트리 기반 모델의 Feature Importance

1. 의사결정나무, CART

- 의사결정나무 알고리즘에는 ID3, C4.5, CART등 존재

- CART : 각 노드가 2개의 child 노드를 가지는 binary tree를 적절한 불순도 지표를 기준으로 생성해 나가는 알고리즘
  - 목적이 분류 일 때에는 불순도 지표로 Gini 계수 및 엔트로피를 이용
  - 목적이 회귀 일 때에는 MSE 등 이용
- 불순도를 가장 크게 감소시키는 변수의 중요도가 가장 크게 된다

2. Gini Importance

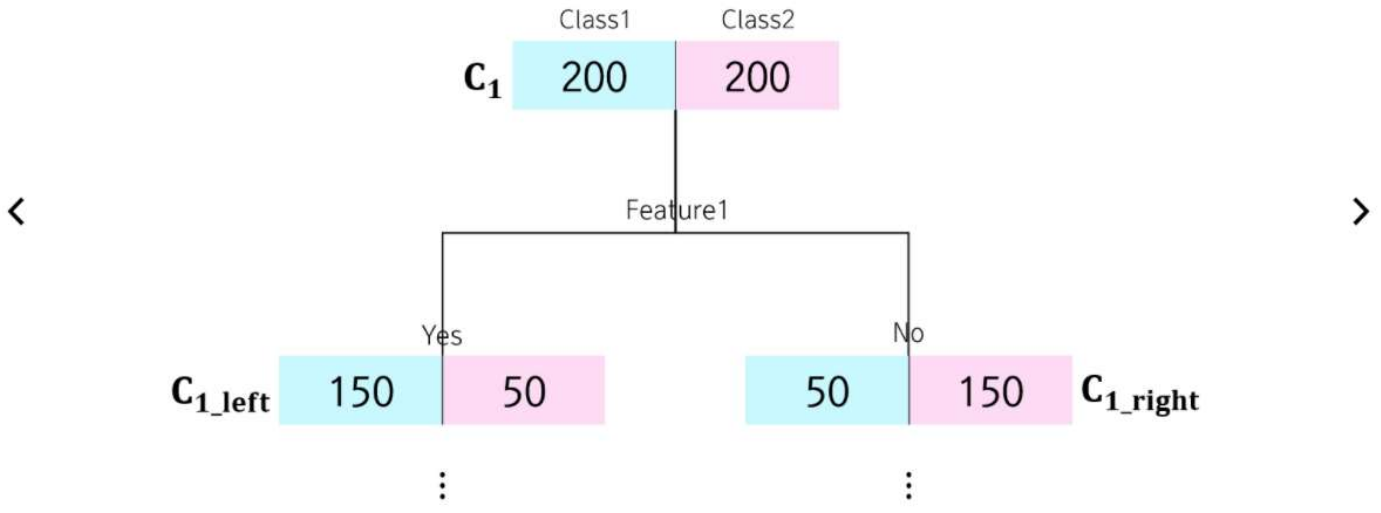
- Scikit-learn에서는 지니 중요도를 이용해서 각 feature 중요도를 측정한다

- 지니 불순도

$$G(N_j) = \sum_{i=1}^{K} p_i(1-p_i) = 1 - \sum_{i=1}^{K} p_i^2$$

해당 노드에서 샘플들이 이질적으로 구성되어 있을수록, 다시 말해서 모든 Class에 골고루 분포되어 있을수록 지니 불순도(impurity)는 높아지게 됩니다. 의사결정나무는 이렇게 불순도를 감소시키는 방향으로 노드를 생성하고 분류를 진행합니다. 이제 아래 간단한 예시를 한번 보겠습니다!

총 샘플이 400개이고, Class가 두 종류, 즉 이항 분류하는 경우입니다. 이 때, 아래 세가지 노드 $C_1, C_{1_left}, C_{1_right}$의 지니 불순도 G를 계산하면 다음과 같습니다.



$$G(C_1) = 1 - \{(\frac{200}{400})^2 + (\frac{200}{400})^2\} = 0.5$$
$$G(C_{1\_left}) = 1 - \{(\frac{150}{200})^2 + (\frac{50}{200})^2\} = 0.25$$
$$G(C_{1\_right}) = 1 - \{(\frac{50}{200})^2 + (\frac{50}{200})^2\} = 0.25$$

- 노드 중요도 (=Information Gain)

- Feature Importance

전체 노드의 중요도 합 대비 i번째 feature에 의해 째진 노드들의 중요도를 합이 바로 i번째 feature 중요도

3. 불순도 기반 Feature Importance의 한계

- Scikit-learn의 디폴트 랜덤 포레스트 Feature Importance는 다소 biased 하다
- 특히 랜덤포레스트는 연속형 변수 또는 카테고리 개수가 많은 변수, 즉 high cardinality 변수들의 중요도를 더욱 부풀릴 가능성이 높다

- 불순도를 기반으로 한 변수 중요도는 데이터를 학습하는 과정에서 얻은 결과이다
- train 데이터셋으로 부터 얻은 통계량으로 계산된 중요도이기 때문에, test 데이터셋에서는 이 변수 중요도가 어떻게 변하는지 알 수 없다. 즉 실제로 test 데이터셋에서는 안 중요한 변수가 학습 과정에서는 가장 중요한 변수로 계산 될 수 있다

- Permutation Importance 와 함께 비교해서 보는 방법이 더 좋다