

방학세미나 후기

정해줘요 학회장팀

김나현 박상훈

INDEX

1. 출제 의도

2. 학회장팀이 짚고 싶은 점


3. 피드백

4. 1등팀 발표

1

출제 의도

DATA


SUPPORT2
External

Linked on 9/14/2023

This dataset comprises 9105 individual critically ill patients across 5 United States medical centers, accessioned throughout 1989-1991 and 1992-1994. Each row concerns hospitalized patient records who met the inclusion and exclusion criteria f...

↓

Dataset Characteristics Tabular, Multivariate	Subject Area Health and Medicine	Associated Tasks Classification, Regression, Other
Feature Type Real, Categorical, Integer	# Instances 9105	# Features 42

환자 기록으로부터 2개월 뒤 기능장애 수준 예측

목적

불균형 클래스에 대한 접근 연습

결측치가 많은 다변량 데이터셋에서의 전처리와 모델 선택 방법 연습

평가지표

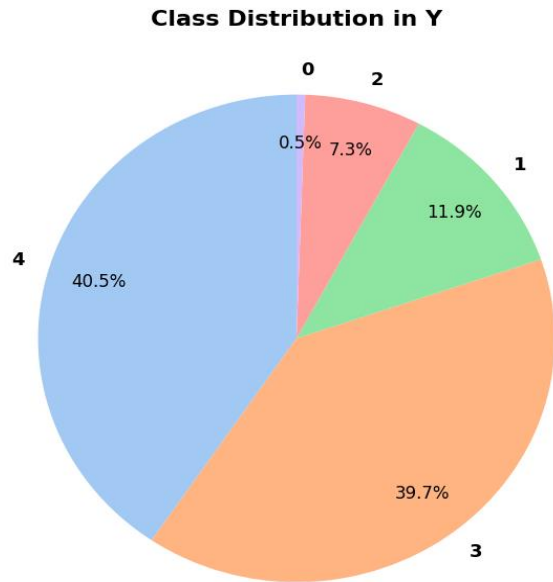
Score Function

: 클래스마다 다른 점수(score) 발생

$$\begin{aligned} \text{Score Function} = \\ -10 \times \text{FN}(\text{Class 4}) + 5 \times \text{TP}(\text{Class 3}) + 100 \times \text{TP}(\text{Class 2}) \\ + 50 \times \text{TP}(\text{Class 1}) + 300 \times \text{TP}(\text{Class 0}) \end{aligned}$$

분류 모델의 성능을 평가하는 지표로 사용

평가지표



Y는 기능장애(functional disability) 수준
(기능장애 심각도는 클래스 번호 순서와 무관)

Class 0 : 삽관 혹은 혼수 상태

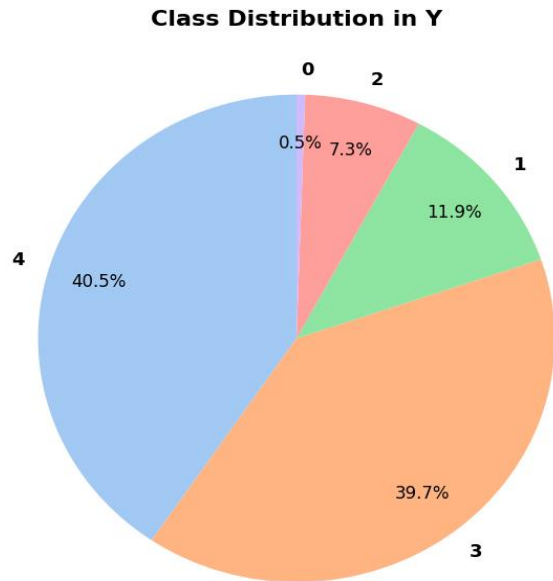
Class 1 : 일상생활 활동 4개 이상 수행 불가능

Class 2 : Sickness Impact Profile 점수 30 이상

Class 3 : 중증도 이상의 기능장애 없음

Class 4 : 사망

평가지표





클래스의 비율이 약 1:12:7:40:40인 수준
Imbalanced Data의 다중 분류 문제

불균형 데이터에 대한 성능 평가에는
일반적으로 Accuracy보다는
(Weighted) F1 Score 등을 사용



Class 0 : 사망
Class 1 : 일상생활 활동 4개 이상 수행 불가능
Class 2 : Sickness Impact Profile 점수 30 이상
Class 3 : 중증도 이상의 기능장애 없음
Class 4 : 사망

평가지표

	실제로 암에 걸린 경우	실제로 암에 걸리지 않은 경우
암으로 진단	제대로 진단하였음	<p>(예시) 병원비를 낭비함, 시간을 낭비함 등등 ...</p>  <p>그래도 건강해서 다행이다!</p>
암으로 진단하지 않음	<p>(예시) 치료를 하지 못해 암이 크게 번짐</p> 	제대로 진단하였음

위 예시처럼, 암인데 암이라고 판단하지 않은 경우 **리스크 ↑**

평가지표

	실제로 암에 걸린 경우	실제로 암에 걸리지 않은 경우
암으로 진단	제대로 진단하였음	<p>(예시) 병원비를 낭비함, 시간을 낭비함 등등 ...</p>  <p>그래도 건강해서 다행이다!</p>
암으로 진단하지 않음	<p>(예시) 치료를 하지 못해 암이 크게 번짐</p> 	제대로 진단하였음

따라서 죽음을 예측하지 못한 경우 리스크가 가해지도록 설정한
Score Function을 구상

분석 과제

불균형 클래스

클래스 0과 1, 2, 3, 4가
1:12:7:40:40 의 비율로 분포
Minor Class에 대한 예측

모델 선택 및 과적합 방지

데이터의 특성에 따른
적절한 모델 선택 및 과적합 방지

NA값 처리

NA값을 대체 또는 제외 필요
EDA를 통해 적절한 대체 필요

EDA의 다양성

대회 중간에 제공된 변수 관련
정보를 활용한 논리적 접근

결측 처리

ph	glucose	bun	urine	adlp	adls	adlsc	death	hospdead	sfdm2	
7.459961				7	7	7	0	0		
7.25					1	1	1	1	<2 mo. follow-up	
7.459961				1	0	0	1	0	<2 mo. follow-up	
				0	0	0	1	0	no(M2 and SIP pres)	
7.509766					2	2	0	0	no(M2 and SIP pres)	
7.65918					1	1	1	1	<2 mo. follow-up	
7.479492				0	1	1	1	0	no(M2 and SIP pres)	
7.509766					0	0	1	0		

원본 데이터에는 이미 결측치가 존재했으며
대회에서 제공된 데이터셋에 없었던 변수들 또한 존재했음

결측 처리

ph	glucose	bun	urine	adlp	adls	adlsc	death	hospdead	sfdm2	
7.459961				7	7	7	0	0		
7.25					1	1	1	1	<2 mo. follow-up	
7.459961				1	0	0	1	0	<2 mo. follow-up	
				0	0	0	1	0	no(M2 and SIP pres)	
7.509766					2	2	0	0	no(M2 and SIP pres)	
7.65918					1	1	1	1	<2 mo. follow-up	
7.479492				0	1	1	1	0	no(M2 and SIP pres)	
7.509766					0	0	1	0		

반응변수로 지정한 Sfdm2를 예측하는 데 사용되기 적절하다고 판단한
설명변수 34개만을 남긴 후 활용

기타

데이터 외적인 background 없이 데이터 자체만으로 인사이트를 얻길 바랐기
때문에, 원본 연구에서 자체적으로 제작한 변수는 모두 삭제



원본 연구 시작 시점에서 알 수 있는 정보로만 환자의 상태를 예측하길 바랐음

기타

데이터 외적인 background 없이 데이터 자체만으로 인사이트를 얻길 바랐기
때문에, 원본 연구에서 자체적으로 제작한 변수는 모두 삭제



- *SUPPORT* 모델 6개월 생존 예측

원본 연구 시작 시점에서 알 수 있는 형질과 환자의 상태를 예측하길 바랐음
등의 컬럼 삭제

기타

- 사망 여부 삭제

- 'dnr' 컬럼 (연명치료포기)의 original value

: {연구 후 DNR, 연구 전 DNR, DNR 없음}

→ 연구 전 DNR만 DNR을 진행한 것으로 간주



원본 연구 시작 시점에서 알 수 있는 정보로만 환자의 상태를 예측하길 바랐음

2

학회장팀이 짚고 싶은 점

불균형 처리

Data-level approaches

Over-sampling

Under-sampling

Hybrid sampling (up + down)

Algorithmic-level approaches

One-class learning

Cost-sensitive learning

Alternate probability cut-off

Ensemble-based approaches

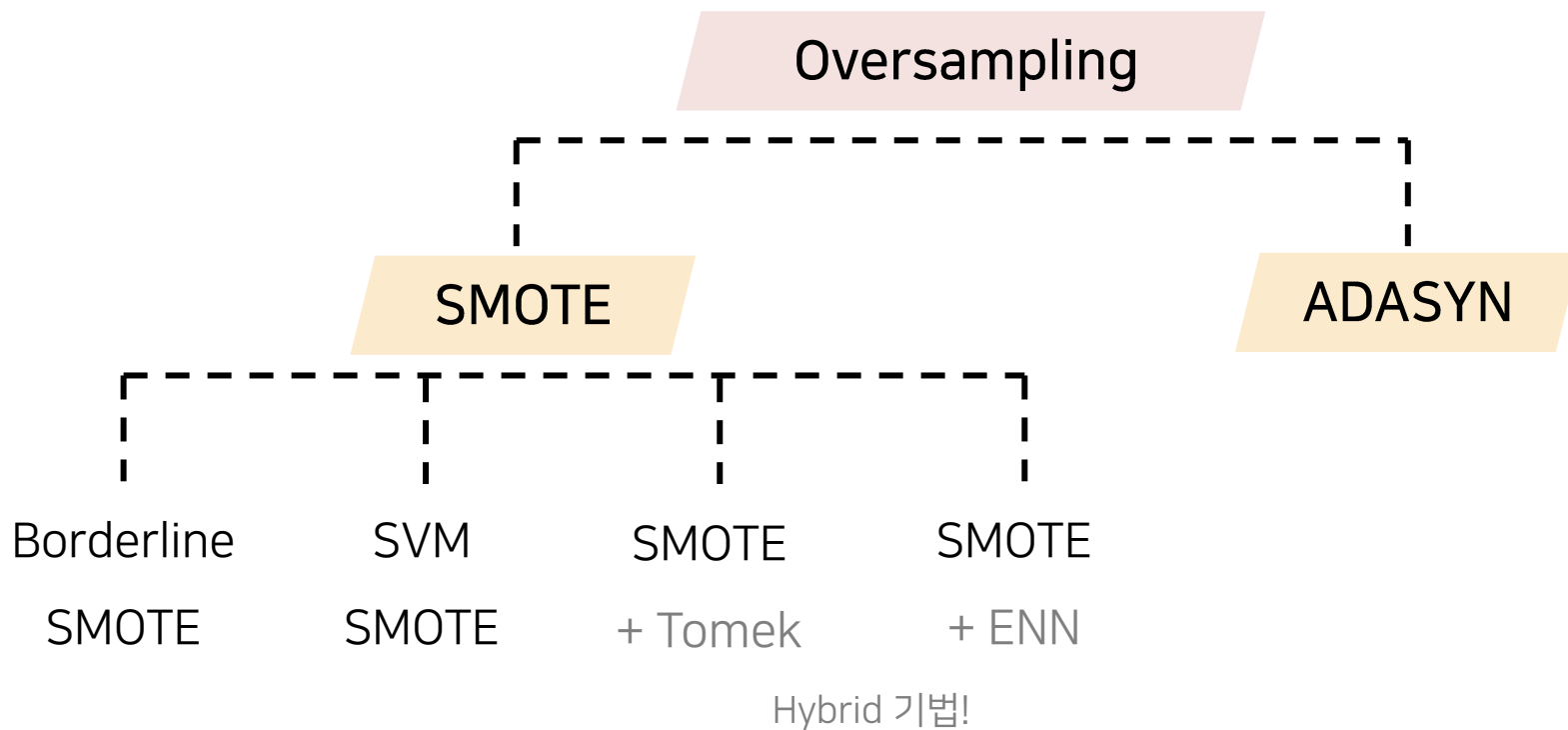
(Boosting) + (Cost-sensitive learning)

이번 데이터셋과 같은 불균형
데이터의 경우, 불균형 처리에
따른 성능 차이가 유의미함



적절한 처리 기법을 선택하는 것이 중요!

불균형 처리



Imblearn 패키지는 SMOTE / ADASYN과 관련하여
다양한 방법들을 제공

불균형 처리 (Highly imbalanced case)

클래스 0 예측의 관점에서 클래스 비율은 매우 불균형하며, Y값 0인 샘플 수 또한 매우 적음



클래스 0 예측을 위해 과도한 Oversampling을 한다면 과적합의 가능성이 커짐

불균형 처리 (Highly imbalanced case)

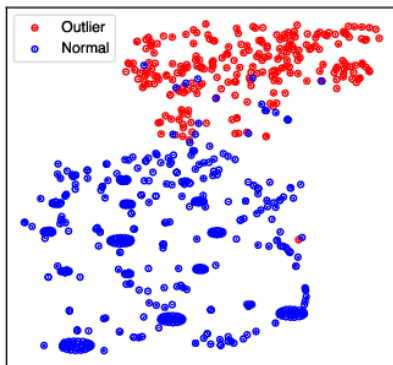
클래스 0 예측의 관점에서 클래스 비율은 매우 불균형하며, Y값 0인 샘플 수 또한 매우 적음



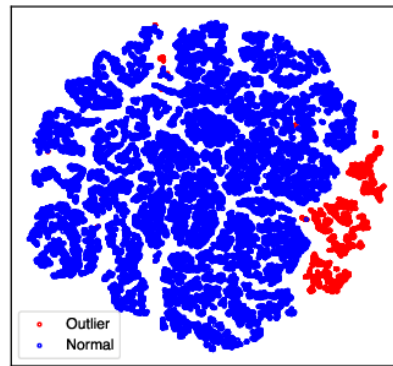
극단적인 클래스 불균형 상황에 맞는 이상치 탐지로 접근해보거나
양상블(0 분류에 더 완화된 기준을 적용하는 보팅)을 사용해보는 것도 좋은 접근 방법

이상치 탐지

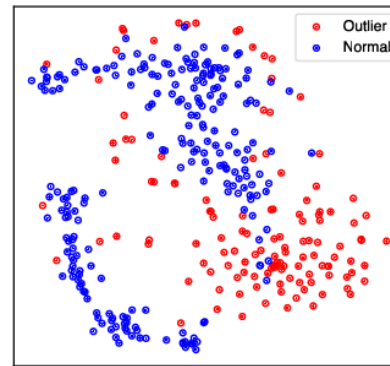
Li, Zheng, et al. "Ecod: Unsupervised outlier detection using empirical cumulative distribution functions." (2022).



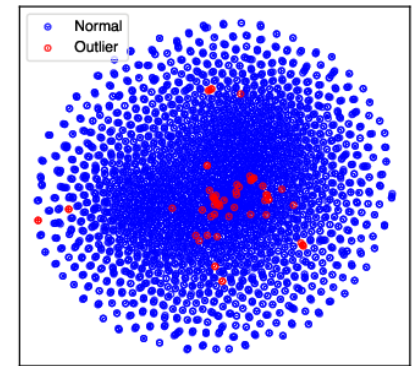
(a) *Breastw (mat)*



(b) *Shuttle (mat)*



(c) *Ionosphere (mat)*

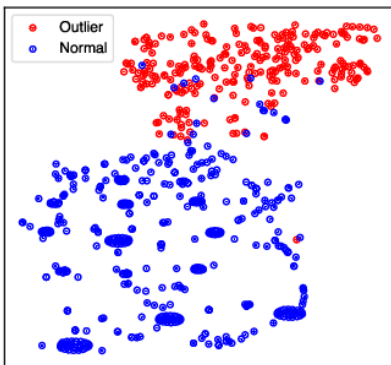


(d) *Speech (mat)*

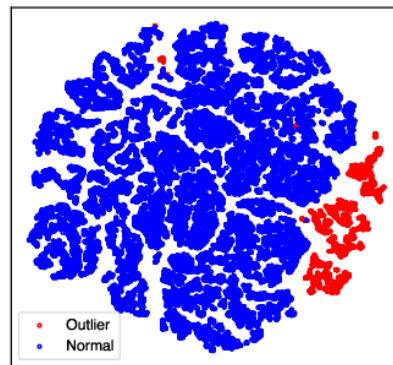
KNN, Isolation Forest, SVDD, Autoencoder,
그리고 2022년 발표된 기법인 ECOD 등 여러 방법 존재

이상치 탐지

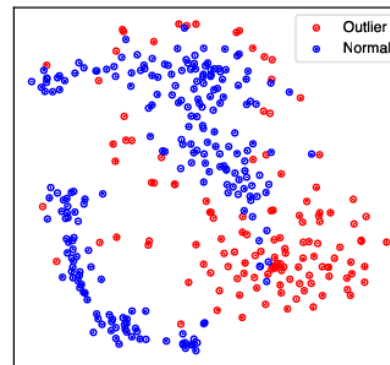
Li, Zheng, et al. "Ecod: Unsupervised outlier detection using empirical cumulative distribution functions." (2022).



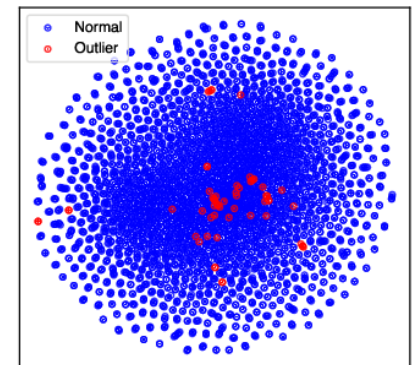
(a) *Breastw (mat)*



(b) *Shuttle (mat)*



(c) *Ionosphere (mat)*



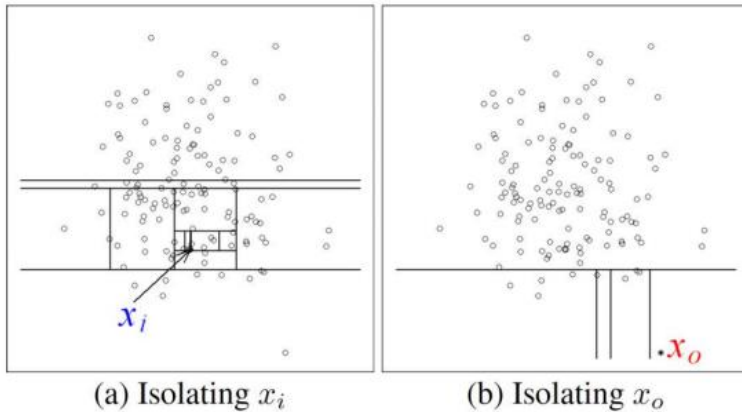
(d) *Speech (mat)*

KNN, **Isolation Forest**, SVDD, Autoencoder,
그리고 2022년 발표된 기법인 ECOD 등 여러 방법 존재

이상치 탐지

Isolation Forest

트리 모델 구조를 가지는 비지도학습 기반 이상치 탐지 모델



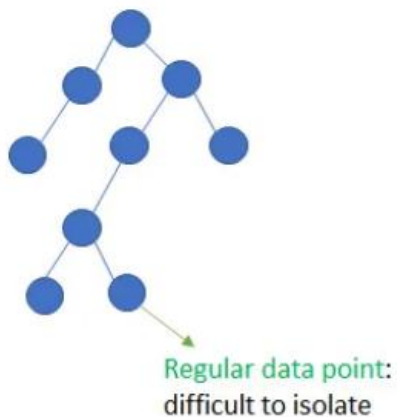
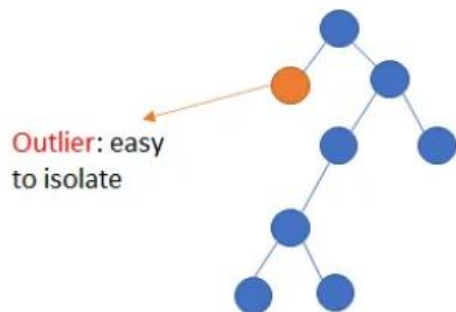
지정된 샘플을 완벽하게 분리하기 위해
필요한 random split의 수를 카운트

이상치 샘플의 경우 **그 split의 수가**
유의미하게 더 적을 것이라는 가정

이상치 탐지

Isolation Forest

트리 모델 구조를 가지는 비지도학습 기반 이상치 탐지 모델



즉 이상치가 속한 **트리 내 노드까지의 평균 path length**가 더 작을 것

이를 점수화하여 경계값을 기준으로
이상치(Anomaly) 판별 가능

비지도학습이므로 경계값은 사용자 지정!

이상치 탐지

Isolation Forest

트리 모델 구조를 가지는 비지도학습 기반 이상치 탐지 모델



즉 이상치가 속한 트리 내 노드까지의
평균 path length가 더 작을 것

이를 점수화하여 경계값을 기준으로

이상치(Anomaly) 판별 가능

비지도학습이므로 경계값은 사용자 지정!

Score Function

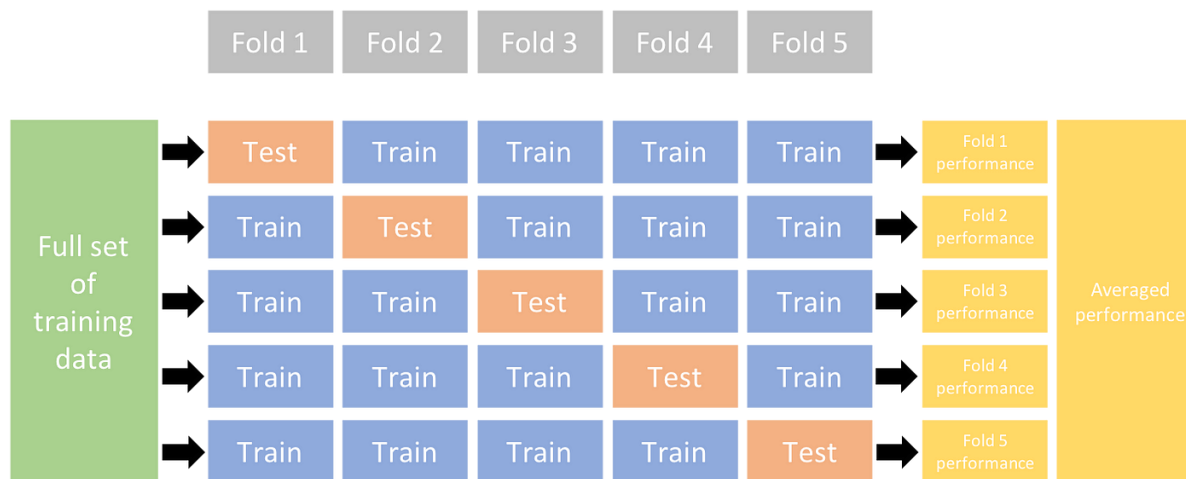
Accuracy, F1-Score 등으로 성능을 어림짐작하는 것은 위험한 선택



반드시 문제에서 주어진 Score-Function을 통해서 optuna 구성할 것

Validation

랜덤 시드가 고정된 상황에서 단순 holdout validation을 신뢰할 것인가?



과적합 방지를 위해 K-CV, LooCV등 다양한 전략을 사용하여 모델 구성하기

모델 선택

항상 머신러닝 or 딥러닝 중 택1 해야하는 것은 아님!

Model Name	YearPrediction	MSLR	Epsilon	Shrutime	Blastchar
XGBoost	77.98 ± 0.11	$55.43 \pm 2e-2$	$11.12 \pm 3e-2$	13.82 ± 0.19	20.39 ± 0.21
NODE	76.39 ± 0.13	$55.72 \pm 3e-2$	$10.39 \pm 1e-2$	14.61 ± 0.10	21.40 ± 0.25
DNF-Net	81.21 ± 0.18	$56.83 \pm 3e-2$	$12.23 \pm 4e-2$	16.8 ± 0.09	27.91 ± 0.17
TabNet	83.19 ± 0.19	$56.04 \pm 1e-2$	$11.92 \pm 3e-2$	14.94 ± 0.13	23.72 ± 0.19
1D-CNN	78.94 ± 0.14	$55.97 \pm 4e-2$	$11.08 \pm 6e-2$	15.31 ± 0.16	24.68 ± 0.22
Simple Ensemble	78.01 ± 0.17	$55.46 \pm 4e-2$	$11.07 \pm 4e-2$	13.61 ± 0.14	21.18 ± 0.17
Deep Ensemble w/o XGBoost	78.99 ± 0.11	$55.59 \pm 3e-2$	$10.95 \pm 1e-2$	14.69 ± 0.11	24.25 ± 0.22
Deep Ensemble w XGBoost	76.19 ± 0.21	$55.38 \pm 1e-2$	$11.18 \pm 1e-2$	13.10 ± 0.15	20.18 ± 0.16

NODE

New datasets

Table 2: **Test results on tabular datasets.** Presenting the performance for each model. MSE is presented for the YearPrediction and Rossman datasets, and cross-entropy loss (with 100X factor) is presented for the other datasets. The papers that used these datasets are indicated below the table. The values are the averages of four training runs (lower value is better), along with the standard error of the mean (SEM)

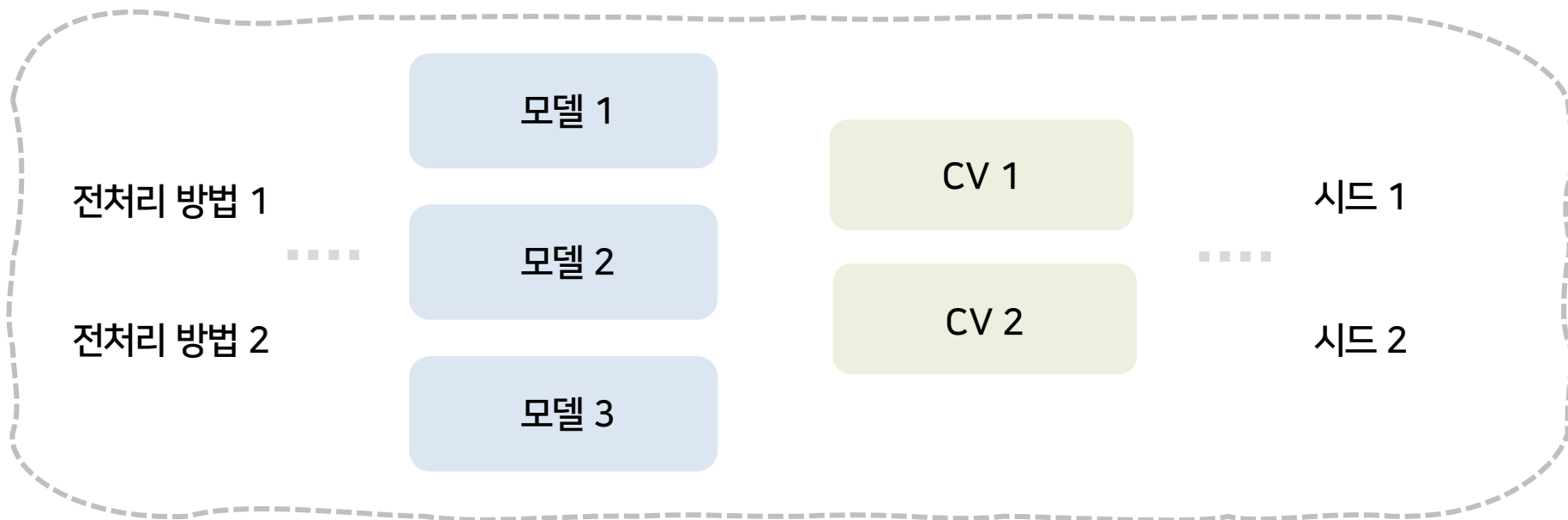
DL + ML이 대표적인 좋은 앙상블의 예시임

출처 : TabNet: Attentive Interpretable Tabular Learning

2

학회장팀이 짚고 싶은 점

앙상블



앙상블을 적용하여 **일반화 성능을 향상을 기대하여,**

Public score와 Private Score의 차이를 줄일 수 있음

불균형 데이터의 경우, 10개의 모델 중 3,4개 이상이 1인 경우 1로 두는 등

다양한 임계값 방법을 적용하여 성능을 올릴 수 있음!

3

피드백

공통

Test 데이터는 편의를 위해 Set으로 구성된 것이지만,
실제로는 개별적으로 다가오는 것입니다!

Test 데이터에 대해 결측값 대체나 scaling 등의 전처리를 진행할 경우
Test 셋에서 최댓값이나 분위수를 이용하는 것, `fit_transform()` 을 한번에 적용하는 것은
Test의 분포를 이용하는 것이므로 **Data Leakage**



만약 Test 데이터에 대한 전처리를 진행할 경우,
Train 데이터로 학습을 시킨 후 따로 모델을 대체하거나
Train 데이터에서 해당 통계량을 저장한 뒤 Test의 전처리에 사용했어야 함!

공통

'논리성' 항목 → 변수명을 공개하는 의의가 더 반영이 되었다면 어땠을까..!

데이터 분석 직무 / 공모전의 경우,
모델의 성능도 중요하겠지만 성능이 다소 낮더라도
분석 과정에서의 논리가 중요함

이근백 교수님+선배님들의 조언..

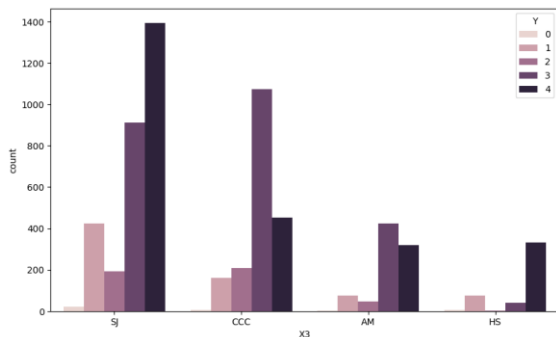
"... 적어도 성능이 절대적인 핵심이 아닌 주제였다는게 드러난 것 같습니다."

"... 도메인 지식에 대한 사고를 필요로 하는 질문을 많이 하셨습니다. "

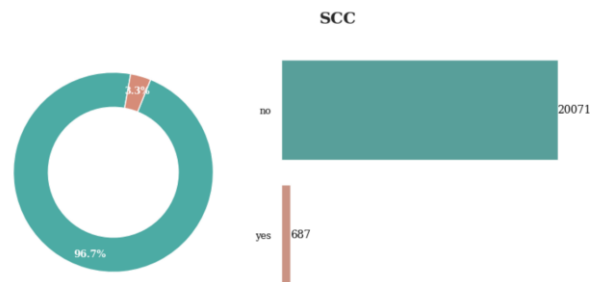
"분석 결과에 대해 어떻게 활용할 수 있는지"

공통

'논리성' 항목 → 변수명을 공개하는 의의가 더 반영이 되었다면 어땠을까..!



범주형 변수의 x값에 따른 그룹화



EDA에서의 도메인 관련 추가 설명

☆ 96.7% don't bother monitoring calorie consumption.
Counting calories? Only folks who don't truly appreciate the art of savoring food would do that, right? 🤔

공통

일부 마스킹된 데이터의 어려움에도 불구하고,
Label과 범주형 자료의 특성에 대해 파악하고 고민한 과정들,
NA를 처리하기 위해 여러가지 시도를 하신 과정들이 두 팀 모두 정말 인상 깊었습니다!

또 클래스 불균형에 대해서
치열하게 고민하신 과정도 잘 드러났다고 생각합니다 ㅎㅎ
앞으로 모델링할 때 방학 세미나 경험이 많은 도움이 되었으면 좋겠습니다~!~!

코드 역시 마크다운을 이용해 깔끔히 정리해주신 덕분에 채점이 편했어요 ㅎㅎ 감사합니다.
그리고 무엇보다 같은 기수끼리 친해지자는 목적에서 보너스 점수를 넣었는데,
팀 구분 없이 모두 가까워지신 거 같아 정말정말 뿌듯합니다 ^__^

4

1등 팀 발표

두구두구두구





1팀



김수진 김준영 이채은 김재원 이나연

축하드립니다~~!



다들 한 주 동안 너무
고생 많으셨습니다!

