Predict Future Sales

BOAZ mini project

권강미 김나현 손형락 유승희 조수연

목차

첫째,

1. 주제

Part 1,

주제

Kaggle Predict Future Sales.

상품 판매량 예측

둘째,

2. EDA 및 전처리

File Description

주어진 파일에 대한 정보.

001 >> sales_train.csv

2013년 1월~2015년 10월까지의 훈련용 일간 데이터

- Data Field

Date

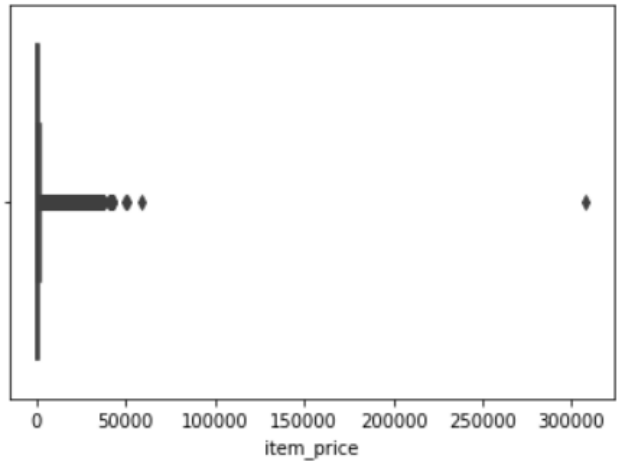
Date_block_num : 날짜 중 Month를 숫자로 나타냄. (ex. January 2013 = 0, February 2013 =1,..., October 2015 = 33)

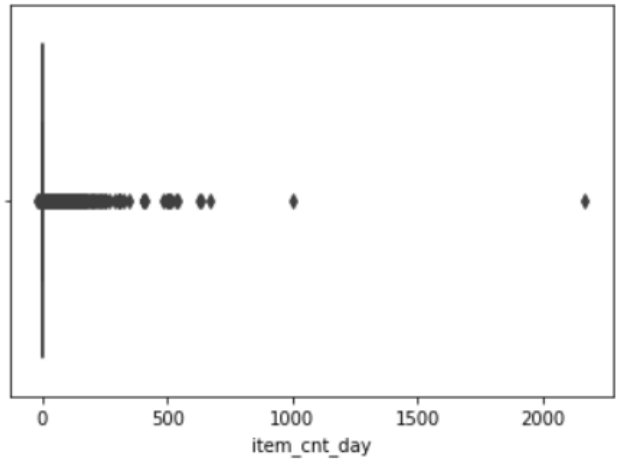
Shop_id

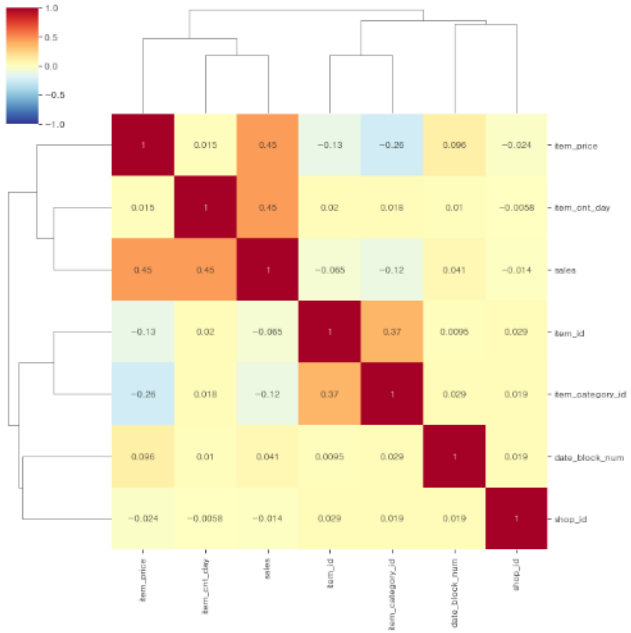
Item_id

Item_price

Item_cnt_day : item이 팔린 개수







Part 1,

File Description

주어진 파일에 대한 정보.

002 >> test.csv

2015년 11월의 판매량을 예측하는 데에 사용될 데이터

- Data Field

Id: shop과 item 튜플을 나타내는 id.

Shop_id

Item_id

003 >> items.csv

상품에 대한 추가 정보

- Data Field

Item_name

Item_id

Item_category_id

Part 1,

File Description

주어진 파일에 대한 정보.

004 >> item_categories.csv

상품 카테고리에 대한 추가 정보

- Data Field

Item_category_name

Item_category_id

005 >> shops.csv

가게에 대한 추가 정보

- Data Field

shop_name

shop_id

Part 2,

전처리

기존 train.csv 전처리.

중복 데이터

중복된 데이터 총 6개 존재

→ 제거해줌.

이상치

Item_price와 item_cnt_day에서 이상치가 존재했지만 따로 제거해주지 않았다.

▷

Kaggle의 평가기준에서 최종 결과값의 범위에 제한을 두어 변환하여 제출하라고 함.

∴ 훈련단계에서부터 target값을 제한.

Test, train 데이터셋 비교

Test셋에만 존재하거나 train셋에만 존재하는 shop_id와 item_id가 있는지 확인.

▷

Test셋에만 존재하는 item_id가 363개 존재.

∴ train셋에 추가.

Part 2,

전처리

기존 train.csv 전처리.

Item_price
Item_cnt_day

상품 가격과 팔린 상품의 개수가 음수인 경우 존재.

Item_price : 단 한 개 존재. → 제거

Item_cnt_day: 총 7356개 존재.

해당 상품이 반품/환불된 경우일수도 있다고 생각해 제거하지 않음.

그룹화

Date_block_num, shop_id, item_id를 기준으로 월별 판매량을 구하고 그룹화.

∵ 결국 예측해야 하는 것이 월별 판매량

Part 2,

전처리

주어진 다른 데이터들을 이용해 새로운 데이터 필드 추가.

Item_category_id

Train셋에 존재하지 않았던 상품에 대한 카테고리 아이디를 추가.

러시아어로 되어 있던 데이터를 selenium을 이용해 영어로 번역.

Shop_city

가게에 대한 추가 정보를 이용해 가게의 위치 정보(도시)를 추가.

	shop_name	shop_id
0	! Yakutsk Ordzhonikidze, 56 fran	0
1	! Yakutsk shopping center "Central" Fran	1
2	Adygea shopping center "Mega"	2
3	Balashiha TRK "October-Kinomir"	3
4	Volzhsky shopping center "Volga Mall"	4

가게 이름의 첫번째 단어가 도시인 것을 이용.

Sub_category

상품 카테고리에 대한 추가 정보를 이용해 상품에 대한 정보를 추가.

	item_category_name	item_category_id
0	PC - Headset / Headphones	0
1	Accessories - PS2.	1
2	Accessories - PS3.	2
3	Accessories - PS4.	3
4	Accessories - PSP.	4

앞 단어와 뒤 단어를 분할, 새로운 컬럼 생성.

Part 2,

전처리

주어진 다른 데이터들을 이용해 새로운 데이터 필드 추가.

Item_cnt_lag 1,2	지난 달의 상품 판매 수량 추가.	Item_cnt_lag1: 1달 전 판매 수량 Item_cnt_lag2: 2달 전 판매 수량
Item_price_lag 1,2	지난 달의 상품의 평균 가격 추가.	Item_price_lag1: 1달 전 상품의 평균 가격 Item_price_lag2: 2달 전 상품의 평균 가격
month, year	Date_block_num 기준으로 그룹화하여 사라진 month, year 값 추가.	

Part 2,

전처리

주어진 다른 데이터들을 이용해 새로운 데이터 필드 추가.

최종 형태

date_block_num	shop_id	item_id	item_cnt_day	item_category_id	shop_city	sub_category_1	sub_category_2	month	year	item_cnt_lag1	item_price_lag1	item_cnt_lag2	item_price_lag2
2	2	32	0	40	1	3	18	3	2013	0	0	0	0
3	2	32	0	40	1	3	18	4	2013	0	0	0	0
4	2	32	0	40	1	3	18	5	2013	0	0	0	0
5	2	32	0	40	1	3	18	6	2013	0	0	0	0
6	2	32	0	40	1	3	18	7	2013	0	0	0	0
7	2	32	0	40	1	3	18	8	2013	0	0	0	0
8	2	32	0	40	1	3	18	9	2013	0	0	0	0
9	2	32	0	40	1	3	18	10	2013	0	0	0	0
10	2	32	0	40	1	3	18	11	2013	0	0	0	0
11	2	32	0	40	1	3	18	12	2013	0	0	0	0
12	2	32	1	40	1	3	18	1	2014	0	0	0	0
13	2	32	0	40	1	3	18	2	2014	1	119	0	0
14	2	32	1	40	1	3	18	3	2014	0	0	1	119
15	2	32	0	40	1	3	18	4	2014	1	149	0	0
16	2	32	0	40	1	3	18	5	2014	0	0	1	149
17	2	32	0	40	1	3	18	6	2014	0	0	0	0
18	2	32	1	40	1	3	18	7	2014	0	0	0	0
19	2	32	0	40	1	3	18	8	2014	1	149	0	0
20	2	32	2	40	1	3	18	9	2014	0	0	1	149

셋째,

3. 모델링

Part 2,

모델링 개요

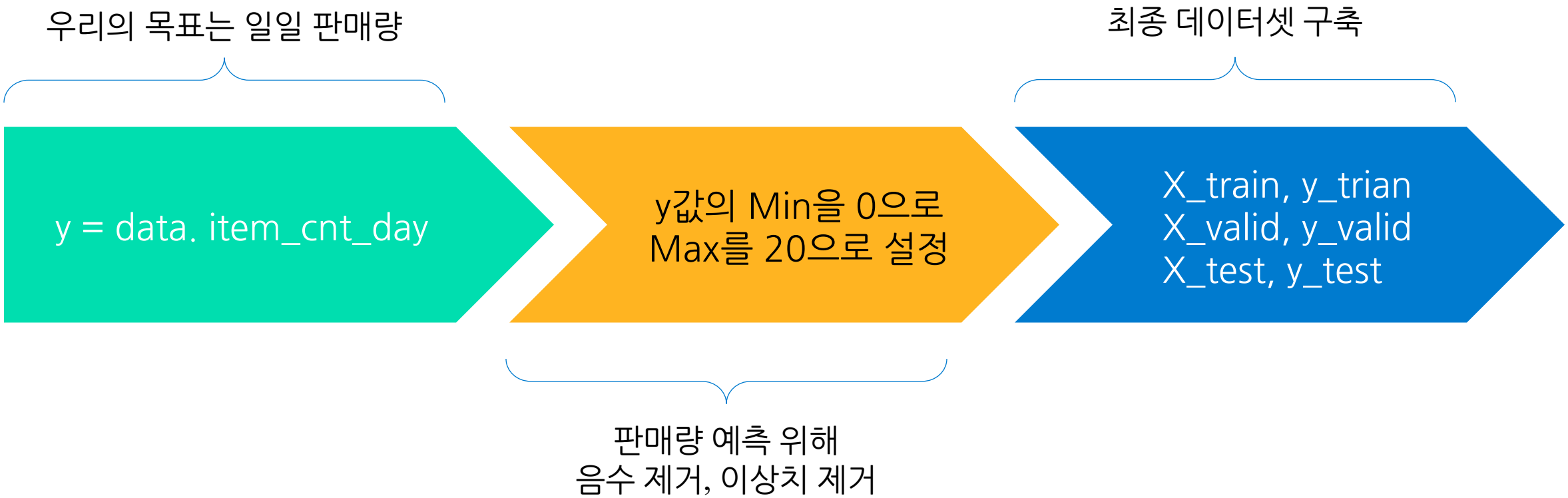
2013-01~2015-09 로 모델링하여 2015-11 대상으로 예측 후 평가하는 방향

Step 1	Step 2	Step 3
2013-01 ~ 2015-09 (date_block_num[0~32]) 모델링	2015-11 (date_block_num[34]) 대상으로 예측	평가

Part 2,

Train, Valid, Test 세트로 분리

데이터 세트 구축



Part 2,

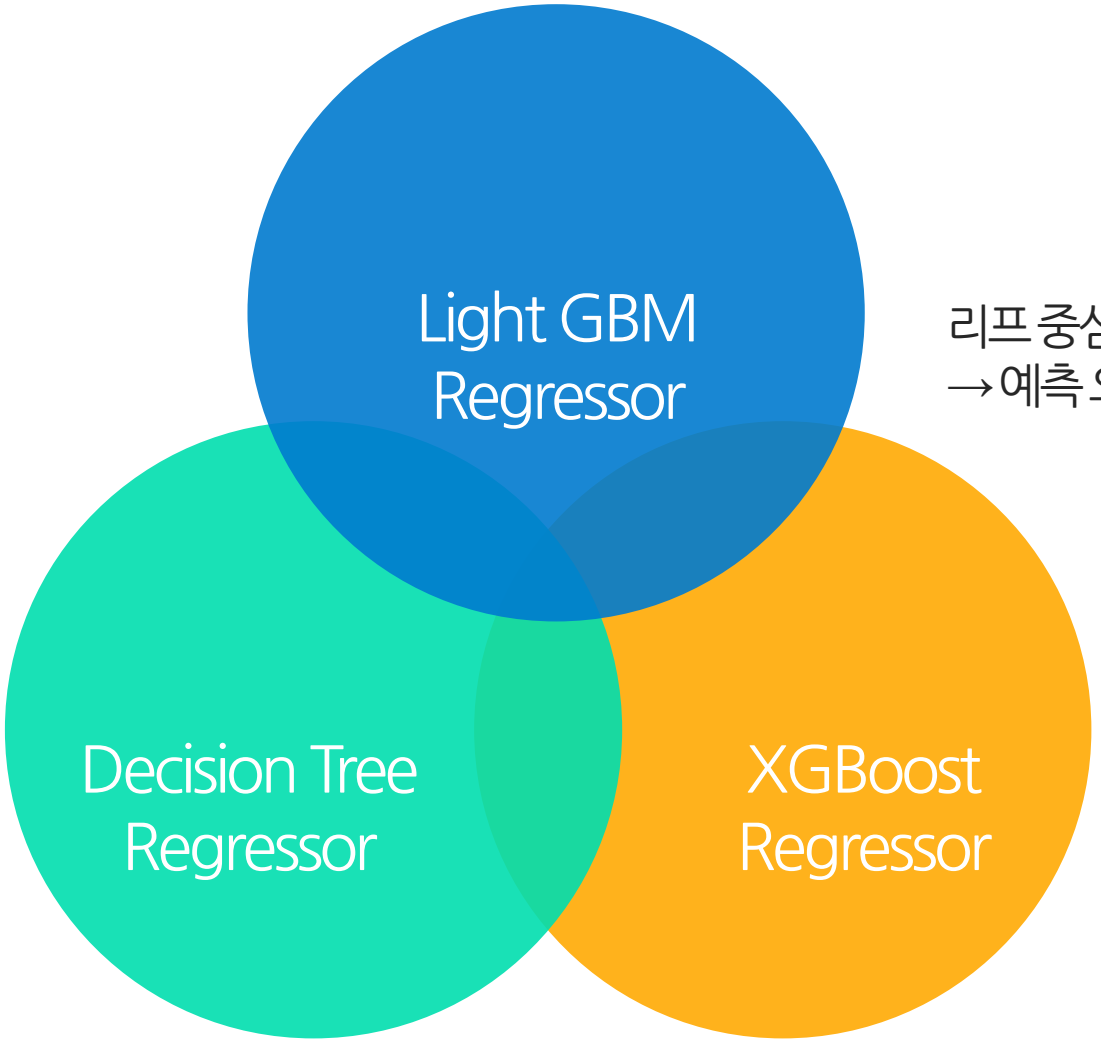
Train, Valid, Test 세트로 분리

X_test 세트

	date_block_num	shop_id	item_id	item_category_id	shop_city	sub_category_1	sub_category_2	month	year	item_cnt_lag1	item_price_lag1	item_cnt_lag2	item_price_lag2
0	2	2	32	40	1	3	18	3	2013	0	0	0	0
1	3	2	32	40	1	3	18	4	2013	0	0	0	0
2	4	2	32	40	1	3	18	5	2013	0	0	0	0
3	5	2	32	40	1	3	18	6	2013	0	0	0	0
4	6	2	32	40	1	3	18	7	2013	0	0	0	0
5	7	2	32	40	1	3	18	8	2013	0	0	0	0
6	8	2	32	40	1	3	18	9	2013	0	0	0	0
7	9	2	32	40	1	3	18	10	2013	0	0	0	0
8	10	2	32	40	1	3	18	11	2013	0	0	0	0
9	11	2	32	40	1	3	18	12	2013	0	0	0	0
10	12	2	32	40	1	3	18	1	2014	0	0	0	0
11	13	2	32	40	1	3	18	2	2014	1	119	0	0
12	14	2	32	40	1	3	18	3	2014	0	0	1	119
13	15	2	32	40	1	3	18	4	2014	1	149	0	0
14	16	2	32	40	1	3	18	5	2014	0	0	1	149
15	17	2	32	40	1	3	18	6	2014	0	0	0	0
16	18	2	32	40	1	3	18	7	2014	0	0	0	0
17	19	2	32	40	1	3	18	8	2014	1	149	0	0
18	20	2	32	40	1	3	18	9	2014	0	0	1	149
19	21	2	32	40	1	3	18	10	2014	2	149	0	0
20	22	2	32	40	1	3	18	11	2014	2	149	2	149

회귀 분석 시도

다양한 기법 적용



리프 중심 트리분할
→ 예측 오류 손실 최소화

Part 2,

LightGBM 파라미터 최적화

파라미터	값	설명
n_estimators	300	Weak learern의 개수
learning_rate	0.08	학습률.,Weak learner가 순차적으로 오류 값을 보정해 나가는 데 적용하는 계수
max_depth	8	트리의 깊이
early_stopping_round	30	조기 중단할 수 있는 최소 반복 횟수
colsample_bytree	0.7	트리 생성에 필요한 피처를 임의로 샘플링하는 비율
subsample	0.7	과적합 방지를 위해 데이터를 샘플링하는 비율
objective	rmse	손실함수

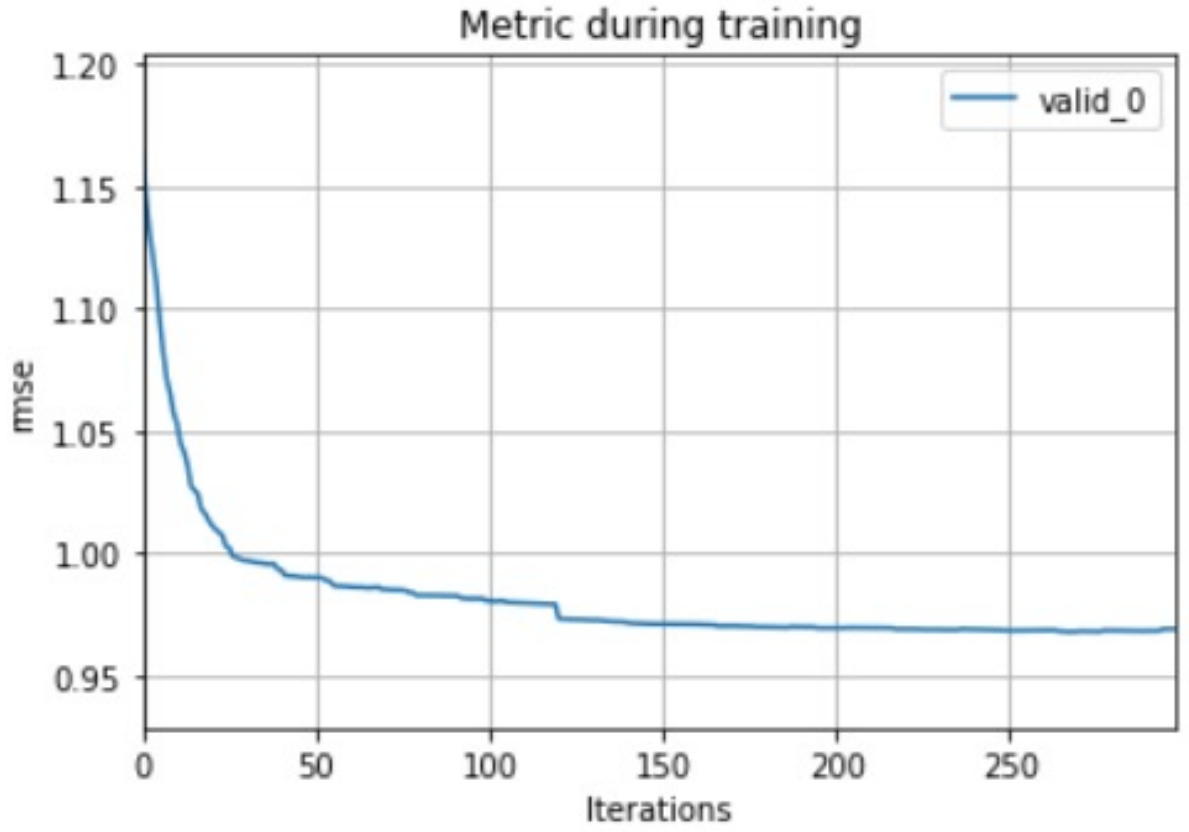
Part 4,

성능(RMSE) 변동 과정

001 >> 수행 시 성능 평가 지표 (metric)

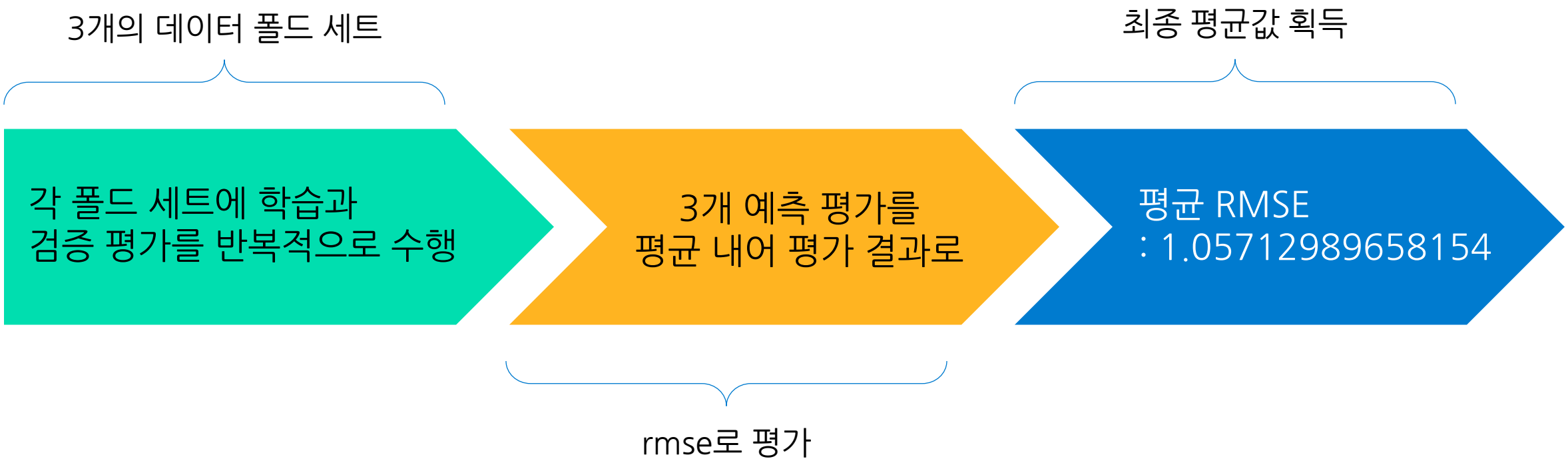
rmse 적용

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$



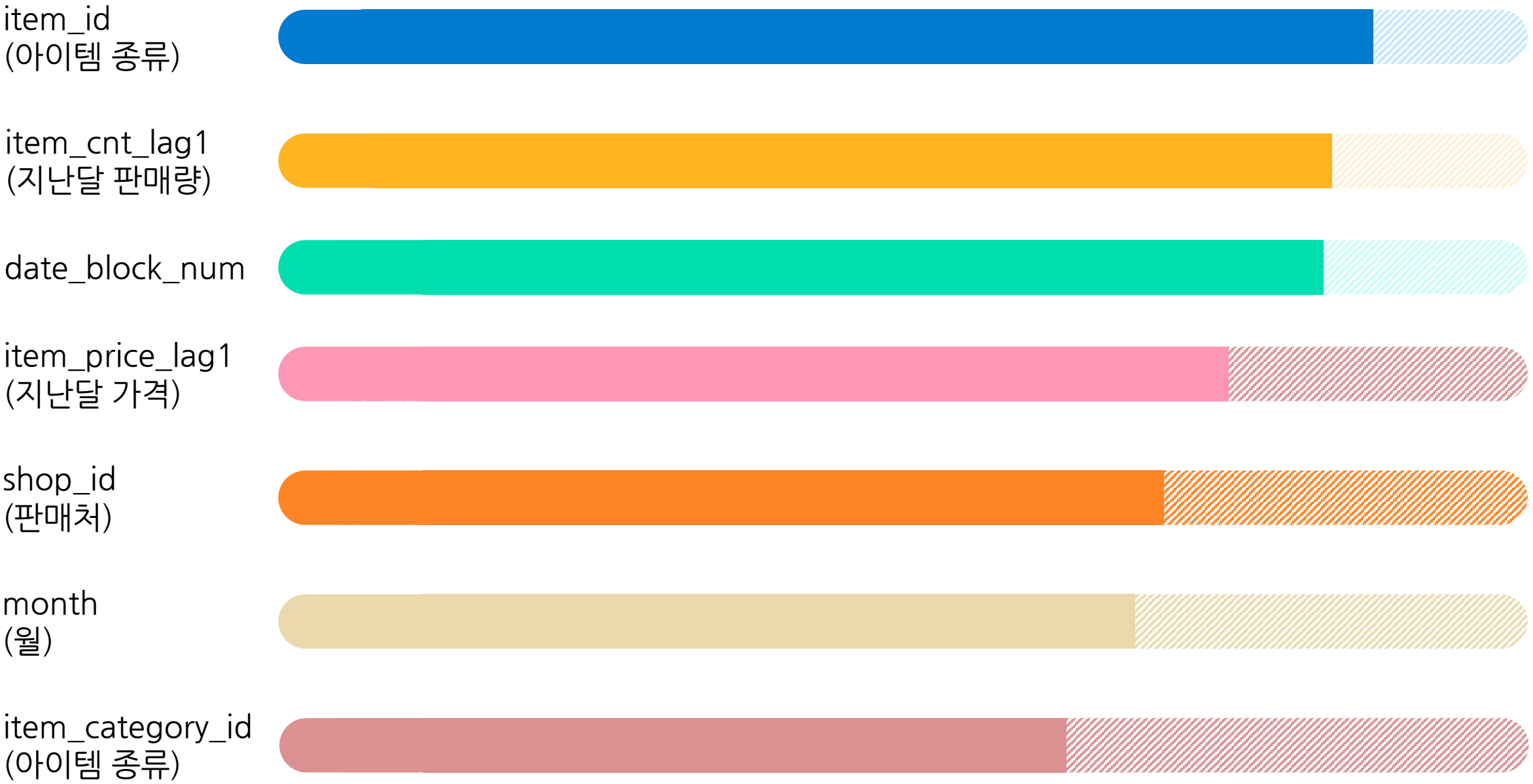
Part 2,

K 폴드 교차검증



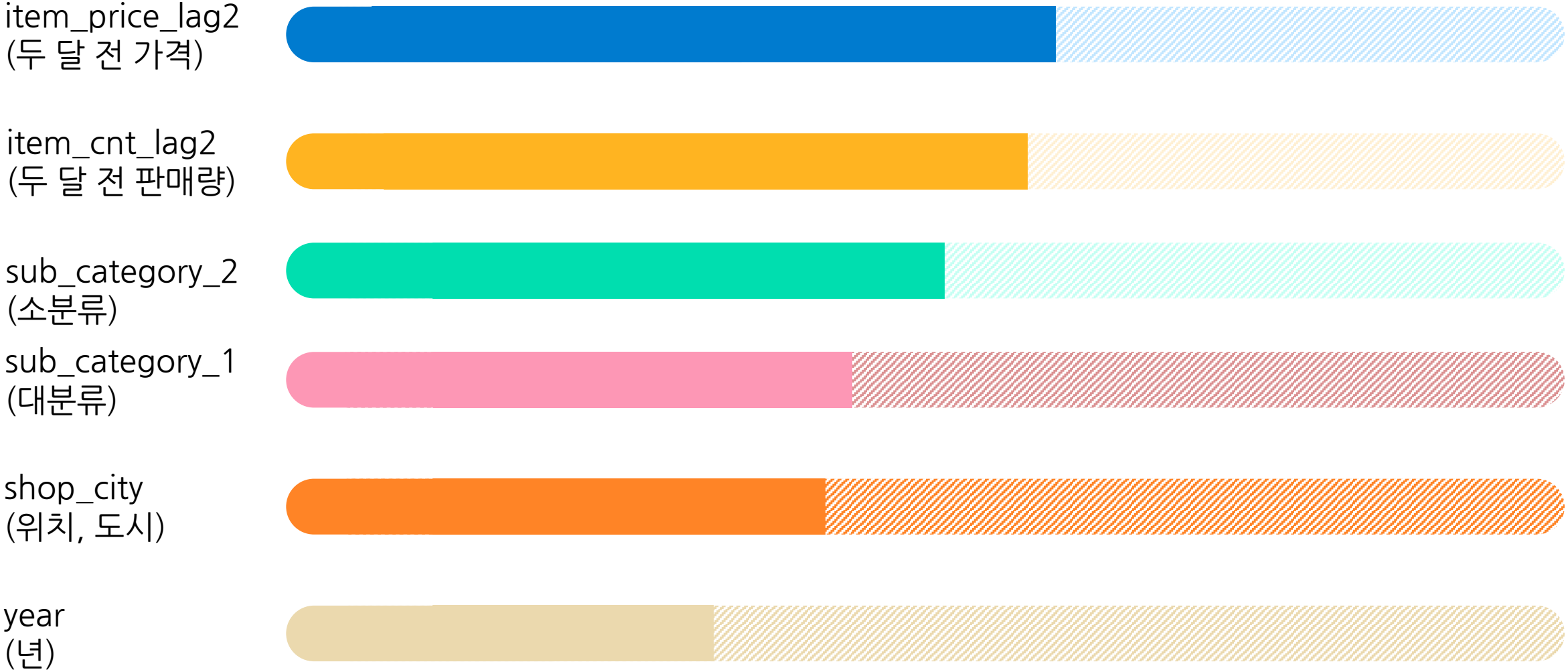
Part 1,

Feature importance



Part 1,

Feature importance



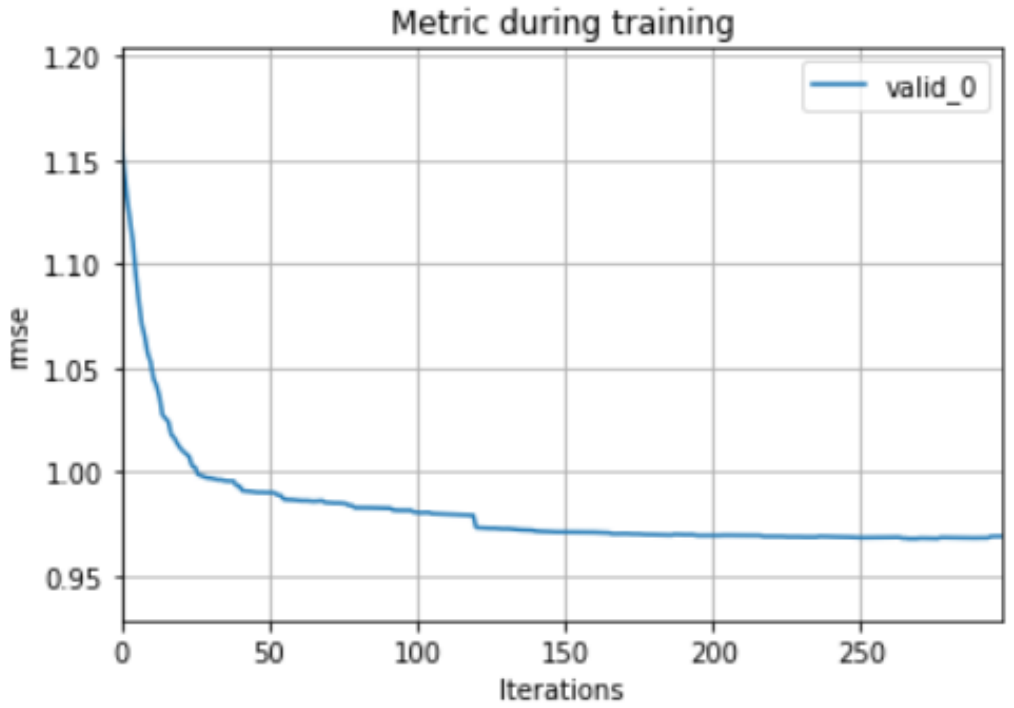
셋째,

결과 분석

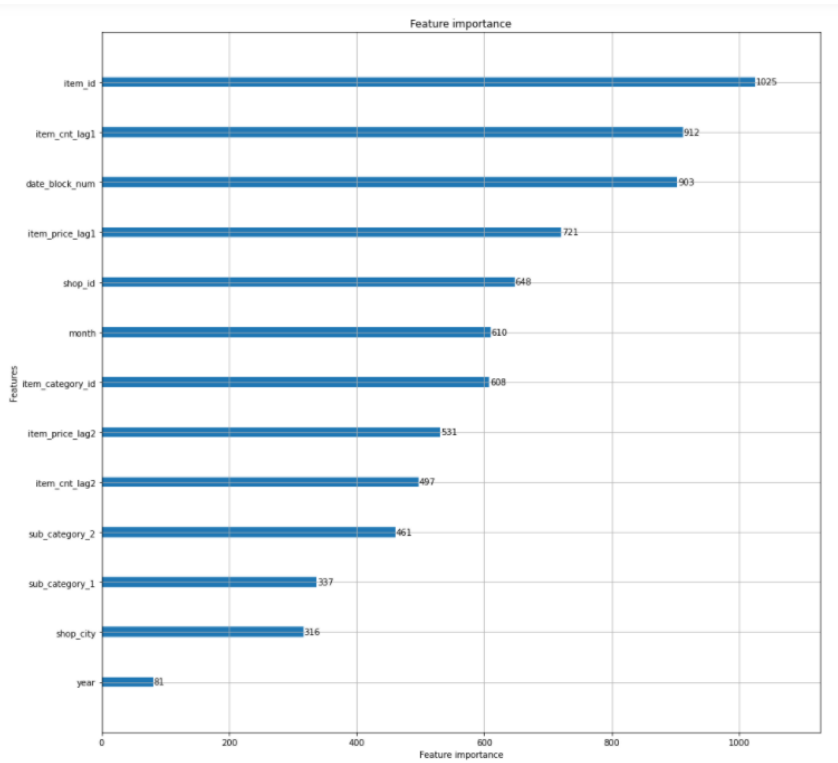
Part 3,

LightGBM Model

RMSE 변동과정



Feature Importance



Kaggle 제출

Kaggle Leader Board Score

1 Test 셋 불러오기

	ID	shop_id	item_id
0	0	5	5037
1	1	5	5320
2	2	5	5233
3	3	5	5232
4	4	5	5268
...	...	...	...
214195	214195	45	18454
214196	214196	45	16188
214197	214197	45	15757
214198	214198	45	19648
214199	214199	45	969

214200 rows × 3 columns

2 X_train 셋

	date_block_num	shop_id	item_id	item_category_id	shop_city	sub_category_1	sub_category_2	month	year	item_cnt_lag1	item_price_lag1
0	2	2	32	40	1	3	18	3	2013	0.0	0.0
1	3	2	32	40	1	3	18	4	2013	0.0	0.0
2	4	2	32	40	1	3	18	5	2013	0.0	0.0
3	5	2	32	40	1	3	18	6	2013	0.0	0.0
4	6	2	32	40	1	3	18	7	2013	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...
7970463	7	30	6723	18	11	7	33	8	2013	0.0	0.0
7970464	6	31	3761	18	11	7	33	7	2013	0.0	0.0
7970465	27	35	6662	18	13	7	33	4	2015	0.0	0.0
7970466	3	26	6669	10	11	8	32	4	2013	0.0	0.0
7970467	18	38	17703	51	15	2	14	7	2014	0.0	0.0

7539217 rows × 13 columns

Kaggle 제출

Kaggle Leader Board Score

3 새로운 데이터프레임 생성

1) Y_pred 내의 element 값에 대해서 0값 보다 작은 값들을 0으로 바꿔주고
20보다 큰 값들을 20으로 바꿔주는 함수

```
y_pred = lgbm_r.predict(X_test)
y_pred = np.clip(y_pred,0,20)
```

2) 데이터프레임 결합 : 열방향(좌우)으로 데이터프

```
new = pd.concat([X_test.reset_index(),
                 pd.DataFrame(y_pred, columns =["item_cnt_month"])],axis = 1)
```

3) 데이터프레임 결합 : new 데이터를 ID 기준으로 합치기

```
submission = tt.merge(new, on = ["shop_id", "item_id"])[["ID","item_cnt_month"]]
```

Score : 1.00558

4 Submission 저장

	ID	item_cnt_month
0	0	0.495192
1	1	0.131651
2	2	1.092706
3	3	0.302912
4	4	0.407871
...	...	...
214195	214195	0.406972
214196	214196	0.080488
214197	214197	0.096448
214198	214198	0.134142
214199	214199	0.134474

214200 rows × 2 columns

“

감사합니다