

Feature_Selection(4)

데이터분석전처리적용반

Boruta algorithm

Permutation

V1	V2	V3	V4	V'1	V'2	V'3	V'4
10	100	52.3	1	20	99	36.3	1
20	99	36.3	0	15	100	28.9	1
10	96	28.9	1	10	96	52.3	0
15	99	44.4	0	10	99	44.4	0

원래 데이터

쉐도우 변수 데이터

- 모든 변수를 복사 → shadow features or permuted copies.
 - 원본 데이터의 독립 변수가 5개 미만인 경우 기존 변수를 복사본으로 만들어 5개 이상 만들.
- 복사한 변수를 타겟 변수에 uncorrelated 하게 만들기 위해 랜덤하게 섞고 원 자료와 결합.
- 결합된 데이터와 원 데이터에 대해 랜덤포레스트 모델을 생성하고, Z-score를 계산.
 - $[(\text{각 트리에 대한 정확도 손실값} - \text{전체 트리의 정확도 손실의 평균}) / \text{정확도 손실 표준편차}]$
- shadow 변수들 중 가장 높은 Z-score를 찾는다. (MSZA, Max Z-score among shadow attributes)
- 원 자료에 대한 Z-score > MSZA인 경우 Hit +1 (이는 MZSA보다 클 때 중요한 변수를 표시하는 수단)
 - 통계적으로 유의수준에서 Z-score < MSZA인 경우, 해당 피처를 중요하지 않은 피처로 드랍한다.
 - 통계적으로 유의수준에서 Z-score > MSZA인 경우, 해당 피처를 중요한 변수로 둔다.
- 위의 과정을 랜덤포레스트가 수행되는 횟수만큼 또는 모든 변수들이 중요한 변수와 중요하지 않은 변수로 tagged 될 때까지 반복.

V1	V2	V3	V4
10	100	52.3	1
20	99	36.3	0
10	96	28.9	1
15	99	44.4	0

원래 데이터



V'1	V'2	V'3	V'4
20	99	36.3	1
15	100	28.9	1
10	96	52.3	0
10	99	44.4	0

쉐도우 변수 데이터



랜덤 포레스트

원래 변수

	V1	V2	V3	V4
중요도	0.2	0.5	0.01	0.01

쉐도우 변수

	V'1	V'2	V'3	V'4
중요도	0.02	0.03	0.04	0.01

변수가 중요하다고 얘기하려면
아무 상관도 없는 변수의 중요도보다는
높아야되지 않겠냐?

나보다 낮은 것들은 뭐야?
그리고도 변수라는
타이틀을 달고 있는거야?

랜덤 포레스트

나무 1

나무 2

...

나무 100

변수	선택	중요도
V1	O	1.3
V2	X	-
V'1	O	0.6
V'2	O	0.2

변수	선택	중요도
V1	O	1.1
V2	O	2.0
V'1	X	-
V'2	X	-

변수	선택	중요도
V1	O	0.9
V2	O	1.6
V'1	X	-
V'2	O	0.1

변수	중요도 평균	중요도 표준편차	Z-score
V1	1.6	0.6	2.67
V2	1.5	0.9	1.67
V'1	0.3	0.9	0.33
V'2	0.25	0.1	2.5

변수	중요도 평균	중요도 표준편차	Z-score
V1	1.6	0.6	2.67
V2	1.5	0.9	1.67
V'1	0.3	0.9	0.33
V'2	0.25	0.1	2.5

구분	V1	V2	V'1	V'2
Z-score	2.67	1.67	0.33	2.5
중요인자 여부	Hit	X		Max

V2는 V'2보다 Z-score가 작으므로
중요하지 않다는 뜻으로 X 표시

V1은 V'2보다 Z-score가 크므로 중
요하다는 뜻으로 Hit 표시

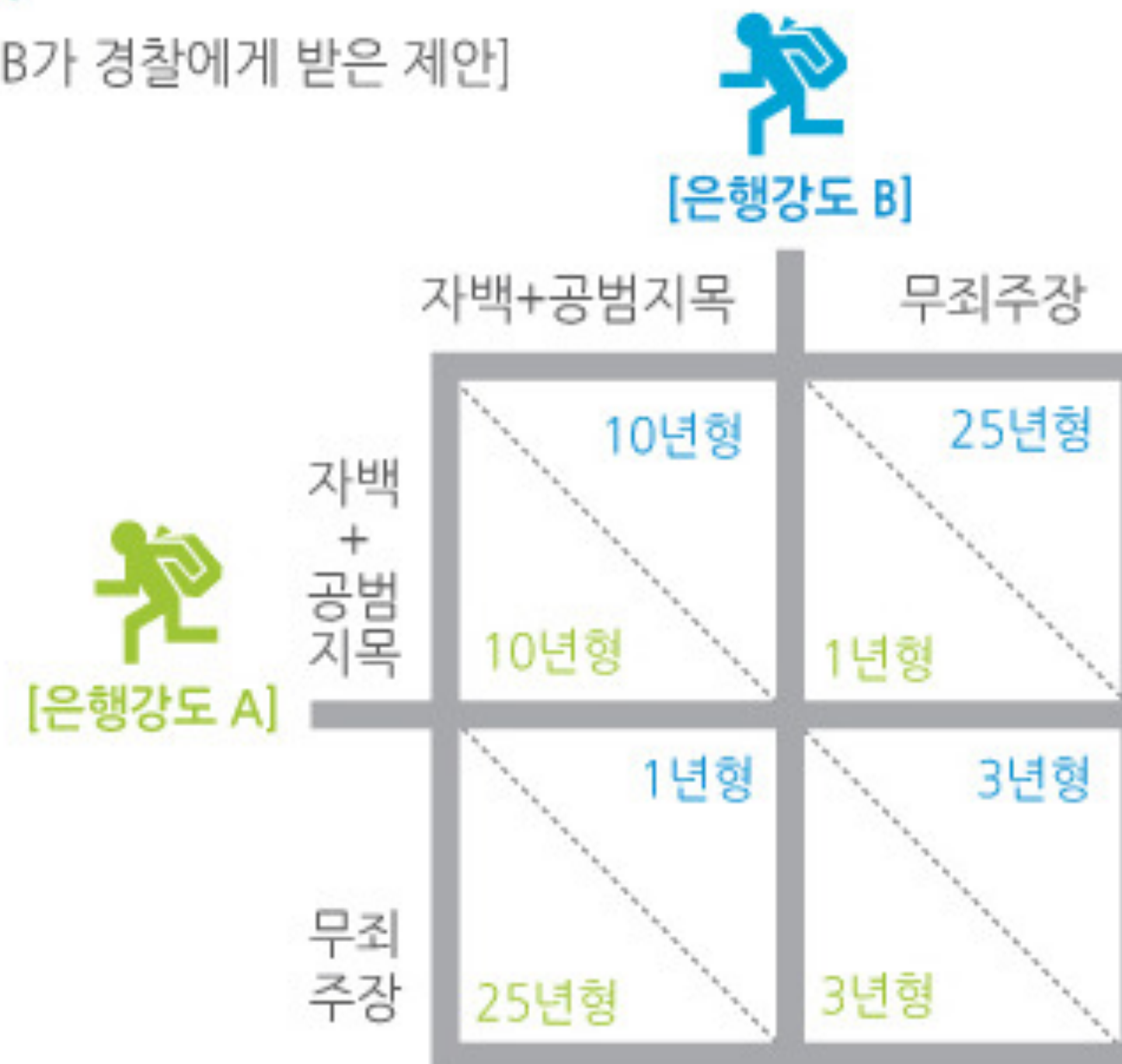
이 과정을 반복한다

스텝	V1	V2
1	Hit	X
2	Hit	Hit
3	X	X
...
50	Hit	X
Hit 합계	40	12

SHAP (SHapley Additive exPlanations)

그림 01 게임이론 사례 1 - 죄수의 딜레마

[A와 B가 경찰에게 받은 제안]



The diagram illustrates a Prisoner's Dilemma game between two players, A and B. Player A is represented by a green running figure icon and is labeled "[은행강도 A]". Player B is represented by a blue running figure icon and is labeled "[은행강도 B]". The matrix shows the possible outcomes based on whether each player confesses or remains silent. The top row represents Player B's strategies: "자백+공범지목" (Confess + Accuse) and "무죄주장" (Claim Innocence). The left column represents Player A's strategies: "자백 + 공범지목" (Confess + Accuse) and "무죄 주장" (Claim Innocence). The payoffs are given in years of prison, with blue text for Player B's payoff and green text for Player A's payoff. Dashed diagonal lines separate the two payoffs in each cell.

		[은행강도 B]	
		자백+공범지목	무죄주장
[은행강도 A]	자백 + 공범지목	10년형 / 10년형	25년형 / 1년형
	무죄 주장	1년형 / 25년형	3년형 / 3년형

하나의 특성에 대한 중요도를 알기 위해 → 여러 특성들의 조합을 구성하고 → 해당 특성의 유무에 따른 평균적인 변화를 통해 값을 계산합니다.

SHAP

SHAP(Shapley Additive exPlanations)는 모델 예측을 해석하는 데 사용되는 방법론입니다. SHAP는 게임 이론의 샐플리 값(Shapley values)을 기반으로 하여, 각 피처(feature)가 모델 예측에 얼마나 기여하는지를 정량적으로 평가합니다. 이는 모델의 "블랙 박스" 특성을 줄이고, 예측 결과를 이해하는 데 도움을 줍니다.

샐플리 값 수식

샐플리 값 ϕ_i 는 다음과 같이 계산됩니다:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! \cdot (|N| - |S| - 1)!}{|N|!} \cdot (v(S \cup \{i\}) - v(S))$$

1. 모든 가능한 피처 조합 생성:

- 피처의 모든 가능한 순열을 고려합니다.

2. 기여도 계산:

- 각 피처가 추가될 때마다 기여도를 계산합니다.
- 기여도는 피처가 추가된 후와 추가되기 전의 모델 예측 차이로 정의됩니다.

3. 평균 기여도 계산:

- 각 피처의 기여도를 모든 가능한 순열에 대해 평균화하여 샐플리 값을 계산합니다.

여기서:

- N 은 모든 피처의 집합입니다.
- S 는 피처 i 를 제외한 피처들의 부분집합입니다.
- $v(S)$ 는 부분집합 S 에 대한 모델 예측 값입니다.
- $v(S \cup \{i\})$ 는 부분집합 S 에 피처 i 를 추가한 후의 모델 예측 값입니다.
- $|S|$ 는 부분집합 S 의 크기입니다.

Additive Feature Attribution Method

Additive Feature Attribution Methods have an explanation model that is a linear function of binary variables :

$$g(z') = \phi_0 + \sum_{i=1}^M \phi_i z'_i,$$

where $z' \in \{0, 1\}^M$, M is the number of simplified input features, and $\phi_i \in \mathbb{R}$.

- g : explanation model , $g(z') \approx f(h_x(z'))$
- f : original prediction model
- z' : simplified input , $z' \approx x'$
- h_x : mapping function , $x = h_x(x')$
- ϕ_i : attribution value

Property 1 (Local accuracy)

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^M \phi_i x'_i \quad (5)$$

The explanation model $g(x')$ matches the original model $f(x)$ when $x = h_x(x')$, where $\phi_0 = f(h_x(\mathbf{0}))$ represents the model output with all simplified inputs toggled off (i.e. missing).

Property 2 (Missingness)

$$x'_i = 0 \implies \phi_i = 0 \quad (6)$$

Missingness constrains features where $x'_i = 0$ to have no attributed impact.

Property 3 (Consistency) Let $f_x(z') = f(h_x(z'))$ and $z' \setminus i$ denote setting $z'_i = 0$. For any two models f and f' , if

$$f'_x(z') - f'_x(z' \setminus i) \geq f_x(z') - f_x(z' \setminus i) \quad (7)$$

for all inputs $z' \in \{0, 1\}^M$, then $\phi_i(f', x) \geq \phi_i(f, x)$.

Theorem 1 Only one possible explanation model g follows Definition 1 and satisfies Properties 1, 2, and 3:

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)] \quad (8)$$

where $|z'|$ is the number of non-zero entries in z' , and $z' \subseteq x'$ represents all z' vectors where the non-zero entries are a subset of the non-zero entries in x' .

LIME

(Local Interpretable Model-agnostic Explanation)

