

# 방학세미나 후기

스마일 학회장팀

김보근 이정환

# INDEX

---

1. 출제 의도
2. 방법론 참고
3. 피드백
4. 1등팀 발표

1

출제 의도

## DATA



### Taiwanese Bankruptcy Prediction

Donated on 6/27/2020

The data were collected from the Taiwan Economic Journal for the years 1999 to 2009. Company bankruptcy was defined based on the business regulations of the Taiwan Stock Exchange.

#### Dataset Characteristics

Multivariate

#### Subject Area

Business

#### Associated Tasks

Classification

#### Feature Type

Integer

#### # Instances

6819

#### # Features

96

Taiwan Economic Journal로부터 수집된 Taiwan 회사의 파산 여부 데이터

**여러 Feature로부터 회사의 파산 여부 예측**

1 : 파산 / 0 : 파산 X

## 목적

**불균형 클래스**에 대한 접근 연습

변수가 많고 결측치가 존재하는 데이터에 대한 **Feature Engineering** 연습

## 평가지표

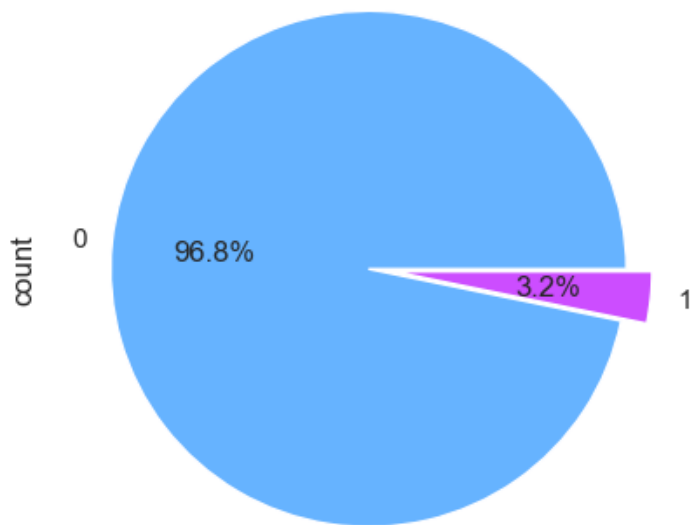
### Cost Function

: 클래스마다 다른 비용 발생

$$\text{Cost Function} = 300 \times FN + 10 \times FP$$

분류 모델의 성능을 평가하는 지표로 사용

## 평가지표



클래스의 분포가 약 30:1인

**Imbalanced Data**의 분류 문제

불균형 데이터에 대한 성능 평가에는



= 300 일반적으로 Accuracy보다는

F1-Score 등 다른 평가 지표 사용

여러 평가 지표에 대해서는 범주팀 클린업 3주차 참고!

분류 모델의 성능을 평가하는 지표로 사용

## 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

그러나 위 예시처럼, 회사가 파산하는 경우는 **흔치 않지만 예측 실패 시 비용 ↑**

Cost-sensitive 문제로도 볼 수 있음

따라서 FN에 대한 비용을 높게 설정한 Cost Function을 통해 성능 평가

## 분석 과제

### 불균형 클래스

0과 1이 30:1의 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

### 차원 축소 문제

95개의 변수가 존재  
변수 필터링, 변수 선택 등 이용



## 분석 과제

Main!

불균형 클래스

0과 1이 30:1의 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

차원 축소 문제

95개의 변수가 존재  
변수 필터링, 변수 선택 등 이용

## 결측 처리



### Taiwanese Bankruptcy Prediction

Donated on 6/27/2020

The data were collected from the Taiwan Economic Journal for the years 1999 to 2009. Company bankruptcy was defined based on the business regulations of the Taiwan Stock Exchange.

#### Dataset Characteristics

Multivariate

#### Subject Area

Business

#### Associated Tasks

Classification

#### Feature Type

Integer

#### # Instances

6819

#### # Features

96

원본 데이터는 결측치가 없는 문제...

→ 직접 만들었습니다!

EDA를 통해 확인한 변수 간 관계나 모델의 특성을 이용해

논리적으로 결측값을 처리할 수 있는지 보고 싶었음

## 결측 처리



### Taiwanese Bankruptcy Prediction

Donated on 6/27/2020

The data were collected from the Taiwan Economic Journal for the years 1999 to 2009. Company bankruptcy was defined based on the business regulations of the Taiwan Stock Exchange.

#### Dataset Characteristics

Multivariate

#### Subject Area

Business

#### Associated Tasks

Classification

#### Feature Type

Integer

#### # Instances

6819

#### # Features

96

원본 데이터는 결측치가 없는 문제...

→ 직접 만들었습니다!

EDA를 통해 확인한 변수 간 관계나 모델의 특성을 이용해  
논리적으로 결측값을 처리할 수 있는지 보고 싶었음



# 2

방법론 참고

## 불균형 클래스

클래스가 30:1로 매우 불균형한 상황이며, Target 값이 1인 개수 자체도 적은 상황



이처럼 클래스 불균형이 매우 심각하고 데이터 수 자체도 적은 상황에서는  
과도한 Oversampling을 통해 증강하면, 데이터가 과적합될 가능성 ↑

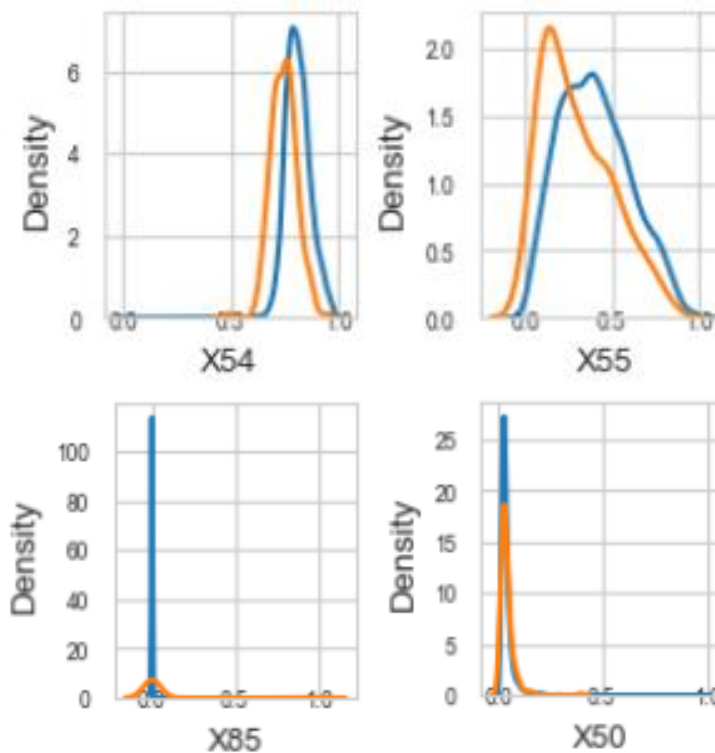
## 불균형 클래스

클래스가 30:1로 매우 불균형한 상황이며, Target 값이 1인 개수 자체도 적은 상황



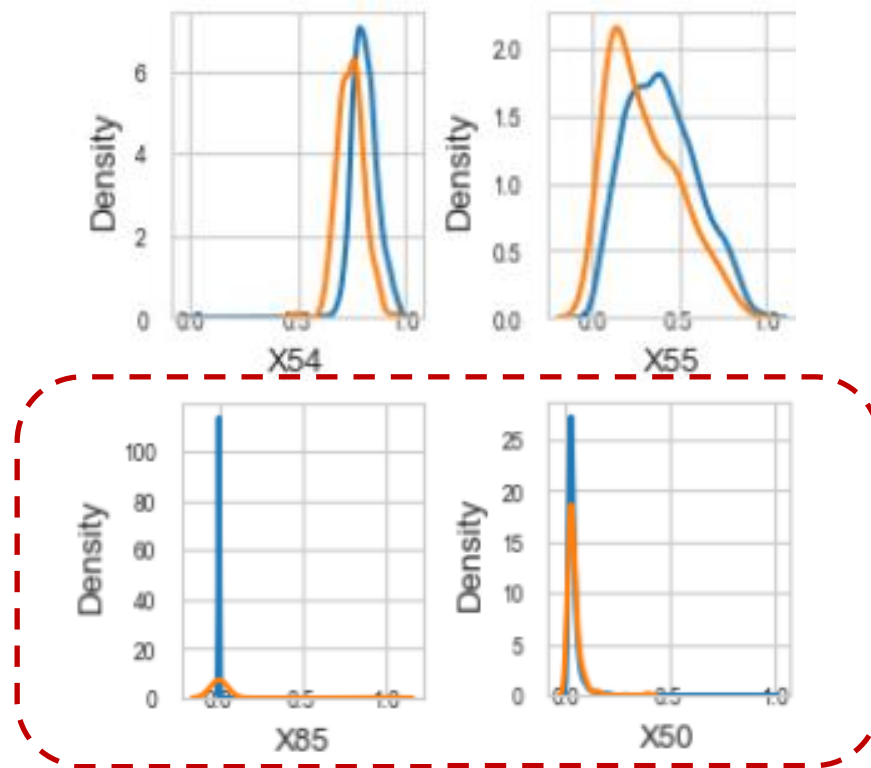
이상치 탐지로 접근해보거나 evaluation metric 또는 threshold 변경,  
양상블(1 분류에 더 완화된 기준을 적용하는 보팅)을 사용해보는 것도 좋은 접근 방법

## 변수 선택



Target에 따른 변수의 분포를 확인했을 때  
대체로 Target 값에 따라 분포가 달라 보이지만,  
Target값에 관계없이 비슷해 보이는 분포가 나타남을 확인

## 변수 선택



따라서, 동질성 검정을 진행하여  
Target 값에 관계없이 비슷한 분포를 제거 가능



## 변수 선택 - 동질성 검정

### 범주형 : 카이 제곱 검정

$H_0$ : 비교하는 두 분포가 동질적이다

$P\text{-value} < \alpha$

→ 비교하는 분포 이질성 존재

### 연속형 : Kolmogorov Smirnov 검정

$H_0$ : 비교하는 두 분포가 동질적이다

$P\text{-value} < \alpha$

→ 비교하는 분포 이질성 존재



귀무가설 기각 X : Target의 0/1에 대한 두 분포가 **동질적**

→ Target을 결정짓는 데에 중요한 요인으로 작용하지 않을 것임을 예상할 수 있음

이번 데이터 셋에서는 ['X50', 'X51', 'X85', 'X94'] 가 해당 ( $\alpha = 0.15$  기준)

※ Target 1의 개수가 매우 적은 상황이기 때문에 기준을 높인 0.15를 선택했습니다!

## 결측 처리

### Single Imputation

상수, 평균, 중앙값 등으로 대체  
Imputation 불확실성 존재하지만  
연산이 매우 빠름

### Multiple Imputation

여러 imputed dataset의 평균으로 대체  
Imputation 불확실성 설명 가능하지만  
연산 시간이 오래 걸림

Multiple Imputation의 대표적인 방법인 MICE는 성능이 좋지만  
오랜 시간이 걸린다는 한계가 존재함

→ EDA를 통해 데이터의 특징을 확인하고, 그에 맞는 Imputation 고려 가능!

## 결측 처리

### Single Imputation

상수, 평균, 중앙값 등으로 대체

### Multiple Imputation

여러 imputed dataset의 평균으로 대체

상관관계 확인 결과, 'X1' 과 'X7' 변수는 상관관계가 매우 높은(0.99x) 변수가 존재  
이 점을 이용해 결측을 보간하거나, 아예 제거하는 것도 좋았다고 판단됨

'X51' 변수의 경우에는 KS-test 에서 Target에 대해 동질적인 분포로 판단되었고,  
결측 비율이 약 0.8 정도로 높기 때문에 제거하는 것이 실제 성능 향상에 도움이 되었음

→ EDA를 통해 데이터의 특징을 확인하고, 그에 맞는 Imputation 고려 가능!

다 의도가 있었던 결측 생성이었습니다 ㅎㅎㅎ ..!

## 결측 처리

### Single Imputation

상수, 평균, 중앙값 등으로 대체  
Imputation 불확실성 존재하지만  
연산이 매우 빠름

### Multiple Imputation

여러 imputed dataset의 평균으로 대체  
Imputation 불확실성 설명 가능하지만  
연산 시간이 오래 걸림

Multiple Imputation의 대표적인 방법인 MICE는 성능이 좋지만

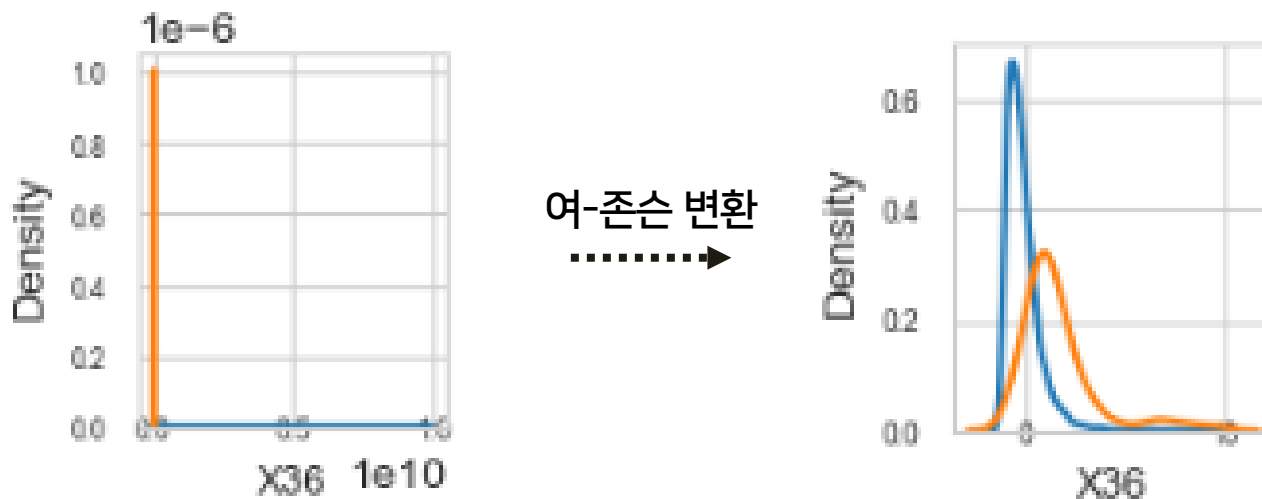
오랜 시간이 걸린다는 한계가 존재함

보간 전후의 분포를 확인해 분포가 왜곡되지 않는지 점검도 필요

→ 데이터의 특징에 맞는 Imputation 고려 가능!

## 결측 처리

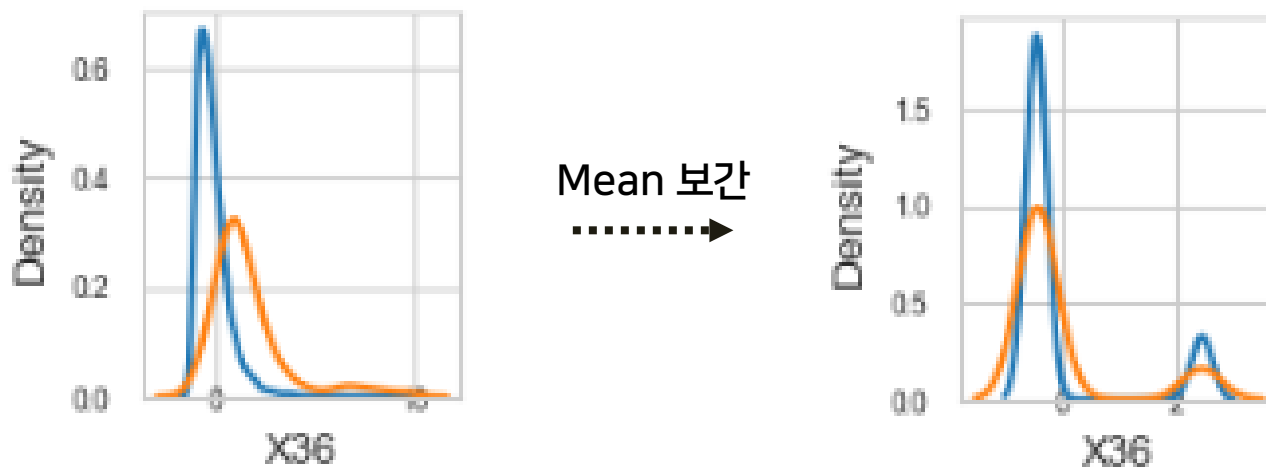
EXAMPLE



우선 변수에 존재하는 extreme한 값들 때문에 분포가 잘 보이지 않을 경우,  
Yeo-Johnson 등의 scaling으로 분포를 확인할 수 있을 것임

## 결측 처리

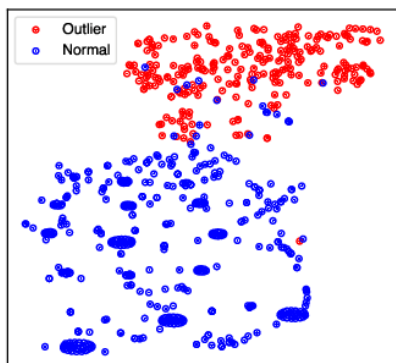
EXAMPLE



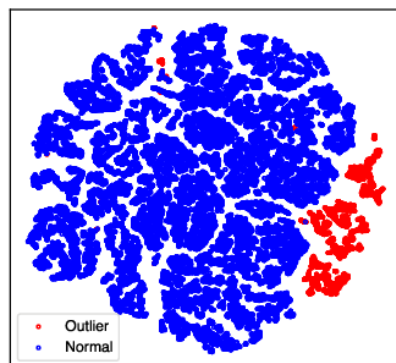
'X36' 변수의 경우, 극단적 값으로 인해 평균으로 보간 시, 분포가 왜곡되는 결과  
→ Median 등으로 대체하는 것이 더 적절한 보간법이었을 것

## 이상치 탐지

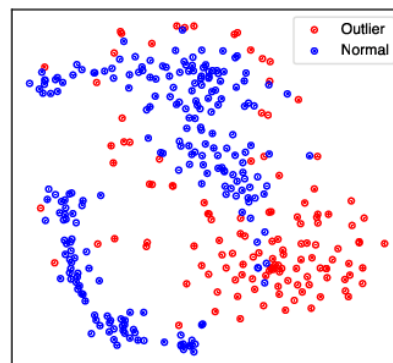
Li, Zheng, et al. "Ecod: Unsupervised outlier detection using empirical cumulative distribution functions." (2022).



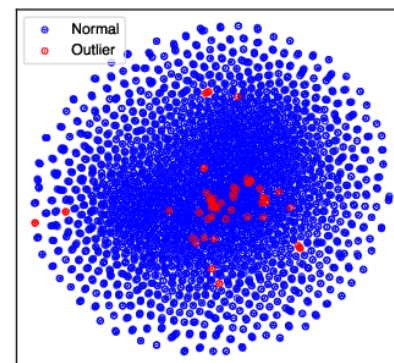
(a) *Breastw (mat)*



(b) *Shuttle (mat)*



(c) *Ionosphere (mat)*

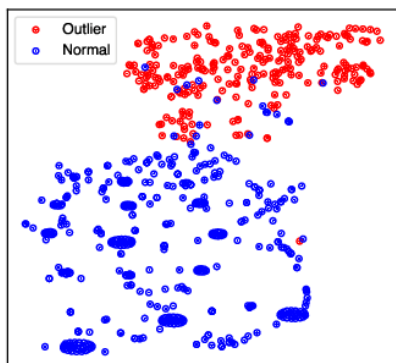


(d) *Speech (mat)*

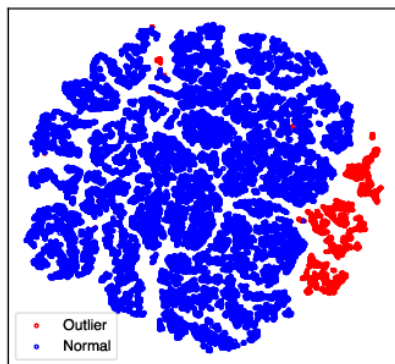
이상치 탐지는 극단적인 클래스 불균형 상황에 사용하는 방법으로  
비율만 봤을 때는 우리 데이터에도 충분히 적용 가능해 보임

## 이상치 탐지

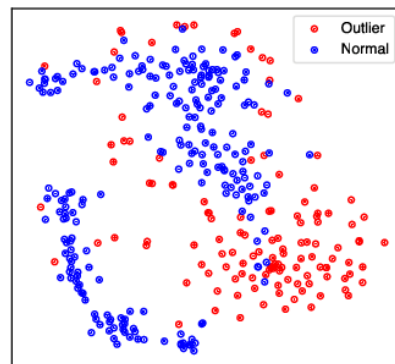
Li, Zheng, et al. "Ecod: Unsupervised outlier detection using empirical cumulative distribution functions." (2022).



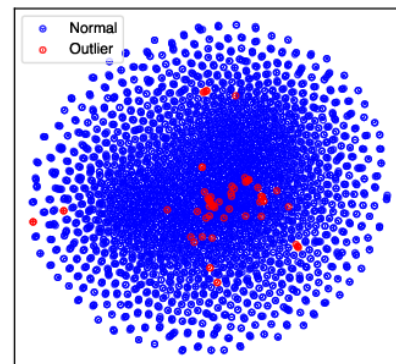
(a) *Breastw (mat)*



(b) *Shuttle (mat)*



(c) *Ionosphere (mat)*



(d) *Speech (mat)*

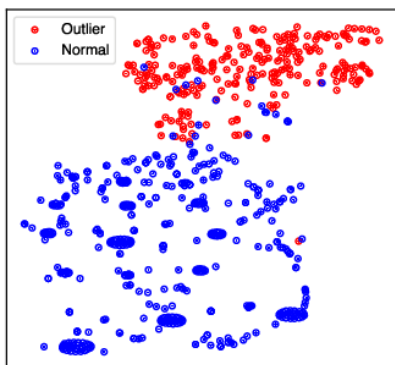
KNN, Isolation Forest, SVDD, Autoencoder,  
그리고 2022년 발표된 논문인 ECOD 등 여러 방법 존재

→ Pyod 패키지를 통해 이상치 탐지를 수행하는 다양한 모델 시도 가능!

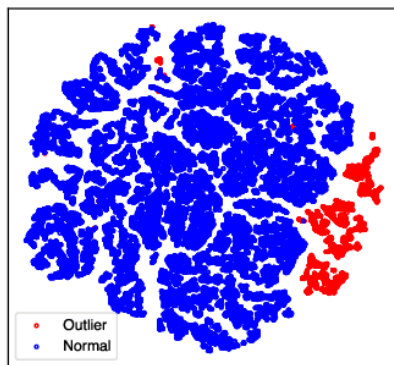


## 이상치 탐지

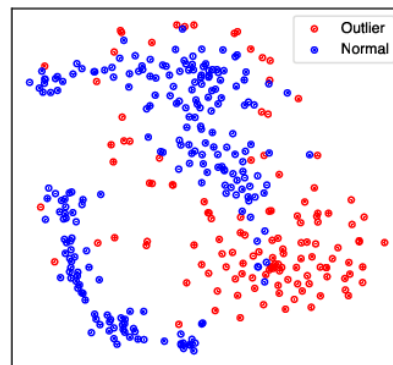
Li, Zheng, et al. "Ecod: Unsupervised outlier detection using empirical cumulative distribution functions." (2022).



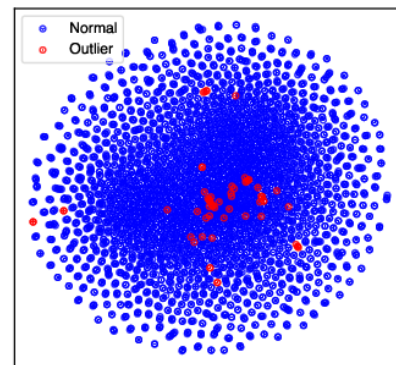
(a) *Breastw (mat)*



(b) *Shuttle (mat)*



(c) *Ionosphere (mat)*



(d) *Speech (mat)*

대부분 비지도 학습이기 때문에, 지도학습에 비해 상대적인 성능은 좋지 않았음

그럼에도 이번 데이터에서는 괜찮은 성능을 보여주었기 때문에

anomaly 여부를 변수로 추가하는 방법론도 고려 가능

## 추가적 성능 향상 방법: 앙상블

단일 모델들을 세운 후, **앙상블** 사용

여러 모델의 예측값들을 함께 이용하여 로버스트한 예측 가능



각 모델의 단일 예측률이 높을 수록

모델 간 **상관관계가 낮을 수록** (다양성)

앙상블 통한 error 보완력 ↑

## 앙상블 대표적 예시

### Weighted Voting Ensemble

성능이 좋은 모델에 가중치를 주며 **다수결**로 예측값을 결정하는 형식

EXAMPLE

6개의 모델이 있는 경우

‘가장 잘 예측하는 모델’에게 3표를 주고,  
나머지 5개의 모델에게 한 표 씩 주었을 때  
가장 좋은 모델의 몇 개의 오분류를  
바로잡을 가능성 부여

0에 대한 예측력은 높지만 1에 대해서  
50%로 찍고 있는 경우 더욱 잘 작동

## 앙상블 대표적 예시

Weighted Voting Ensemble

예측값이 연속적인 경우

### Weighted Average Ensemble

여러 모델의 예측 값들의 **가중평균**을 사용

6개의 모델이 있는 경우

EXAMPLE  
LGBM & XGB 모델의 예측력이 높다는 사실 발견

'가장 잘 예측하는 모델'에게 3표를 주고,

나머지 5개의 모델에게 한 표씩 주었을 때

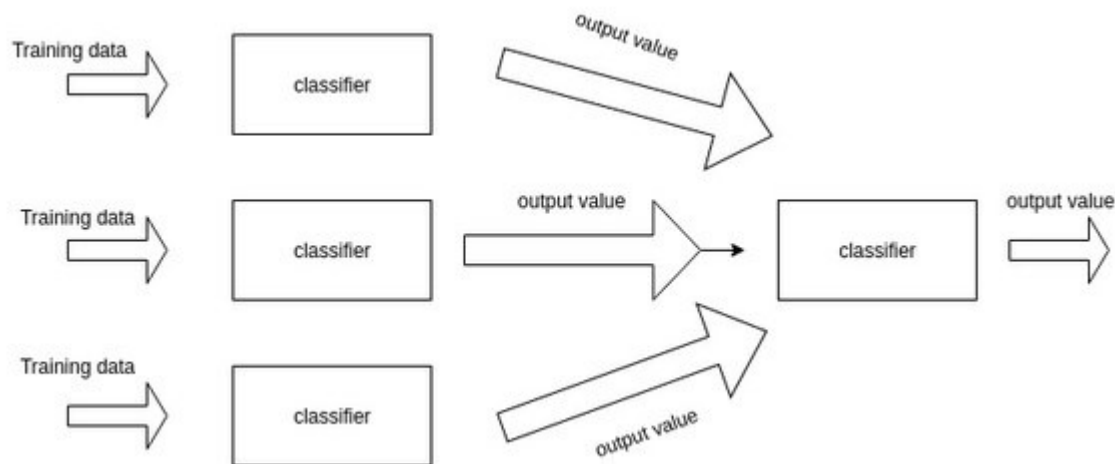
LGBM, XGB 모델의 예측 값을

ex) 6:4 의 비율로 평균내어 사용

## 앙상블 대표적 예시

### Stacking Ensemble

개별 모델의 예측 결과들로 메타 데이터셋을 만들고  
최종 모델에 적용하여 예측을 수행하는 방법

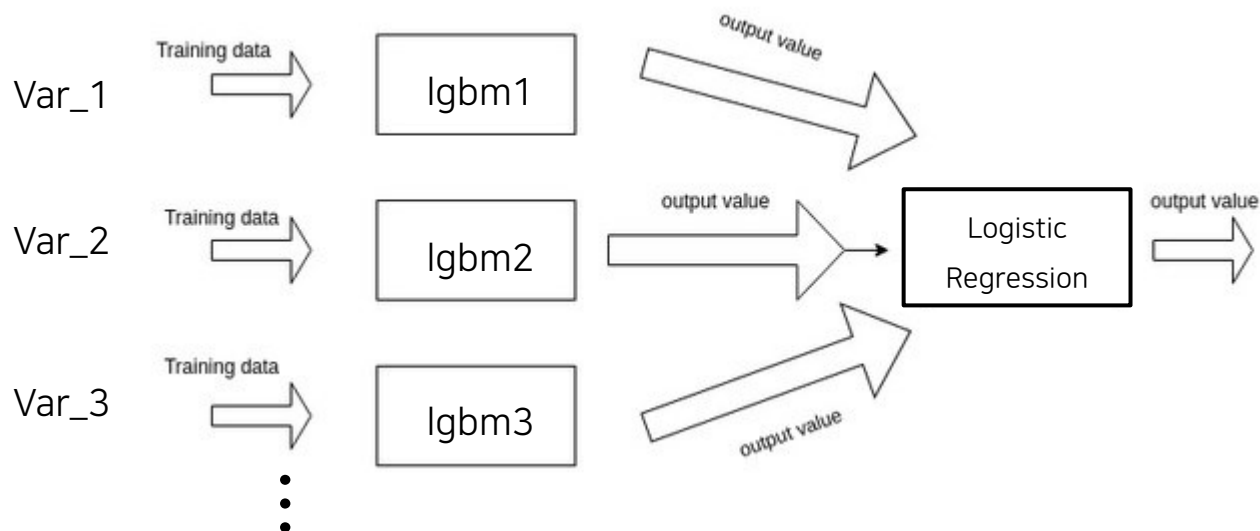


## 앙상블 대표적 예시

### Stacking Ensemble

각 변수별로 학습을 진행하여 200개의 모델로 예측을 한 후

최종 로지스틱 회귀모델로 최종 예측 진행



## 앙상블 대표적 예시

스태킹 기법 - 다중 계층 스태킹, 메타 특징 변수 생성,  
홀드아웃 데이터를 이용한 블렌딩 기법 등  
이외에도 다양한 앙상블 기법 존재



데이터/경진대회 성격, 평가지표 등에 따라 효과가 달라짐  
→ 상황에 따라 적용

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

마스킹된 데이터의 어려움에도 불구하고, 높은 상관관계를 지니는 변수들과 NA를 처리하기 위해 EDA와 Feature Engineering을 수행한 과정이 두 팀 모두 정말 인상 깊었습니다!!

또 극단적인 클래스 불균형, 그리고 적은 데이터에 대한 과적합을 방지하기 위해서 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



2팀



김동희 김민주 방건우 박채원 이동기 진재언

축하드립니다~~!



다들 한 주 동안 너무  
고생 많으셨습니다!

