[투빅스 Week 3 정규 세션] 앙상블 모델 (Ensemble)

0. 분산 / 편향
   - 분산 : 개별적인 모델링이 퍼진 정도 (여러 번 진행 시 얼마나 결과가 서로 비슷하게 나오는지)
   - 편향 : 정답과 먼 정도
   - 분산 높고 편향 낮음 -> 모델의 복잡도가 낮음
   - 분산 낮고 편향 높음 -> 모델의 복잡도가 높음
1. 앙상블 모델 목적
   - 모델을 만들었을 때 분산 또는 편향이 너무 높은 경우 효과적인 모델이라 할 수 없음
   - 분산 줄이기 -> Bagging, 랜덤 포레스트
   - 편향 줄이기 -> AdaBoost
2. 결합 : 성능이 일정한 수준 + 다양한 모델
3. 배깅
   - 데이터셋 (bootstrap) 복원추출
   - 모델 복잡도가 낮을 때 적합
4. 랜덤 포레스트
   - 결정트리 기반 + 랜덤화
   - 변수 중요도 산출 가능
5. AdaBoost
   - 학습 후 잘못 분류된 데이터를 골라 다시 학습시킴 (랜덤 추출 X)
6. GBM (Gradient Boosting Machine)
   - 경사하강법 + 부스팅 기반
   - 잘못 분류된 데이터는 잔차를 학습시킴
7. XGBoost
   - Approximate algorithm
   - Aware split finding
8. 배깅 vs. 부스팅
   - 배깅은 다시 학습 시 모든 데이터의 추출 확률이 같다
   - 부스팅은 잘못 분류된 데이터 위주로 다시 추출해서 학습시킨다.