* **Feature Selection**

**모델을 구성하는 주요 피처들을 선택 :**

불필요한 다수의 피처들로 인해 모델 성능을 떨어뜨릴 가능성 제거

설명 가능한 모델이 될 수 있도록 피처들을 선별

* **Feature Selection 유형**

피처 값의 분포, Null, 피처간 높은 상관도, 결정 값과의 독립성 등을 고려

모델의 피처 중요도(Feature importance) 기반

* **사이킷런 Feature Selection 지원**

**RFE(Recursive Feature Elimination) :**

모델 최초 학습 후 Feature 중요도 선정

feature 중요도가 낮은 속성들을 차례로 제거하며 학습/평가를 수행하여 최적 feature 추출

수행시간이 오래 걸리고, 낮은 속성을 제거하는 것이 Feature Selection을 올바른 목표가 아닐 수 있음.

**SelectFromModel :**

모델 최초 학습 후 선정된 Feature 중요도에 따라 평균/중앙값의 특정 비율 이상인 Feature 선택

* **Permutation(순열) importance 개요**

학습데이터를 제거, 변조하면 다시 학습을 하므로 오래 걸림.

일반적으로 테스트 데이터(검증 데이터)에 **특정 피처들을 반복적으로 변조한 뒤**

모델 성능이 얼마나 저하되는지 기준으로 해당 피처의 중요도를 평균적으로 산정

테이블이(가) 표시된 사진

자동 생성된 설명

해당 과정을 ftr1, ftr2…. 모든 피처 진행

* **왜 feature importance는 절대적인 feature selection 기준이 될 수 없나?**

Feature importance는 최적 tree 구조를 만들기 위한 피처들의 impurity가 중요기준임.

결정 값과 관련이 없어도 feature importance가 높아질 수 있음.

Featrue importance는 학습 데이터를 기반으로 생성 🡪 테스트 데이터에서는 달라질 수 있음

Feature importance는 number형의 높은 cardinality feature에 biased 되어 있음 **ex) ID같은경우**