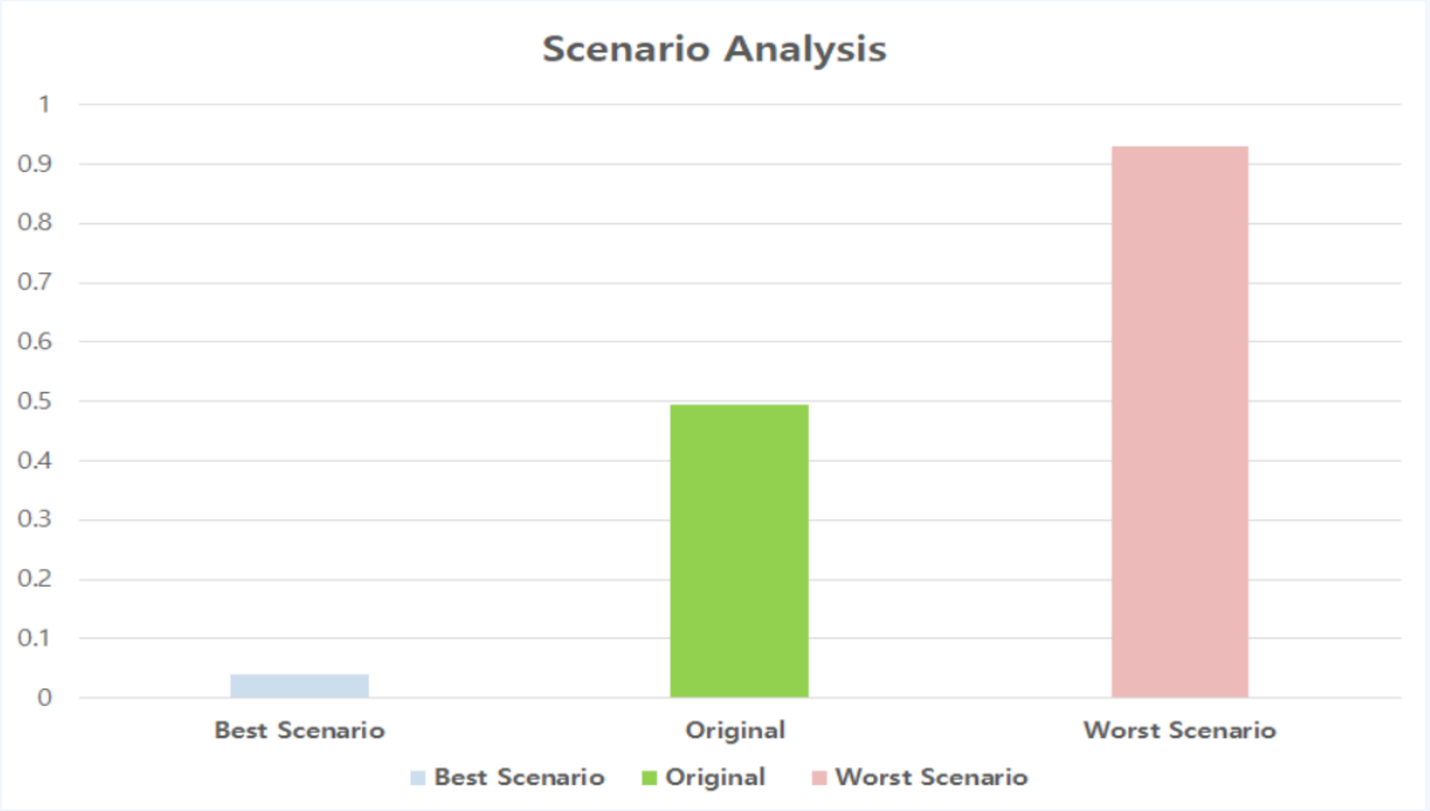
※시나리오 : 1개의 row = 환자 1명에 대한 2500개 조합 중 1가지

연속형 feature는 최대/최소값을 고려한 구간 → 5개의 random sample
이산형 feature는 모든 unique 값





중요 5개 feature의 값들을 변화시켜 조합한 인공 데이터



	Drg_severity	Lactate	Bicarbonate	pCO2	White Blood Cells	Probability
Worst	4	1.78	19.46	42.03	13.59	0.93
Original	4	1.02	26.73	37.26	15.12	0.49
Best	1	0.55	29.02	20.51	5.32	0.04

특정 Feature 값이 변화함에 따라 사망 확률의 변동성 예측 가능

결론 및 한계점

- 결론
  - 환자의 기본적인 의료 데이터와 의료 검사를 통한 예측 모델링
  - DT를 활용한 Rule Extraction으로 인사이트 도출
  - 환자의 사망률에 영향을 많이 미치는 feature들과 패턴 확인 가능

- 한계점
  - 리소스 부족으로 다양한 모델 선택의 제약
  - 다양한 조합의 hyperparameter tuning 어려움

QnA

Thank You

QnA

Q1. DT를 제외한 모델들은 성능 비교에만 사용 된 것인가요?

A. Tree based model인 DT, XGB가 다른 DL model보다 성능이 더 잘 나왔습니다. 특히, 가장 좋은 성능을 보여준model은 XGB였습니다. DT의 경우는 타 모델대비 가장 복잡도가 낮음에도 XGB와 유사한 성능을 보였습니다. 그래서 저희는 DT model의 장점을 살려 rule extraction을 진행했고, 예측 성능이 가장 우수한 XGB를 사용하여, 시나리오 분석을 진행했습니다. 따라서 DT, XGB를 이용했습니다. 나머지 두 모델 MLP, TabNet은 성능 비교에만 사용되었습니다.

QnA

Q2. 환자 개개인이 갖는 데이터들이 다르고 예외적인 모습을 보일 수 있는데 그런 경우에는 본 모델이 어떻게 적용될 수 있을까요?

A. 그렇기 때문에, 시나리오 분석을 진행했습니다. 환자 1명에 대해 주요 변수별 모든 경우를 고려한 dataset을 구축하여, model을 통한 사망 예측을 진행했습니다. 여기서 주요변수는 단순히 feature importance기준으로 선별했고, domain knowledge가 있다면, 해당 변수를 다르게 추출할 수 있습니다. 환자 개개인이 가지는 모든 경우를 고려하여 의사에게 제공될 수 있고, feature와 target관의 관계가 모두 선형적인 특성을 가지지 않기 때문에, 의사가 놓칠 수 있는 부분에 대하여 인사이트를 제공해 줄 수 있습니다.

예외적인 패턴을 탐지하지 못하는 문제는 원래 저희가 구축한 모델 뿐 아니라, data driven approach가 갖는 일반적 한계점이라고 볼 수 있을 것 같습니다. 저희도 그런 한계를 인지하고 있었기 때문에 tree-based 모델을 학습하는 과정에서 pruning을 진행하거나, 상대적으로 복잡도가 낮은 모델을 사용하는 등 오버피팅을 방지하여 최대한 일반화 성능이 높은 모델을 구축하려고 노력하였고, 시나리오 분석을 진행할 때도 예측능력이 상대적으로 우수한 XGBoost를 이용하는 등 다양한 모델을 활용하여 예외적인 패턴들에 대해서도 최대한 대비하려고 하였습니다. (이어서)

QnA

Q2. 환자 개개인이 갖는 데이터들이 다르고 예외적인 모습을 보일 수 있는데 그런 경우에는 본 모델이 어떻게 적용될 수 있을까요?

A. 그러나 앞서 언급했던대로, 학습 시에 보지 못하였던 패턴에 대해서도 강건한 성능을 내기를 기대하는 것은 data driven approach에서는 굉장히 어려운 일이고, 의료데이터의 특성 상 환자의 생명과 직결된만큼, 이 모델만을 절대적으로 맹신하기 보다는 실제로 현장에서 도메인 지식이 있는 전문가들에게 보조적으로 도움을 줄 수 있도록 활용하는 것이 가장 적절한 모델의 활용방법일 것 같습니다.