텍스트, 도표, 라인, 스크린샷이(가) 표시된 사진

자동 생성된 설명**random forest**

회귀 및 분류 문제를 해결하는 데 사용할 수 있는 머신 러닝 기법이다. 여러 개의 의사 결정 트리로 구성되어 있으며 각각의 의사 결정 트리는 출력 결과를 내놓는다. 이때 여러 개의 의사 결정 트리가 내놓은 출력을 앙상블 기법 중 하나인 배깅을 사용하여 최종 결과를 출력한다. 배깅은 여러 개의 모델이 독립적이다. 랜덤 포레스트 역시 여러 개의 독립적인 의사 결정 트리로 이루어져 있다고 볼 수 있다.

장점 :

1. 앙상블이기 때문에 과적합이 잘 일어나지 않는다.
2. 앙상블이기 때문에 결측치나 이상치에 강하다.
3. 회귀 알고리즘이 아닌 의사결정나무 알고리즘에 기반한 기법이기 때문에 scaling, 정규화 과정이 아예 필요 없다.
4. 비선형적 데이터에 강하다. 이 역시 회귀 알고리즘이 아니라 의사결정나무이기 때문이다.
5. 앙상블이기 때문에 새로운 데이터가 들어와도 크게 영향을 받지 않는다.

단점 :

1. 노드가 많아질수록 학습 시간과 계산 연산량이 높다.

텍스트, 스크린샷, 도표, 폰트이(가) 표시된 사진

자동 생성된 설명**decision tree**

특정 기준에 따라 데이터를 분리하는 모델을 의사 결정 트리라고 한다. 사진과 같이 구성된 모델을 의사 결정 트리라고 이야기한다. 데이터를 분리할 수 있는 질문을 던지는 노드를 의사결정 노드라고 하고 그 질문으로 분리되는 노드를 리프 노드라고 한다. 그리고 트리의 꼭대기에 루트 노드가 존재한다. 리프 노드는 더 이상 데이터가 분리되지 않는 노드이다.

**도표, 텍스트, 라인, 스크린샷이(가) 표시된 사진

자동 생성된 설명gradient boosting**

도표, 라인, 스크린샷, 원이(가) 표시된 사진

자동 생성된 설명주로 결정 트리 기반이며 boosting 방법을 사용하는 모델이다. 학습이 진행될수록 전의 모델을 개선하는 방향으로 나아가기 때문에 예측력이 좋아진다. 즉 loss값이 줄어든다. 내가 가진 모델로 loss값을 최소화하기 위해서는 미분해서 기울기가 0인 부분으로 가까워져야한다. 이런 과정을 부스팅 방법과 결합하여 반복학습을 하는 것을 gradient boosting이라고 한다

장점

1. 머신러닝 계열의 모델 중 성능이 좋은 편이다.
2. tree 모델의 장점들을 가지고 있다. 수치형, 범주형 모두 예측이 가능하고 결측 데이터 처리, 피처 중요도 선택이 쉽다.

단점

1. 배깅과 비교할 경우 속도가 훨씬 느리다.
2. 잔차를 줄여나가는 학습 방법 상 과적합 가능성이 높아진다. 이를 방지하기 위한 방법은 subsampling을 잘 조절하여 1보다 작은 값을 설정한다.
3. 하이퍼파라미터 튜닝 시간이 오래 걸린다.