# 데이터

Otto Group Product Classification Challenge는 상품의 속성을 보고 어떤 카테고리인지 예측하는 문제이다.

해당 대회에서 제공된 데이터셋은 93개의 feature와 200,000개 이상의 상품 데이터를 포함하고 있다. 목표는 주요 제품 카테고리들을 구별할 수 있는 예측 모델을 구축하는 것이다.

# 코드 흐름

스태킹+feature engineering+모델 앙상블 결합

1단계

1. 33개의 서로 다른 모델 학습
2. 모두 5-fold-cross-validation index로 학습
3. 데이터 전처리 방식: 로그 변환, 스케일링, 루트 변환 등

2단계

1. 1단계의 33개 메타 특성 + 1단계에서 생성된 7개의 특성을 입력
2. 3개의 모델 학습, 각 4-fold-cross-validation
3. 교차검증 결과를 통해 성능 추정 및 무의미한 모델 제거

3단계

1. 2단계 3개 모델의 예측값 결합
2. XGBoost에 비중을 많이 두고 NN과 ExtraTrees 예측도 함께 반영

# 차별점, 배울점

* 수작업으로 만든 8개의 engineered features를 포함하였다는 것이 인상적이었다.
* 교차검증 결과를 통해 무의미한 모델을 제거하는 과정도 적절한 타이밍에 필요하다는 것을 느꼈다.
* 여러 모델의 예측값을 결합하는 과정에서 특정 모델에 가중치를 두어 정확도를 높이는 과정이 중요하다고 느꼈다.