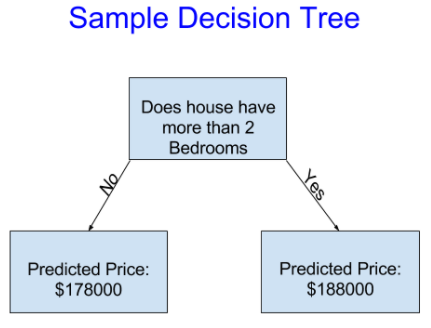
머신러닝\_20220119

# 1.모델 작동 방식(How Models Work)

소개

기계 학습 모델의 작동 방식과 사용 방법에 대한 개요

이전에 통계 모델링이나 기계 학습을 해본 적이 있다면 이는 기본이라고 느낄 수 있으나, 곧 강력한 모델을 만들 것임.  
  
이 코스에서는 다음과 같은 시나리오에 따라 모델을 제작함.  
  
당신 사촌은 부동산 투기로 수백만 달러를 벌었다. 데이터 과학에 대한 당신의 관심 때문에 그는 당신과 비즈니스 파트너가 되겠다는 제안을 받았다. 그는 돈을 공급하고, 당신은 다양한 집들의 가치가 얼마인지를 예측하는 모델을 공급하게 될 것이다.  
  
ex) 사촌에게 과거에 부동산 가치를 어떻게 예측했냐고 물으면, 사촌은 그냥 직관이라고 말한다. 그러나 더 많은 의문이 제기되면서 그는 과거에 보았던 집들로부터 가격 패턴을 알아냈고, 그는 그 패턴을 이용하여 그가 고려하고 있는 새로운 집들에 대한 예측을 하고 있다.  
  
기계 학습도 같은 방식으로 작동한다. 우리는 의사 결정 트리라고 불리는 모델부터 시작할 것이다. 좀 더 정확한 예측을 할 수 있는 화려한 모델들이 있다. 그러나 의사 결정 트리는 이해하기 쉬우며 데이터 과학에서 가장 우수한 모델 중 일부를 위한 기본 구성 요소이다.  
  
단순화를 위해 가장 간단한 의사 결정 트리부터 시작.

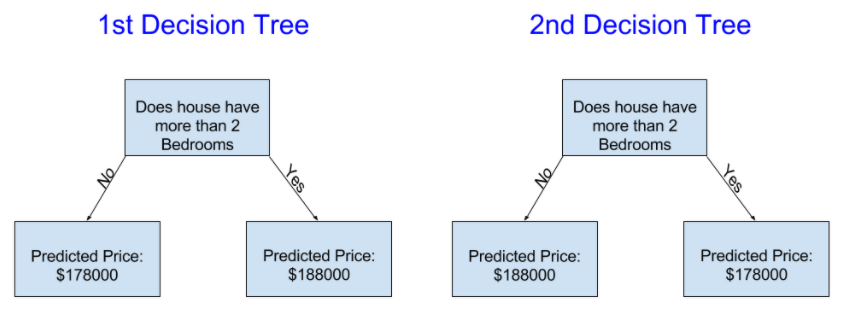


집을 단지 두 개의 범주로 나눈다. 검토 중인 주택의 예측가격은 동일한 범주의 주택의 역사적 평균가격이다.  
  
데이터를 사용하여 집을 두 그룹으로 나눈 다음, 다시 각 그룹의 예측 가격을 결정한다. 데이터에서 패턴을 캡처하는 이 단계를 **fitting** or **training**라고 한다.

\*모형을 **fitting**하는데 사용되는 데이터: **training data**  
  
모델이 적합한 방법(예: 데이터를 분할하는 방법)에 대한 세부 사항은 나중으로 남겨놓을만큼 충분히 복잡하다. 모형이 fitting된(적합된) 후에는 이 모형을 새 데이터에 적용하여 추가 주택의 가격을 예측할 수 있다.

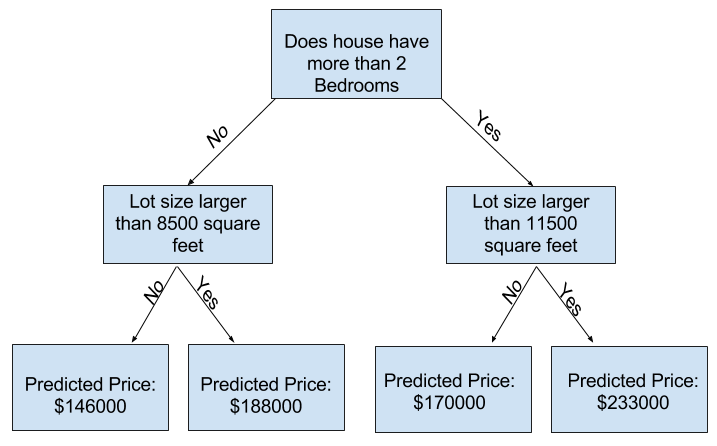
Improving the Decision Tree

다음 중 부동산 교육 데이터를 적합시킴으로써 발생할 가능성이 더 높은 두 가지 의사결정 트리는?



(1st Decision Tree)는 침실이 적은 집보다 침실이 많은 집이 더 높은 가격에 팔리는 경향이 있는 현실을 포착했기 때문에 더 말이 될 것이다. 이 모델의 가장 큰 단점은 욕실 수, 필지 크기, 위치 등 집값에 영향을 미치는 대부분의 요소를 포착하지 못한다는 것이다.

더 많은 "splits(분할)"이 있는 트리를 사용하여 더 많은 요인을 포착할 수 있다. 이것들은 "deeper" 트리라고 불린다. 각 주택의 총 부지 크기를 고려하는 의사 결정 트리는 다음과 같을 수 있다.



우리는 의사 결정 트리를 추적하여 주택의 가격을 예측하고, 항상 그 주택의 특성에 맞는 경로를 선택한다. 그 집의 예상 가격은 나무 밑바닥(맨 아래)에 있다. 우리가 예측을 하는 아래쪽의 지점을 leaf(잎)이라고 한다.