앙상블 모델 세션 과제

XGBoost와 LightGBM 기법 조사 및 요약

6기 2조 장윤태

1. XGBoost

Ensemble Boosting model들의 단점들을 보완하기 위해 탄생

그 중에서도 특히 Gradient boosting 모델의 순차적인 계산으로 인한 느린 속도, 잔차를 계속 줄임으로써 과적합 발생과 같은 단점을 보완

보완하는 방법으로는 느린 속도 문제점을 직렬식(순차적) 수행이 아닌 병렬식 수행으로 보완, 과적합 발생 문제점을 과도한 학습이 이뤄지지 않도록 정규화 식을 추가해 모델 복잡도가 상승할수록 손실값에 페널티를 부여하는 방식 구현으로 보완이 존재

이 외에도 결측치를 내부적으로 처리하여 정확도 상승, 가지치기를 통해 분할 수를 축소하여 효율적 분석, 조기 중단과 같은 특징이 존재

사용은 사이킷런 래퍼 XGBoost 클래스인 XGBClassifier 함수나 xgboost 패키지를 import하여 사용

주요 부스터 파라미터에는 learning\_rate, n\_estimator, max\_depth, min\_split\_loss, min\_child\_weight 등이 존재

1. LightGBM

Boosting의 단점을 보완한 XGBoost의 성능을 향상시키기 위해 탄생

XGBoost에서 병렬식 수행으로 속도 문제점을 어느정도 보완하였으나 여전히 긴 학습 시간이라는 단점 존재

보완하는 방법으로는 기존 트리 분할 방식을 리프 중심의 분할 방식을 채택함으로써 보완이 존재

기존에는 모든 depth가 동일하다는 가정하에 모든 node를 분할하여 균형 잡힌 형태의 트리를 생성하나 LightGBM은 상대적으로 분할이 더 필요한 node(손실이 큰 트리)에 대해서만 분할이 이뤄져 보다 효율적

깊이가 깊어져 과적합이 발생할 수도 있기에 max-depth 파라미터를 설정하여 조절이 필요

사용은 파이썬 패키지인 lightgbm에서 LGBMClassifier 함수를 import하여 사용

주요 부스터 파라미터에는 n\_estimator, max\_depth, learing\_rate, min\_child\_samples, num\_leaves, boosting\_type 등이 존재