* **XGBoost(eXtra Gradient Boost)**

**뛰어난 예측 성능**

**GBM 대비 빠른 수행 시간 :** CPU 병렬 처리, GPU지원

**다양한 성능 향상 기능 :** 규제(Regularization) 기능 탑재, Tree Pruning

**다양한 편의 기능 :** 조기중단(Early Stopping), 자체 내장된 교차 검증, 결손 값 자체 처리

|  |  |  |
| --- | --- | --- |
| **항목** | **파이썬 Wrapper** | **사이킷런 Wrapper** |
| **사용 모듈** | from xgboost as xgb | from xgboost import XGBClassifier |
| **학슴용과 테스트용**  **데이터세트** | DMatrix를 별도로 생성  train = xgb.DMatrix(data = X\_train, label=y\_train) | 넘파이나 판다스 이용 |
| **학습 API** | Xgb\_model = **xgb.train()** | XGBClassifier.fit() 바로 학습 |
| **예측 API** | Xgb\_model.predict()  반환결과는 예측 결과를 추정하는  **확률 값 반환** | XGBClassifier.predict()  **예측 결과값 반환** |
| **피처 중요도 시각화** | plot\_importance() | plot\_importance() |

* **XGBoost 조기 중단 기능 (Early Stopping)**

|  |  |
| --- | --- |
|  | XGBoost는 특정 반복 횟수 만큼 **더 이상 비용함수가 감소하지 않으면** 지정된 반복횟수를 다 완료하지 않고 **즉시** **수행 종료 가능.**  너무 반복횟수를 단축하면 예측 성능 최적화가 안된 상태에서 학습이 중단될 수도 있음 |

* **LightGBM 개요**

**XGBoost 대비 장점 :** 더 빠른 학습과 예측 수행 시간

더 작은 메모리 사용량

카테고리형 피처의 자동 변환과 최적 분할

|  |  |
| --- | --- |
| * **GBM, XGboost 의 트리 분할** | * **LightGBM의 트리 분할** |
| 최소 depth로 구성하기 위함. |  |

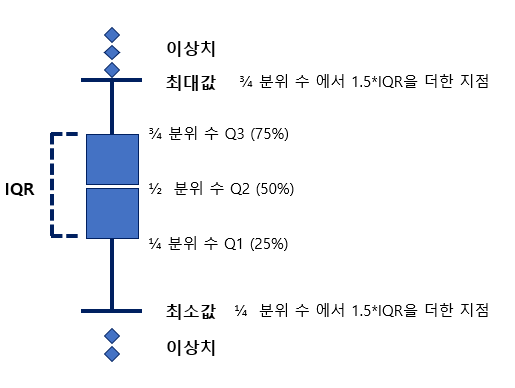
* **파이썬 래퍼와 사이킷런 래퍼 하이퍼 파라미터 비교**

|  |  |  |  |
| --- | --- | --- | --- |
| **유형** | **파이썬 래퍼 LightGBM** | **사이킷런 래퍼 LightGBM** | **사이킷런 래퍼 XGBoost** |
| **파라미터명** | num\_iterations | n\_estimators | n\_estimators |
| learning\_rate | learning\_rate | learning\_rate |
| max\_depth | max\_depth | max\_depth |
| min\_data\_in\_leaf | min\_child\_samples | N/A |
| bagging\_fraction | subsample | subsample |
| feature\_fraction | colsample\_bytree | colsample\_bytree |
| lambda\_I2 | reg\_lambda | reg\_lambda |
| lambda\_I1 | reg\_alpha | reg\_alpha |
| early\_stopping\_round | early\_stopping\_rounds | early\_stopping\_rounds |
| num\_leaves | num\_leaves | N/A |
| min\_sum\_hessian\_in\_leaf | min\_child\_weight | min\_child\_weight |

* **Log 변환**

왜곡된 분포도를 비교적 정규 분포에 가깝게 변환해줌

* **IQR(Inter Quantile Range) 를 이용한 Outlier Removal**

****

* **언더 샘플링(Undersampling)**

많은 레이블을 가진 데이터 세트를 적은 레이블을 가진 데이터 세트 수준으로 감소

* **오버 샘플링(Oversampling)**

적은 레이블을 가진 데이터 세트를 많은 레이블을 가진 데이터 세트 수준으로 증식

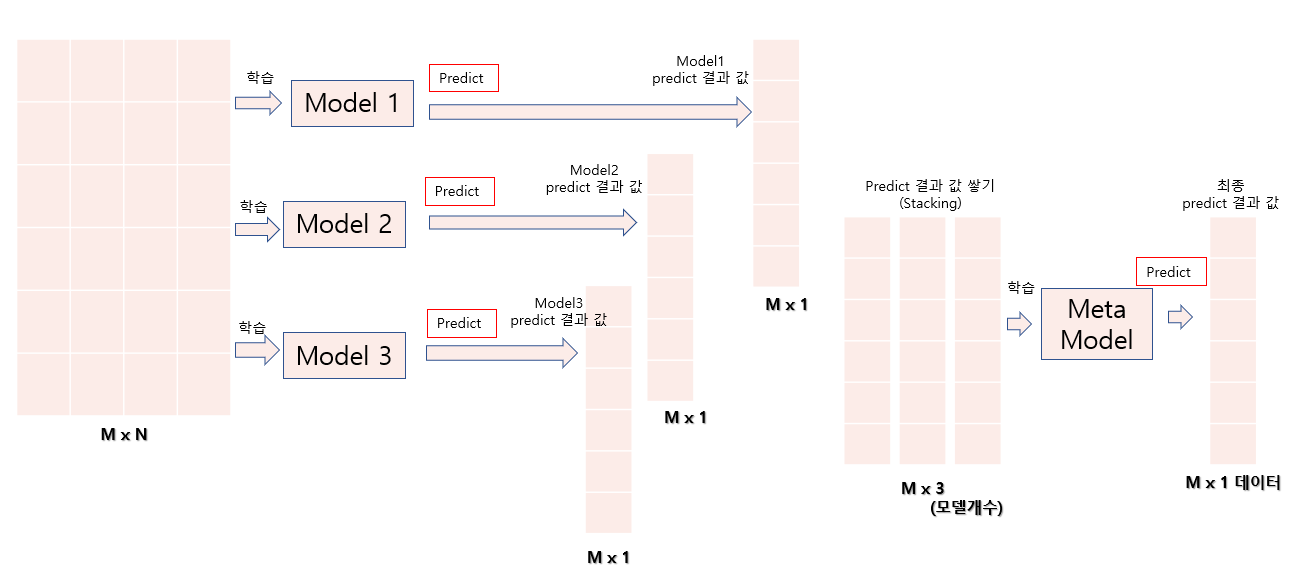
* **SMOTE(Synthetic Minority Over-Sampling Technique) 개요**

원본 데이터에서 근접 이웃의 데이터를 신규 증식하여 오버 샘플링.

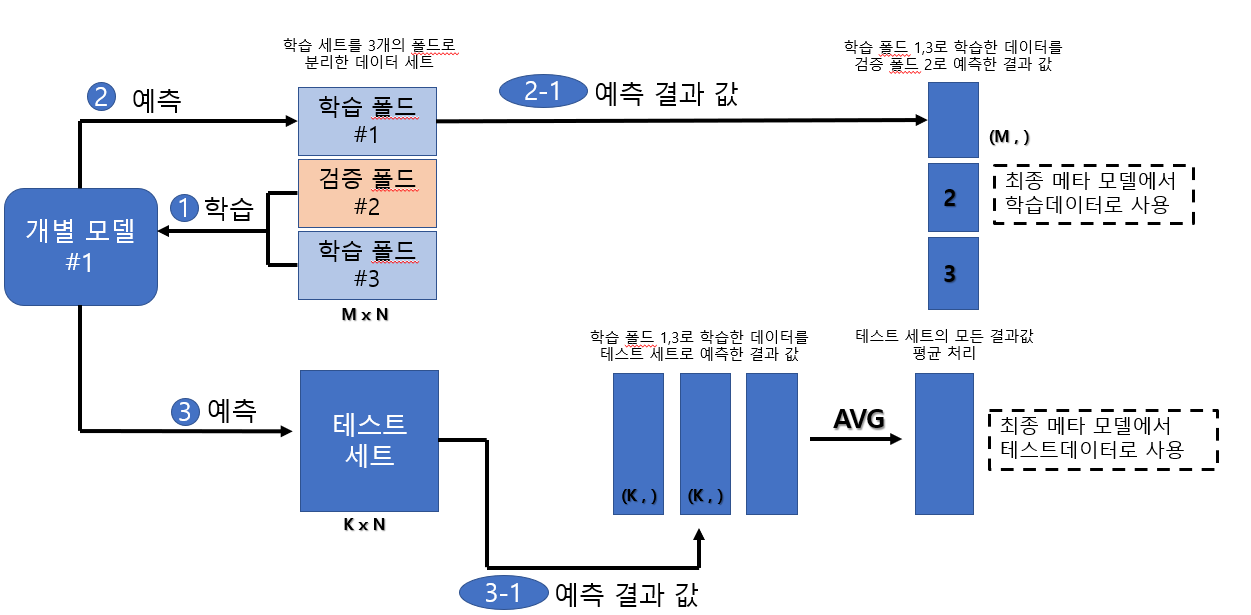
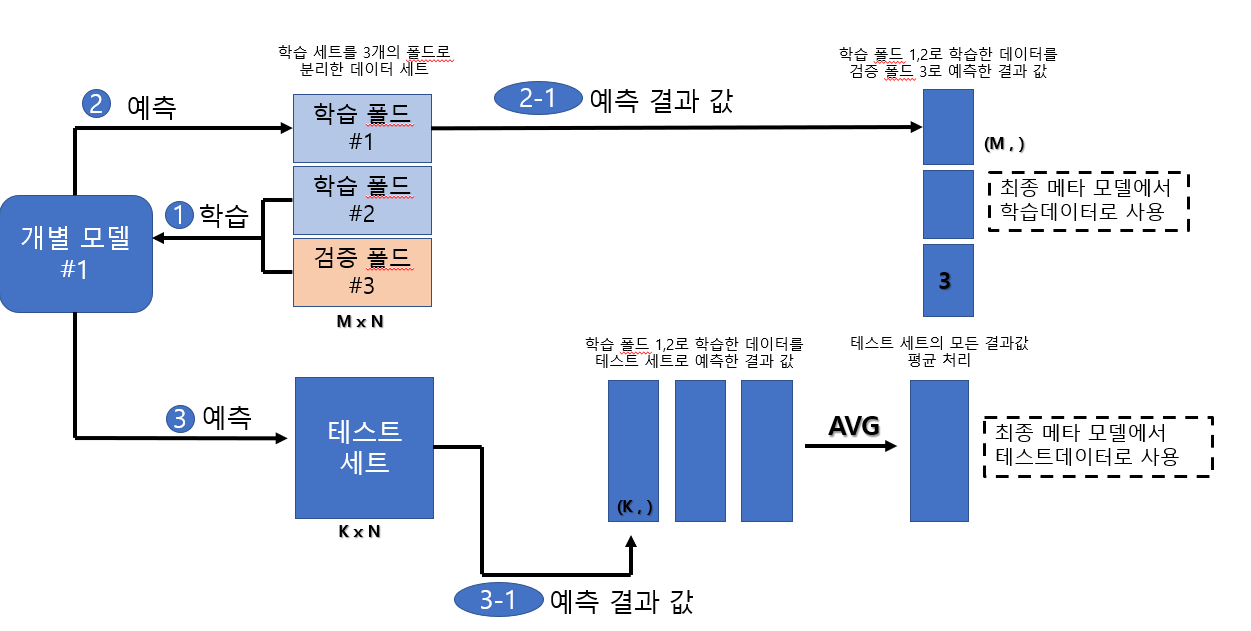
* **Basic Stacking Model – Diagram**

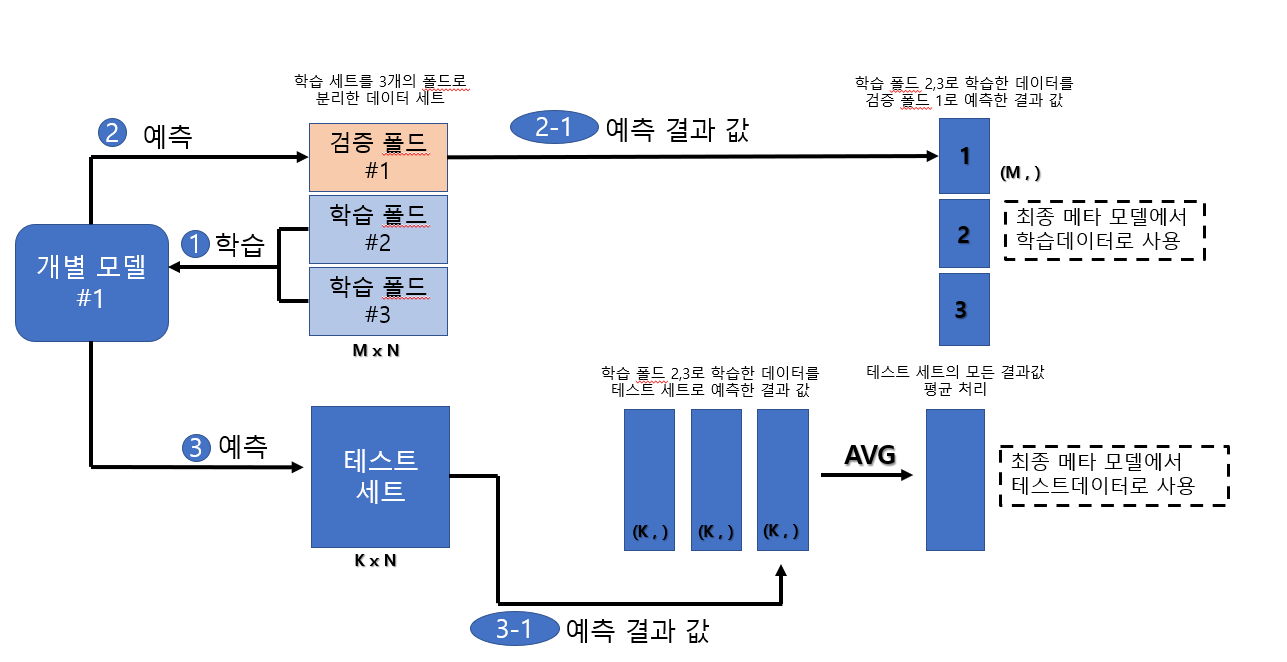
기반 모델들이 예측한 값을 Stacking하여, 메타 모델이 다시 학습하여 예측하는 모델

반드시 개발성능이 좋아진다고 보장되지는 않음.



* **교차 검증 세트 기반의 Stacking**

****



이를 개별모델 #1, 개별모델 #2, 개별모델 #3, 개별모델 #4 각각 학습데이터와 예측 데이터들을 모두 합쳐 최종 메타 모델이 학습하고 예측.

이때, 최종 평가는 원본 테스트 레이블과 비교하여 진행함.