데이터 분류분석
- 스토어의 일일 판매량 예측 -

임웅빈

1. Introduction

소비자 물가는 원유가격의 영향을 받는다. 일일 원유가격 지수의 변화가 소비자의 구매에 영향을 미칠것이다. 이 분석은 마켓 상품 판매 데이터, 일일 유가 지수 데이터, 공휴일 데이터를 결합하여, 일일 유가를 입력하면 일일 판매량의 예측을 목표로 한다.

2. 데이터 수집 및 전처리

Store Sales data		WTI Oil Index data		Store Holiday data	
date	날짜 및 시간 정보	date	날짜 및 시간 정보	date	날짜 및 시간 정보
store_nbr	체인점 지점 번호	dcoilwtico	원유가격	type	공휴일 종류
family	품목명				
sales	스토어 당일 판매액				
onpromotion	프로모션 여부				

Table1. Store Sales data　　Table2. WTI Oil Index data　　Table3. Store Holiday data

3개의 데이터 세트중 공통 feature인 date 열을 기준으로 결합하였고, 원유가격, 공휴일 종류, 날짜를 features로 스토어 당일 판매액을 target으로 분류 분석을 진행하였다.

	oil_price	type	sales	year	month	day
0	93.14	5	496092	2013	1	2
1	92.97	5	361461	2013	1	3
2	93.12	5	354460	2013	1	4
3	93.20	5	336123	2013	1	7
4	93.21	5	318348	2013	1	8
...	...	...	...	...	...	...
1173	49.59	5	734140	2017	8	9
1174	48.54	3	651387	2017	8	10
1175	48.81	4	826374	2017	8	11
1176	47.59	5	760922	2017	8	14
1177	47.57	3	762662	2017	8	15

1178 rows × 6 columns

범주형 데이터로 전환

	oil_price	type	year	month	day	target
0	93.14	5	2013	1	2	2
1	92.97	5	2013	1	3	1
2	93.12	5	2013	1	4	1
3	93.20	5	2013	1	7	1
4	93.21	5	2013	1	8	1
...	...	...	...	...	...	...
1173	49.59	5	2017	8	9	4
1174	48.54	3	2017	8	10	3
1175	48.81	4	2017	8	11	4
1176	47.59	5	2017	8	14	4
1177	47.57	3	2017	8	15	4

1178 rows × 6 columns

Target Description		
범주	판매량 기준	분포 갯수
1	400k ↓	326
2	400k~600k	321
3	600k~700k	289
4	700k ↑	242

Fig1. Converting to Linear Categorical

일일 판매액이 선형 데이터이기 때문에 임의의 기준을 주어 범주형 데이터로 전환하였다.

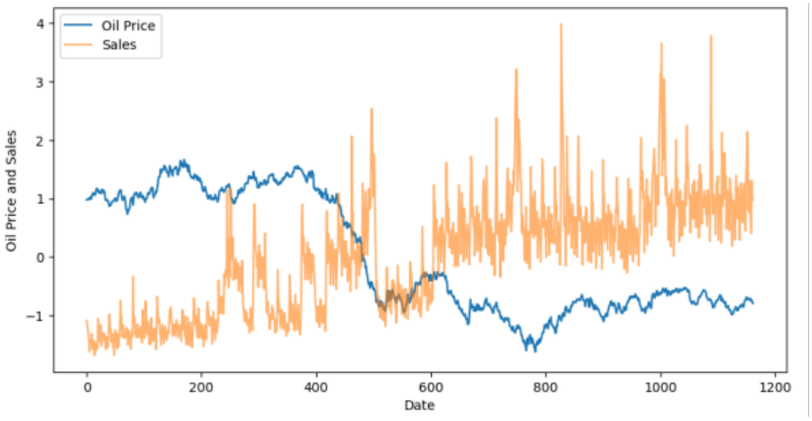

Fig2. 일자별 유가지수와 일일 판매량 Plot 차트

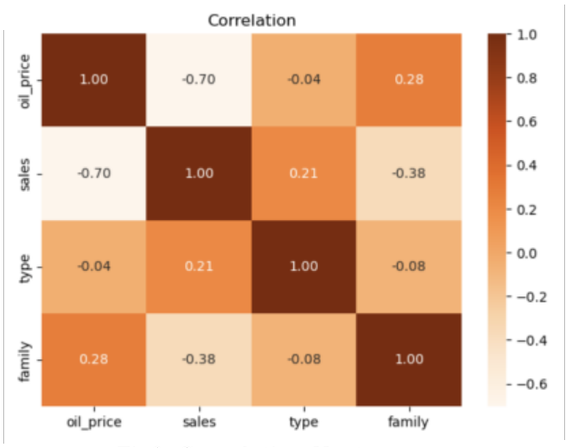

Fig3. Correlation Heatmap

Fig2의 plot 차트에서 Oil Price와 Sales 데이터가 반비례 관계임 확인하였고,
Fig3의 히트맵에서 sales데이터가 oil_price데이터에 가장 크게 영향을 받고 있음을 확인하였다.

3. 평가

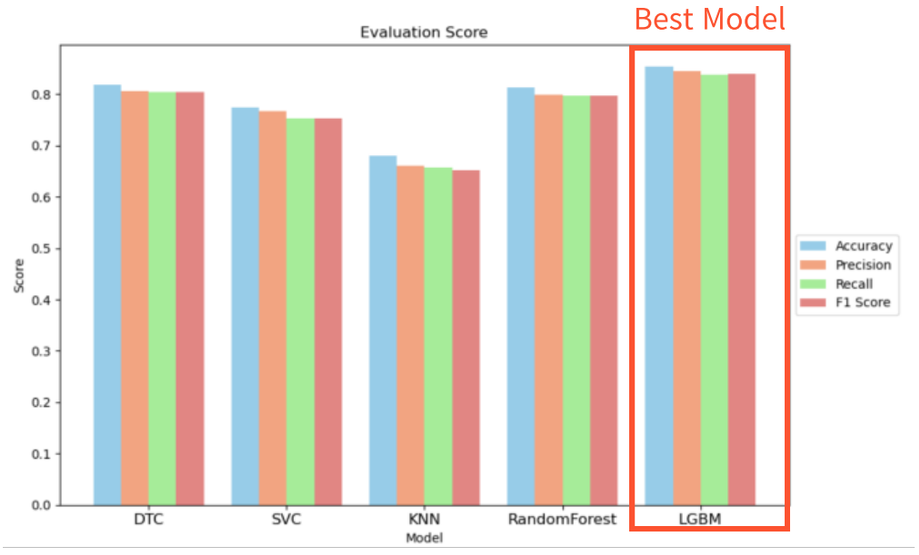

Fig4. Model별 성능 점수

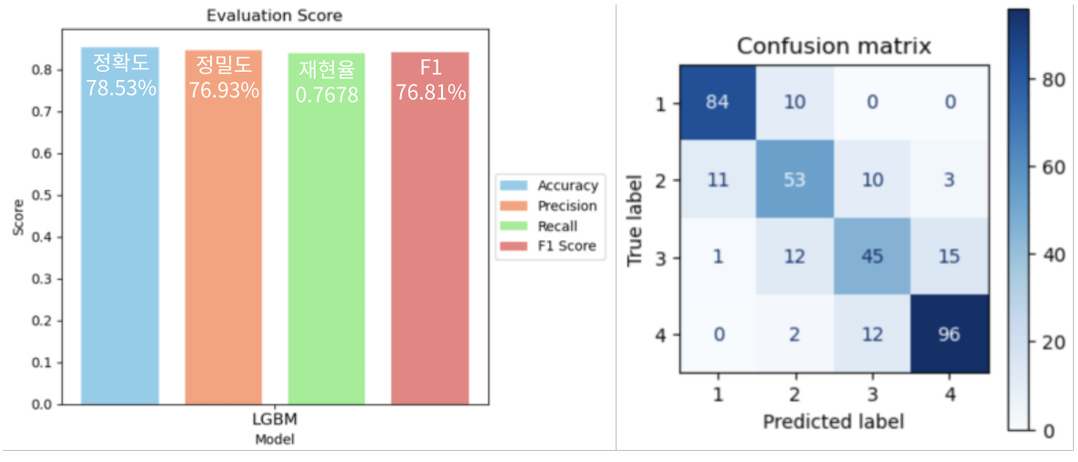

Fig5. LGBM 모델의 성능

모델별로 주요 지표 4가지를 출력하여 성능을 확인하였다.
LGBM 성능이 가장 좋고, KNeighborsClassifier 성능은 상대적으로 매우 좋지 않다.

4. 고찰

이번 모델의 목표는 당일 유가를 입력하여 일일 판매량을 예측하는것이 목표였다.
- 수집된 데이터세트에 가장 높은 성능결과를 도출한 모델은 LGBM모델이었다.
- target의 데이터가 선형적이기 때문에 LGBM모델이 비교적 높은 성능을 나타냈다.
- target을 분류해야 했기 때문에 선형적인 일일 판매량의 데이터의 범주를 설정해주고, 이산적으로 변환 해주어야 하는 번거로움이 있었다.
- target의 분류 범위 설정에 대한 기준은 주관적이기 때문에 범용성이 떨어진다.
- 시계열 데이터를 범주형데이터로 변환해서 학습하다 보니 모델 성능이 떨어졌다고 판단된다.
- 해당 데이터 세트는 분류보다 회귀 모델에 적합한 것을 알 수 있었다.