

# 수요예측 모델(코르카 모델 내제화) Serving 및 refactoring 작업 공감

## \* Neural Ode Model(판매량 예측 모델) Serving 목적

- 1) 행사고도화 PO: MGS recommendor를 적용한 최적행사제안 리스트 산출
- 2) 구색최적화 PO: 점포-SKU의 판매량 예측

→ 목표: Model Training 및 Serving 과정의 Pipeline화

참고자료: [행사 고도화 설명 자료](#)

## \*Pipeline 작업을 위한 코드 Refactoring 작업 진행

### \*Refactoring의 목적

#### 1) 직관적이지 않은 코드, 간소화 및 구조화 필요

-여러개의 Script안에 Function들이 연결되어있는 구조로 Feature 추가 및 파라미터 변경등 코드 변경작업이 어려움

→모델 Train 및 Inference 과정까지 Function을 쪼개서 일부분만 수정/변경하여 손쉽게 학습하도록 진행 (Pytorch Lightning 모듈적용)

```
1 [In]: trainer = pl.Trainer(**trainer_args)
      executed in 93ms, finished 17:07:09 2021-10-28

      GPU available: True, used: True
      TPU available: False, using: 0 TPU cores
      IPU available: False, using: 0 IPU's

2 [In]: data_module = DataModule(**params) #data load & make data_loader
      executed in 9.25s, finished 17:07:19 2021-10-28

      Trainer v1.0
      **Loading train dataset...
      ['date': ['2020-01-01', '2021-09-30'], 'str_cd': ['1117'], 'prdt_cd': ['310']]
      train_data_loading_done.
      **Loading validation dataset...
      ['date': ['2021-10-01', '2021-10-10'], 'str_cd': ['1117'], 'prdt_cd': ['310']]
      validation_data_loading_done.
      train_data_loader start
      train_data_loader complete.
      valid_data_loader start
      valid_data_loader complete.
      data_loader done.

3 [In]: model = OdeNetLight(**params)
      executed in 5ms, finished 17:07:25 2021-10-28

4 [In]: trainer.fit(model, data_module) #model fitting
      execution queued 17:07:25 2021-10-28
      LOCAL_RANK: 0 - CUDA_VISIBLE_DEVICES: [0,1,2,3]

      | Name | Type | Params
      |-----|-----|-----
      0 | ode | SimpleVF | 44.7 K
      44.7 K | Trainable params
      0 | Non-trainable params
      44.7 K | Total params
      0.17B | Total estimated model params size (MB)

      Validation sanity check: 0%
```

- 모델학습과정: data\_loader→ trainer 생성→ 모델생성→ Fitting의 4단계 구조로 간소화
- \*Feature 추가시: data\_loader.py에 변수 추가
- \*파라미터 및 경로 변경시: main.py에 arg\_dict파일에서 변경

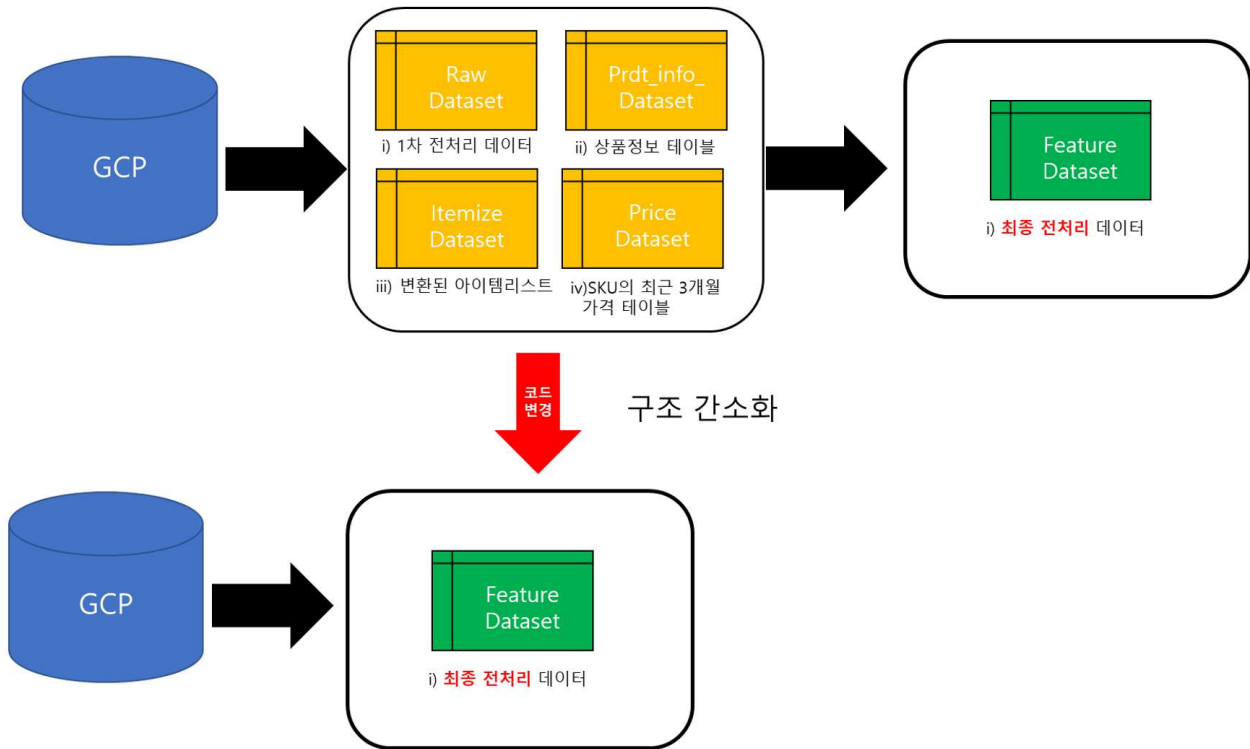
※Pytorch Lightning 공식사이트

#### 2)Serving을 위한 Data 전처리 및 Model 모델 구동시간 단축필요

- 변경전: GCP → RawDataset 생성시 3개월의 1개 점포 기준 약 3시간의 소요시간 필요

- 변경후: GCP → Feature Dataset으로 바로 생성된 테이블을 Load하여 1분이내의 소요시간 단축

**\*변경된처리 과정**



### 3) Serving Model 시연

(\*weekly때 Serving 과정 시연 예정)

-validation loss 기준 가장 좋은 Top3 모델 checkpoint pt 파일로 저장후 가장 loss가 낮은 model을 자동 load하여 판매량 inference 진행

-점포-상품의 예상판매량이 조회 가능함

-예시: 2021-10-27일 왕십리점의 샘표간장의 예측판매량

```
1aa=serving.inference(date='2021-10-27', str_cd='1117', prdt_cd='8801052971131')
```

executed in 2.16s, finished 13:23:38 2021-11-01

B0에서 요청사항 조회 중...

{'date': '2021-10-27', 'str\_cd': '1117', 'prdt\_cd': '8801052971131'}

처리 완료

예측한 데이터 수: 1개

데이터 로드: 2.15s

모델 Inference: 0.00s

```
1aa
```

executed in 6ms, finished 13:23:38 2021-11-01

meta정보(예측기간의 목요일)	예측값	실제값(존재시)
-------------------	-----	----------

0 2021-10-21_1117_8801052971131	34.179489	8.0
---------------------------------	-----------	-----

#### 4)향후계획

-모델 성능고도화

-Airflow pipeline 연동

-Tensorboard 연동