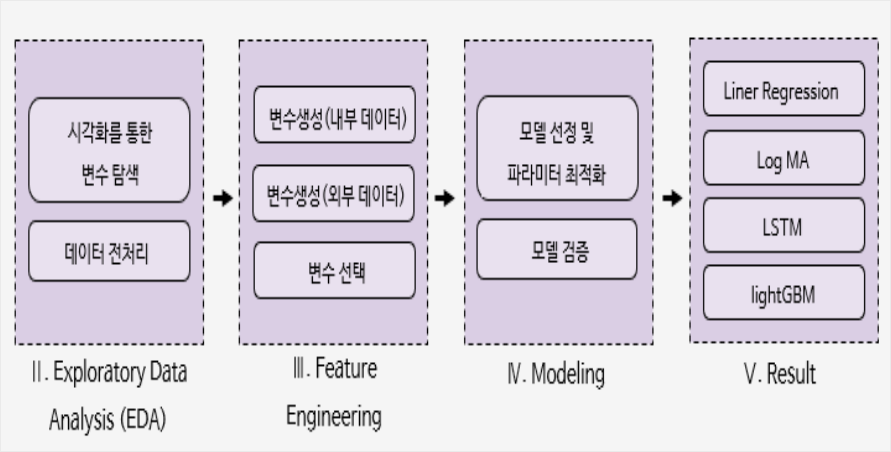
1. 김현우 (19940729)
2. choco_9966@naver.com
3. 김현우/박주연/ 이주영

전통적인 수요 예측 모델과 최신 모델 사이의 성능비교

Favorita Grocery Sales Data를 통해

유통업체의 경우 정확한 수요예측은 과거부터 중요한 이슈 중 하나이다. 과거에는 linear regression, logMA 모델이 많이 쓰였지만 최근에는 LSTM, LightGBM등의 모델을 사용하는 추세이다. 이에 대해 전통적인 수요예측 모델과 최신 모델 간의 성능에 비교 분석을 진행하였다.

다음은 분석의 진행 방향이다.



정교한 데이터 전처리 과정을 거친 데이터를 기반으로 모델의 성과를 분석하였을 때, 다음의 표와 같은 결과를 얻을 수 있었다.

Model 분류	모델	결과	최초 등장 연도	순위
전통적 기법	Linear regression	0.591	1903년도	4
	logMA	0.575	1938년도	3
현대적 기법	LSTM	0.527	1997년도	2
	LightGBM	0.523	2016년도	1

이를 통하여 예측의 정확도 측면에서는 LightGBM, LSTM, logMA, Linear regression 순으로 모델의 성능이 우월하다고 판단할 수 있었다.

Data Science Competition 2018

2018. 7. 29

목차

김현우
choco_9966@naver.com
박주연
이주영

I. Introduction

1. 문제 제기

최근에 편의점 폐기 음식에 대한 알바생과 점주의 상반된 입장이 사회적 이슈다. 다들 편의점 아르바이트 하는 친구들에게 폐기 음식을 얻어먹어 본 경험이 한번씩은 있지 않은가.

(편의점 아르바이트생의 폐기 음식 인증샷. 한국일보 2018.07.12)

하지만 편의점 입장에서 폐기는 나와도 문제, 없어도 문제인 손실의 주범이다. 폐기가 나오면 그만큼 손실을 본 것이고 폐기가 없다면 매출 상승의 기회를 놓친 것이기 때문이다. 점주들 사이에서 '적정 발주량은 신도 못 맞힌다'는 이야기가 돌 정도로 수요예측이 어렵다. 이처럼 유통업체에서는 과거부터 수요예측이 중요한 이슈인 동시에 해결하기 어려운 문제이다.

그렇기 때문에, 우리는 kaggle에 올라와 있는 에콰도르 대형마트 체인의 매출 데이터를 이용해 매출을 예측하고 전통적 모델과 최신 모델 간의 비교분석을 진행하기로 하였다.

비교 대상 모델은 전통적인 수요예측 기법인 1)linear regression, 2)logMA와 최신 모델인 3)LSTM, 4)LightGBM이다. 이와 동시에, 정교한 데이터 전처리가 데이터 분석에 있어서 필수적이라는 사실을 반영하여 데이터 전처리 작업도 진행하기로 하였다.

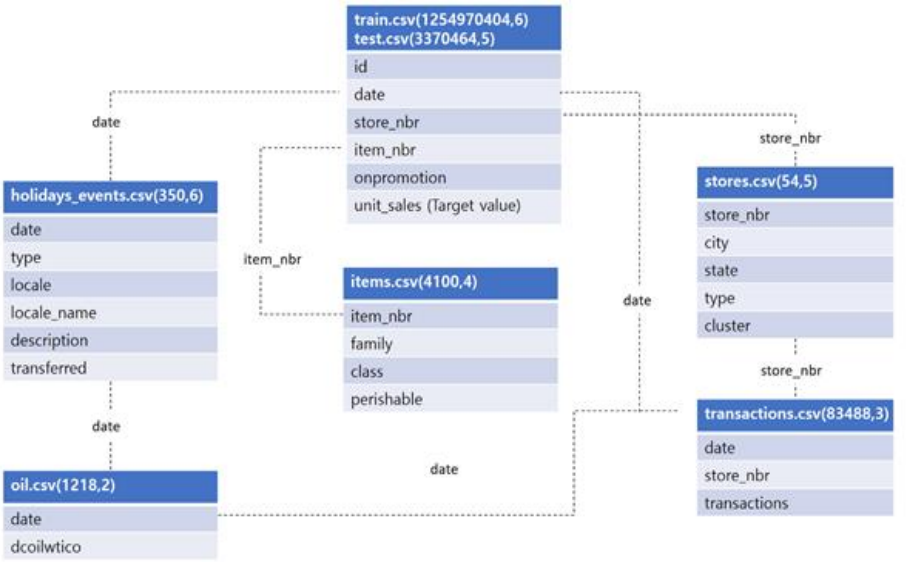
즉 우리의 목표는 정교한 데이터 전처리 과정을 거친 후, 대표적인 4가지 모델 간의 성능 비교이다. 이를 통해 미래의 수요예측 기법에 있어서의 방향성을 예측 해 볼 수 있다.

2. 데이터 소개

분석에는 kaggle에서 진행된 Favorita Grocery Sales Forecasting 대회의 데이터를 이용하였다. 식료품점의 재고관리를 위해 판매량을 예측하는데 필요한 데이터가 제공 되어있다. 기본 제공 데이터는 총 7개의 파일로 구성되어 있다.

데이터 출처 : https://www.kaggle.com/c/favorita-grocery-sales-forecasting/data

[그림 1] Data Schema



train.csv / test.csv : 해당 지점(store)의 해당 품목(item)이 얼만큼 팔렸는가(unit_sales)를 나타내는 기본 데이터.
onpromotion은 해당기간에 판촉행사의 유무에 대한 논리 변수이다.

items.csv : 4100가지의 품목의 개별적인 특징, 특히 perishable 변수는 상품이 상할 수 있는 품목인지에 대한 자료로서 특별한 처리가 요구됨.

stores.csv : 54개 매장에 대한 정보로서, 어느 지역(city), 어느 종류의 매장(type)인지에 대한 데이터

holidays_events.csv : 기간, 지역에 휴일이 있을 경우 휴일의 종류(type)와 자세한 설명을 기록한 데이터

oil.csv : WTI 평균유가를 나타낸 시계열 데이터

transactions.csv : 기간(date), 매장(store_nbr)에 따른 총 거래량(transactions)에 대한 데이터

II. Exploratory Data Analysis(EDA)

1. 시각화를 통한 변수 탐색

시각화를 통한 변수탐색을 통해 전체 데이터에 대한 이해를 높임과 동시에 아이디어를 데이터 전처리 과정에 반영하여 분석의 완성도를 높이고자 한다.

1) Overview

변수 이해의 첫 단계로 데이터의 변수 종류, 변수 타입, 기술 통계량을 살펴보았다. train 데이터는 가장 많은 정보를 가지고 있는 동시에 중요한 아이디어를 제공한다.

<표 1> train.csv 변수설명

train.csv	변수 예시
observations	125,497,040
variables	6
id	<int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 ...
date	<chr> "2013-01-01", "2013-01-01"
store_nbr	<int> 25, 25, 25, 25, 25, 25, 25, 25, 25, 25 ...
item_nbr	<int> 103665, 105574, 105575, 108079 ...
onpromotion	<lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...

자료: Kaggle competition – favorite stores

합리적으로 observation의 수를 줄일 수 있는 방법이 없을까?

<표 1>을 보면 observations은 1억2천만개에 달하는 큰 파일(4.7GB)임을 알 수 있다.

date, store_nbr, item_nbr의 데이터 타입을 수정할 필요가 있고, unit_sales가 정수형이 아닌 double인것을 봐선 소수점의 단위로 표현된 값이 존재한다. onpromotion 변수에는 결측치(NA)가 있다.

onpromotion의 결측치 분포를 살펴보자

<표 2> train.csv 변수 기술 통계량

train.csv	Min.	1st Quarter	Median	Mean	3rd Quarter	Max.
id	0	31374260	62748520	62748520	94122779	125497039
store_nbr	1	12	28	27.46	43	54. 00
item_nbr	96995	522383	959500	972769	1354380	2127114
unit_sales	-15372	2	4	8.55	9	89440

자료: Kaggle competition – favorite stores

<표 3> train.csv 변수 기술 통계량

	Mode	TRUE	FALSE	NA's
date	Character			
onpromotion	Logical	7810622	96028767	21657651

자료: Kaggle competition – favorite stores

unit_sales는 부호에 따라 별도의 처리가 필요하다.

unit_sales의 비정상적인 극단치의 이유를 분석해보자.

onpromotion 결측치 처리가 필요하다.

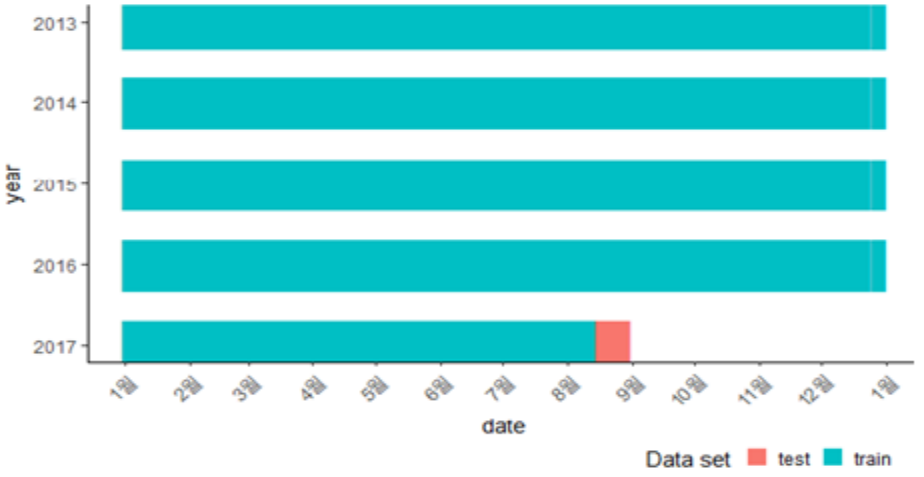
<표 2>와 <표 3>을 보면 unit_sales에 음수가 존재하는 것을 확인할 수 있다. 이것은 반품된 수량을 의미한다. 즉 부호에 따라 다른 의미를 가지고 있으므로 다른 처리가 필요하다. 또한 3분위수와 Max값의 차이가 약 10000배인 것으로 보아 이상치가 의심스럽다. onpromotion의 결측치(NA)값이 21,657,651개이고 이는 전체 데이터의 관측개수의 16%에 해당한다.

2) 개별 특징 시각화

변수 이해의 두번째 단계로 개별 데이터 시각화를 진행한다. 예측 값인 unit_sales에 영향을 미치는 주요 변수 및 데이터를 기반으로 분석하였다.

(1) 예측해야 하는 Test기간 탐색

[그림 2] 학습기간과 예측기간



자료: Kaggle competition – favorite stores

예측 기간이 주어진 기간에 비하여 짧다.

과거의 정보가 예측에 큰 도움이 안될 가능성이 있다.

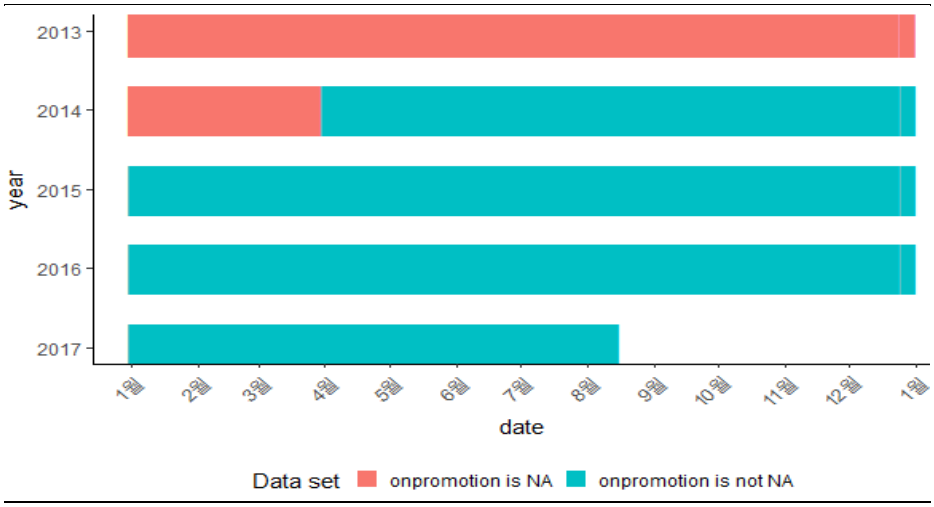
[그림 2]에서 train 데이터는 2013년 1월 1일부터 2017년 8월 14일까지의 store별, item별 unit_sales의 정보를 담고 있다. 반면에 test 데이터는 2017년 8월 15일부터 2017년 8월 30일까지로 train 데이터에 비해 상당히 짧은 기간의 정보를 담고 있다.

분석 대상인 <Favorita Grocery>의 판매량은 대형마트 특성상 고객들의 소비 패턴에 많은 영향을 받는다. 이를 고려하면 먼 과거의 데이터가 최근의 데이터에 비해 설명력이 떨어질 것으로 예측할 수 있다.

(2) onpromotion 변수 탐색

train 데이터의 onpromotion변수에는 train 데이터의 전체 관측치의 16%에 해당하는 21,657,651개의 결측치가 존재하는 것을 <표 1>에서 파악하였다.

[그림 3] onpromotion변수 결측치 분포



자료: Kaggle competition – favorite stores

onpromotion 결측치를 다룰 때 결측치가 특정구간에 몰려 있는 것을 고려하자.

[그림 3]을 보면 결측치가 2013년부터 2014년 3월까지 분포하는 것을 파악할 수 있다.

test 데이터는 train 데이터의 마지막 날로부터 15일(전체 train 데이터의 1.7%)에 해당하는 것을 고려하였을 때 결측치가 생긴 2013년 ~ 2014년 3월의 데이터를 제거하는 것이 모델의 예측력을 높여 줄 것이다.

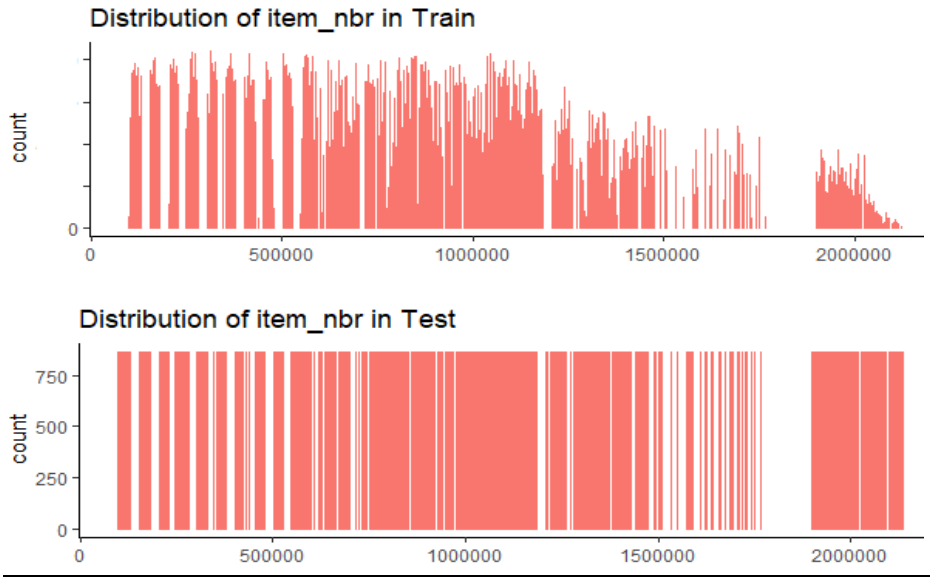
(3) store_nbr, item_nbr 변수 탐색

[그림 4] 지점 개수 분포 히스토그램

자료: Kaggle competition – favorite stores

시사점 및 아이디어

[그림 5] 취급 품목 개수 분포 히스토그램



자료: Kaggle competition – favorite stores

train 데이터의 비대칭으로 인하여 test 데이터 예측에 부정적 영향을 미칠 가능성이 있다.

[그림 4]와 [그림 5]을 보면 test 데이터의 경우 store_nbr, item_nbr 모두 고르게 분포한 반면 train data는 일정하지 않은 분포를 보인다.

train 데이터가 고르게 분포하지 않은 이유는 관측 기간내에 추가된 새롭게 생긴 store와 item이 존재하기 때문이라고 유추할 수 있다.

2015년 이후로 store와 item의 분포가 다르다.

실제로 2015년을 기준으로 store_nbr, item_nbr의 관측 수 분포를 비교해보면 아래의 <표 4>와 같다.

<표 4> 2015년 전후 store/item 변수 변화

	2015년 이전	2015년 이후
store의 수	48	54
item의 수	2928	4067

자료: Kaggle competition – favorite stores

이를 test기간으로 확장시키면 store수는 그대로이지만 195개의 새로운 item이 나타난다.

2015년도를 기준으로 onpromotion, store, item의 관측 수가 많이 달라지는 것을 확인 할 수 있다. 정확한 이유는 확인이 어렵지만 앞서 train 데이터의 onpromotion의 변수의 결측치도 2013년과 2014년에 집중적으로 존재하는 것을 보았을 때 2013년과 2014년 데이터에 문제가 있음을 의심해 볼 수 있다.

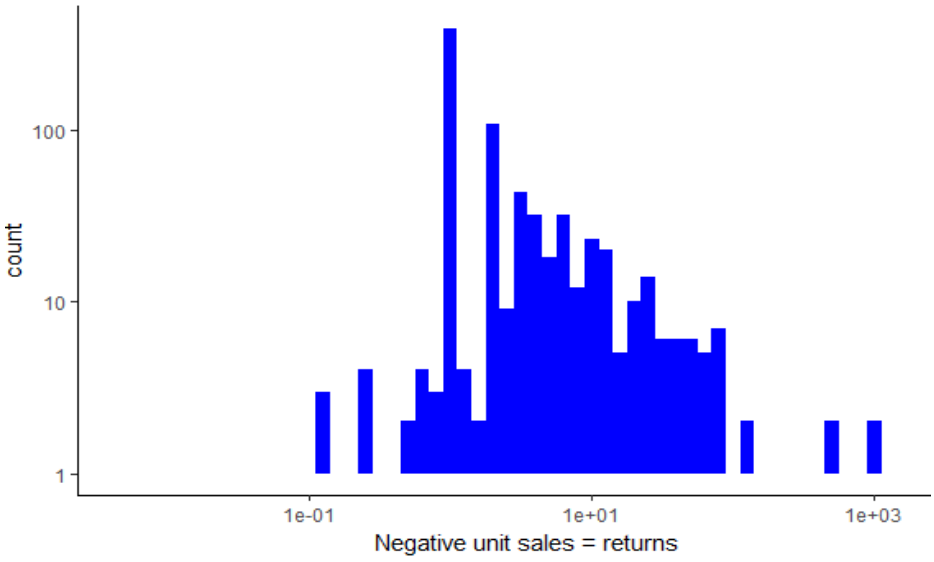
test 데이터의 기간에는 store와 item의 수는 일정하기 때문에 모델의 예측력을 높이기 위하여 train 데이터의 store_nbr과 item_nbr의 수를 조정해주어야 할 필요성이 있음을 확인하였다.

(4) unit_sales 변수 탐색

unit_sales의 음수 값은 반품을 의미하므로 양수인 경우와 음수인 경우, 두가지로 나누어서 분석하였다.

먼저 음수의 경우, <표 2>에서 unit_sales 변수에 음수가 존재하고 1분위수와 최솟값이 비이상적으로 차이가 나는 것을 확인했다.

[그림 6] 음의 unit_sales 분포 히스토그램



자료: Kaggle competition – favorite stores

[그림 6]에서 음의 unit_sales 즉, 반품된 물품의 수가 대부분 1과 100사이에 분포되어 있는 것을 확인할 수 있다. 하지만 음수의 관측치들은 전체 관측개수의 0.0001%로 아주 작은 부분을 차지하고 있다. 또한 우리는 반품된 상품의 수량보다는 매출에 더 관심을 두고 있다.

따라서 이러한 값들은 이상치로 취급하여 학습시키지 않는 것이 모델 설명력 향상에 도움이 될 것이라고 판단하였다.

양수의 경우, 앞의 <표 2>를 통해 unit_sales의 최댓값 89440과 3분위수인 9는 10000배 가까이 차이가 나는 것을 확인하였다. 이 최댓값이 이상치일 가능성이 존재하므로 unit_sales를 내림차순으로 정리하여 극단치의 발생 원인을 탐색하였다.

<표 5> unit_sales 이상치 분석

id	date	store_nbr	item_nbr	unit_sales	onpromotion
9318956	2016-10-07	39	1976284	89440	FALSE
76939364	2016-04-21	20	841842	44142	FALSE
77960441	2016-05-02	2	1162932	30000	FALSE
76693277	2016-04-18	45	559870	20748	FALSE
77959454	2016-05-02	2	305227	20000	FALSE

먼저, id 9318956 관측치는 Cuenca라는 지역으로 10월 7일 특별한 축제가 없었다. 그 날의 에콰도르의 정보를 찾아본 결과 에콰도르가 축구 월드컵 예선경기에서 칠레를 3:0으로 이긴 날이었다. 또한 item.csv 파일을 통해 탐색한 결과, 그 날 가장 많이 팔린 항목은 고기(item_nbr : 1976284)였다. 이는 축구경기라는 이벤트가 극단치의 원인일 수 있음을 시사한다.

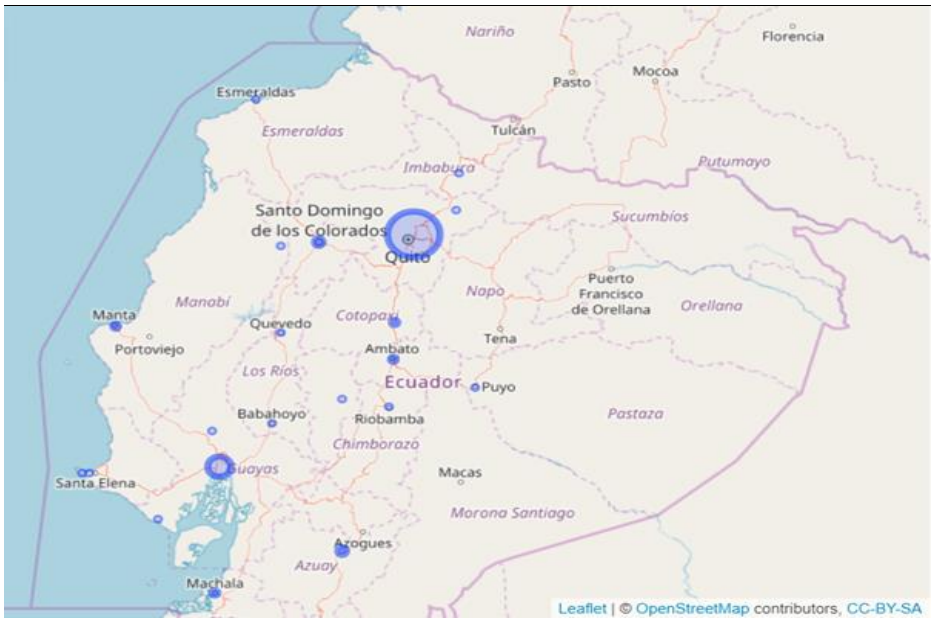
다음으로, 나머지 4개의 극단치는 2016년도 4월 중후반부터 5월초 사이에 모여 있는 것을 알 수 있다. 이는 2016년 4월 16일에 에콰도르에 발생한 규모 7.8의 지진의 영향임을 유추할 수 있다. 비록 store_nbr : 2, 20, 45는 지진이 발생한 장소에서 멀리 떨어진 곳에 위치하고 있지만 구호품을 보내기위해 sales가 크게 증가했음을 예상할 수 있다.

이러한 극단치들은 이상치로 모델의 설명력을 저하시키기 때문에 모델 학습시에는 제거할 필요가 있다.

(5) store data 탐색

먼저 도시 별 store 수와 전체 unit_sales를 살펴보았다.

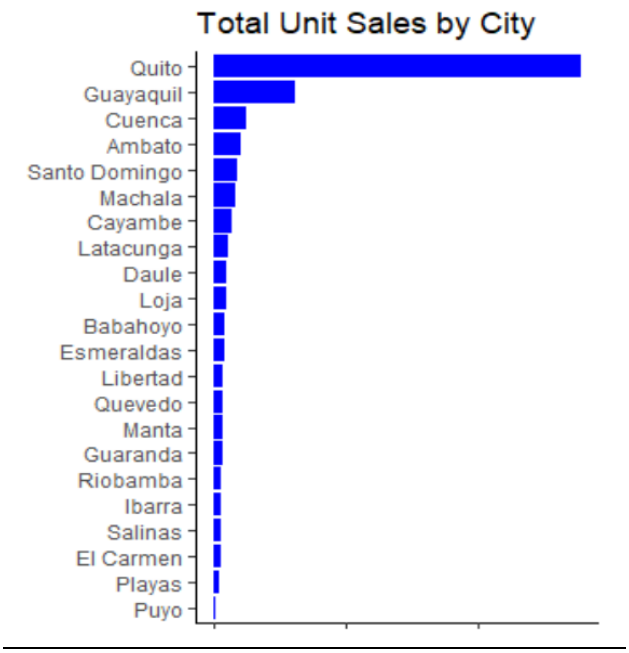
[그림 7] 에콰도르 도시 별 매장 분포



자료: Kaggle competition – favorite stores

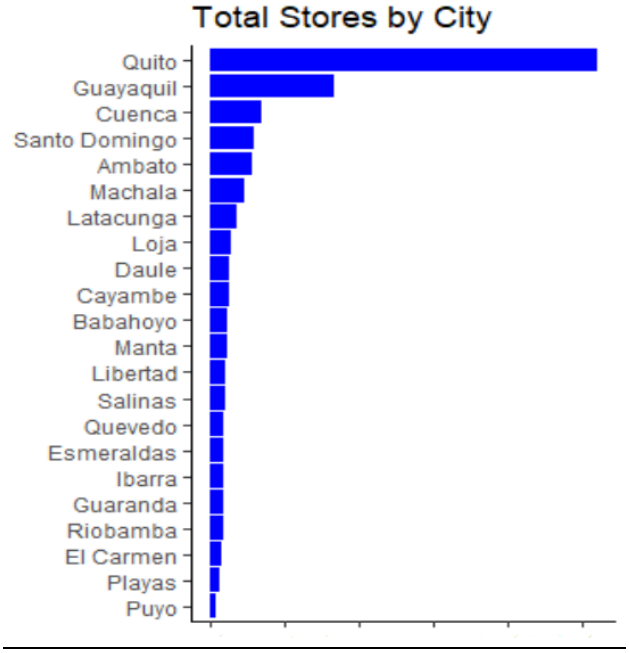
<Favorita Grocery>는 Quito지역에 가장 많고(동그라미의 크기가 큼.) 그 다음으로는 Guayaquil지역에 많이 위치해있다. total unit sales와 total stores를 비교하면 [그림 8], [그림 9]과 같다.

[그림 8] 도시 별 판매량 분포



자료: Kaggle competition – favorite stores

[그림 9] 도시 별 매장 수 분포



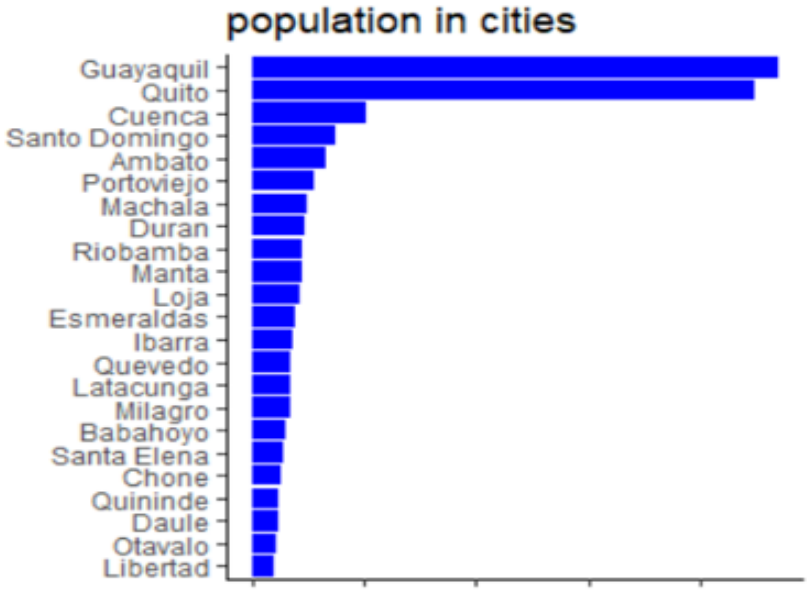
자료: Kaggle competition – favorite stores

시사점 및 아이디어

도시 별 total unit sales는 Quito, Guayaquil순으로 높다. 즉, 도시 별 store수와 total unit sales이 비례한다는 것을 확인 할 수 있다. 하지만 지역 별 매출 2위인 Guayaquil은 store 당 total unit sales가 Quito에 비해 적어 보인다.

먼저, store 수와 관계없이 도시의 인구밀도가 낮아 store 당 total_sales가 적을 수 있기 때문에 도시 별 인구수를 분석하였다.

[그림 10] 도시 별 인구수 분포



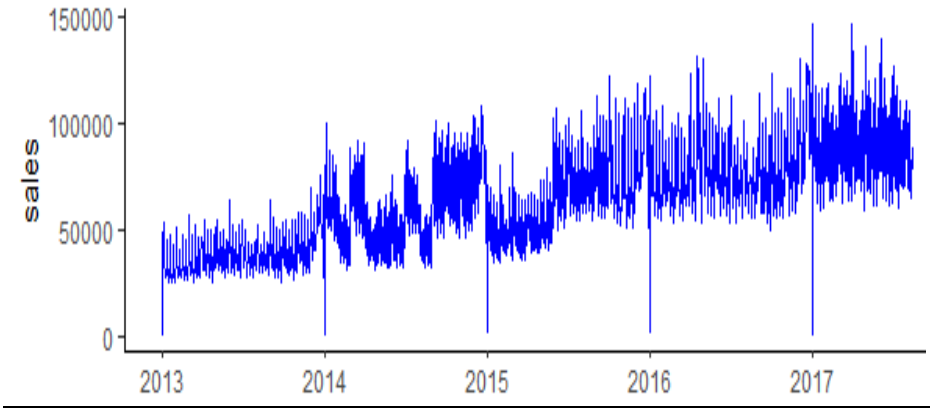
자료: Kaggle competition – favorite stores

[그림 10]을 통해 Guayaquil가 에콰도르 전체 인구수 1위인 것을 통해 인구밀도가 낮아서 store 당 total_sales가 적은 것이 아님을 확인 할 수 있다. 오히려 Guayaquil은 매출 1위인 Quito보다 인구수 자체는 크다. 그렇다면 인구수와 대형마트 store수가 비례하지 않는 이유는 무엇일까?

<Favorita Grocery>는 Quito의 상권을 장악하고 있다. 반면에 Guayaquil의 상권은 여러 대형마트 경쟁사들이 경쟁하고 있어 store수 대비 높은 매출을 올리지 못하고 있음을 확인할 수 있었다.

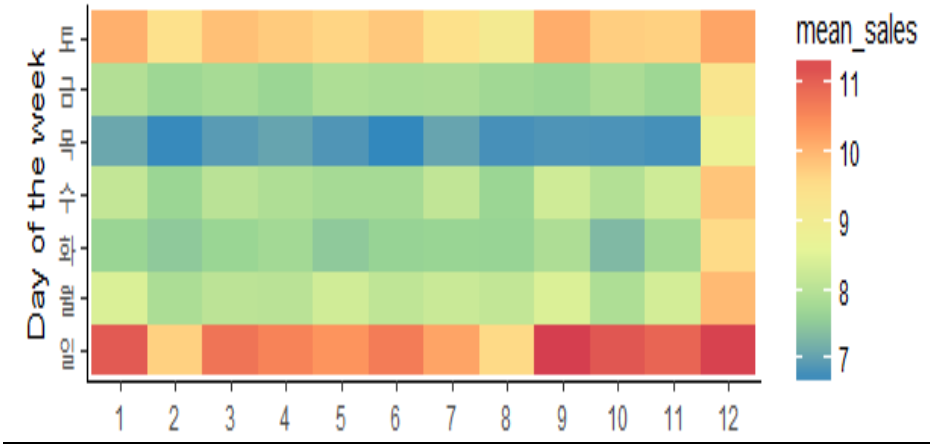
(6) date 변수 탐색

[그림 11] 시간별 unit_sales 분포



자료: Kaggle competition – favorite stores

[그림 12] 요일 별 / 월별 unit_sales 분포



자료: Kaggle competition – favorite stores

[그림 11]를 통해 1월1일 unit_sales가 0으로 떨어지는 것을 확인할 수 있다. 전체적으로 시간의 흐름에 따라 우상향의 추세를 보인다. 하지만 2014년~2015년도 사이는 sales가 이러한 우상향의 추세에서 벗어나는데 이 부분에 대한 분석이 필요해 보인다.

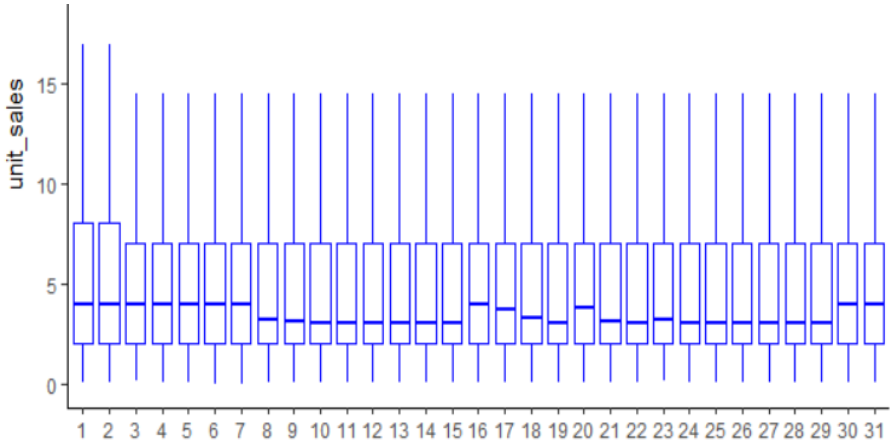
다음으로 [그림 12]의 요일 별 차이를 보면 평일보다는 주말에 상대적으로 sales가 높다는 것을 확인 할 수 있다. 그리고 목요일은 평일 중에서도 가장 낮은 평균 판매량을 보이는 것을 확인 할 수 있다.

[그림 12]의 월별 차이를 보면 1년 중에 12월이 가장 높은 평균 판매량을 갖고 2, 8월은 주변에 비하여 현저히 낮은 평균판매량을 갖는다. 이유를 생각해보면 12월에 가장 높은 평균판매량을 갖는 이유는 크리스마스에 사람들의 소비가 증가하기 때문일 가능성이 높다.

시사점 및 아이디어

2, 8월에 매출이 감소하는 이유는 2월, 8월은 휴가철이라는 사실과 <Favorita grocery>가 대도시인 Quito에서 가장 큰 매출을 보이고 있다는 점을 동시에 고려한다면 2월, 8월에 사람들이 휴가를 위해 대도시를 떠나면서 Quito의 매출이 상대적으로 하락해 결국 전체 <Favorita grocery>의 평균매출에 부정적인 영향을 미쳤을 가능성이 높다.

[그림 13] 월중 unit_sales 분포



자료: Kaggle competition – favorite stores

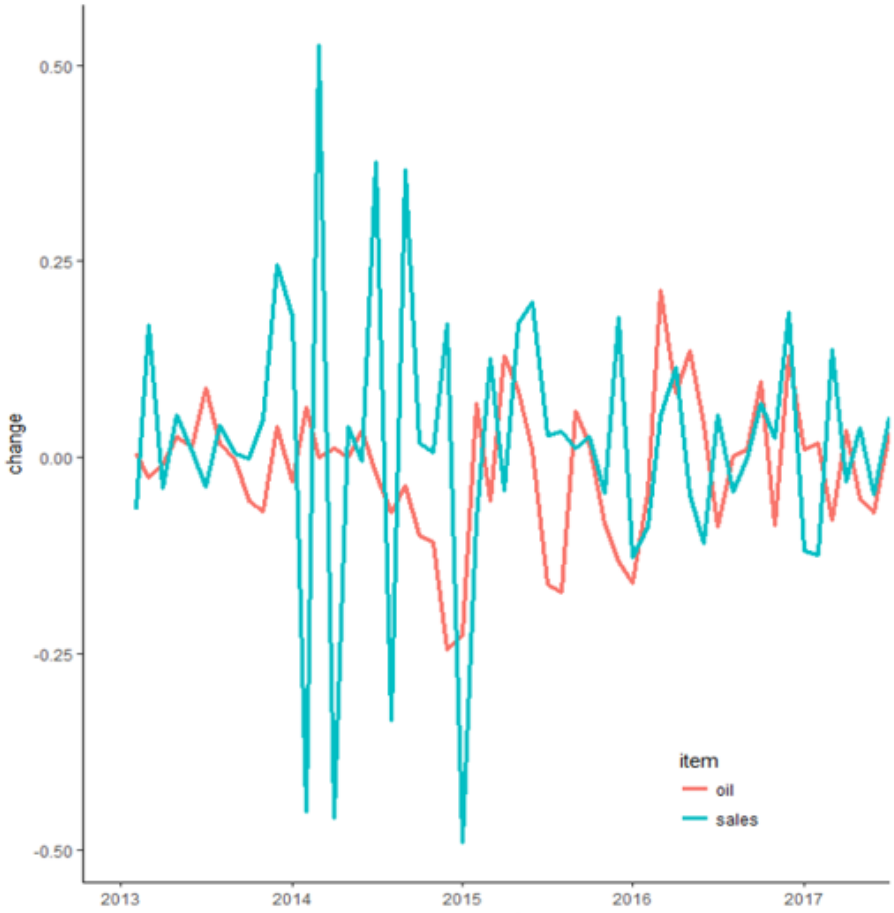
월급날이 unit_sales에 영향을 미치는 것을 확인하였다. Feature engineering 단계에서 변수를 추가하자

[그림 13]를 보면 15일에서 16일로, 29일에서 30일로 넘어가는 시점에서 unit_sales의 median값이 오르는 것을 확인할 수 있다. 에콰도르 또한 대한민국과 비슷하게 대부분 월말과 15일에 임금을 받기 때문에 소비가 증가한 것으로 예측할 수 있다.

(7) oil data 탐색

에콰도르는 남미 국가 중 세번째로 높은 석유매장량을 보이고 석유산업이 GDP의 15%, 총 수출의 50%를 차지할 정도로 석유에 대한 의존도가 높은 국가이다. 그렇기에 세계적인 유가에 의해 사람들의 소비심리가 바뀔 것이라고 예측할 수 있다. 따라서 석유의 가격 변화율과 total sales의 변화율이 어떠한 관계를 가지는지 비교해 보았다.

[그림 14] 유가와 unit_sales의 변화율 비교



자료: Kaggle competition – favorite stores

[그림 14]를 육안으로 보았을 때에는 oil가격의 변화율과 unit_sales의 변화율이 상관관계를 갖는지 파악하기 힘들다. 따라서 두 변수의 상관관계를 구해본 결과, <표 6>과 같이 0.04으로 비교적 낮은 상관관계계수를 갖는 것을 확인할 수 있다

시간 차이를 변화시킴에 따라 상관관계 비약적으로 상승하는 경우도 존재하지만 이는 높은 양의 상관관계를 가질 것이라는 처음 예상에 충분히 부합하지 않는 결과이다.

시간차이가 6개월 날 경우 유가가 가장 강한 상관관계를 보이는 것을 확인하였다. Feature engineering 단계에서 변수를 추가하자

<표 6> 2015년 이후 유가와 판매량 변화율 간의 상관관계

Time lag	0month	1month	3month	6month	12month
Correlation	0.04	0.17	−0.05	0.39	0.17
Corr_Tstat	0.18	0.84	−0.22	1.84	0.61

자료: Kaggle competition – favorite stores

<Favorita grocery>가 의식주와 밀접한 관계를 가진 산업에 속하기 때문에 비교적 유가의 경제상황에 영향을 덜 받을 것으로 유추할 수 있다. 즉 아무리 석유 의존국가여도 석유가격이 동시적으로 국민들은 기본적인 의식주에 대한 소비에 큰 영향을 미치지는 않는 것으로 해석할 수 있다.

하지만, 우리는 이 결론을 두 가지 이유로 주의해야한다. 첫째, 우리는 단지 5년 동안의 월간 데이터만 가지고 상관관계 분석을 진행하였으며 이는 상관 관계를 확립하는 데 충분하지 않은 관측 개수일 수 있다.

둘째, <표 6>에서처럼 유가의 변화가 소비자 구매에 영향을 미칠 때까지 시간이 걸릴 수 있으므로 시간 지연 정도에 따라 상관관계가 강하게 나타날 수 있다. 따라서 시간 차이 정도를 다양하게 변화하여 oil price와 현재의 total_sales를 비교해봐야 한다.

2. 데이터 전처리

1) 2013-2014년 데이터 삭제

[그림 2]에서 확인했듯이 긴 전체 관측기간(약 1600일)에 비해 test 데이터 기간은 15일에 불과하다.

사람들의 라이프 스타일에 큰 영향을 받는 대형마트임을 고려하였을 때, 과거 데이터는 현재의 unit_sales 예측에 있어서 효과적이지 못하다.

[그림 3]에서 onpromotion은 train 데이터의 16%에 해당하는 결측치를 가지고 있고 이는 2013~2014년에 집중되어 있다.

<표 4>에서는 2015년 전 후로 item과 store가 급격하게 증가하는 것을 확인할 수 있다.

*Hair et al.(2006)*에 따르면 10~20%의 결측치는 채워주는 것을 권장한다. 하지만 위에서 확인한 onpromotion의 결측치는 관측기간과 상관성을 띄는 비무작위적인 결측치로 imputate를 하기 어렵다. 따라서 위 4가지 사실을 근거로 2013~2014년도 데이터를 제거하였다.

2) 업다운샘플링을 통한 store, item 수의 비대칭 해결

[그림 4]와 [그림 5]을 통하여 test 데이터와 train 데이터의 store_nbr과 item_nbr 빈도수의 불균형을 확인했다.

이를 해결하기위해 store 52와 같이 관측수가 부족한 데이터는 up-sampling을 해주고 빈도수가 많은 데이터는 down-sampling을 해줌으로써 분포의 균형을 맞추어 주었다.

3) unit_sales의 극단치 및 음수 처리

<표 2>에서 확인했듯 unit_sales는 판매량임에도 불구하고 음수 값과 비이성적 극단치를 가지고 있다. 또한, [그림 7]을 보면 unit_sales의 음수값은 1과 100사이에 분포하고 전체 0.0001%의 비율로 작은 부분을 차지하고 있다.

예측할 때, 반품은 가정하지 않고 있기 때문에 음수값을 이상치로 취급해 해당 데이터를 제거해 주었다.

<표 5>에서 unit_sales의 극단치들의 날짜(date)를 근거로 극단치가 발생한 이유를 추론해 본 결과 에콰도르의 축구 우승, 지진 등의 특수한 이벤트로 인한 극단치였기 때문에 이를 이상치로 간주하고 제거하여 모델의 성능을 높였다.

III. Feature Engineering

1. 변수 생성(내부 데이터)

EDA에 기반하여 내부 데이터를 이용해 새로운 기본 변수를 생성하였다.

is_salary : 16일 30일은 월급일로 위의 [그림 13]에서 봤듯이 unit_sales의 median값이 증가하는 것을 볼 수 있다. 그래서 월급날 여부를 0, 1으로 표시하는 변수.

is_salary_last : 월급을 받은 날로부터 며칠이 지났는지를 표시해줄 변수.

oil_lag_6 : [그림 14]에서 봤을 때, 6달 뒤의 유가와 unit_sales 사이에 상관관계가 높은 것을 확인할 수 있음.

last_sales_day : 7,14,30 : 7,14,30일 전의 sales의 평균

last_promo_day : 마지막 promotion으로부터 며칠이 경과했는가를 나타내는 변수

2. 변수 생성(외부 데이터)

EDA에 기반하여 모델의 성능을 향상시킬 수 있는 외부 데이터를 추가해 새로운 변수를 생성하였다.

weather : 날씨에 따라 소비 심리나, 외출 빈도가 다르기 때문에 외부 데이터를 이용해 추가하였다. 에콰도르의 날씨를 도시별로(city)를 기준으로 생성하였다.

cpi : 물가지수에 따라 소비자들의 소비 패턴이 변화할 수 있기 때문에 에콰도르의 월별 CPI(Consumer Price Index) 지수를 생성하였다.

world_banana_price : 에콰도르는 바나나수출국 1위국가로 oil과 마찬가지로 전세계 바나나가격에 민감하기 때문에 OECD 바나나 상품 가격을 변수로 추가하여 unit_sales에 대한 설명력을 높인다.

3. 변수 선택

과적합(overfitting)의 문제를 피하기 위해 feature selection을 진행하였다. 기본적으로 하나의 변수를 추가해가면서 validation의 cost값을 구하는 것이 가장 합리적인 방법이지만 부족한 시간 여건상 scikit-learn에서 제공하는 SelectKBest method를 이용하였다.

IV. Modeling

1. 모델 선정 및 파라미터 최적화 / 모델 검증

1) Linear regression Model

(1) 모델 전처리

선형회귀분석은 연속형의 target변수를 예측하는 가장 대표적인 예로 여러 조건과 가정을 만족하였을 때 강력한 설명력을 보여주는 모델이다. 가우스 마르코프 정리에 의한 5가지 조건은 아래와 같다.

1. 설명변수 x(연속형,이산형)와 종속변수 y(연속형)는 선형의 관계를 가진다.
2. 샘플은 모수로부터 편향없이 무작위로 추출된다.
3. 모든 설명변수 x에 대해 잔차의 조건부 기댓값은 0이다.
4. 설명변수 x들 사이에 완벽한 다중공산성(colinearity)이 없어야 한다.
5. 모든 설명변수 x에 대해 잔차의 조건부 분산은 일정하다.

이를 유의하여 log전처리를 취해주고, onpromotion과 같이 범주형 변수에 대해서는 one-hot-encoding을 적용하여 모델 전처리를 진행하였다.

(2) 모델 검증

<표 7> Linear regression 검증

OLS Regression Results			
Dep. Variable	unit_sales	R-squared	0.324
Model	OLS	Adj. R-squared	0.324
Method	Least Squares	F-statistic	3.224e+05
Date	Sat, 28 jul 2018	Prob (F-statistic)	0.00
No.Observations	3370464	AIC	2.974e+07
Df. Residuals	3370464	BIC	2.974e+07

자료: Kaggle competition - favorite stores

비록 모델의 32%밖에 설명하지 못하지만 모델 자체는 유의미한 결과를 얻었다.

(3) 결과

NWRMSLE = 0.591

2) Log MA

(1) 모델 전처리

전통적으로 시계열 데이터는 규칙적인 패턴과 불규칙적인 패턴의 결합으로 여겨져 왔고, 이는 곧 과거의 규칙적인 패턴이 현재에 미치는 영향을 나타내는 자기상관성(auto - correlation)과 불규칙적인 패턴의 평향성을 초래하는 이동평균현상(moving average)으로 구분하였다.

본 분석에서는 여러 시계열 모델 중 가장 높은 예측력을 가지는 logMA 모형을 사용하였다. unit_sales에 log(+1)값을 취하여 정규화 분포에 가깝게 변환시킨 후 MA 모형을 적용한다. 위의 세가지 모델과는 다르게 변수로 1,3,7,14,28,56,112일 전의 평균 sales를 구해서 평균 낸 값을 예측에 사용한다.

(2) 결과

NWRMSLE = 0.575

3) LSTM

(1) 모델 전처리

RNN에서 파생된 딥러닝 모델로서 이용하는 과거 정보의 시점과 사용 시점의 거리가 멀 경우 그 영향을 제대로 받아들이지 못하는 것을 해결하기 위해 cell-state를 추가한 모델이다.

다변량회귀분석과 마찬가지로 설명변수들간의 상관관계가 낮을수록 모델의 설명력이 높아지므로 다중공산성이 없어야한다.

(2) 하이퍼파라미터 최적화

마찬가지로 LSTM 내부에 쓰이는 여러가지 하이퍼파라미터는 베이지안 최적화를 이용하여 찾아주었다.

(3) 모델 검증

K-fold를 이용하여 진행하였다.

(4) 결과

NWRMSLE = 0.527

4) LightGBM

(1) 모델 전처리

현재 kaggle에서 가장 많이 사용되는 모델로 Gradient boosting decision tree라는 머신러닝 알고리즘을 이용하는 방법이다. 그 이유는 이상치와 결측치에 덜 민감하고 무엇보다도 빠른속도로 좋은 결과를 내기 때문이다. Lightgbm을 사용하기 위해 먼저 데이터를 Matrix형태로 조정해주고 범주형 변수들은 따로 저장해서 모델에 인식시켜주었다. 그리고 모델에 쓰이는 하이퍼파라미터는 LightGBM Github의 설명을 참고하여 기본적인 default값을 설정해 주었다.

(2) 하이퍼파라미터 최적화

하이퍼파라미터 최적화는 최근 kaggle 및 다양한 연구에서 좋은 성과를 보이고 있는 베이지안 최적화를 이용하여 진행하였다. 다른 최적화 기법인 그리드서치나 랜덤서치는 매 시행이 독립적이어서 과거의 정보를 활용하지 못하고 많은 비용이 든다는 단점이 있다.

(3) 모델 검증

검증은 K-fold를 이용하여 진행하였다. 그 이유는 시계열데이터는 예측시점과 가까울수록 상관성이 높기 때문이다.

(4) 결과

NWRMSLE = 0.523

V. Result

1. 요약 및 결론

<표 8> 모델간 결과 비교

Model 분류	모델	결과	최초 등장 연도	순위
전통적 기법	Lineaer regression	0.591	1903년도	4
	Log MA	0.575	1938년도	3
현대적 기법	LSTM	0.527	1997년도	2
	LightGBM	0.523	2016년도	1

전통적인 수요예측 기법과 최근에 쓰이는 수요예측 기법의 성능 비교를 위해 <Favorita Grocery>의 데이터를 가지고 분석을 진행하였다. 먼저 EDA과정을 통해서 데이터를 정제가 필요한 부분과 모델의 성능을 높일 변수를 탐색하였다. 이를 통해 데이터 전처리와 feature Engineering을 수행하고 마지막으로 4가지 모델(linear regression, logMA, LSTM, LightGBM)로 예측을 진행하였다.

모델의 성능 측면에서 LightGBM(2016), LSTM(1997), log MA(1938), Linear regression(1903) 순으로 최근에 나온 모델일수록 결과가 좋았다. 예상했던 대로 최근에 나온 기법이 전통적 수요예측 기법보다 성능측면에서 우월했다.

Linear regression 과 logMA의 경우 갑작스러운 변동에 반응하지 못하고 새롭게 추가되는 관측치를 대처하지 못하는 한계가 있기 때문에 item이 새롭게 추가되는 본 분석에서는 충분한 성능을 보여주지 못했다.

2. 한계

위의 <Favorita Grocery> 데이터에서는 최신 수요예측 모형이 전통적 수요예측 모형보다 좋은 성능을 보였다. 하지만 위 분석의 과만으로 최신 수요예측 모형이 무조건 우위에 있다는 결론을 내리기는 어렵다.

먼저, 사례가 적어서 위 분석의 결과만으로 결론을 내리기에는 신뢰도가 떨어진다.

두번째로, 위의 사례에서는 모델 성능의 척도를 '예측의 정확도'만 이용하였지만, 데이터를 수집하는데 드는 비용, 분석을 진행하는 시간, 메모리 등의 다른 변수들 또한 고려하여 모델의 성능을 판단하여야 한다.

세번째로, logMA의 경우 단순히 unit_sales의 이동평균을 기간을 달리하여 평균 낸 결과로 예측하였다. 만일 promotion이나 월급일, 주말, 휴일과 같은 정보에 가중치를 부여했다면 더 좋은 결과를 얻었을 것이다.

마지막으로, 최신기법의 경우, 특히 LightGBM, LSTM의 경우 대표적인 black-box모형으로 결과가 좋더라도 그 과정을 설명하고 해석하는 데에 한계가 있다.

위의 4가지를 고려하면 단순히 최근에 나온 모델일수록 우월하다고 단정짓고 사용하기보다는 충분히 사례를 수집하고, 상황에 맞는 모델을 사용해야한다. 또한 '예측의 정확성' 측면 뿐 아니라 설명력, 데이터를 수집하는데 쓰이는 비용, 분석 시간, 자원, 목적 등 여러가지 변수를 다양한 측면에서 고려해야 한다.