Week 4 스터디 발표

Feature Selection: SHAP

SHAP

SHAP은 머신러닝에서 feature의 영향/효과를 계산할 때 사용하는 방법으로, 게임이론에 등장하는 **shapley value**를 바탕으로 목적변수(target)에 대한 각 feature의 영향이 측정된다.

Permutation importance와 다른 점은, permutation importance는 모델의 성능이 얼마나 떨어지는지에 따라 변수 중요도를 측정하고 feature을 선택하지만 SHAP은 변수가 목적변수(target)에 미치는 영향을 기준으로 변수를 선택한다.

Shapley Value

호텔 가격을 책정하는데 다음과 같은 요소들이 고려된다고 가정 → 주변에 공원 여부, 면적 100, 3층, 반려동물 허용 여부

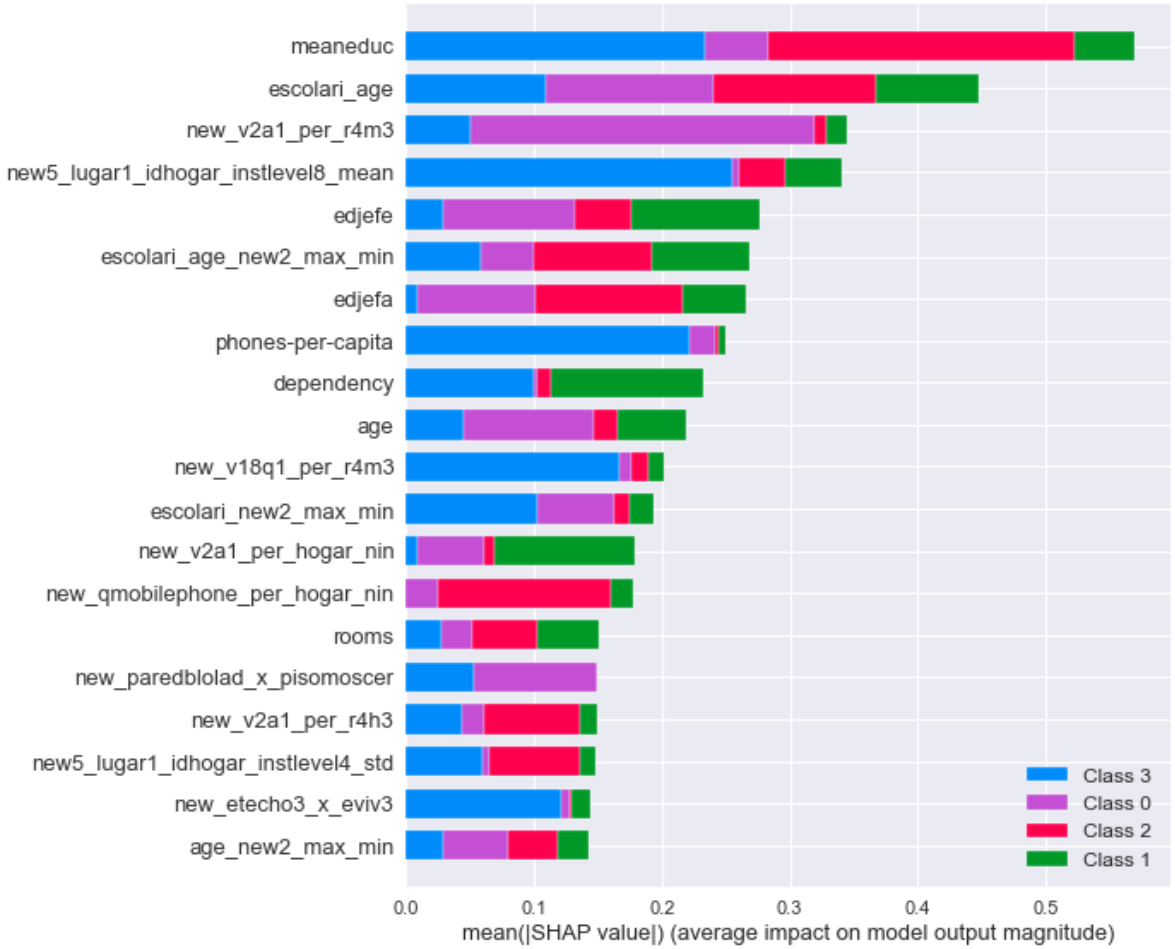
여기서 우리는 **반려동물 허용여부**가 목적변수인 호텔 가격에 미치는 영향을 측정하는 것이 목표이다.
'반려동물 허용여부'를 제외한 다른 변수들은 다음과 8개의 경우로 조합될 수 있다.

- no feature
- 주변 공원여부
- 면적 100
- 3층
- 주변 공원여부, 면적 100
- 주변 공원여부, 3층
- 면적 100, 3층
- 주변 공원여부, 면적 100, 3층

8가지의 변수 조합으로부터 반려동물 허용여부가 호텔 가격에 미치는 marginal contribution을 계산하여 가중 평균을 구하면 Shapley value 추출 가능

Global interpretation

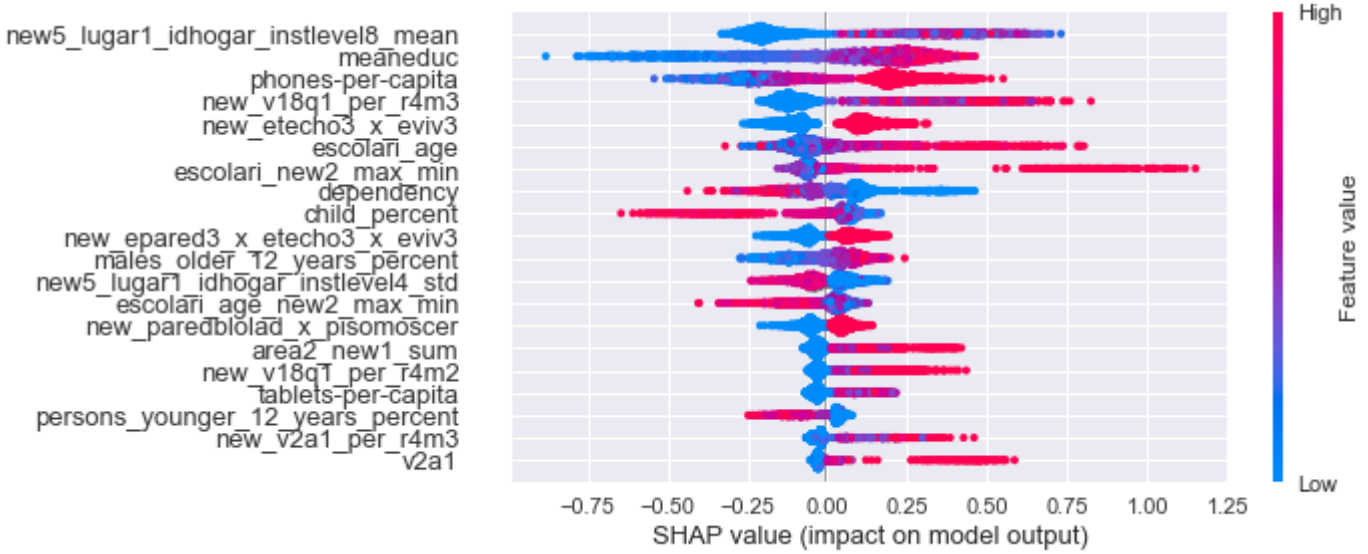
(1) Feature Importance plot



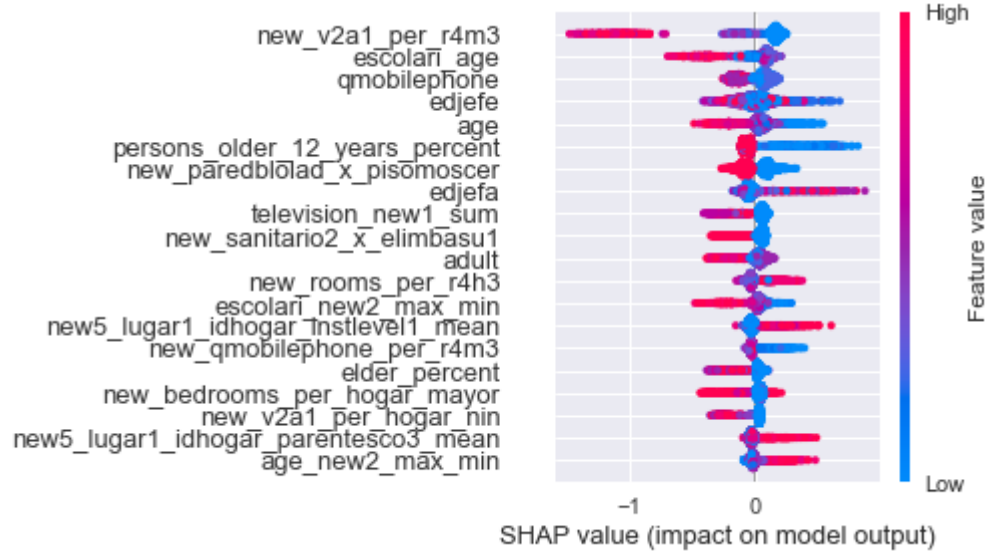
Feature의 각 클래스에 미치는 영향을 나타내는 plot

(2) Summary plot

Class 3: Non-vulnerable



Class 0: Extreme poverty



각 featur의 shapely value가 표시되어있는 plot

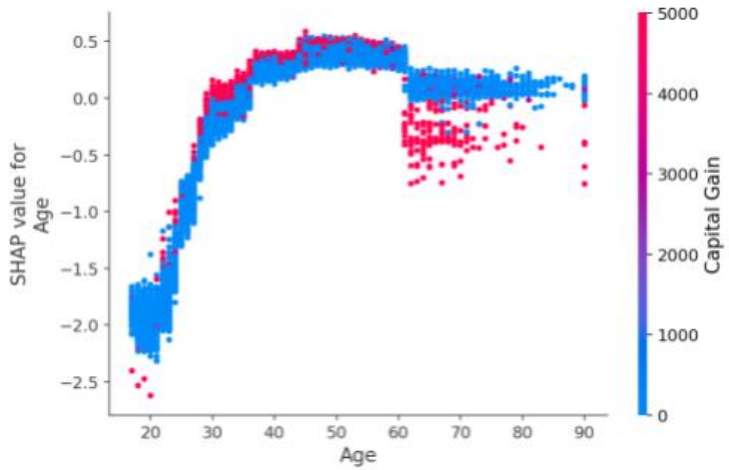
→ 변수 중요도 순으로 위에서부터 나열
→ 각 feature와 target value사이의 관계성 파악 가능

(3) Dependence plot

```
shap.dependence_plot('meaneduc', shap_imp[3], X_train)
```

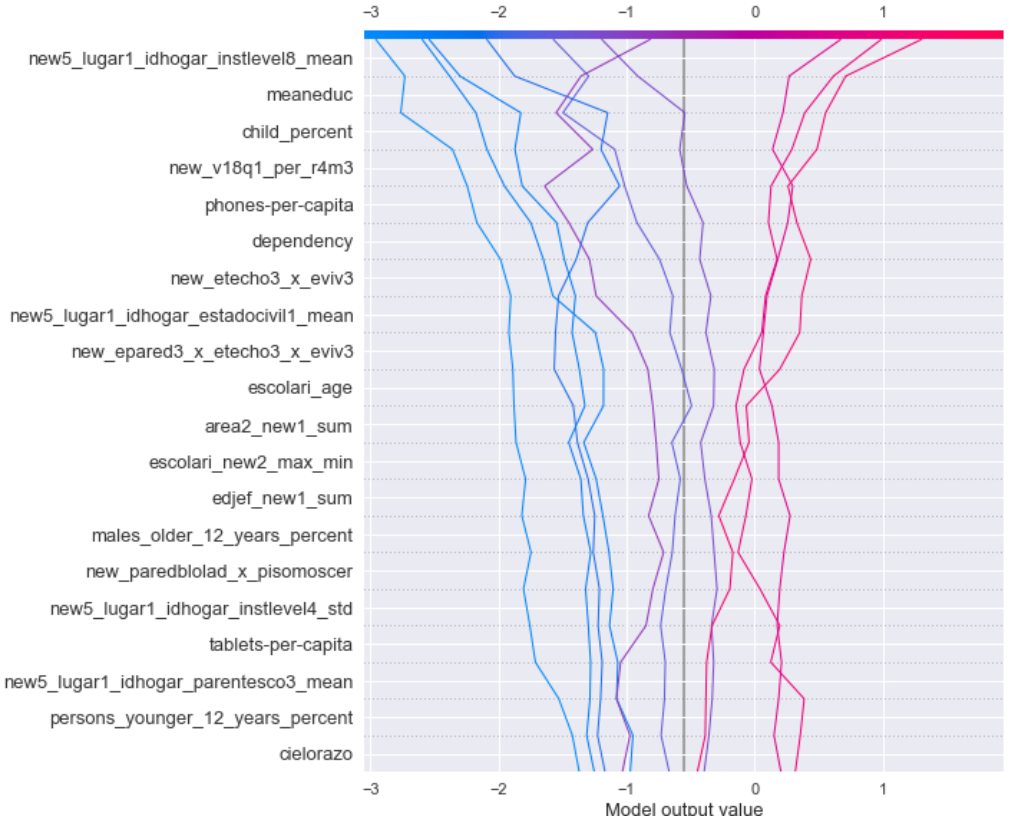
1. 특정 feature와 target value의 관계 파악 가능
2. 특정 feature와 interaction이 가장 큰 변수와의 관계 (자동적으로 선택)

Better example



Local interpretation

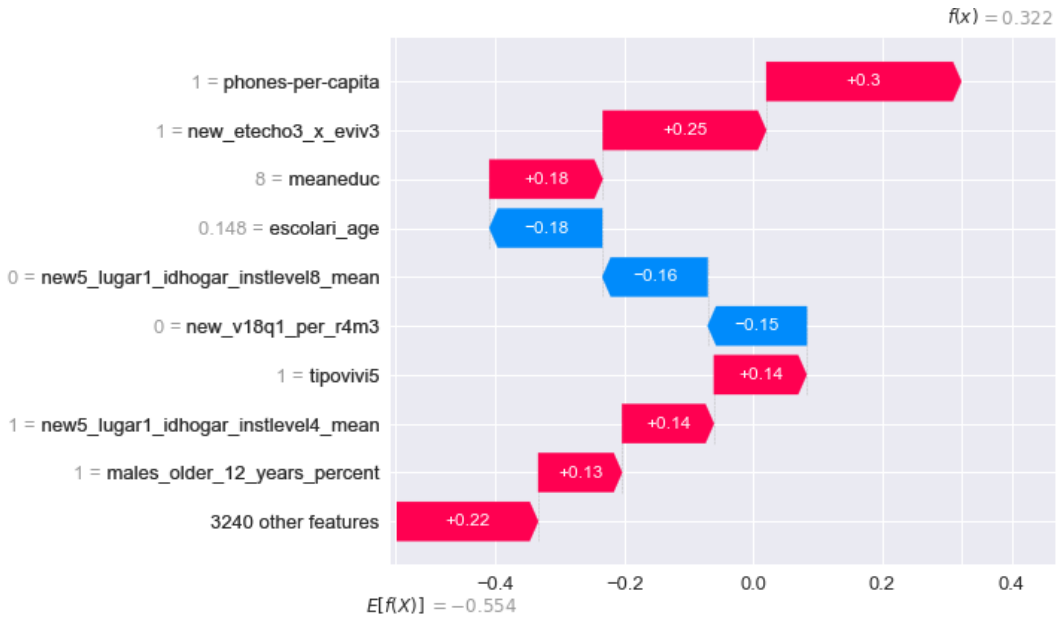
(1) Decision plot



개별 row에서 각 feature value가 target value에 어떤 영향을 미치는지 해석 가능

Local interpretation

(2) Waterfall plot



개별 row에서 각 feature value가 target value에 어떤 영향을 미치는지 해석 가능