PORTFOLIO

Name
김영훈 YoungHoon Kim

Phone
010-9004-6355

Email
genholy77@gmail.com

INTRODUCTION

안녕하세요, 김영훈입니다.

저는 회계팀에서 3년 5개월 동안 재무데이터를 다루었고,
이후 KT프로젝트에서 언어데이터 수집 업무를 2년 4개월간 수행하였습니다.

대학원에서 인공지능학을 전공하여, Python을 활용한 인공지능 알고리즘 및 데이터 분석 기술을 익혔습니다. 또한 개인적인 스터디 활동으로 정교한 데이터 추출을 위해 SQL 기술을 익혔습니다.

여러 가지 도구를 활용한 데이터 분석으로 비즈니스적 의사결정에 필요한 인사이트를 제공하도록 하겠습니다.

EDUCATION

학력 사항

2023.08 숭실대학교 정보과학대학원 졸업

인공지능학 전공, 2021.03 - 2023.08

4.39

평균 학점(4.5 기준)

2013.02 한국외국어대학교 졸업

경제학 전공, 2005.03 - 2013.02

3.40

평균 학점(4.5 기준)

WORK EXPERIENCE

경력 사항

Defulx C&C KT DS 협력사
프리랜서, 2023.07-2023.12

- KT NH투자증권 음성자산화 사업
- 텍스트 분석(TA, Text Anlaysis) 엔진 데이터 구축 업무
- 성과 : 추출율 89% → 94%

Yesmanpower KT DS 파견업체
사원, 2021.08-2023.07

- KT AICC B2B / B2G 구독형 서비스
- 언어모델(LM, Language Model) 데이터 고도화 업무
- 성과 : 인식률 75% → 97%

천랩(CJ Bioscience) 재경팀
대리, 2017.11-2021.03

- 재무데이터 분석
- 분기/반기/연 회계결산 업무
- 성과 : IPO 심사 통과

PERSONAL SKILLS

핵심 역량

프로그래밍 언어 활용 능력

Python, SQL

문서 작성 능력

Excel, PowerPoint, Word

데이터 분석력

재무회계 데이터
언어 데이터

PROJECT

제주 특산물 가격 예측

Index

PROJECT

1. 개요

목적

제주도에는 다양한 특산물이 존재하고 그 중 양배추, 무, 브로콜리, 감귤은 제주도의 대표적인 특산물 중 일부. 특산물들의 안정적이고, 효율적인 수급을 위해 해당 특산물들의 가격에 대한 정확한 예측이 필요.

데이터 출처

데이콘, 제주테크노파크

URL

https://dacon.io/competitions/official/236176/overview/description

평가

심사 기준: RMSE

Public Score: 2023.03.04~2023.03.17의 데이터로 측정

Private Score: 2023.03.04~2023.03.31의 데이터로 측정

작업 방식

Python을 활용하여 탐색적 데이터 분석(EDA) 후, 가격 예측에 필요한 데이터 전처리 후 테스트셋에 대한 가격 예측

PROJECT

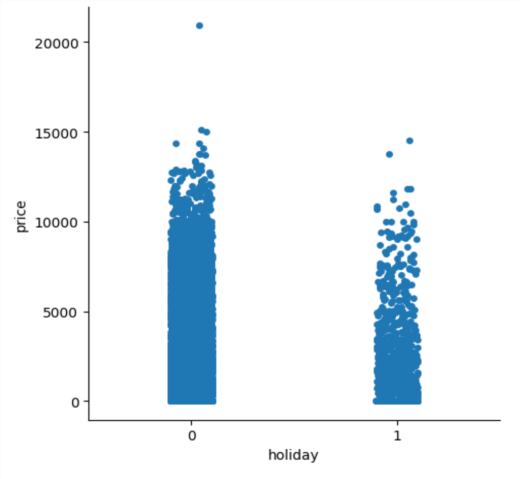
2. Data Set

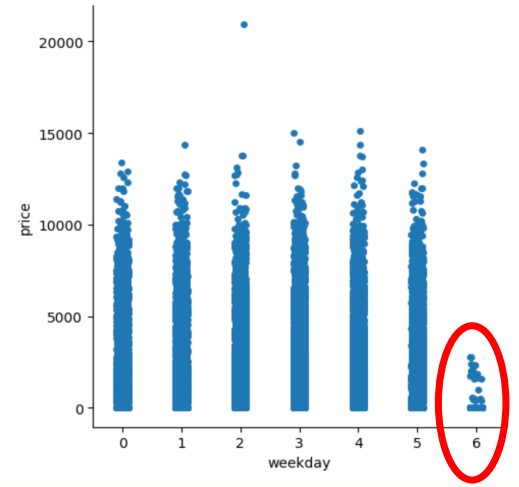
1) **train.csv :** 2019년 01월 01일부터 2023년 03월 03일까지의 유통된 품목의 가격 데이터
   - **item:** 품목 코드
     - TG : 감귤
     - BC : 브로콜리
     - RD : 무
     - CR : 당근
     - CB : 양배추
   - **corporation :** 유통 법인 코드
     - 법인 A부터 F 존재
   - **location :** 지역 코드
     - J : 제주도 제주시
     - S : 제주도 서귀포시
   - **supply(kg) :** 유통된 물량, kg 단위
   - **ID :** 품목(item)_법인(corporation)_지역(location)_날짜(timestamp) 조합 ex) TG_A_J_20230304
   - **timestamp :** 년-월-일 ex) 2019-01-01
   - **price(원/kg) – Target** : 유통된 품목들의 kg 마다의 가격, 원 단위
2) **test.csv :** 2023년 03월 04일부터 2023년 03월 31일까지의 데이터(ID, timestamp, item, corporation, location)

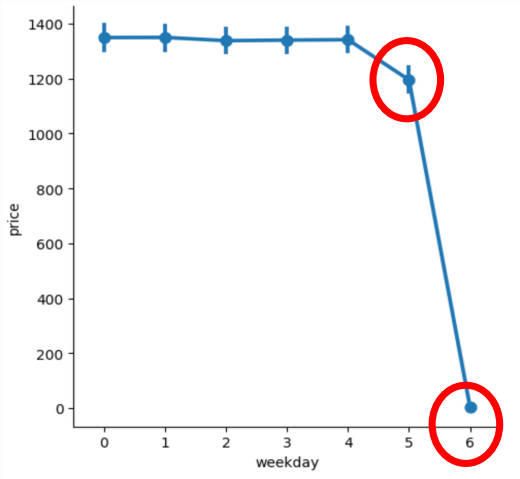
PROJECT

3. EDA & 변수 전처리

- **Timestamp 날짜**
  - 전처리 : timestamp 1개의 컬럼 (year-month-day) 포맷을 학습을 위해 3개의 컬럼으로 분리 → **year, month, day**
  - 공휴일 holiday 분석 : 예측할 데이터 기간(23.03.04~23.03.31)에는 공휴일 존재하지 않음으로 배제
  - 주말 weekday 분석 : 일요일에는 대부분 가격이 0인 것으로 확인됨 → ***토(5)/일요일(6) 변수 추가**하여 모델 성능 테스트



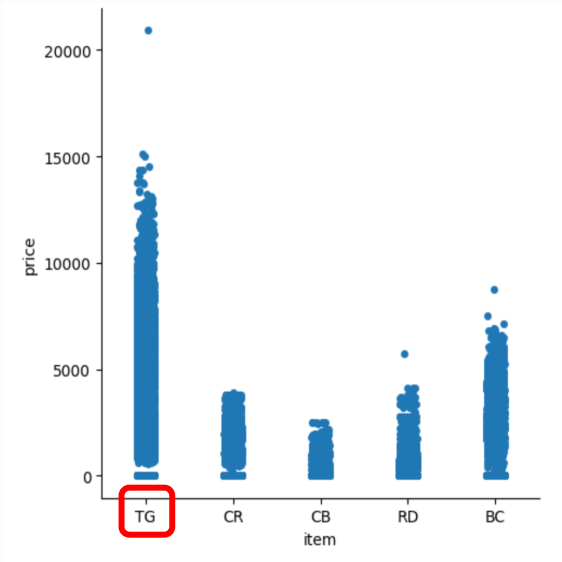




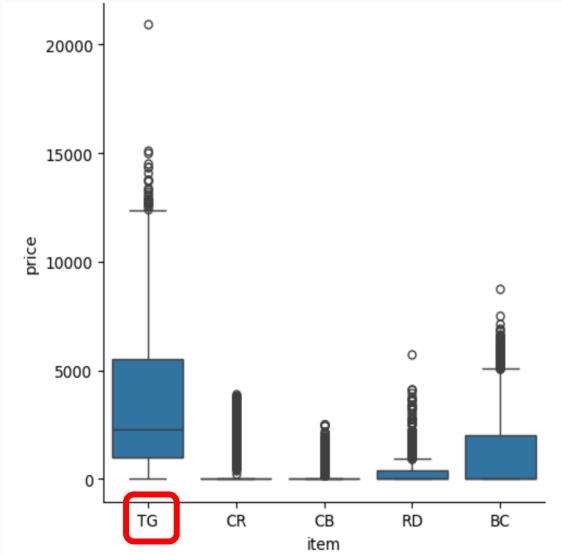
PROJECT

3. EDA & 변수 전처리

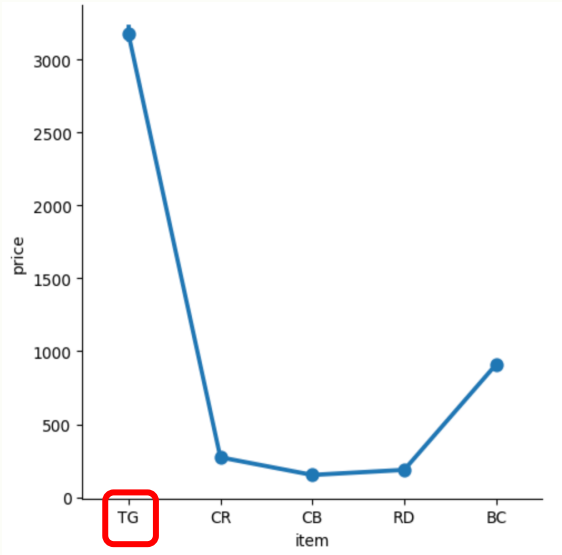
- **Item 상품 (TG 감귤, CR 당근, CB 양배추, RD 무, BC 브로콜리)**
  - 전처리 : 학습을 위해 예측 방법에 맞게 item별 One Hot Encoding
  - TG(감귤)와 그 외 item들간의 가격대 형성 차이가 큰 것으로 확인됨 → ***TG와 TG 아닌 상품들**로 분리하여 모델 성능 테스트
    → ***아이템별**로 분리하여 모델 성능 테스트



가격 분포 1



가격 분포 2

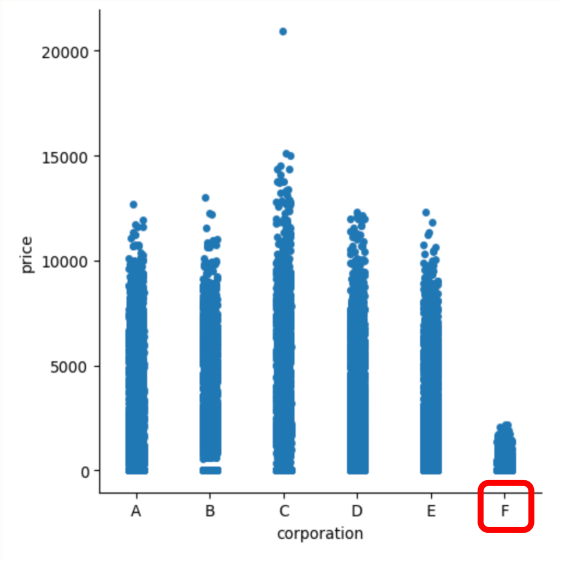


평균 가격

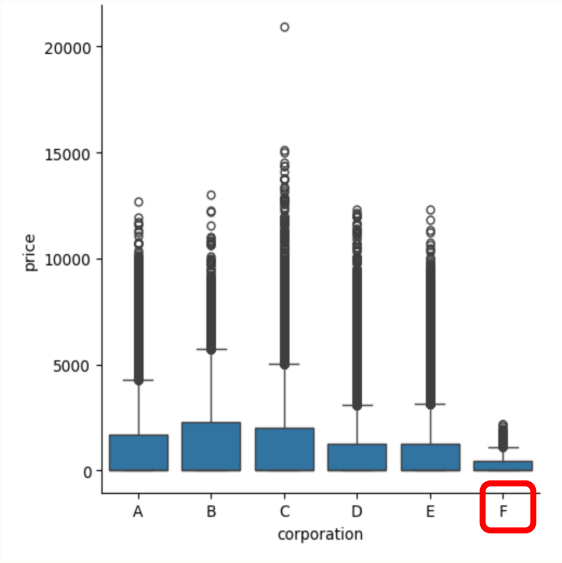
PROJECT

3. EDA & 변수 전처리

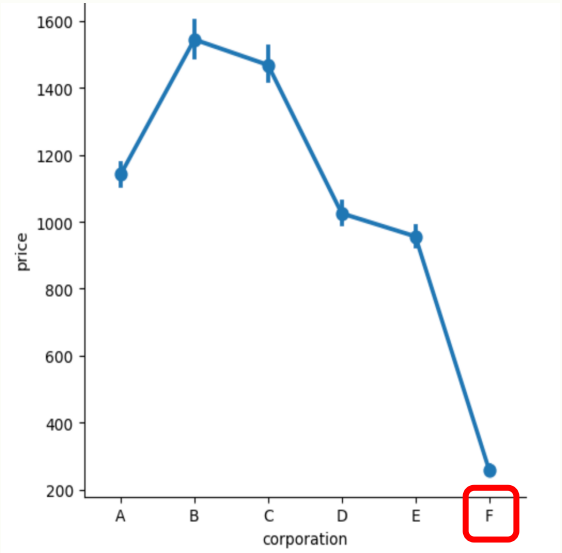
- **Corporation 법인 (A, B, C, D, E, F)**
  - 전처리 : 학습을 위해 corporation별 One Hot Encoding
  - F 법인과 그 외 법인간의 가격대 형성 차이가 큰 것으로 확인됨 → ***F법인 별도로 추가하여 성능 테스트**



가격 분포 1



가격 분포 2

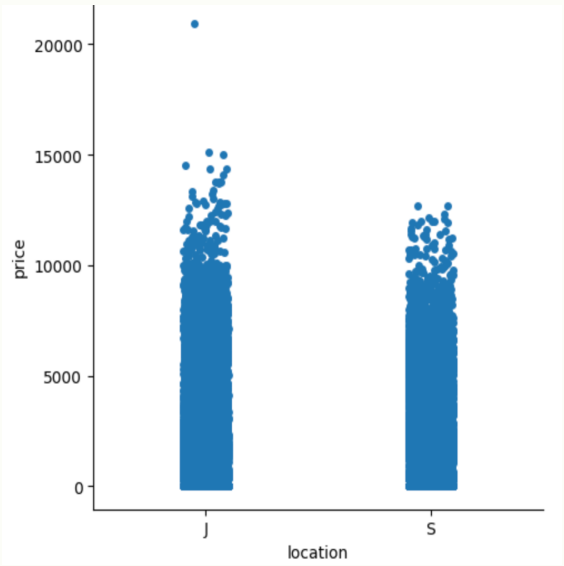


평균 가격

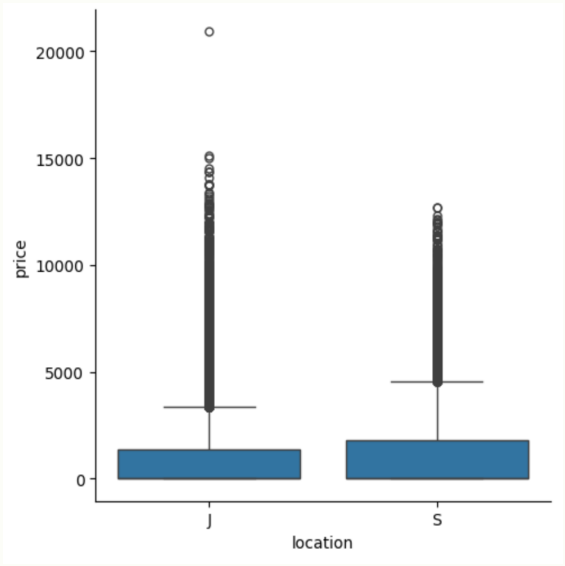
PROJECT

3. EDA & 변수 전처리

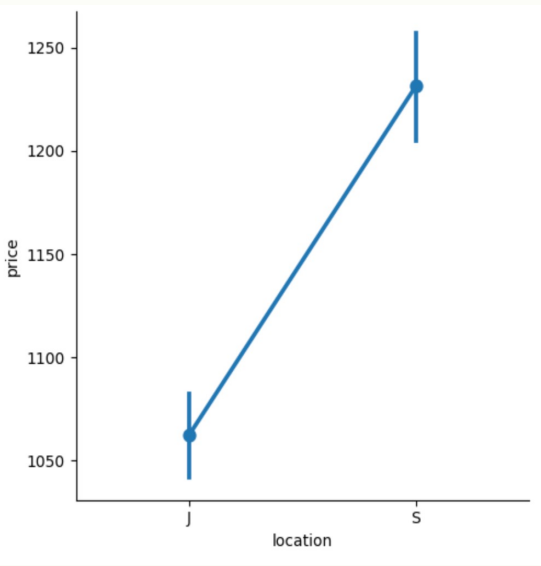
- **Location 위치 (J 제주시, S 서귀포시)**
  - 전처리 : 학습을 위해 location별 One Hot Encoding
  - 위치 간의 가격 차이가 크지 않음



가격 분포 1



가격 분포 2



평균 가격

PROJECT

3. EDA & 변수 전처리

- **Train data : 그 외 변수(ID, Timestamp, Supply) 전처리**
  - ID : 학습을 위해 제거
  - Timestamp : year, month, day 별도 칼럼 추가로 삭제
  - Supply : test 데이터에 없는 관계로 삭제

- **Test data : 그 외 변수(Timestamp) 전처리**
  - Timestamp : year, month, day 별도 칼럼 추가로 삭제

PROJECT

4. Baseline : Model(RandomForest) & Data

- Data
  - One hot encoding : item, corporation, location
  - Train data : year, month, day, item(BC~TG), corporation(A~F), location(J, S)
  - Test data : ID, year, month, day, item(BC~TG), corporation(A~F), location(J, S)

	year	month	day	BC	CB	CR	RD	TG	A	B	C	D	E	F	J	S
0	2019	1	1	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
1	2019	1	2	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
2	2019	1	3	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
3	2019	1	4	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
4	2019	1	5	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
59392	2023	2	27	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
59393	2023	2	28	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
59394	2023	3	1	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
59395	2023	3	2	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
59396	2023	3	3	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0

59397 rows × 16 columns

Train data

	ID	year	month	day	BC	CB	CR	RD	TG	A	B	C	D	E	F	J	S
0	TG_A_J_20230304	2023	3	4	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
1	TG_A_J_20230305	2023	3	5	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
2	TG_A_J_20230306	2023	3	6	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
3	TG_A_J_20230307	2023	3	7	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
4	TG_A_J_20230308	2023	3	8	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1087	RD_F_J_20230327	2023	3	27	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
1088	RD_F_J_20230328	2023	3	28	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
1089	RD_F_J_20230329	2023	3	29	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
1090	RD_F_J_20230330	2023	3	30	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
1091	RD_F_J_20230331	2023	3	31	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0

1092 rows × 17 columns

Test data

PROJECT

4. Baseline : Model(RandomForest) & Validation

- Model & Validation 검증
  - Model : RandomForest
  - K-Fold validation, n-splits : 10회
  - Average RMSE : 1197.28

```
KFold(n_splits=10, random_state=None, shuffle=True)
MSE : 1343209.589, RMSE : 1158.969
MSE : 1467121.658, RMSE : 1211.248
MSE : 1447040.574, RMSE : 1202.930
MSE : 1417953.904, RMSE : 1190.779
MSE : 1617910.046, RMSE : 1271.971
MSE : 1492477.497, RMSE : 1221.670
MSE : 1390400.919, RMSE : 1179.153
MSE : 1371705.528, RMSE : 1171.198
MSE : 1559911.451, RMSE : 1248.964
MSE : 1245246.933, RMSE : 1115.906
Average RMSE :  1197.2788188560464
```

Result

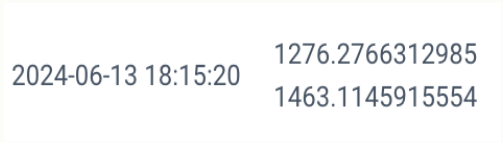
```python
kf = KFold(n_splits=10, random_state=None, shuffle=True)
kf.get_n_splits(X)

print(kf)


list_RMSE = []

for i, (train_index, valid_index) in enumerate(kf.split(X)) :

    X_train, X_valid = X.iloc[train_index], X.iloc[valid_index]
    Y_train, Y_valid = Y[train_index], Y[valid_index]

    model = RandomForestRegressor()
    model.fit(X_train, Y_train)

    val_predict = model.predict(X_valid)
    val_predict

    mse = mean_squared_error(Y_valid, val_predict)
    rmse = np.sqrt(mse)

    list_RMSE.append(rmse)

    print('MSE : {0:.3f}, RMSE : {1:.3f}'.format(mse, rmse))

mean = sum(list_RMSE) / len(list_RMSE)
print('Average RMSE : ', mean)
```

Code

PROJECT

4. Baseline : Model(RandomForest) & Prediction

- Model Test Data 예측 결과
  - Public Score (23.03.04~23.03.17) : **1276.28**
  - Private Score (23.03.04~23.03.31) : **1463.11**



Dacon 제출 결과

```
model = RandomForestRegressor()
model.fit(X, Y)

T_predict = model.predict(T)
print(T_predict)

pred_all['answer'] = T_predict
print(pred_all)
```

```
[3558.3  3801.39  887.88 ...  452.38  441.49  434.26]
                  ID   answer
0     TG_A_J_20230304  3558.30
1     TG_A_J_20230305  3801.39
2     TG_A_J_20230306   887.88
3     TG_A_J_20230307  3193.51
4     TG_A_J_20230308  3125.82
...               ...      ...
1087  RD_F_J_20230327   292.80
1088  RD_F_J_20230328   450.83
1089  RD_F_J_20230329   452.38
1090  RD_F_J_20230330   441.49
1091  RD_F_J_20230331   434.26

[1092 rows x 2 columns]
```

Code & Prediction Result

PROJECT

5-1. 개선 : Modle(RandomForest) & Data

- Data : **토/일요일(weekday) 변수 추가**
  - One hot encoding : item, corporation, location
  - Train data : is sat, is sun, year, month, day, item(BC~RD), corporation(A~F), location(J, S)
  - Test data : ID, is sat, is sun, year, month, day, item(BC~RD), corporation(A~F), location(J, S)

	is_saturday	is_sunday	year	month	day	BC	CB	CR	RD	TG	A	B	C	D	E	F	J	S
0	0	0	2019	1	1	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
1	0	0	2019	1	2	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
2	0	0	2019	1	3	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
3	0	0	2019	1	4	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
4	1	0	2019	1	5	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
59392	0	0	2023	2	27	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
59393	0	0	2023	2	28	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
59394	0	0	2023	3	1	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
59395	0	0	2023	3	2	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
59396	0	0	2023	3	3	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0

59397 rows × 18 columns

Train data

	ID	is_saturday	is_sunday	year	month	day	BC	CB	CR	RD	TG	A	B	C	D	E	F	J	S
0	TG_A_J_20230304	1	0	2023	3	4	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
1	TG_A_J_20230305	0	1	2023	3	5	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
2	TG_A_J_20230306	0	0	2023	3	6	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
3	TG_A_J_20230307	0	0	2023	3	7	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
4	TG_A_J_20230308	0	0	2023	3	8	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1087	RD_F_J_20230327	0	0	2023	3	27	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
1088	RD_F_J_20230328	0	0	2023	3	28	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
1089	RD_F_J_20230329	0	0	2023	3	29	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
1090	RD_F_J_20230330	0	0	2023	3	30	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
1091	RD_F_J_20230331	0	0	2023	3	31	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0

1092 rows × 19 columns

Test data

PROJECT

5-1. 개선 : Model(RandomForest) & Validation

- Model & Validation 검증
  - Model : RandomForest / K-Fold validation, n-splits : 10회
  - Average RMSE : 862.59

```
KFold(n_splits=10, random_state=None, shuffle=True)
MSE : 620841.223, RMSE : 787.935
MSE : 836327.857, RMSE : 914.510
MSE : 745720.318, RMSE : 863.551
MSE : 774668.357, RMSE : 880.152
MSE : 768942.163, RMSE : 876.893
MSE : 802182.929, RMSE : 895.647
MSE : 767190.674, RMSE : 875.894
MSE : 712988.813, RMSE : 844.387
MSE : 645194.974, RMSE : 803.240
MSE : 780790.625, RMSE : 883.624
Average RMSE :  862.5832728793141
```

Result

PROJECT

5-1. 개선 : Model(RandomForest) & Prediction

- Model Test Data 예측 결과
  - Public Score (23.03.04~23.03.17) : **811.99**
  - Private Score (23.03.04~23.03.31) : **1003.35**

2024-06-26 13:09:12	811.9891873238 1003.3502324417	

Dacon 제출 결과

```
[3769.9     0.   3461.05 ...  462.91  451.34  454.5 ]
                 ID   answer
0     TG_A_J_20230304  3769.90
1     TG_A_J_20230305     0.00
2     TG_A_J_20230306  3461.05
3     TG_A_J_20230307  3307.21
4     TG_A_J_20230308  3125.73
...               ...      ...
1087  RD_F_J_20230327   469.04
1088  RD_F_J_20230328   474.42
1089  RD_F_J_20230329   462.91
1090  RD_F_J_20230330   451.34
1091  RD_F_J_20230331   454.50

[1092 rows x 2 columns]
```

Prediction Result

PROJECT

5-2. 개선 : Modle(RandomForest) & Data

- Data : 토/일요일 변수 추가, **아이템 TG와 TG아닌 상품들로 데이터 분리**
  - One hot encoding : TG외 item(TG_etc), corporation, location
  - Train data : is sat, is sun, year, month, day, corporation(A~F), location(J, S), TG외 item(BC~RD), TG(item 열 삭제)
  - Test data : ID, is sat, is sun, year, month, day, corporation(A~F), location(J, S), TG외 item(BC~RD), TG(item 열 삭제)

	is_saturday	is_sunday	year	month	day	A	B	C	D	E	F	J	S	BC	CB	CR	RD
0	0	0	2019	1	1	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0
1	0	0	2019	1	2	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0
2	0	0	2019	1	3	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0
3	0	0	2019	1	4	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0
4	1	0	2019	1	5	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
44162	0	0	2023	2	27	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0
44163	0	0	2023	2	28	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0
44164	0	0	2023	3	1	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0
44165	0	0	2023	3	2	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0
44166	0	0	2023	3	3	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0

44167 rows × 17 columns

ex) Train data TG외 item

	ID	is_saturday	is_sunday	year	month	day	A	B	C	D	E	F	J	S	BC	CB	CR	RD
0	CR_A_J_20230304	1	0	2023	3	4	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0
1	CR_A_J_20230305	0	1	2023	3	5	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0
2	CR_A_J_20230306	0	0	2023	3	6	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0
3	CR_A_J_20230307	0	0	2023	3	7	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0
4	CR_A_J_20230308	0	0	2023	3	8	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
807	RD_F_J_20230327	0	0	2023	3	27	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0
808	RD_F_J_20230328	0	0	2023	3	28	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0
809	RD_F_J_20230329	0	0	2023	3	29	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0
810	RD_F_J_20230330	0	0	2023	3	30	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0
811	RD_F_J_20230331	0	0	2023	3	31	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0

812 rows × 18 columns

ex) Test data TG외 item

PROJECT

5-2. 개선 : Model(RandomForest) & Validation

- Model & Validation 검증
  - Model : RandomForest / K-Fold validation, n-splits : 10회
  - Average RMSE : 684.28

```
KFold(n_splits=10, random_state=None, shuffle=True)
tg
MSE : 2612816.198, RMSE : 1616.421
MSE : 2372449.733, RMSE : 1540.276
MSE : 2502382.902, RMSE : 1581.892
MSE : 2739680.966, RMSE : 1655.198
MSE : 2181305.250, RMSE : 1476.924
MSE : 2482109.532, RMSE : 1575.471
MSE : 2681517.875, RMSE : 1637.534
MSE : 2742449.721, RMSE : 1656.034
MSE : 2757528.569, RMSE : 1660.581
MSE : 2308008.349, RMSE : 1519.213
Average RMSE_tg : 1591.954476061176
------------------------------------------------------
etc
MSE : 157896.370, RMSE : 397.362
MSE : 141984.939, RMSE : 376.809
MSE : 127123.625, RMSE : 356.544
MSE : 148945.928, RMSE : 385.935
MSE : 122864.571, RMSE : 350.520
MSE : 141230.786, RMSE : 375.807
MSE : 141771.587, RMSE : 376.526
MSE : 154497.546, RMSE : 393.062
MSE : 110638.254, RMSE : 332.623
MSE : 135191.563, RMSE : 367.684
Average RMSE_etc : 371.28718434509994
------------------------------------------------------
RMSE =  [1591.95447606  371.28718435]
weights =  [0.25641026 0.74358974]
Weight Average =  684.2787976056322
```

Result

```python
from sklearn.model_selection import KFold

kf = KFold(n_splits=10, random_state=None, shuffle=True)
print(kf)

items = ['tg', 'etc']

list_Result = []

for i in items :

  X_KF = data[f'X_{i}']
  Y_KF = data[f'Y_{i}']
  mean_KF = f'mean_{i}'

  kf.get_n_splits(X_KF)

  print(i)

  list_RMSE = []

  for j, (train_index, valid_index) in enumerate(kf.split(X_KF)) :

      X_KF_train, X_KF_valid = X_KF.iloc[train_index], X_KF.iloc[valid_index]
      Y_KF_train, Y_KF_valid = Y_KF[train_index], Y_KF[valid_index]

      model = RandomForestRegressor()
      model.fit(X_KF_train, Y_KF_train)

      val_predict = model.predict(X_KF_valid)
      val_predict

      mse = mean_squared_error(Y_KF_valid, val_predict)
      rmse = np.sqrt(mse)

      list_RMSE.append(rmse)

      print('MSE : {0:.3f}, RMSE : {1:.3f}'.format(mse, rmse))

  mean_KF = sum(list_RMSE) / len(list_RMSE)

  print('Average RMSE_{} :'.format(f'{i}') , mean_KF)

  print('------------------------------------------------------------------------')
```

Code

PROJECT

5-2. 개선 : Model(RandomForest) & Prediction

- Model Test Data 예측 결과
  - Public Score (23.03.04~23.03.17) : **810.81**
  - Private Score (23.03.04~23.03.31) : **1010.90**

2024-06-25 19:42:16	810.8132633135 1010.904635023

Dacon 제출 결과

	ID	answer
0	TG_A_J_20230304	3727.89
1	TG_A_J_20230305	0.00
2	TG_A_J_20230306	3486.40
3	TG_A_J_20230307	3365.65
4	TG_A_J_20230308	3122.93
..	...	...
275	TG_E_S_20230327	4794.31
276	TG_E_S_20230328	4869.63
277	TG_E_S_20230329	4825.91
278	TG_E_S_20230330	4655.96
279	TG_E_S_20230331	4220.99

[280 rows x 2 columns]

	ID	answer
0	CR_A_J_20230304	573.63
1	CR_A_J_20230305	0.00
2	CR_A_J_20230306	2846.68
3	CR_A_J_20230307	2853.34
4	CR_A_J_20230308	2850.81
..	...	...
807	RD_F_J_20230327	480.32
808	RD_F_J_20230328	472.75
809	RD_F_J_20230329	455.72
810	RD_F_J_20230330	441.90
811	RD_F_J_20230331	453.03

[812 rows x 2 columns]

Prediction Result

```
data1 = {
    'X_tg': X_tg,
    'Y_tg': Y_tg,
    'T_tg': T_tg,
    'X_etc': X_etc,
    'Y_etc': Y_etc,
    'T_etc': T_etc,
    'pred_tg': pred_tg,
    'pred_etc': pred_etc
}

items1 = ['tg', 'etc']


for i in items1 :

  print(i)

  X_tst = data1[f'X_{i}']
  Y_tst = data1[f'Y_{i}']
  T_tst = data1[f'T_{i}']
  pred_tst = data1[f'pred_{i}']
  model_tst = f'model_{i}'
  T_predict_tst = f'T_predict_{i}'

  model_tst = RandomForestRegressor()
  model_tst.fit(X_tst, Y_tst)

  T_predict_tst = model_tst.predict(T_tst)
  print(T_predict_tst)

  pred_tst['answer'] = T_predict_tst
  print(pred_tst)

  print('------------------------------------------------------------------------')
```

Code

PROJECT

5-3. 개선 : Model(RandomForest) & Data

- Data : 토/일요일 변수 추가, **아이템 상품별로 데이터 분리**
  - One hot encoding : corporation, location
  - Train data : is sat, is sun, year, month, day, corporation(A~F), location(J, S), item 열 삭제
  - Test data : ID, is sat, is sun, year, month, day, corporation(A~F), location(J, S), item 열 삭제

	is_saturday	is_sunday	year	month	day	A	B	C	D	E	F	J	S
0	0	0	2019	1	1	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
1	0	0	2019	1	2	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
2	0	0	2019	1	3	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
3	0	0	2019	1	4	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
4	1	0	2019	1	5	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
15225	0	0	2023	2	27	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
15226	0	0	2023	2	28	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
15227	0	0	2023	3	1	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
15228	0	0	2023	3	2	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
15229	0	0	2023	3	3	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0

15230 rows × 13 columns

ex) Train data TG

	ID	is_saturday	is_sunday	year	month	day	A	B	C	D	E	F	J	S
0	TG_A_J_20230304	1	0	2023	3	4	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
1	TG_A_J_20230305	0	1	2023	3	5	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
2	TG_A_J_20230306	0	0	2023	3	6	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
3	TG_A_J_20230307	0	0	2023	3	7	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
4	TG_A_J_20230308	0	0	2023	3	8	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
275	TG_E_S_20230327	0	0	2023	3	27	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
276	TG_E_S_20230328	0	0	2023	3	28	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
277	TG_E_S_20230329	0	0	2023	3	29	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
278	TG_E_S_20230330	0	0	2023	3	30	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
279	TG_E_S_20230331	0	0	2023	3	31	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0

280 rows × 14 columns

ex) Test data TG

PROJECT

5-3. 개선 : Model(RandomForest) & Validation

- Model & Validation 검증
  - Model : RandomForest / K-Fold validation, n-splits : 10회
  - Average RMSE : 648.80

```
TG
MSE : 2302418.849, RMSE : 1517.372
MSE : 2728409.114, RMSE : 1651.790
MSE : 2621879.505, RMSE : 1619.222
MSE : 2952461.017, RMSE : 1718.273
MSE : 2423900.892, RMSE : 1556.888
MSE : 2034108.321, RMSE : 1426.222
MSE : 2818383.728, RMSE : 1678.804
MSE : 2356127.744, RMSE : 1534.968
MSE : 2506496.928, RMSE : 1583.192
MSE : 2103386.200, RMSE : 1450.306
Average RMSE_TG : 1573.7036624804143
```

```
RD
MSE : 26006.275, RMSE : 161.265
MSE : 26867.369, RMSE : 163.913
MSE : 45990.748, RMSE : 214.455
MSE : 18115.255, RMSE : 134.593
MSE : 30056.154, RMSE : 173.367
MSE : 28797.359, RMSE : 169.698
MSE : 36387.842, RMSE : 190.756
MSE : 25155.746, RMSE : 158.606
MSE : 18586.169, RMSE : 136.331
MSE : 29674.987, RMSE : 172.264
Average RMSE_RD : 167.5246717438412
```

```
BC
MSE : 271020.332, RMSE : 520.596
MSE : 296062.955, RMSE : 544.117
MSE : 263279.281, RMSE : 513.107
MSE : 325121.279, RMSE : 570.194
MSE : 373465.165, RMSE : 611.118
MSE : 315264.177, RMSE : 561.484
MSE : 326488.465, RMSE : 571.392
MSE : 314749.074, RMSE : 561.025
MSE : 381661.785, RMSE : 617.788
MSE : 248725.051, RMSE : 498.723
Average RMSE_BC : 556.9544169836182
```

```
CR
MSE : 101746.602, RMSE : 318.977
MSE : 75777.972, RMSE : 275.278
MSE : 94954.537, RMSE : 308.147
MSE : 108601.778, RMSE : 329.548
MSE : 112876.752, RMSE : 335.971
MSE : 88211.223, RMSE : 297.004
MSE : 116615.888, RMSE : 341.491
MSE : 106007.660, RMSE : 325.588
MSE : 96967.640, RMSE : 311.396
MSE : 95955.924, RMSE : 309.768
Average RMSE_CR : 315.3167968828101
```

```
CB
MSE : 47445.918, RMSE : 217.821
MSE : 41626.180, RMSE : 204.025
MSE : 42228.060, RMSE : 205.495
MSE : 47865.025, RMSE : 218.781
MSE : 41183.598, RMSE : 202.937
MSE : 31161.489, RMSE : 176.526
MSE : 45159.712, RMSE : 212.508
MSE : 34175.868, RMSE : 184.867
MSE : 33689.972, RMSE : 183.548
MSE : 42293.665, RMSE : 205.654
Average RMSE_CB : 201.21626537289384
```

```
RMSE =  [1573.70366248  167.52467174  315.31679688  201.21626537  556.95441698]
weights =  [0.25641026 0.20512821 0.17948718 0.12820513 0.23076923]
Weight Average =  648.7980168372198
```

Result

PROJECT

5-3. 개선 : Model(RandomForest) & Prediction

- Model Test Data 예측 결과
  - Public Score (23.03.04~23.03.17) : **817.75**
  - Private Score (23.03.04~23.03.31) : **979.84**

2024-06-25 19:30:29	817.7534573048 979.8373629522

Dacon 제출 결과

TG

	ID	answer
0	TG_A_J_20230304	3718.15
1	TG_A_J_20230305	0.00
2	TG_A_J_20230306	3489.47
3	TG_A_J_20230307	3428.24
4	TG_A_J_20230308	3178.00
..	...	...
275	TG_E_S_20230327	4814.34
276	TG_E_S_20230328	4862.73
277	TG_E_S_20230329	4812.10
278	TG_E_S_20230330	4562.07
279	TG_E_S_20230331	4193.86

[280 rows x 2 columns]

RD

	ID	answer
588	RD_A_J_20230304	508.76
589	RD_A_J_20230305	0.00
590	RD_A_J_20230306	604.62
591	RD_A_J_20230307	608.79
592	RD_A_J_20230308	566.44
...	...	...
1087	RD_F_J_20230327	520.04
1088	RD_F_J_20230328	481.87
1089	RD_F_J_20230329	460.58
1090	RD_F_J_20230330	468.48
1091	RD_F_J_20230331	466.69

[224 rows x 2 columns]

CR

	ID	answer
280	CR_A_J_20230304	438.57
281	CR_A_J_20230305	0.00
282	CR_A_J_20230306	2720.59
283	CR_A_J_20230307	2722.46
284	CR_A_J_20230308	2718.12
..	...	...
471	CR_E_S_20230327	0.00
472	CR_E_S_20230328	0.00
473	CR_E_S_20230329	0.00
474	CR_E_S_20230330	0.00
475	CR_E_S_20230331	0.00

[196 rows x 2 columns]

CB

	ID	answer
476	CB_A_J_20230304	509.71
477	CB_A_J_20230305	0.00
478	CB_A_J_20230306	651.48
479	CB_A_J_20230307	634.13
480	CB_A_J_20230308	351.94
...	...	...
1059	CB_F_J_20230327	827.75
1060	CB_F_J_20230328	814.48
1061	CB_F_J_20230329	838.45
1062	CB_F_J_20230330	830.38
1063	CB_F_J_20230331	753.79

[140 rows x 2 columns]

BC

	ID	answer
784	BC_A_J_20230304	2768.96
785	BC_A_J_20230305	0.00
786	BC_A_J_20230306	2654.07
787	BC_A_J_20230307	2481.67
788	BC_A_J_20230308	2464.84
...	...	...
1031	BC_E_S_20230327	2621.23
1032	BC_E_S_20230328	2764.70
1033	BC_E_S_20230329	1650.27
1034	BC_E_S_20230330	2764.55
1035	BC_E_S_20230331	2684.52

[252 rows x 2 columns]

Prediction Result

PROJECT

5-4. 개선 : Model(RandomForest) & Data

- Data : 토/일요일 변수추가, 특이점이 있는 아이템 TG와 기업F만 표시하고, item/corporation열 삭제
  - One hot encoding : location
  - Train data : is sat, is sun, is TG, is F, year, month, day,location(J, S)
  - Test data : ID, is sat, is sun, is TG, is F, year, month, day,location(J, S)

	is_saturday	is_sunday	is_TG	is_F	year	month	day	J	S
0	0	0	1	0	2019	1	1	1.0	0.0
1	0	0	1	0	2019	1	2	1.0	0.0
2	0	0	1	0	2019	1	3	1.0	0.0
3	0	0	1	0	2019	1	4	1.0	0.0
4	1	0	1	0	2019	1	5	1.0	0.0
...	...	...	...	...	...	...	...	...	...
59392	0	0	0	1	2023	2	27	1.0	0.0
59393	0	0	0	1	2023	2	28	1.0	0.0
59394	0	0	0	1	2023	3	1	1.0	0.0
59395	0	0	0	1	2023	3	2	1.0	0.0
59396	0	0	0	1	2023	3	3	1.0	0.0

59397 rows × 9 columns

Train data

	ID	is_saturday	is_sunday	is_TG	is_F	year	month	day	J	S
0	TG_A_J_20230304	1	0	1	0	2023	3	4	1.0	0.0
1	TG_A_J_20230305	0	1	1	0	2023	3	5	1.0	0.0
2	TG_A_J_20230306	0	0	1	0	2023	3	6	1.0	0.0
3	TG_A_J_20230307	0	0	1	0	2023	3	7	1.0	0.0
4	TG_A_J_20230308	0	0	1	0	2023	3	8	1.0	0.0
...	...	...	...	...	...	...	...	...	...	...
1087	RD_F_J_20230327	0	0	0	1	2023	3	27	1.0	0.0
1088	RD_F_J_20230328	0	0	0	1	2023	3	28	1.0	0.0
1089	RD_F_J_20230329	0	0	0	1	2023	3	29	1.0	0.0
1090	RD_F_J_20230330	0	0	0	1	2023	3	30	1.0	0.0
1091	RD_F_J_20230331	0	0	0	1	2023	3	31	1.0	0.0

1092 rows × 10 columns

Test data

PROJECT

5-4. 개선 : Model(RandomForest) & Validation

- Model & Validation 검증
  - Model : RandomForest / K-Fold validation, n-splits : 10회
  - Average RMSE : 1244.65

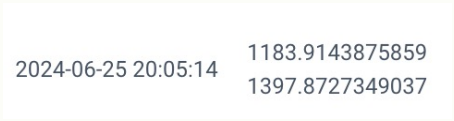
```
KFold(n_splits=10, random_state=None, shuffle=True)
MSE : 1577769.056, RMSE : 1256.093
MSE : 1602659.955, RMSE : 1265.962
MSE : 1652743.851, RMSE : 1285.591
MSE : 1440094.820, RMSE : 1200.040
MSE : 1485806.242, RMSE : 1218.937
MSE : 1447156.759, RMSE : 1202.978
MSE : 1550723.954, RMSE : 1245.281
MSE : 1502885.187, RMSE : 1225.922
MSE : 1721811.000, RMSE : 1312.178
MSE : 1521560.680, RMSE : 1233.516
Average RMSE : 1244.6496401818035
```

Result

PROJECT

5-4. 개선 : Model(RandomForest) & Prediction

- Model Test Data 예측 결과
  - Public Score (23.03.04~23.03.17) : **1183.91**
  - Private Score (23.03.04~23.03.31) : **1397.87**



Dacon 제출 결과

```
[2573.95462193    0.         2906.18792857 ...  717.104       670.89183333
  615.142     ]
                  ID       answer
0     TG_A_J_20230304  2573.954622
1     TG_A_J_20230305     0.000000
2     TG_A_J_20230306  2906.187929
3     TG_A_J_20230307  3042.197425
4     TG_A_J_20230308  2812.599536
...               ...          ...
1087  RD_F_J_20230327   723.668167
1088  RD_F_J_20230328   674.515333
1089  RD_F_J_20230329   717.104000
1090  RD_F_J_20230330   670.891833
1091  RD_F_J_20230331   615.142000

[1092 rows x 2 columns]
```

Prediction Result

PROJECT

5-5. 개선 : Model(LightGBM) & Data, Validation, Prediction

- Data : 토/일요일 변수 추가, **아이템 TG와 TG아닌 상품들로 데이터 분리**

- Model & Validation 검증
  - Model : LightGBM / K-Fold validation, n-splits : 10회
  - Average RMSE : 704.17

```
RMSE =  [1537.66834167  416.7604027 ]
weights =  [0.25641026 0.74358974]
Weight Average =  704.172694740452
```

- Model Test Data 예측 결과
  - Public Score (23.03.04~23.03.17) : **822.59**
  - Private Score (23.03.04~23.03.31) : **955.88**

Validation Result

2024-06-24 18:28:11	822.5870891896 955.8783279211

Dacon 제출 결과

PROJECT

5-6. 개선 : Model(LightGBM) & Data, Validation, Prediction

- Data : 토/일요일 변수 추가, **아이템별로 데이터 분리**

- Model & Validation 검증
  - Model : LightGBM / K-Fold validation, n-splits : 10회
  - Average RMSE : 650.47

- Model Test Data 예측 결과
  - Public Score (23.03.04~23.03.17) : **824.01**
  - Private Score (23.03.04~23.03.31) : **947.50**

```
RMSE =  [1535.42291335  179.99116893  306.46693093  200.87312507  602.70433006]
weights =  [0.25641026 0.20512821 0.17948718 0.12820513 0.23076923]
Weight Average =  650.4649127511104
```

Validation Result

2024-06-26 12:41:49	824.0092077201 947.5025980457

Dacon 제출 결과

PROJECT

6. Validation & Prediction Score 비교

- Validation K-Fold 10회 검증 / Public Score 23.03.04~23.03.17 예측 / ***최종 평가 지표 Private Score 23.03.04~23.03.31 예측**

No	Model	Validation	Public Score	***Private Score**
1	Base Line RandomForest	1197.28	1276.28	1463.11
2	*Weekday RanmdomForest	862.58	811.99	1003.35
3	*Weekday LightGBM	887.95	837.55	973.45
4	Weekday & *TG-TG ETC RandomForest	684.28	810.81	1010.90
5	Weekday & *TG-TG ETC *LightGBM	704.17	822.59	**955.58**
6	Weekday & *Item별 RandomForest	648.80	817.75	979.84
7	Weekday & Item별 *LightGBM	650.47	824.01	**947.50**
8	Weekday & *is TG / is F RandomForest	1244.65	1183.91	1397.87

PROJECT

7. 결론

1) Basline(기준지표) 학습모델 RandomForest에서 토/일요일 여부를 독립변수로 추가하여 학습한 결과 검증(+334.7) 및 예측 단계(+464.29, +459.76) 에서 점수가 대폭 상승하였음

2) 위 1)모델에서 추가적으로 TG와 TG아닌 상품들을 분리하여 학습/검증/예측한 경우
   - 검증/Public 예측 : RandomForest 우세
   - 최종평가지표 Private 예측 : LightGBM 우세

3) 아이템별로 분리하여 학습/검증/예측한 경우
   - 검증/Public 예측 : RandomForest 우세
   - 최종평가지표 Private 예측 : LightGBM 우세

4) LightGBM 학습모델 최종평가지표 Private Score기준
   - TG와 TG아닌 상품들을 분리한 경우 : 955.58
   - 아이템별로 분리한 경우 : 947.50
   - 아이템별로 분리한 경우가 +8.08 의 차이로 우세

(결론) 토/일요일 반영 및 아이템별로 분리하여 LightGBM 모델로 학습/검증/예측한 경우, 최종평가지표 Private Score 기준으로 성능이 가장 우수함(947.50)

Young Hoon Kim

SNS
https://github.com/realhoon

Phone
+82-10-9004-6355

Email
genholy77@gmail.com