5등 약식분석보고서

20202670 하서경

1. EDA

1차 Competition의 상위 팀이 사용한 Feature들이 제공받은 소스코드와 동일하여 상이한 Feature를 만들 필요성을 느꼈다. 고객이 구매한 상품 정보에 의미가 많이 담겨있다고 판단해 Group 또는 성별과 goodcd, brd_nm, corner_nm 등 상품 정보간의 다중분석을 진행하여 Target별 차이를 보이는 상품을 저장하고 이를 Feature Generation에 반영하였다.

2. Data Cleansing

1차 약식분석보고서에서 서술한 바와 동일하다.

3. Feature Engineering

1) Feature Generation

EDA에서 찾은 상품들의 주구매매장, 주구매브랜드, 평균액 등을 Feature로 생성하였다. 또한 1차 때는 사용하지 않았던 Word2Vec과 BOW를 생성하였다. 제공받은 파일 그대로가 아닌 Word2Vec의 벡터의 차원수를 train data의 mode, test data의 mode로 설정하고 Spare한 BOW는 PCA를 진행해 의미를 가질 몇 개의 Feature만을 활용하였다.

Generation을 거쳐 1차 Competition에서 사용한 Feature, EDA를 기반으로 생성한 Feature, Word2Vec, BOW 4가지의 Feature와 4가지 중 1차 Competition에서 사용한 Feature, Word2Vec, BOW 조합, EDA를 기반으로 생성한 Feature, Word2Vec, BOW 조합 2가지까지 총 6개의 Feature Set을 사용했다.

2) Feature Transformation

1차 약식분석보고서와 Outlier, Category Feature Encoding 등의 작업은 동일하게 진행하였으나 1차 Competition 상위팀이 Scaler만 사용하거나 Transformer만 사용한 점에서 최적의 Scaler와 Transformer 조합을 찾는 함수에 추가적으로 최적의 Scaler, 최적의 Transformer을 찾는 함수를 정의하고 적용하였다.

Word2Vec은 단어의 의미를 그대로 담은 채 고차원에서 저차원으로 단어를 표현하는 것으로 Transformation을 할 경우 단어의 의미가 변경될 수 있어 하지 않았다. BOW는 0, 1 그 자체에 의미가 있다 생각하여 Transformation 하지 않았다.

3) Feature Selection

생성한 6조합의 Feature들을 그대로 모델에 넣기엔 Feature수가 너무 많아 학습시간이 오래 걸리며 유의미하지 않은 Feature들이 섞여있어 성능이 좋게 나오지 않을 것이라 가정하였다. 실제로 사용할 모델이자 학습에 사용된 Feature Importance를 확인할 수 있는 Catboost와 LGBM으로 사용할 Feature들을 추리고자 하였다. 각 모델을 Default로 선언하고 전체 Feature를 넣어본 뒤 Importance가 1보다 큰 값들만 다시 넣어보는 과정을 반복해 Logloss가 감소하다 늘어나는 회차 직전에 모델에 넣은 Feature들만 사용하기로 추렸다. 6개 조합에 두 모델을 적용해 추려 생성된 12개의 Feature set 중 Logloss가 1.6 미만인 Feature Set만 사용하기로 했다.

최종적으로 사용하기로 한 Feature Set은 아래와 같다.

- 1차에서 사용한 Feature에 LGBM을 적용해 추린 148개 Feature(logloss: 1.58057)
- 1차에서 사용한 Feature들과 Word2Vec, BOW 조합을 Catboost로 추린 120개 Feature(‘’: 1.57055)
- 1차에서 사용한 Feature들과 Word2Vec, BOW 조합을 LGBM으로 추린 4656개 Feature(‘’: 1.54723)
- EDA를 기반으로 생성한 Feature, Word2Vec, BOW 조합 중 LGBM으로 추린 4135개 Feature(‘’: 1.58197)
- Word2Vec Feature 중 LGBM으로 추린 1530개 Feature(‘’: 1.57240)

4. Catboost

Feature Engineering을 거쳐 사용하기로 한 5개의 Feature Set을 Catboost에 넣어 학습시켰다. 이때 Catboost는 Tunning이 의의를 갖지 않는 모델로 iterations=1000, learning_rate=0.03, bootstrap_type='Bayesian'을 지정해 Catboost가 알아서 학습하도록 설정했다.

5. LGBM

Catboost에 의해 선택된 Feature Set을 제외하곤 Feature수가 많아 모델 Tunning이 작동하지 않았다. Catboost에 의해 선택된 120개의 Feature으로 Bayesian Optimization하고 학습하도록 하였다. 이 외의 Feature set은 아래의

DNN과 성능을 유사하게 하여 Ensemble해 사용하기를 목표로 하여 n_estimators, min_child_samples에 직접 임의의 값을 넣어보며 성능을 1.5 초반으로 맞추었다.

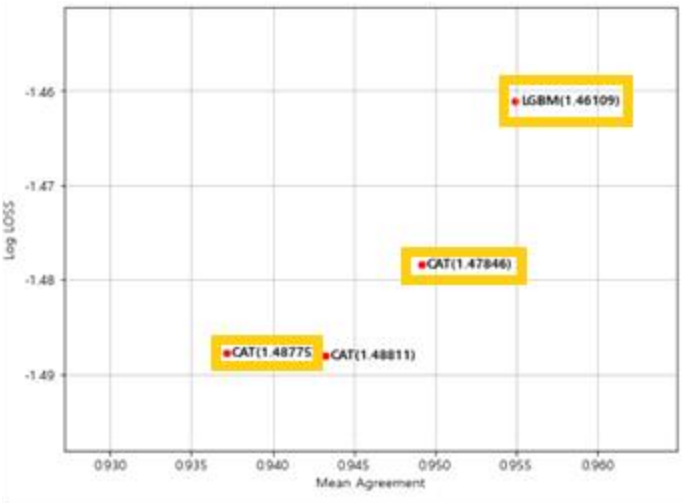
6. DNN

5개의 Feature Set으로 학습을 진행시키며 kt.Hyperband에 factor를 추가하는 등의 작업으로 1.5 초반의 성능을 갖는 DNN을 seed를 바꿔가며 여러 개 생성하였다.

7. Ensemble

LGBM과 DNN의 Ensemble 모델과 Catboost, 상위 팀 모델 등을 Ensemble하거나 LGBM끼리 Ensemble해 다른 모델과 Ensemble 하는 방법 등을 구상하고 시도하였으나 Ensemble 효과를 크게 보는 결과는 없었다. 이에 최종 제출 파일로 선택한 두 조합은 아래와 같다.

• 능선을 따라 택한 1, 2등의 모델과 1차 Competition에서 사용한 Feature들과 EDA를 기반으로 생성한 Feature, Word2Vec, BOW 조합 중 LGBM에 의해 선택된 Feature로 학습시킨 Catboost를 앙상블하였다.



• 능선을 따라 1차 Competition에서 사용한 Feature들과 Word2Vec, BOW 조합 중 LGBM에 의해 선택된 Feature와 Word2Vec Feature 중 LGBM을 통해 선택된 Feature를 사용한 2개의 DNN간의 Ensemble한 모델과 1,2,3등의 모델을 앙상블하였다.