

Flu Shot Prediction

신종플루와 계절독감 백신 접종 여부 예측

범주형자료분석팀

이지연 | 심예진 | 조장희 | 조혜현 | 진효주



CONTENTS



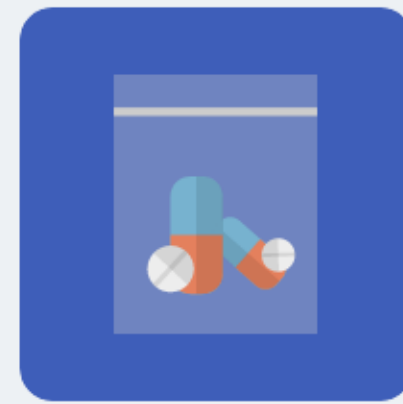
1

2주차 피드백



2

인코딩



3

Multi-label
Classification



4

모델링



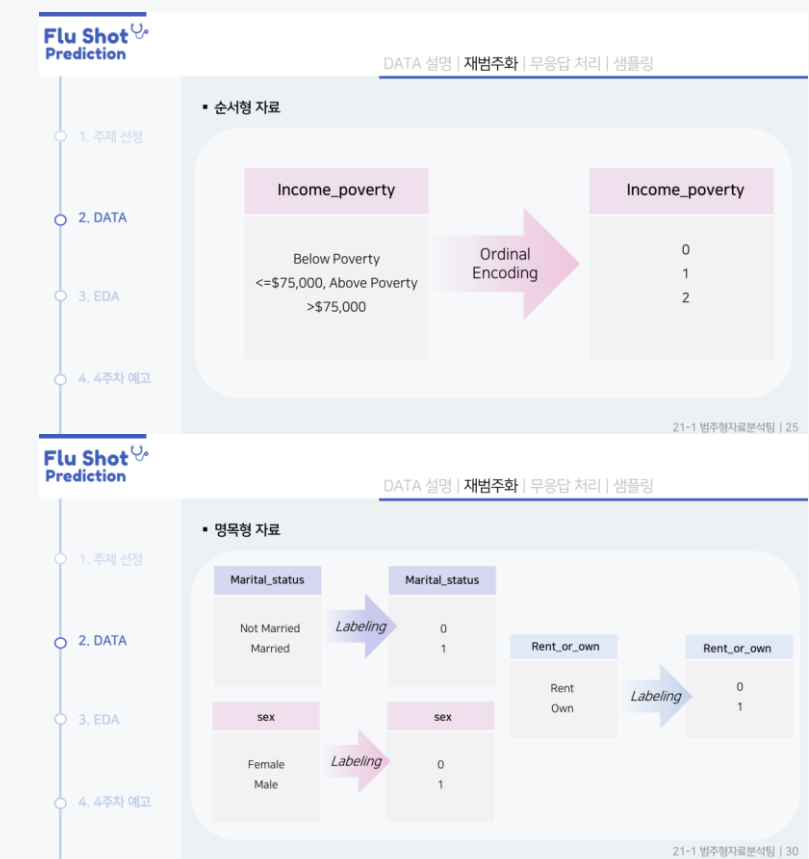
5

한계 및 의의

~ 범주형 변수는 고민이 있어~

범주형 변수는 모델링 전 컴퓨터가 이해할 수 있도록 수치 형태로 인코딩 해주어야 함

우리의 Feature 변수에는 **순서형 자료**와 **명목형 자료**가 혼합되어 있는데...인코딩을 어떻게 하는 것이 좋을까?



치열했던 2주차의 전처리...

- 1. 2주차 피드백
- 2. 인코딩
- 3. Multi Label Classification
- 4. 모델링
- 5. 한계 및 의의

One-Hot Encoding

: 가변수를 만들어 수많은 0과 한 개의 1 의 값으로 범주를 구별하는 인코딩



장점

- 변수내 레벨 값들이 서로 분리되어 있어 거짓 관계나 영향이 생기지 않음



단점

- 차원이 크게 늘어나 비효율적
- 트리기반 모델에 적절하지 않음!



트리 기반 모델에 원-핫 인코딩이 적절하지 않은 이유?

1. 단순히 0과 1로만 결과를 내기 때문에 큰 정보적 이득 없이 tree의 depth가 늘어나게 됨
2. 랜덤포레스트처럼 일부 feature들만 샘플링하여 트리를 만들어나가는 경우, 원핫 인코딩으로 생성된 feature가 많기 때문에 더 많이 선택될 가능성이 존재!

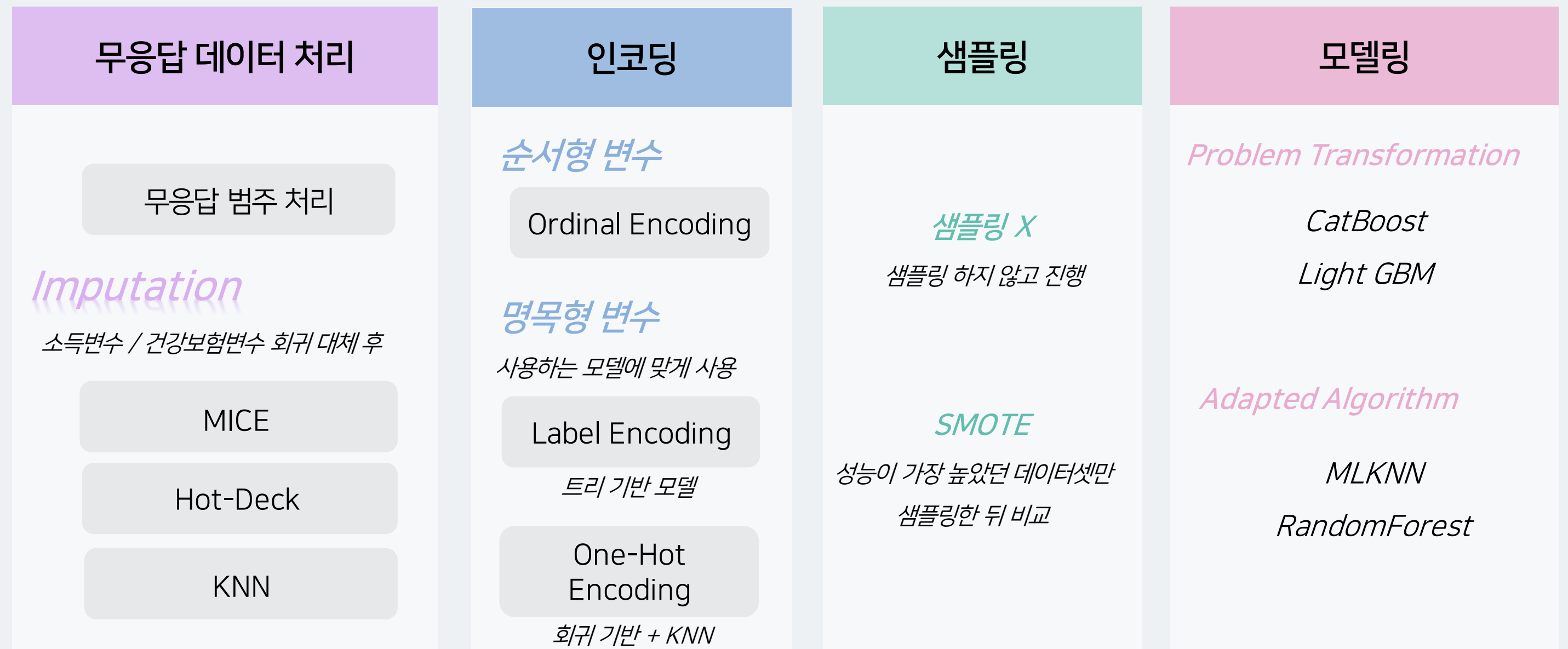


→ 회귀 모형과 KNN 기반 모델링의 경우 원-핫 인코딩 사용!

- 1. 2주차 피드백
- 2. 인코딩
- 3. Multi Label Classification
- 4. 모델링
- 5. 한계 및 의의

~ 범주형의 고민 해결 ~

전처리 - 모델링 흐름 총정리!



** KNN Imputation의 경우 원핫인코딩해 준 후 진행하였기 때문에 트리 기반 모형에는 사용 X

■ Problem Transformation



✓ 다중 라벨 데이터를 단일 라벨 문제로 해결하는 방법

✓ 아무 단일 라벨 분류기 사용 가능

✓ 각 분류와 해당 분류가 아닌 데이터를 토대로 분류기를 학습시켜 문제 해결



Binary Relevance



Chain Classifier



Label Powerset

1. 2주차 피드백

2. 인코딩

3. Multi Label Classification

4. 모델링

5. 한계 및 의의

- Adapted Algorithm

Adapted Algorithm

기존 classification 알고리즘을 바꿔서 다중 라벨 문제를 푸는 방법

ML-KNN

x에서 가장 가까운 k개의 이웃 중에 가장 흔한 라벨을 지정하는 방법
베이저안 추론을 통해 각 클래스에 대한 확률을 계산 가능

Decision Tree

데이터의 불확실성을 측정하는 엔트로피를 계산하여 정보 획득이 가장 큰 노드 생성.
다중 클래스 엔트로피를 통해 트리를 구성하고 각 트리의 라벨 구성을 통해 예측
양상블, 랜덤포레스트가 대표적

1. 2주차 피드백

2. 인코딩

3. Multi Label Classification

4. 모델링

5. 한계 및 의의

1. 2주차 피드백

2. 인코딩

3. Multi Label Classification

4. 모델링

5. 한계 및 의의

Hamming Loss



전체 라벨 중 잘못 분류된 라벨을 의미

$$\frac{1}{NL} \sum_{i=1}^N \sum_{j=1}^L I(\hat{y}_j^i \neq y_j^i)$$

단점

각각 라벨에 대해 계산하기 때문에 최소화 했을 때
라벨에 대한 의존성이 고려되지 않을 수 있음

Exact Match



모든 라벨에 대해 맞은 비율을
나타내는 방법


$$\frac{1}{N} \sum_{i=1}^N I(\hat{y}_i = y_i)$$

단점

모두 맞춘 부분만 고려하기 때문에
부분적으로 맞은 예측을 무시함

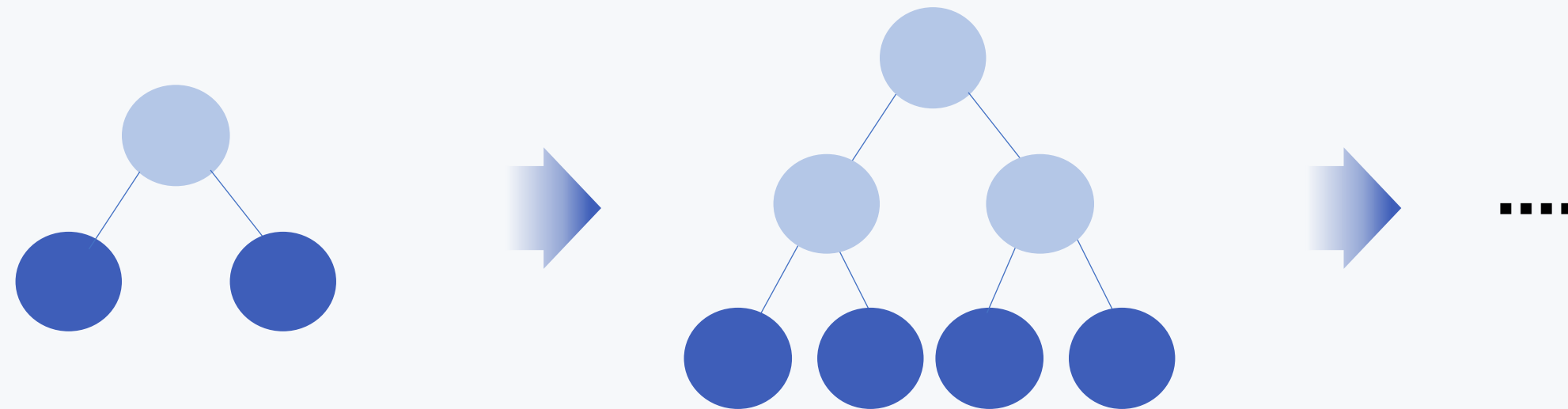
CatBoost | LightGBM

■ Cat Boost

Gradient boosting 기법 중 하나로, 
범주형(categorical) 변수를 처리하는 데 유용한 알고리즘

특징

- Level-wise로 트리를 만들어 나감.
- 순서에 따라 모델을 만들고 예측하는 방식인 Ordered boosting을 활용



Level - wise tree growth

1. 2주차 피드백

2. 인코딩

3. Multi Label Classification

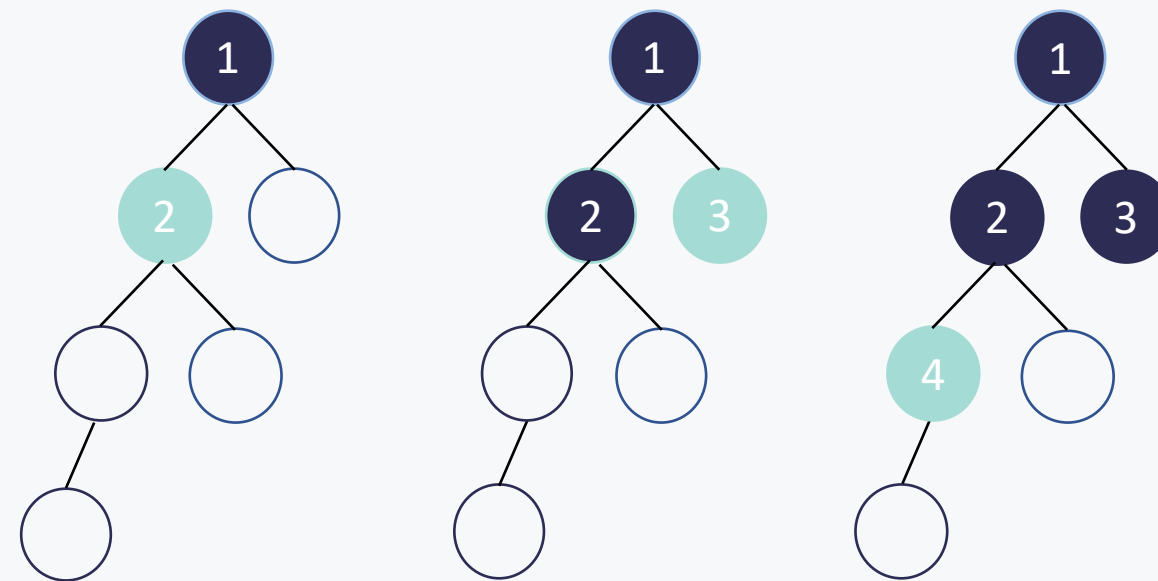
4. 모델링

5. 한계 및 의의

CatBoost | LightGBM

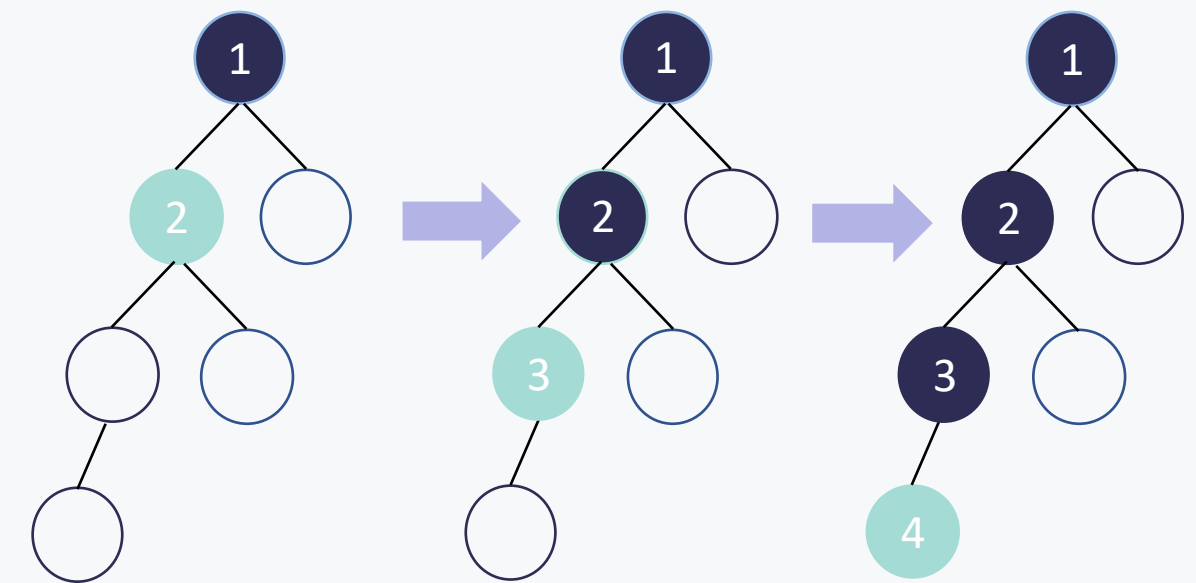
LightGBM

기존 트리 기반 모델



우리 함께 가자!

LGBM



강한 놈만 데려간다

- 균형 잡힌 트리를 유지하면서 분할
 - 트리의 깊이를 최소화
- 균형을 맞추기 위해 시간이 많이 듦

- 최대 손실값을 가지는 리프 노드를 분할
 - 트리의 깊이가 깊어짐
 - 비대칭적인 트리

→ 예측오류를 최소화 하고 속도가 빠름!

LGBM은 왜 이렇게 빠른가요?

다른 트리 기반 모델은 level-wise인 반면

LGBM은 leaf-wise이기 때문

다음 슬라이드에서 자세히 알아보자!

■ ML-KNN

Sunflow091667님이 직접 만드신 로고..^^

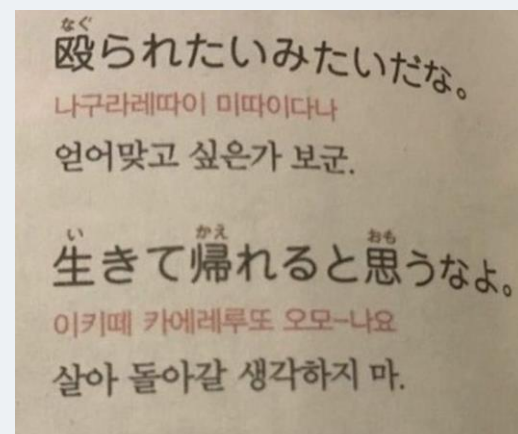
MLKNN

X에서 가장 가까운 K개의 이웃 중에 가장 흔한 레이블을 지정하는 방법

베이지안 추론을 통해 각 클래스에 대한 확률 계산

$$y_j = \begin{cases} 1, & \text{if } P(c_{j,x}|y_j = 1)P(y_j = 1) \geq P(c_{j,x}|y_j = 0)P(y_j = 0) \\ 0, & \text{otherwise} \end{cases}$$

$$P(y_j = 1|c_{j,x})P(c_{j,x}|y_j = 1)P(y_j = 1)$$



수식은 그냥 멋있어 보이려고 넣었어용..^^

■ MLR 패키지



효주야.. 나 mlr이 너무싫어...

지연아 나두 mlr이 너무 싫다 ㅎㅎ



~지연이와 효주는 mlr이 싫어~

특별출연: 황정현

R에서 Multi Label Classification을 할 때 사용



rFern, RandomForestSRC 패키지와 함께 사용 가능

makeLearner 함수를 통해 모델 지정 후

predict.type을 "prob"으로 설정하여

출력되는 결과값 설정



rFern에서는 확률로 결과값을 설정할 수 없어서

RandomForest만 이용

1. 2주차 피드백

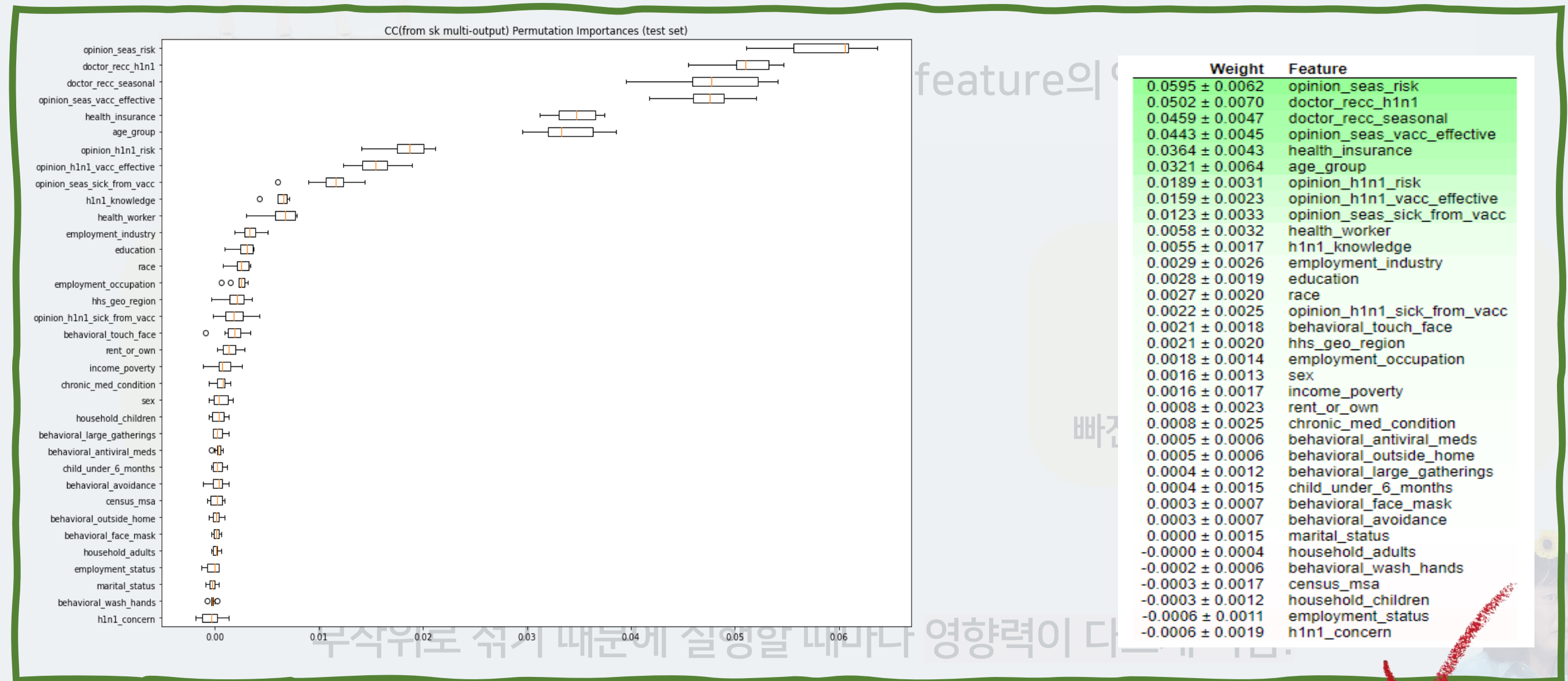
2. 인코딩

3. Multi Label Classification

4. 모델링

5. 한계 및 의의

■ 변수 중요도



Permutation Feature Importance 값이 실행할 때마다 다르기 때문에 범위로 표현!

■ 로지스틱 회귀모형을 통한 변수 해석

백신을 맞지 않고 H1n1/계절독감에 걸릴 걱정을 가지고 있는 사람

H1N1 백신	2	3	4	5
β	0.05	0.08	0.16	0.23
$\exp(\beta)$	1.05	1.08	1.17	1.26

값이 점점 커짐

계절 독감 백신	2	3	4	5
β	0.11	0.22	0.28	0.35
$\exp(\beta)$	1.12	1.25	1.32	1.42

값이 점점 커짐

1. 2주차 피드백

2. 인코딩

3. Multi Label Classification

4. 모델링

5. 한계 및 의의

- 1. 2주차 피드백
- 2. 인코딩
- 3. Multi Label Classification
- 4. 모델링
- 5. 한계 및 의의

■ 범주팀의 최종 결과 ... (두구두구두구두구)



zzang에게 혼나고 있는
sunflower091667

Welcome back!

Login*

Sunflower091667의 4개의 분신 ..

Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines
HOSTED BY DRIVENDATA

GLORY! 10 MONTHS LEFT

Lara123	22	0.8634	2020-06-22 14:28:48		1
uomcse_digdata	23	0.8634	2020-06-22 08:28:15		15
bigfoot	24	0.8633	2020-06-22 15:34:07		11
uomcse_flex	25	0.8633	2020-06-22 13:07:20		19
data_savv	26	0.8633	2020-06-22 11:43:24		2
uomcse_Goblin_Loot	27	0.8633	2020-06-22 07:11:51		2
ConSurv_UOM_CSE	28	0.8633	2020-06-16 15:28:25		56
Akal	29	0.8632	2020-06-21 11:11:58		1
uomcse_code_hunters	30	0.8632	2020-06-24 14:44:20		51
mwitichenor	31	0.8632	2020-09-11 22:00:00		19
Dana_Makov	32	0.8632	2020-12-14 00:00:00		23
Masi	33	0.8632	2021-03-19 18:32:00		1
sunflower091667	34	0.8631	2021-05-07 14:06:04		11
Krish1234	35	0.8631	2020-06-22 12:40:31		3
Yada	36	0.8630	2020-11-19 12:40:15		20
uomcse_code_labs	37	0.8630	2020-06-21 19:43:13		16
adurieu	38	0.8630	2021-02-02 17:26:21		3
uomcse_team	39	0.8630	2020-06-22 14:57:56		17

2111명 중 무려 34등을 한
Sunflower091667 a.k.a. 범주팀

