DACON

Dacon 13회

Mission 13. 2019 Jeju BigData Competition
-퇴근시간 버스승차인원 예측 모델링 경진대회

Dining

목차

1 데이터 전처리

2 모델링

3 CV system

1. 데이터 전처리

1. Cleaning
2. Combine
3. Outlier remove
4. Agg feature
5. Frequency encoding

1. 데이터 전처리

1. Cleaning
   - Bus route id 0000 제거
   - Station name 띄어쓰기 제거
   - 위도, 경도 round해서 사용
2. Combine
   - Bus route id + station code
   - Bus route id + 위도, 경도 정보
3. Outlier remove
   - Train, test data 두개의 데이터 셋에 모두 존재하는 데이터만 사용

4. Agg Feature
   - Day 기준, 평균/80백분위수 데이터로 집계함수 사용
   - Day 기준, (bus route id, station code, station name) 평균/80백분위수 데이터로 집계함수 사용

5. Frequency encoding
   - Bus route id, station code, station name 등의 Categorical 변수의 경우 빈도수를 계산해 줌으로써 모델에 다양한 정보를 주었음

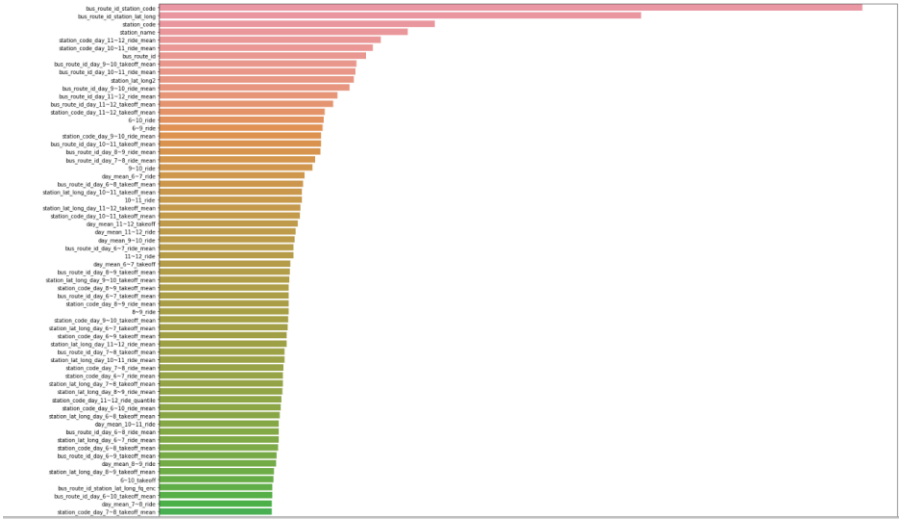
2. 모델링

1. Lgbm 사용
2. 파라미터

```
lgb_params = {
        'objective':'regression',
        'boosting_type':'gbdt',
        'metric':'rmse',
        'n_jobs':-1,
        'learning_rate':0.003,
        'num_leaves': 700,
        'max_depth':-1,
        'min_child_weight':5,
        'colsample_bytree': 0.3,
        'subsample':0.7,
        'n_estimators':50000,
        'gamma':0,
        'reg_lambda':0.05,
        'reg_alpha':0.05,
        'verbose':-1,
        'seed': SEED,
        'early_stopping_rounds':50
    }
```

데이터 다운로드

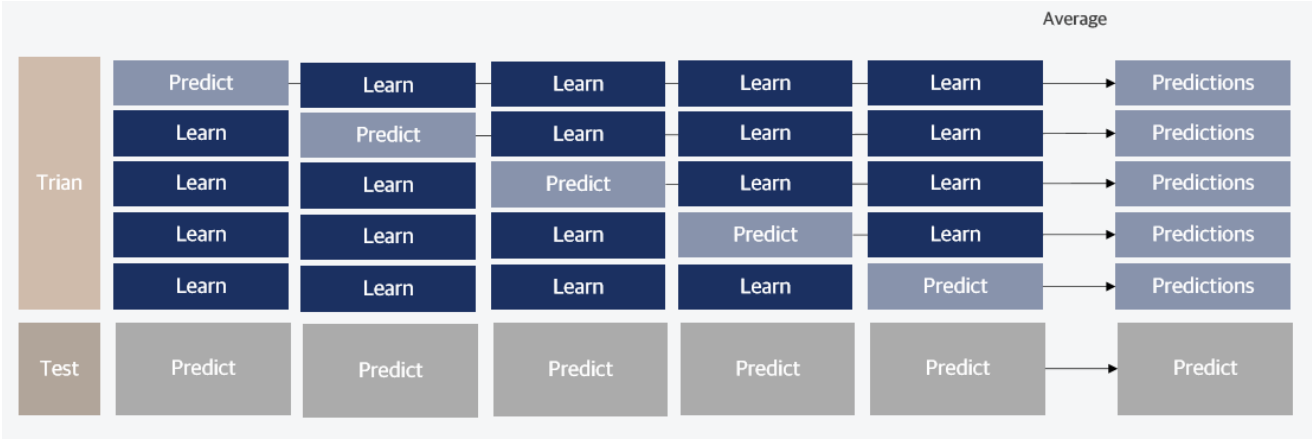
3. Feature importance



4. CV score와 lb score를 맞추기 위해 노력하였고,
마지막의 잘못된 피쳐를 제외하고는 모두 일치하였습니다.

3. CV system

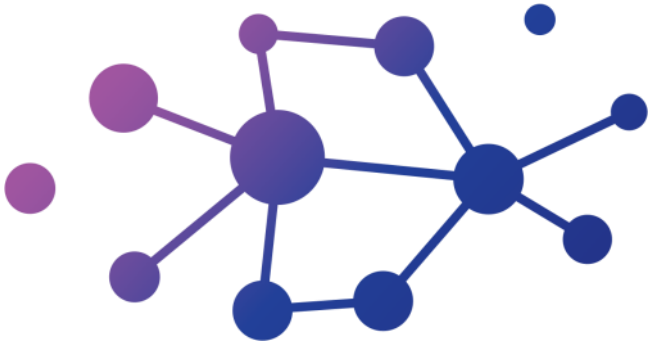
데이터 다운로드



1. Kfold를 이용한 out of folds 이용
2. StratifiedKFold 를 이용

3. 앙상블

THANK YOU



대회 참가해보기