[Week4] 파머완 4장 필사

산탄데르 고객 만족 예측

대회 소개

- 고객 만족도를 높이기 위한 머신러닝 알고리즘
- 고객 데이터를 기반으로 맞춤형 상품 추천 제공
- 만족/불만족 고객의 데이터를 분류하는 이진 분류 문제

XGBoost

- GBM에 기반해서 느린 수행 시간 및 과적합 규제 부재 등의 단점을 보완
- 분류에 있어서 뛰어난 예측 성능을 나타냄

LightGBM

- XGBoost와 함께 부스팅 계열 알고리즘에 속함
- XGBoost보다 빠른 학습과 예측 수행 시간, 더 작은 메모리 사용량을 자랑

early_stopping_rounds

- 사이킷런 래퍼 XGBoost의 조기 중단 관련 파라미터 중 하나
- 평가 지표가 향상될 수 있는 반복 횟수를 정의
- 너무 급격하게 줄이면 충분한 학습이 되지 않아 오히려 예측 성능 저하될 수 있음

ROC 곡선 & AUC

- ROC 곡선 : FPR(실제는 negative지만 positive로 예측한 비율)이 변할 때 TPR(실제값과 예측값 모두 positive인 비율)이 어떻게 변하는지 나타내는 곡선
- AUC : ROC 곡선의 넓이
- ROC 곡선이 가운데 직선에서 멀어질수록 성능 뛰어남
- AUC가 1에 가까울수록 좋은 수치 ⇒ FPR이 작은 상태에서 얼마나 큰 TPR을 얻을 수 있는가

MoA 예측

대회 소개

MultiOutputClassifier

- 신약에 대한 매커니즘을 예측하는 알고리즘
- 약물에 여러 MoA가 있으므로 다중 레이블 분류 문제

다중분류

- 클래스가 3개 이상인 분류
- sklearn.multiclass : Multiclass classification
- sklearn.multioutput : Multilabel classification , Multioutput regression

Multi-label Classification

- 다중 분류 중에서도 하나의 샘플이 하나의 분류에만 속하는 것이 아니라 여러 개의 분류에 속하는 문제
- 분류 체계가 상호 배타적이지 않다는 특징
- 라벨이 0과 1로 이루어진 매트릭스 형태로 나타남

- 타겟 당 하나의 분류기를 학습시키는 전략
- 타게을 예측하기 위해 예측 함수를 측정
- 각 레이블을 독립적으로 처리
- 하이퍼 파라미터는 XGboost와 동일

predict_proba

- 불확실성을 추정하는 함수
- 각 샘플에 대해 어느 클래스에 속할지 그 확률을 0에서 1 사이의 값으로 돌려줌

logloss

- 분류 모델 성능 평가 시 사용 가능한 지표
- 모델이 얼마의 확률을 가지고 예측했는지에 초점
- predict_proba가 계산해 준 확률과 음의 로그함수를 이용하여 값을 계산 → 작을수록 좋은 모델