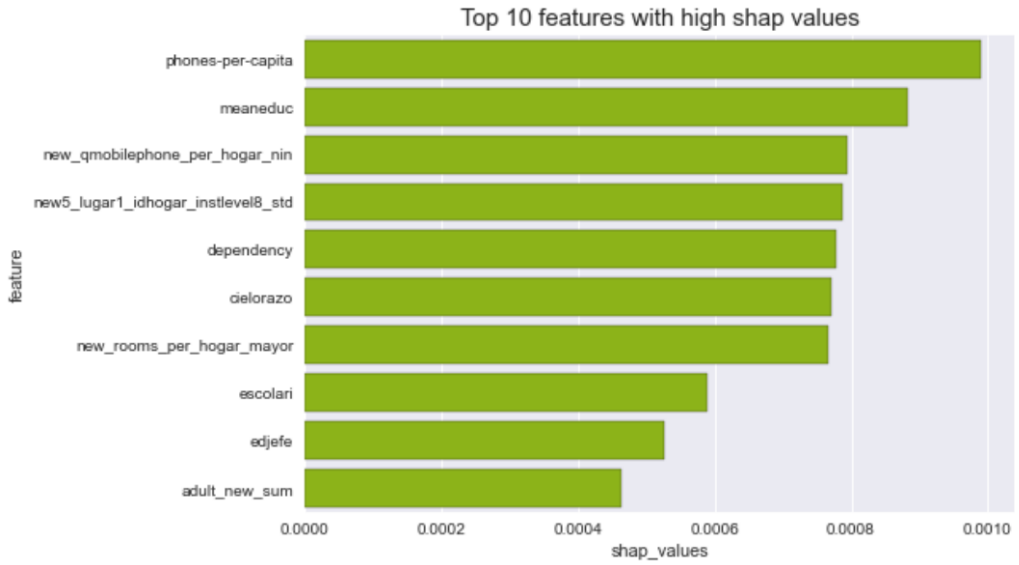
Week 5 스터디 발표

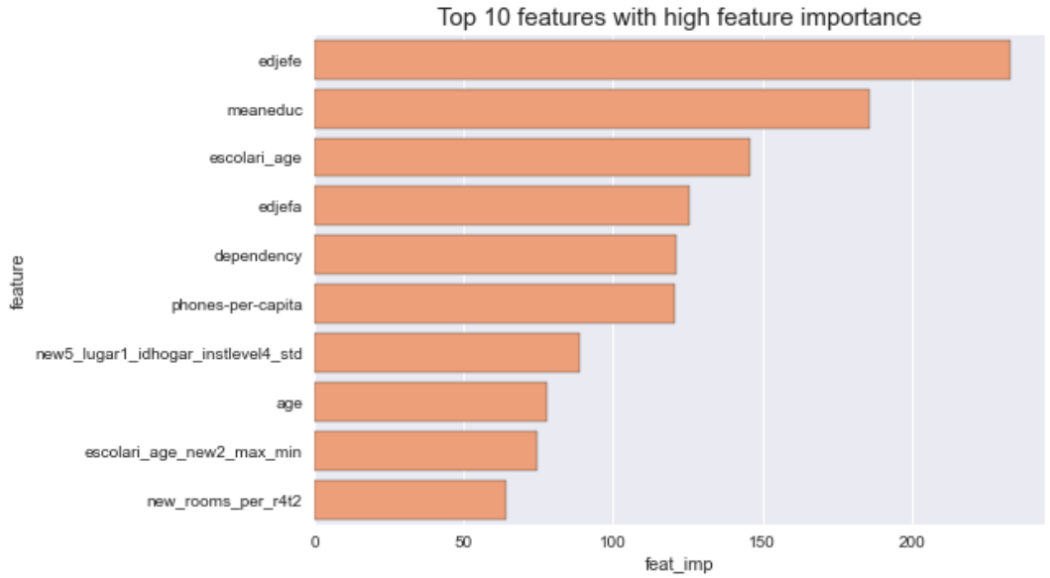
Feature selection using SHAP

SHAP vs Feature importance

(1) SHAP values



(2) Feature importance

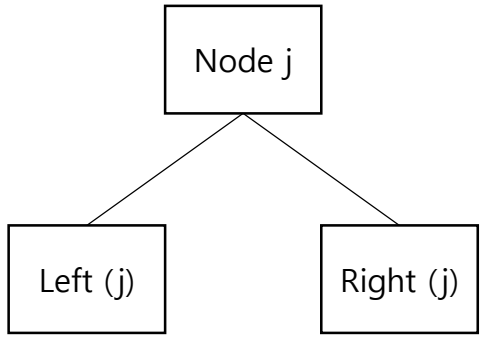


How to calculate feature importance in Scikit-learn

Step 1. Calculate node importance

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)}$$

- ni_j : node j 의 importance
- w: weighted number of samples reaching node j
- C: impurity value(불순도) of node j (Gini-index)



Step 2. Feature importance

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in all\ nodes} ni_k}$$

- **불순도 (Impurity) 측도** $\left(\sum_{k=1}^{K} \hat{p}_{mk} = 1\right)$
  - (기호) $\hat{p}_{mk}$: region R_m에서 class k가 차지하는 비율 ($\hat{p}_{mk(m)}$: region R_m에서 $\hat{p}_{mk}$의 최대값)

①	Gini Index	$I_G(R_m) = \sum_{m=1}^{K} \hat{p}_{mk}(1-\hat{p}_{mk}) = 1 - \sum_{m=1}^{K} \hat{p}_{mk}^2$
②	Cross Entropy	$I_C(R_m) = -\sum_{m=1}^{K} \hat{p}_{mk} \log_2(\hat{p}_{mk})$
③	Misclassification Error	$I_E(R_m) = 1 - p_{mk(m)}$

(K: class의 총 개수)

- **Information Gain**
  - Classification Tree에서 분할 (split) 변수 (x_j) 및 node 선택의 기준
  - 상위 cell R에서 두 개의 영역 R_l (왼쪽)과 R_r (오른쪽)로 나누어질 때, N, N_{R_l}, N_{R_r}을 각각 R, R_l, R_r에 포함된 관측치의 개수라고 하면 ($N = N_{R_l} + N_{R_r}$),

$$IG(R, x_j) = I(R) - \frac{N_{R_l}}{N_R} I(R_l) - \frac{N_{R_r}}{N_R} I(R_r).$$

  - $IG(R, x_j)$가 최대가 되도록 분할변수 (x_j)와 node를 선택

Why SHAP is better?

1. Feature importance: 범주형 변수 중 high cardinality 특징을 가진 변수의 중요도를 더 높게 측정하는 경향이 있음

2. SHAP is consistent: consistency 속성은 모델이 변경되어도 특성값의 marginal contribution이 (다른 특성에 관계없이) 증가하거나 동일하게 유지되는 경우 shaply value도 증가하거나 동일하게 유지된다.

   → 그러나 Feature importance는 모델에 따라서 바뀜