* 1. 분석 모형
     1. XGBoost

부스팅이란 약한 학습기를 여러 개 연결해 강한 학습기를 만드는 앙상블 방법을 의미한다. 앞의 모델을 보완해 나가면서 일련의 예측기를 학습시키는 것이 아이디어이다. 부스팅 방법 중 가장 선호되는 것은 AdaBoost(Adaptive boosting의 줄임말)과 Gradient Boosting이다. Adaboost는 전의 예측기를 보완하는 새로운 예측기를 만들기 위해 이전 모델이 과소적합했던 훈련 샘플의 가중치를 더 높이는데, 이렇게 하면 새로운 예측기는 학습하기 어려운 샘플에 점점 더 맞춰지게 된다. 이런 방식을 에이다부스트에서 사용한다. 에이다부스트처럼 그레디언트 부스팅은 앙상블에 이전까지의 오차를 보정하도록 예측기를 순차적으로 추가하나, 에이다부스트처럼 반복마다 샘플의 가중치를 수정하는 대신 이전 예측기가 만든 잔여 오차에 새로운 예측기를 학습시킨다. XGBoost는 최적화된 그레디언트 부스팅을 구현한 것으로 Extreme Gradient Boosting의 약자이다. 이 패키지는 톈치 천이 개발했고 매우 빠른 속도, 확장성, 이식성을 목표로 한다.

* 1. 분석 자료

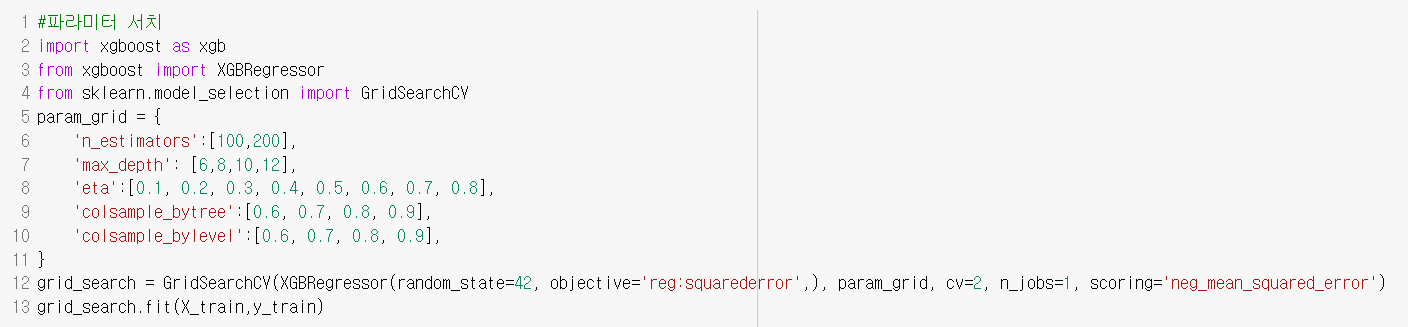
서울시 마포구의 날짜별, 시간별 기상상황과 따릉이 대여 수 데이터를 이용한다. 데이터는 시간, 기온, 비, 풍속, 습도, 시정, 오존, 미세먼지, 시간에 따른 따릉이 대여 수를 포함하고 있다.

* 1. 분석 방법

Xgboost의 따릉이 대여량에 대한 예측을 비교한다.

결측치는 0으로 대체한 후 IQR지수를 이용해 25%부터 75% 내에 들지 않는 이상치들을 제거해 전처리한 데이터를 사용한다. 7:3의 비율로 train 데이터셋과 test 데이터셋을 분리했으며 각 train 데이터들을 표준화함으로써 컬럼마다 달랐던 데이터들의 분포를 일정하게 맞춰준다.

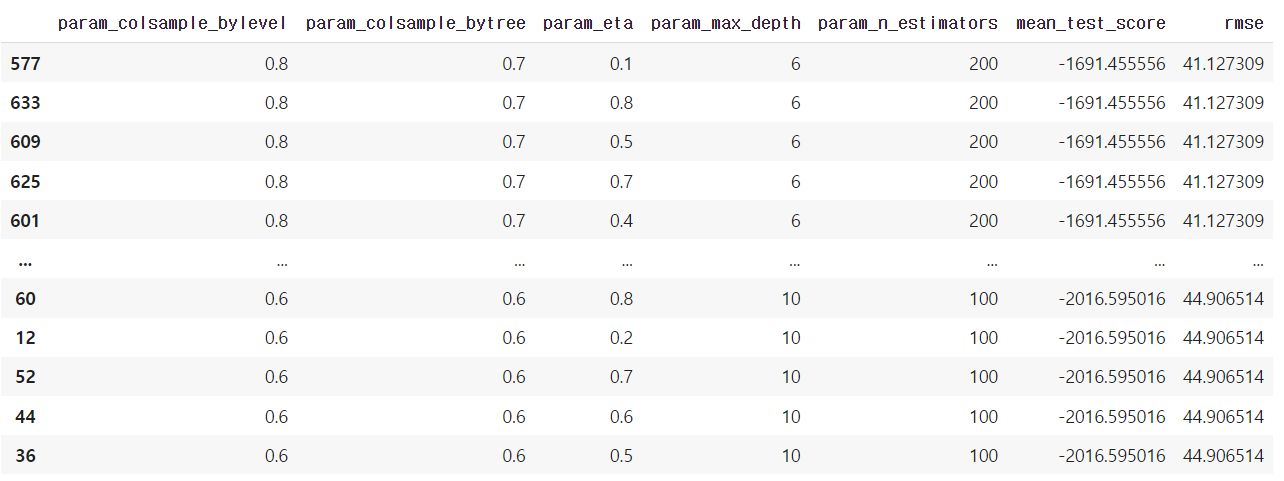
1. 분석
   1. 모형별 적합 결과
      1. XGBoost



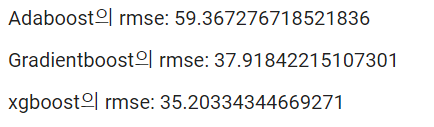
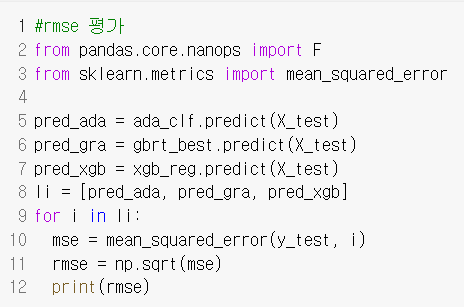
모델의 성능을 개선하기 위해 그리드서치로 최적의 파라미터를 찾은 결과, colsample\_bylevel이 0.8, colsample\_bytree가 0.7, eta가 0.1, max\_depth가 8, n\_estimators가 200일 때 가장 성능이 좋음을 확인할 수 있었다.

각 파라미터들을 살펴보자면 colsample\_bylevel은 트리의 레벨 별로 훈련 데이터의 변수를 샘플링해주는 비율이고, colsample\_bytree는 트리를 생성할 때 훈련 데이터에서 변수를 샘플링해주는 비율이다. Eta는 딥러닝에서의 학습률과 같은 개념으로 값이 너무 높으면 학습이 잘 되지 않을 수 있으며 너무 낮으면 학습이 느릴 수 있다. max\_depth는 트리 모델의 최대 깊이를 의미하고 n\_estimators는 생성할 트리의 갯수를 의미한다.

파라미터의 종류와 개수를 더 늘려 그리드 서치를 해보고 싶었으나 시간관계 상 위 5개의 파라미터만 적용했다.



각 적용된 파라미터 값들을 토대로 train 데이터셋 내에서의 rmse의 결과값들을 비교해 보면, 정해진 범위 내에서 colsample\_bylevel, colsample\_bytree, n\_estimators는 값이 클수록, max\_depth는 값이 작을수록 rmse 값이 줄어드는 것을 확인할 수 있었다. 그러나 학습률 eta는 rmse 값과 어떠한 연관성을 보이지는 않았다.



테스트 데이터셋에서 이 모델을 적용해 평가한 결과 xgboost의 rmse는 35.20334344669271이 나옴을 확인할 수 있었다.

1. 참고문헌

핸즈온 머신러닝 2판, 오렐리앙 제롱, 한빛미디어

Kaggle 우승작으로 배우는 머신러닝 탐구생활, 정권우, 비제이퍼블릭