2025/05/05 (이예지,이지민,정재선,조수빈)

YB 3조 수상작 리뷰 보고서

# **주제**

Otto Group Product Classification Challenge (다중 클래스 분류 문제)

### 주어진 제품 데이터를 바탕으로 해당 제품이 어떤 카테고리에 속하는지 예측하는 문제 (그러므로, 제품 분류 문제를 해결하는 머신러닝 모델 구축이 목적)

### XGBoost 모델의 학습 및 평가 방법 설명

### 모델 해석 (feature importance, tree graph 등)

### R 기반으로 XGBoost의 기능을 이해하고 시각화하는 데 초점

# **데이터**

* 데이터셋 구성
  + train.csv: 학습용 데이터 (61878 x 95, 94개 feature + 1개 target) – 15000개 이상의 제품에 대한 정보가 포함되어 있으며, 각 제품은 93개의 특성으로 설명됨. 목표 변수는 product\_category(0~8)로, 총 9개의 카테고리 중 하나로 분류됨.
  + test.csv: 테스트 데이터 (144368 x 94) – train 데이터와 유사하지만 라벨이 없는 상태로 제공됨. 이 데이터에 대한 예측을 제출해 리더보드 순위를 결정.
* 특징
  + 모든 feature는 정수형 (feat\_1 ~ feat\_93)
  + id 열은 정보 없음 → 삭제
  + target은 문자열 클래스 (Class\_1 ~ Class\_9) → 숫자로 변환

# **코드 분석**

* EDA
  + 각 특성의 통계량을 확인하며 데이터의 분포, 결측값 파악
  + Pandas와 matplotlib를 사용해 각 특성의 분포를 시각화하고 데이터의 특성 이해
  + 상관 행렬 시각화 : 특성 간의 상관 관계를 분석하여 모델에 유용할 수 있는 특성 간의 관계를 파악
* 전처리
  + Xgboost, data, table, magrittr methods
  + Fread( )로 빠르게 데이터 코딩
  + Id 칼럼 제거
  + Target 문자열을 숫자형 벡터로 변환
  + Data.table을 XGBoost의 사용을 위해 matrix 형태로 변환
  + StandarScaler나 MinMaxScaler 등을 사용하여 특성들을 정규화
  + 각 특성의 결측값을 평균 대체, 중앙값 대체 등으로 처리함
* 모델링 & 모델 이해
  + XGBoost 모델을 사용하여, 예측의 정확도를 높이는데 주력
  + Xgb.cv 통한 logloss 검증하여 xgboost( ) 학습을 수행
  + Xgb.importance( )로 gain 기준 feature 중요도 산출
  + Xgb.plot.importance( ) 시각화로 상위 10개 feature 표시
  + Xgb.dump( )을 통해 tree 구조 raw 출력
  + Xgb.plot.tree( )으로 실제 트리 구조 시각화
  + 앙상블 기법으로 여러개의 모델을 조합해 성능 개선
  + GridSearchCV나 RandomizedSearchCV를 활용하여 모델의 하이퍼파라미터를 최적화
  + 5-Fold 교차검증을 사용하여 모델의 성능을 안정적으로 평가하고 오버 피팅을 방지

# **리뷰 (느낀 점)**

* + XGBoost 모델을 사용해 높은 예측 성능을 얻고, Gradient Boosting 기법을 기반으로 하여 여러 모델을 앙상블하는 데 유리함을 배움
  + 앙상블 기법에서 하나의 모델만 사용할 때보다 여러 모델을 결합했을 때 성능이 더 좋아짐을 알 수 있었음
  + 5-fold 교차 검증을 통해 모델을 평가하고, 데이터의 분포에 관계없이 일반화 성능을 높일 수 있었음. 교차 검증 기법을 사용해 모델의 과적합을 방지하고 안정적인 예측이 가능하였음을 확인함.
  + 데이터를 제대로 이해하고 전처리하는 것이 전체 모델 성능에 큰 영향을 미치는 것을 확인함.
  + 데이터 전처리 과정에서 결측값을 평균화하거나 중앙값으로 대체하는 방법들이 보편적으로 사용됨을 알 수 있었음. 또한 평균화, 중앙값 대체 이외의 결측값 처리 방법 학습 및 이에 대한 적용의 필요성을 느낌.