차세대융합기술연구원
ADVANCED INSTITUTE OF CONVERGENCE TECHNOLOGY

트리 회귀 모델

2021.08.03 여지민

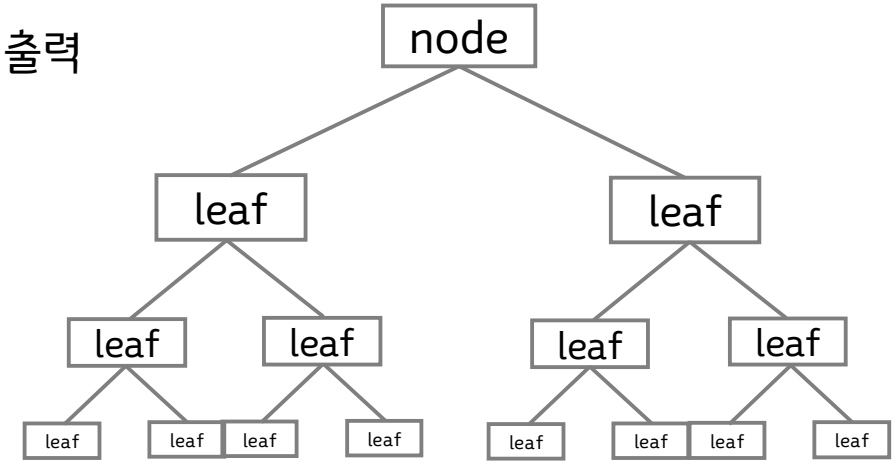
Decision Tree

Decision Tree Regressor

정답에 가장 빨리 도달하는 질문(ex. 특성 i는 a보다 큰가?)을 통해 예측값 출력

가능한 모든 질문에서 타깃값에 대해 가장 많은 정보를 가진 것을 선택

예측변수 평균을 예측값으로 출력



장점

- 쉬운 시각화
- 데이터 스케일링 전처리 불필요

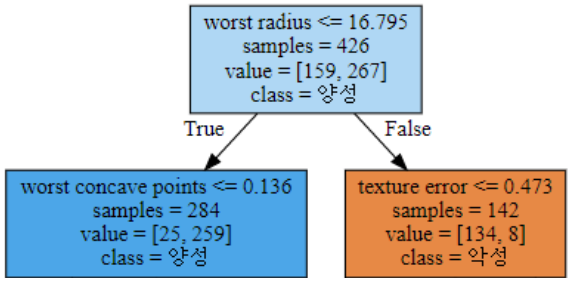
단점

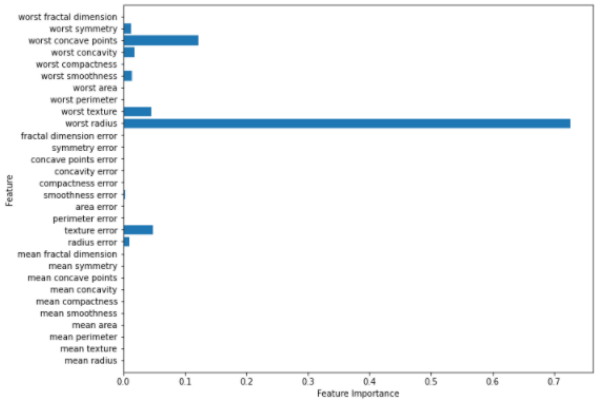
- 과대적합이 발생하기 쉬워 일반화 성능이 좋지 않음

  → 사전 가지치기

  → 사후 가지치기

- 훈련 데이터 범위 밖의 포인트에 대해 예측 불가





Random Forest

Random Forest Regressor

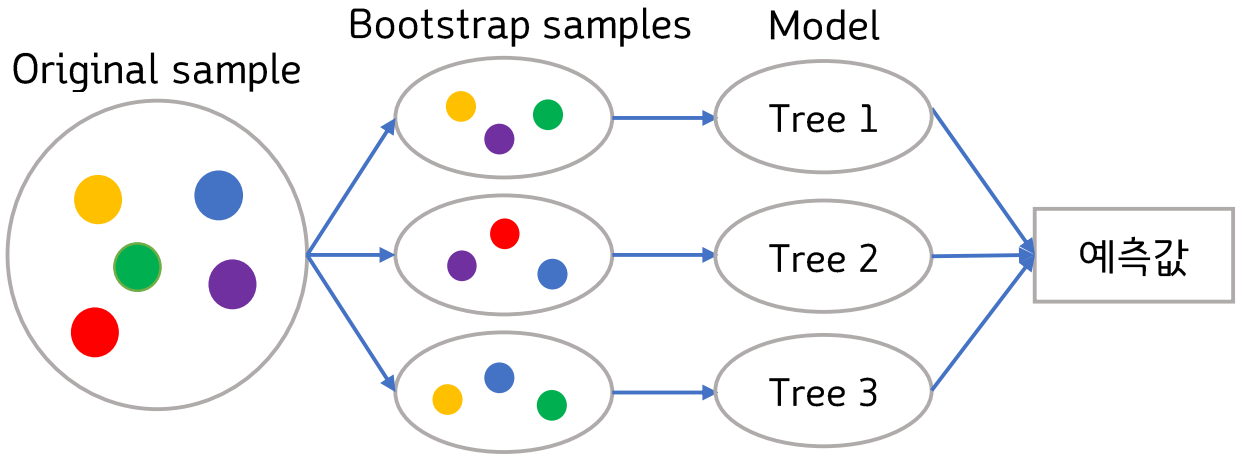
여러 결정 트리 예측값의 평균을 예측값으로 출력

* 무작위 트리 생성이 중요

1. 데이터 포인트 무작위 선택
2. 분할 테스트에서 특성 무작위 선택

각 노드에서 모든 변수에 대해 Information Gain이 가장 높은 방향으로 질문 생성

* scikit learn에서 max_features로 특성 개수 조절



장점

- 데이터 스케일링 전처리 불필요
- 매개변수 튜닝 많이 하지 않아도 성능 우수

단점

- 차원이 높고 희소한 데이터(ex. 텍스트 데이터)에 잘 작동하지 않음 → 선형 모델에 더 적합
- 선형모델보다 많은 메모리 사용하여 훈련 및 예측 느림

Extra Tree

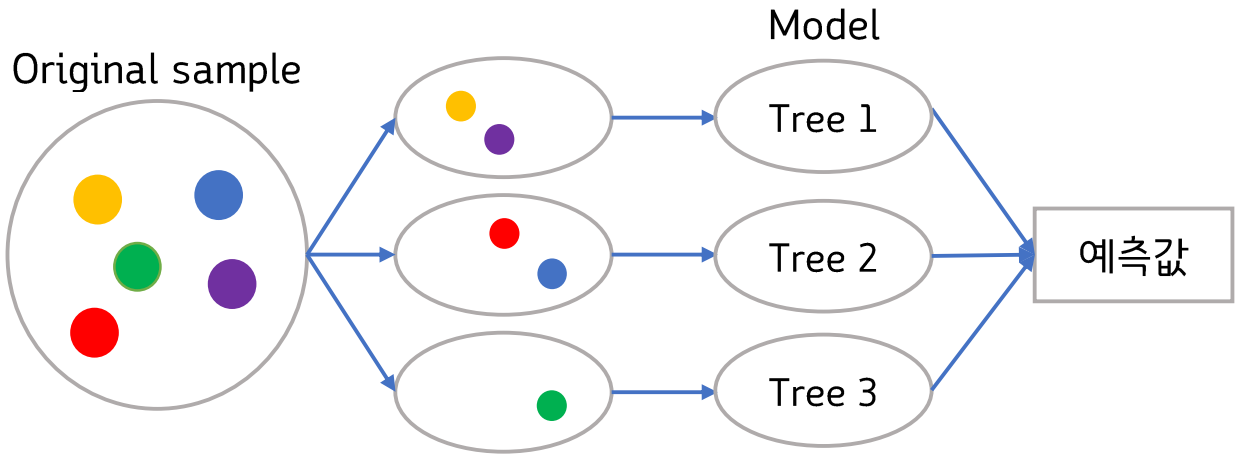
Extra Tree Regressor

전체 훈련 데이터 세트를 사용하여 트리 모델 생성

각 트리의 노드는 무작위로 선택한 변수에 대해서 분할

Bagging 불가

* Bagging: Bootstrap 샘플링을 통해 모델을 학습시켜 결과 집계

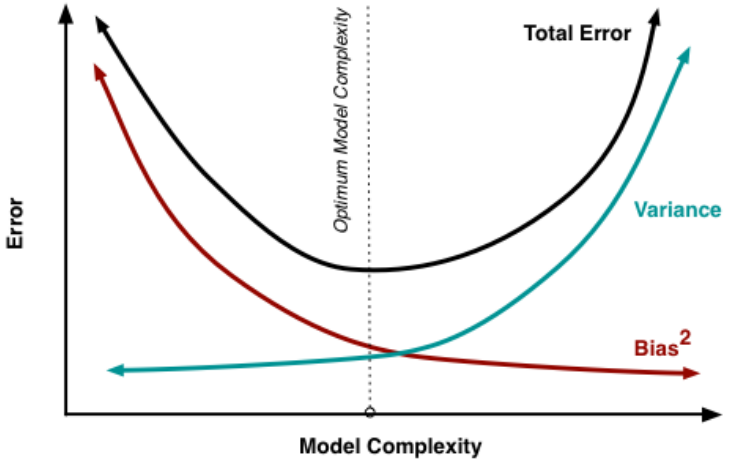


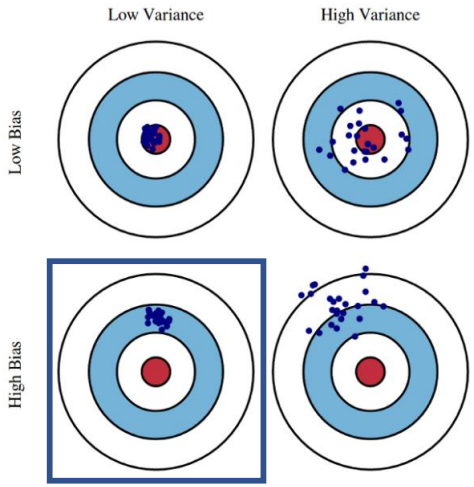
장점

- 계산 속도 빠름(랜덤 포레스트의 약 3배)
- 분산 감소

단점

- 모델의 무작위성이 강해서 더 많은 트리 훈련 필요
- Bias 증가

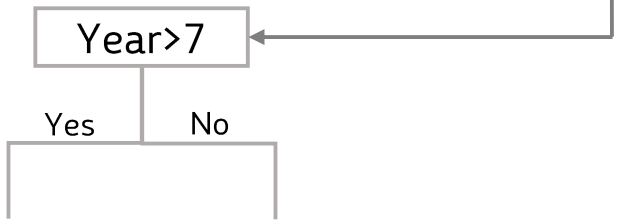




Random Forest vs Extra Tree

Random Forest

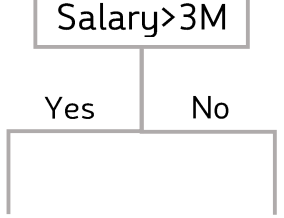
Feature	Information Gain	Selection
Year	72.9	O
Hits	32.1	X
Salary	23.7	X



Extra Tree

Feature	Random Selection
Year	X
Hits	X
Salary	O

Selected Feature	Split point	Information Gain
Salary	>3M	23.7



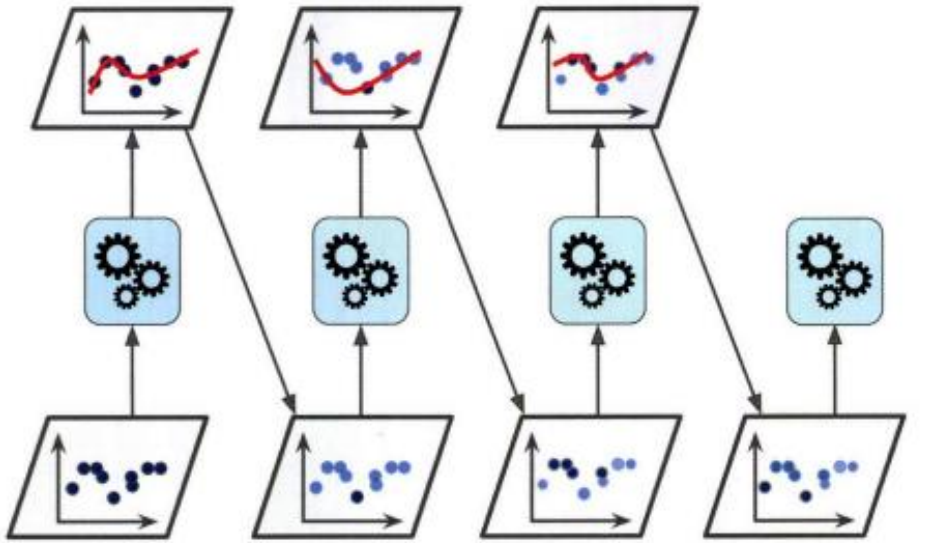
AdaBoost(Adaptive Boosting)

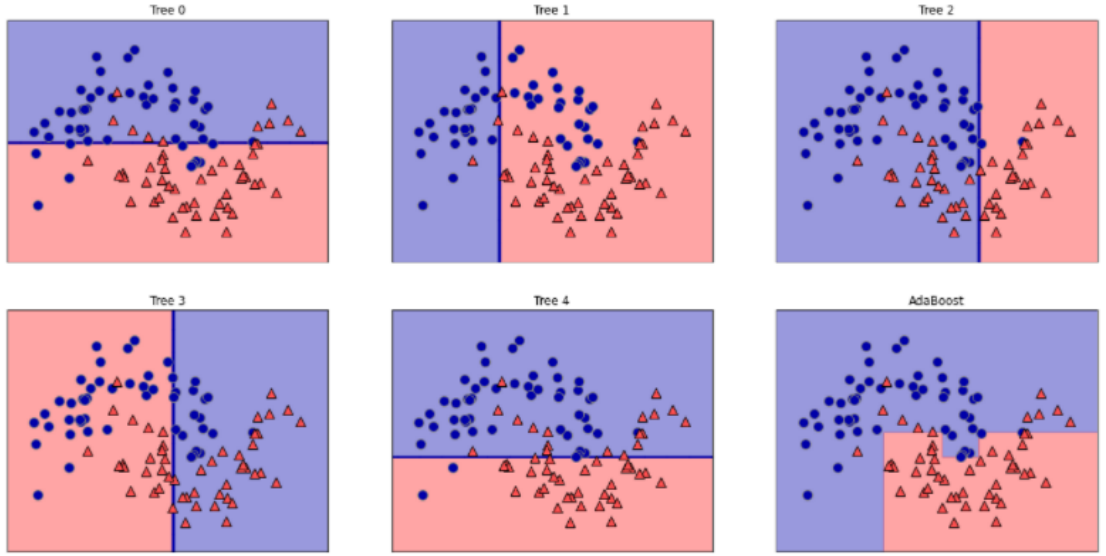
AdaBoost Regressor

*Boosting: 약한 학습기를 여러 개 연결하여 강한 학습기를 만드는 앙상블 방법

순차적으로 트리를 생성함으로써 이전 트리의 오차를 보완하여 전체 예측 오류를 최소화

→ 올바르게 예측되지 못한 샘플에 가중치를 더하고 올바르게 예측된 샘플에 대해서 가중치를 덜함





Grandient Boosting

Gradient Boosting Regressor

순차적으로 트리를 생성함으로써 이전 트리의 오차를 보완하여 전체 예측 오류를 최소화

*Boosting: 약한 학습기를 여러 개 연결하여 강한 학습기를 만드는 앙상블 방법

학습률(learning_rate)로 모델의 복잡도 결정

→ 학습률 낮추면 비슷한 복잡도의 모델 만들기 위해서 많은 트리 추가
→ 학습률 높이면 모델 복잡도 및 과대적합 가능성 상승

장점

- 약한 학습기를 사용하여 메모리 적게 사용하고 예측 속도 빠름

단점

- 랜덤포레스트보다 민감한 매개변수 설정
- 긴 훈련시간