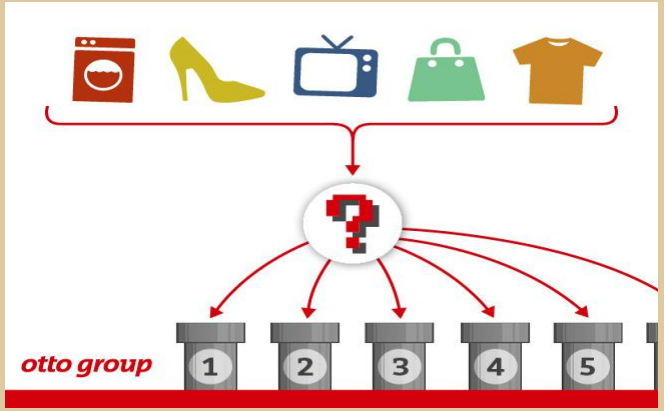
Kaggle Competitions ③

6. Otto group product classification challenge

◯ Data

- train.head() → (61878, 94)
- test.head() → (14368, 93)
- 데이터 설명
  - feature는 1부터 93까지 존재
  - category(target)은 Class1부터 9까지 존재
  - category는 가장 중요한 제품 범주 (패션, 전자제품 등) 중 하나를 나타냄



◯ Environment

- Python 3.73 with `jupyter==1.0.0`
- pandas, numpy, matplotlib, seaborn
- scikit-learn

➢ Perpose

익명화(anonymization)된 상품 정보 데이터를 통해 주어진 상품 카테고리 (target) Class 1~9에 대하여 Classification

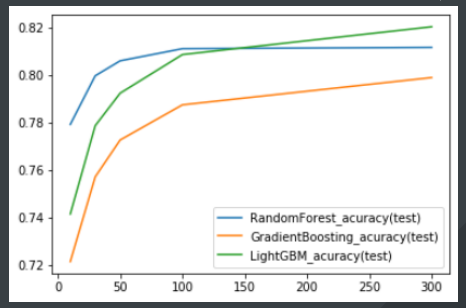
➢ Methodology

- 알고리즘 : RandomForest, GradientBoosting, lightGBM 성능 비교
- 모델평가 : multi-class logarithmic loss
  - N : number of products
  - M : number of class labels

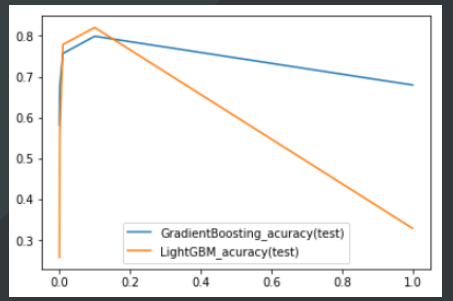
$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}),$$

➢ Solution

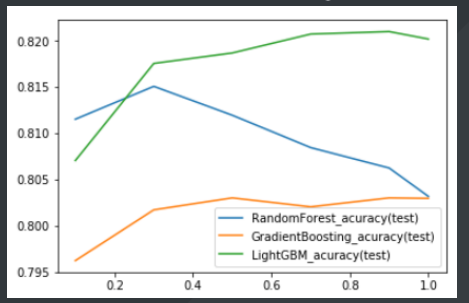
- 머신러닝 모델 성능 비교(Hold-Out Vallidation : test set의 Accuracy 사용)



【x1 = n_estimators】



【x2 = learning rate】



【x3 = max features】

- 스코어 : LightGBM 0.43259 예측(상위9.9% Rank)