제2회 유통데이터 활용 경진대회

데이터를 활용한 중소 유통물류 수요판매량 예측

#유통 꿈나무

목차

1. 분석 개요

- 연구 요약
- 연구 배경 및 필요성

2. 분석 방법 및 절차

- 데이터 탐색
- 활용 데이터 탐색
- 결측치 처리
- 내부 변수 및 외부 변수
- 가중치 부여
- Trend(추세) 변수

3. 분석 모형 적용 및 결과

- 외부 변수 추가 및 선정 방법
- 모델링 (XGBoost, Random Forest, Light GBM, Gradient Boost)
- 결과

4. 활용 방안

5. 활용 데이터 참고 문헌 및 출처

연구 요약

현재 매우 빠른 속도로 성장하고 있는 **유통 물류** 시장에서 정확한 **수요 예측**은 **재고 최적화, 공급망 관리, 고객 서비스 향상** 등 다양한 비즈니스 의사 결정 과정에서 중요한 역할을 한다.

본 연구에서는 유통 물류 판매 데이터와 **시장 환경의 변화**를 반영하기 위한 다양한 소비자, 유통업자, 소매업, 환경 및 경제지수 등의 외부 데이터를 활용하여 **판매량 수요 예측**을 진행하였다. 기존 시계열 분석에 사용 되는 전통적인 알고리즘이 아닌 **XGBoost, Random Forest, Light GBM, Gradient Boost** 등 다양한 회귀 모델을 사용하여 최적의 모델을 선정하였다. 이는 시장의 환경을 반영하여 **오차를 개선한 정확한 수요예측 모델**을 통해, **물류센터 내 다양한 운영 비용을 절감**하고, **고객사별 주문량 예측 서비스를 마케팅/영업에서 활용하는 방안**을 제안한다.

1. 분석 개요

연구 배경 및 필요성

물류 서비스 산업은 빠르게 성장하고 변화하므로 **AI와 데이터 분석**을 활용한 수요 예측과 운송 및 인력 공급 계획을 효율적으로 수립하고, 환경변수를 고려한 운영의 필요성 존재

전략적 의사 결정 및 경쟁 우위 확보

재고와 생산 계획을 효율적으로 조정함으로써 경쟁에서의 우위를 점하기 위함

고객 서비스 향상

처리 및 배송 시간을 줄여 고객 만족을 상시키고, 재고 부족으로 인한 고객 불만을 방지하기 위함

e커머스 풀필먼트

짧은 시간에 배송을 완료 해야하는 사업의 특성을 만족시키기 위함

2. 분석 방법 및 절차

데이터 탐색 - 자료 파악

제공 데이터 (필수 데이터) *2021-01-04 ~ 2022-07-03 (78주)*

구분	판매일	정보
1	구분	매출, 반품
2	우편번호	소매점 위치 정보
3	판매수량	상품이 판매된 수량
4	옵션코드	EA : 최소 단위, CS: 묶음 단위, BX : 박스 단위
5	규격	상품 입고 시 박스에 담겨져있는 수량
6	입수	해당 옵션 코드에 상품이 들어있는 EA 수량
7	상품 바코드	상품에 부여되는 코드 번호
8	상품명	상품 이름

데이터 특성 파악

- **구분**

'판매 수량'을 예측하기 위해 반품 데이터를 제외하고 **매출 데이터**만 사용

- **우편번호**

대부분의 소매점 위치가 **경북 지역**에 위치하고 있음을 확인

- **상품바코드, 상품명**

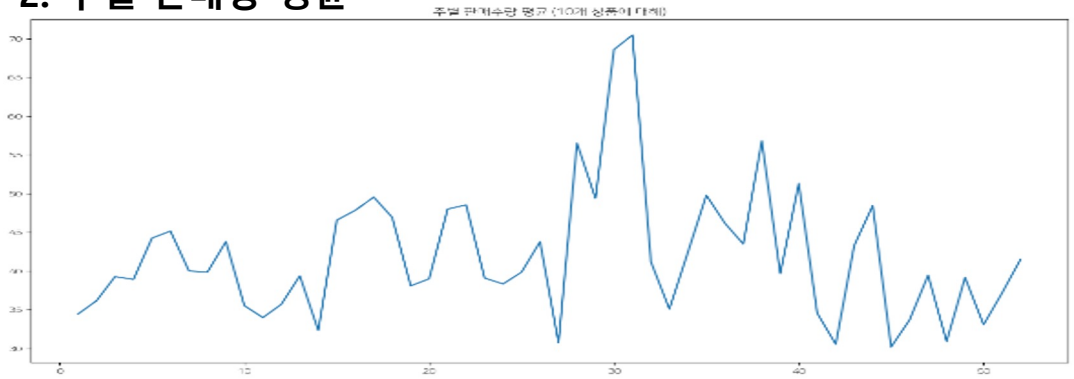
총 8716개의 상품 정보 중, 제품명은 같지만 묶음과 낱개가 다른 상품들이 있다는 것을 확인

2. 분석 방법 및 절차

데이터 탐색 - EDA

1. 일별 총 판매 수량

2. 주별 판매량 평균



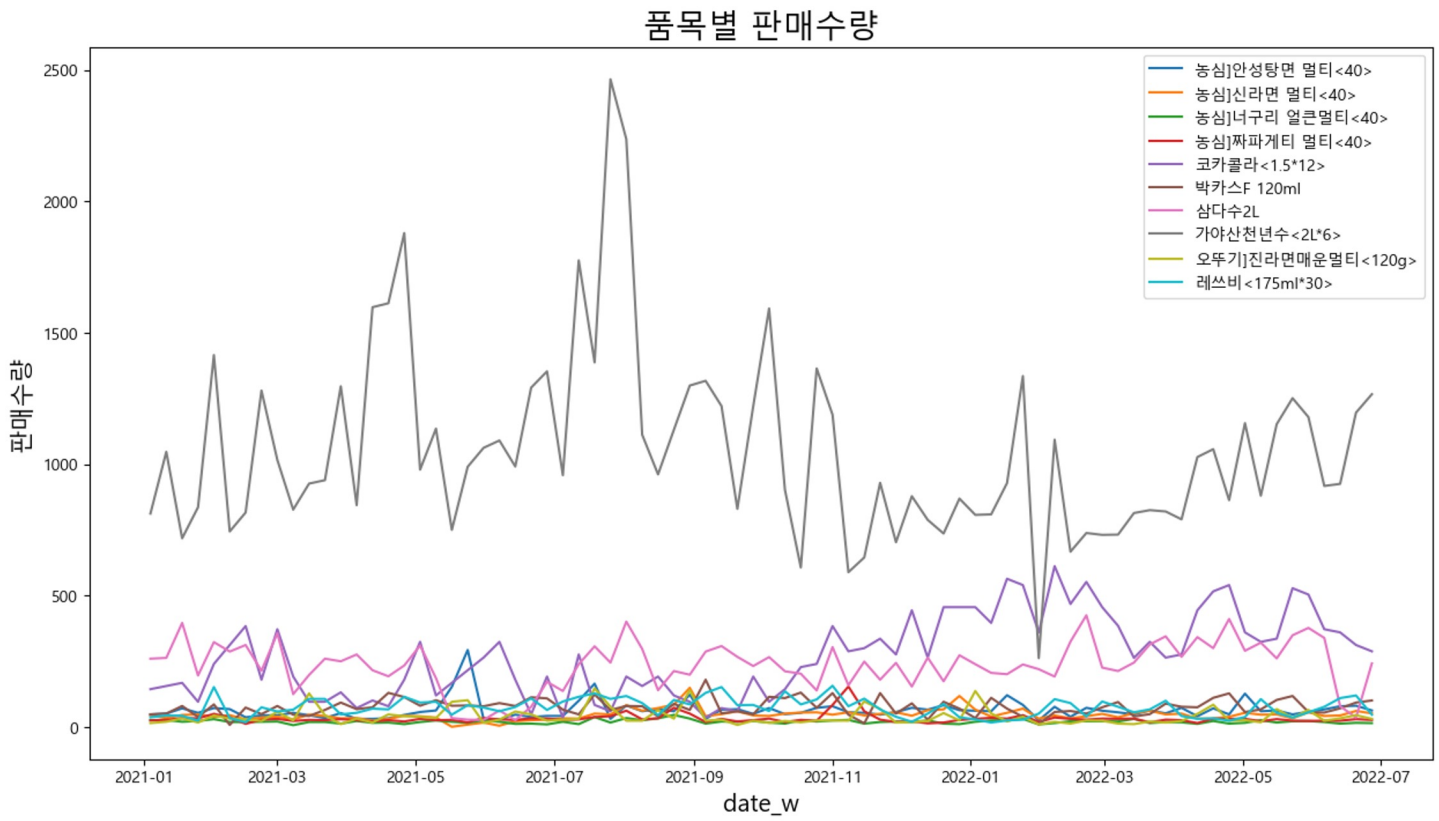
** EDA란?*
탐색적 데이터 분석을 의미하며, 시각화를 통해 데이터의 특성을 파악하는 것

데이터 특성 파악

- 일별 판매수량과 주별 판매수량 모두 **7 - 8월**에 급증하는 것을 확인
- 각 상품의 모든 일자에 대한 판매량이 존재하지 않음
  품목별 판매수량 중 **NaN 값**이 존재하는 일자가 존재
  → 존재하는 일자가 존재
- 일별 판매수량은 변동성이 크기에 **주별 판매수량으로 분석**을 진행하는 것이 더 유의미하다고 판단

2. 분석 방법 및 절차

데이터 탐색 - EDA



데이터 특성 파악

- EA(최소 단위)에 비해 CS나 BX의 판매 수량이 적음
- 삼다수와 가야산천년수와 같이 동일 분류인 상품이더라도 판매수량의 유사성이 떨어져, Train data를 **예측 상품인 10개의 품목만으로 구성**

2. 분석 방법 및 절차

결측치 처리

판매 수량 데이터에
결측치를 갖는 '**주차**' 존재

(2021-01-04 ~ 2022-07-03
(78주)동안 주(일주일)를 나타냄)

누락된 주차에 대해
이전 값과 다음 주차 판매량
평균으로 결측치를 채움

	상품 바코드	상품명	옵션코드	date_w
0	8.801043e+12	농심]안성탕면 멀티<40>	BX	78
1	8.801043e+12	농심]신라면 멀티<40>	BX	77
2	8.801043e+12	농심]너구리 얼큰멀티<40>	BX	78
3	8.801043e+12	농심]짜파게티 멀티<40>	BX	78

	상품 바코드	상품명_x	옵션코드	date_w
0	8.801043e+12	농심]안성탕면 멀티<40>	BX	78
1	8.801043e+12	농심]신라면 멀티<40>	BX	78
2	8.801043e+12	농심]너구리 얼큰멀티<40>	BX	78
3	8.801043e+12	농심]짜파게티 멀티<40>	BX	78
4	8.801094e+12	코카콜라<1.5*12>	EA	78

범주형 데이터 처리 [옵션코드]

머신러닝에서는 범주형 변수를
사용할 수 없기 때문에 옵션코드에
수치형 변수 (0,1)로 변환하는
One hot encoding을 적용함

연속형 데이터 처리

변수들의 값의 범위가 일치하도록
조정하여, 연속형 데이터에 대해
정규화 기법인 **Min-max Scaling**을 사용함

2. 분석 방법 및 절차

내부 변수

변수명
월
주
분기
년도
월_일_년도
주기성 변수
추세 변수 (Trend)

(필수 데이터)

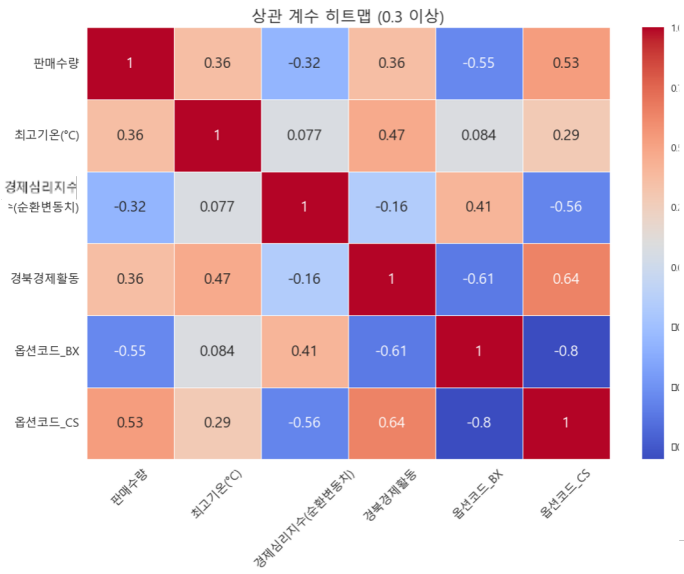
변수명
주별_QUANTITY
주별_AMOUNT
주별_PRICE
월별_QUANTITY
월별_PRICE

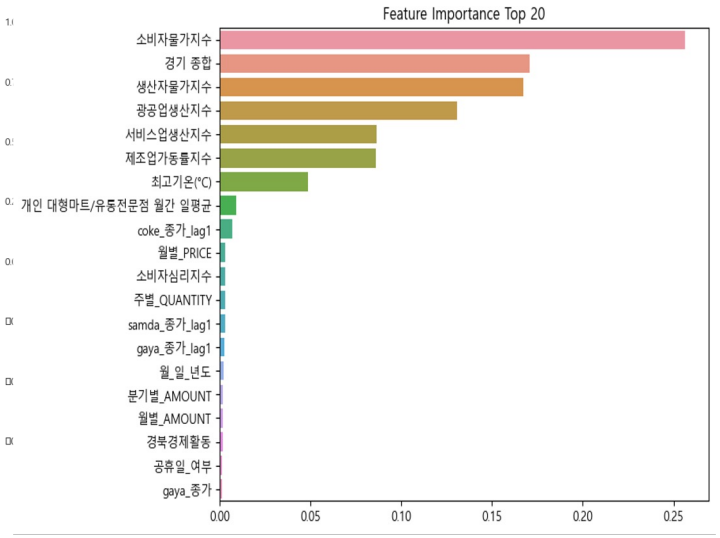
(참고 데이터)

변수명
월별_AMOUNT
분기별_QUANTITY
분기별_AMOUNT
분기별_PRICE

(참고 데이터)

내부변수 중요도





상관분석

확률론과 통계학에서 두 변수 간에 어떤 선형적인 관계를 갖고 있는지분석하는 방법으로 **판매수량과의 상관도** 확인

Feature Importance

모델 학습 과정에서 각 변수들이 판매수량에 미치는 영향를 나타낸 지표

2. 분석 방법 및 절차

외부 변수

대/소형 마트
개인 대형마트/유통전문점 총액
개인 대형마트/유통전문점 월간 일평균
개인 신용카드 슈퍼마켓 총액
개인 신용카드 슈퍼마켓 월간 일평균
대형마트_주별_QUANTITY
대형마트_주별_AMOUNT
대형마트_주별_PRICE

주가
코카콜라_주가
가야산_주가
삼다수_주가

기타 (환경적 요인)
공휴일
공휴일_여부
최고기온(℃)
코로나 일별 확진자수

경제 관련지수	
경제 심리지수(원계열)	
경제심리지수(순환변동치)	소비자
소비자심리지수	
경기 종합	
소비자 물가지수	
생산자 물가지수	
광업생산지수	
서비스업생산지수	생산자
제조업가동률지수	
광공업-생산자확산지수	
경북고용률	
경북 경제활동	환경적

변수 선택 기준

- 유통 판매량은 다양한 요인에 영향을 받기 때문에, 유통 과정의 이해관계자인 **소비자 , 생산자, 소매업체와 관련된 지표**를 외부 변수로 사용함
- 해당 데이터의 탐색 결과를 반영하여 **경북지역 데이터**를 수집
- **주가**에 **지연값(lag)**을 추가하여 예측 상품과 상관분석을 해본 결과, 1,2,3 주차의 높은 상관관계를 보여 **lag값을 3주로 설정**함

2. 분석 방법 및 절차

가중치 부여

ex) 안성탕면 멀티<40>

변수	x 상관계수	가중치 적용된 내부변수
월		월 * 0.0687
주		주 * 0.70139
분기		분기 * 0.086273
…		…
공휴일_여부		공휴일_여부 * 0.079021
최고기온(℃)		최고기온(℃) * 0.082941
코로나 일별 확진자수		코로나 일별 확진자수 * 0.094216

*** 가중치 부여 방법**

1. 각각의 변수와 판매수량 간의 상관계수 계산
2. (품목별 변수) * (상관 계수) = 가중치 적용된 변수

상품명	공휴일_가중치	공휴일_여부_가중치	최고기온_가중치	patient_가중치	년도_가중치	분기_가중치	월_가중치	주_가중치	…	stock_coke_mean_가중치
농심]안성탕면 멀티<40>	0.018935	0.028361	0.154700	-0.058148	0.014938	0.086273	0.068759	0.070139	…	0.094216
농심]신라면 멀티<40>	0.045027	0.074867	-0.007921	0.020643	0.023976	0.325617	0.328592	0.342958	…	0.160113
농심]너구리 얼큰멀티<40>	-0.122665	-0.064674	0.021887	-0.012871	-0.064901	0.046221	0.001135	-0.004910	…	-0.048674
농심]짜파게티 멀티<40>	-0.115347	-0.130636	0.066745	-0.109187	-0.132975	0.224395	0.203391	0.190037	…	-0.036848

[품목별 가중치 예시]

- **모든 변수** (내부변수, 외부변수)에 대해 판매수량과 각 예측 상품에 대한 상관 관계를 구하기 위해 **상관 분석**을 수행
- 전체 판매 수량에 따른 중요도 (Feature Importance)를 통해 변수를 선택하는 대신, **품목별 판매수량에 대한 따른 영향도를 반영하기 위해 상관계수를 가중치로 적용**

**상관 분석: 두 변수 간의 관련성을 평가하는 통계적 기법으로 1에 가까울수록 강한 관련성을 지님*

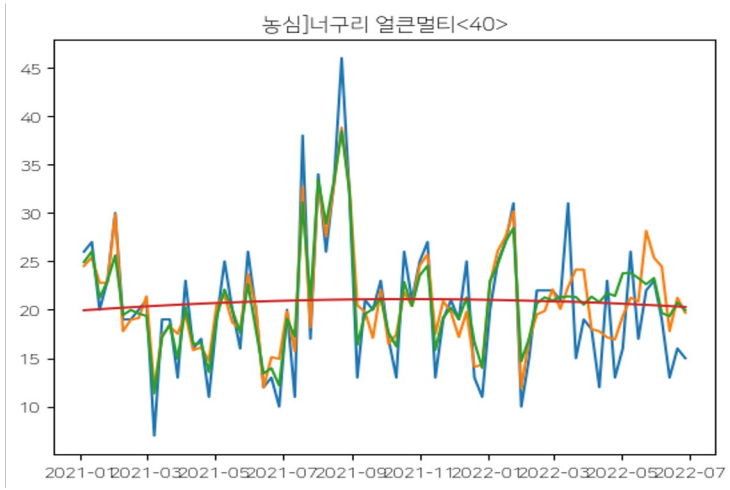
2. 분석 방법 및 절차

Trend(추세) 변수

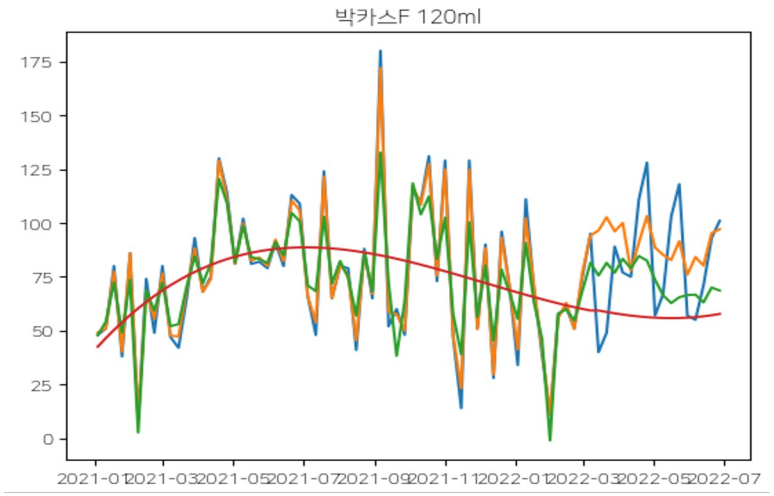
- 예측해야 하는 6개월의 기간에 대한 실제 판매수량 정보를 모르기 때문에, 시계열 예측 모델에 변수로 사용되는 **판매 수량의 이동 평균이나 lag값 등을 사용하기에는 적합하지 않다**고 판단함

- 따라서, 상품별 과거의 판매수량을 바탕으로 **예측이 필요한 기간의 추세를 다항 회귀선을 통해 예측하여 모델이 함께 반영하도록 함**

2차항 Fitting



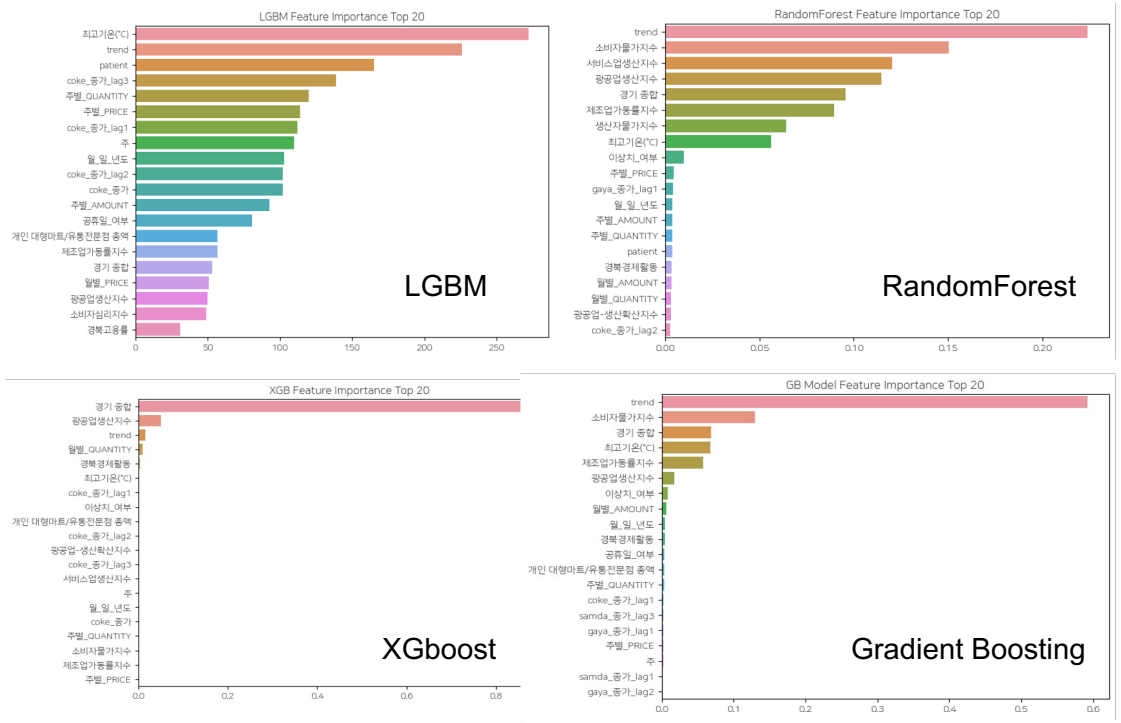
3차항 Fitting



각 상품별로 2차, 3차, 4차 등 다항 회귀 모델을 활용하여 **추세**를 예측한 뒤, **'trend'** 라는 변수로 추가하여 학습에 활용함

3. 분석 모형 적용 및 결과

모델 선정



4가지 모델에 대한 feature importance

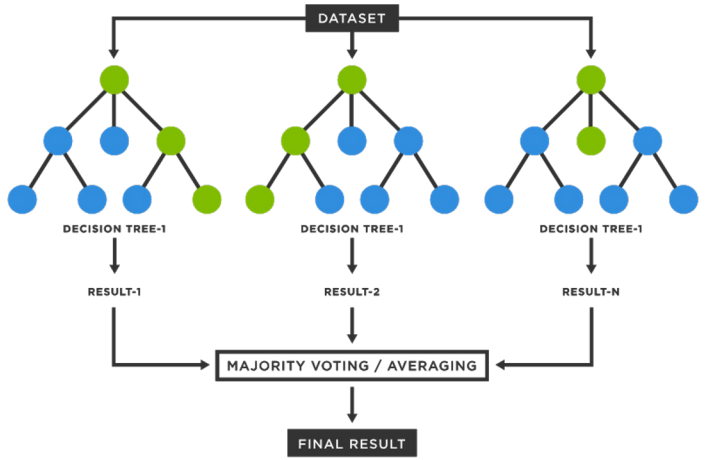
4가지 모델에 대해 feature importance를 진행한 결과,
현재 사용 중인 변수에 LGBM 과 Random Forest Model이 가장 적합

LightGBM

트리 기반 Gradient Boosting 모형
Leaf-wise(리프 중심 트리 분할) 확장 방식으로
빠른 속도 대용량 데이터에 적합

Random Forest

트리 기반 Gradient Boosting 모형
예측 변동성이 작고, 과적합 가능성 낮음



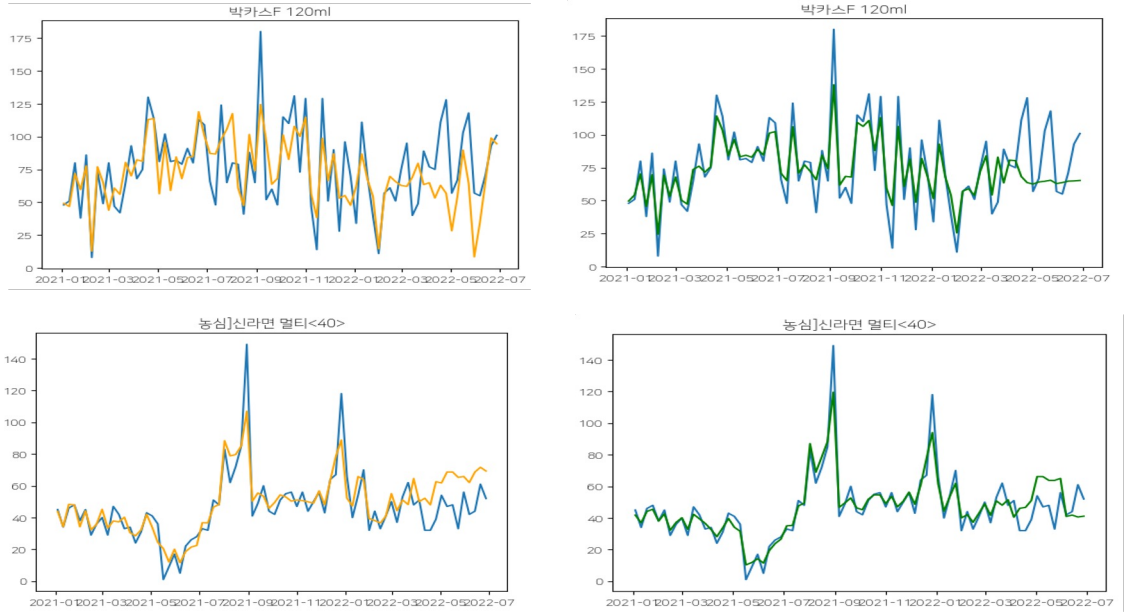
3. 분석 모형 적용 및 결과

모델 비교 및 결과 (RMSE)

RMSE란? 예측 모델에서 예측한 값과 실제 값 사이의 평균 차이

	xg	LGBM	rf	gb
가야산천년수<2L*6>	279.042	142.907	221.346	285.717
삼다수2L	137.934	130.512	114.894	103.528
박카스F 120ml	29.216	33.181	30.855	42.209
레쓰비<175ml*30>	32.004	34.086	30.911	33.950
코카콜라<1.5*12>	284.170	311.480	274.699	258.295
농심]안성탕면 멀티<40>	31.949	24.637	23.728	21.023
농심]신라면 멀티<40>	16.355	18.622	14.623	10.476
농심]짜파게티 멀티<40>	6.431	6.474	9.277	9.871
오뚜기]진라면매운멀티<120g>	22.217	27.803	18.956	18.211
농심]너구리 얼큰멀티<40>	9.106	6.713	6.575	9.175
10개평균	84.843	**73.642**	**74.586**	79.246

Light GBM , Random Forest 예측 그래프



Light GBM 그래프 (주황)

Random Forest 그래프 (초록)

- 전체 변수에 대해 학습한 4개 모델 중 RMSE가 가장 낮은 2개 모델 **Light gbm, Random Forest** 로 검증 진행
- **각 품목 별로 Random Forest** 과 **Light GBM** 보다추세를 반영하고 있음을 확인

3. 분석 모형 적용 및 결과

추가 조정 - 하이퍼 파라미터 조정(Pycaret)

→ **최종 선정 파라미터**

```
RandomForestRegressor
RandomForestRegressor(max_depth=9, min_impurity_decrease=0.1,
                      min_samples_leaf=4, min_samples_split=7, n_jobs=-1,
                      random_state=123)
```

***최적의 예측 모델을 도출하기 위해, Pycaret 라이브러리를 활용하여 하이퍼 파라미터를 조정**

: 하이퍼 파라미터를 조정한 결과, Random Forest의 **RMSE 값**이 **74 → 65**로 개선됨

최종 모델 도출 과정

XGBoost, Random Forest, Light GBM, Gradient Boost 적용

↓

RMSE값 기준으로 Random Forest, Light GBM Model 선택

↓

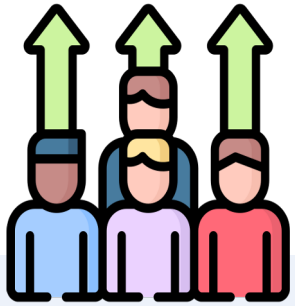
Random Forest Model 하이퍼 파라미터 조정

↓

최종 Random Forest, LGBM 하이브리드 모델

4. 활용 방안

기대효과



풀필먼트 센터 내
최적화된 작업 인력 확보 및
비즈니스 의사 결정을 통해
안정적인 상품 출고 가능

오차 개선을 통해 물류센터 내
다양한 **운영 비용 절감** 가능

고객사별 주문량 예측
서비스를 마케팅/영업 신규
상품 개발 및 프로모션에도
활용 가능

5. 활용 데이터 참고 문헌 및 출처

- *한국은행 경제통계시스템 ECOS (경제심리지수, 소비자심리지수, 개인대형마트 유통전문점, 개인 신용카드 슈퍼마켓 데이터)*
- *인베스팅닷컴 (코카콜라 주가)*
- *KRX 정보데이터시스템 (삼다수, 가야천년수 주가)*
- *한국천문연구원 특일 정보 (공휴일, 기념일)*
- *KDX 한국데이터거래소(경북 코로나 확진자 수)*
- *기상청 기상자료 개방 포털(호미곶 날씨데이터)*
- *KOSIS(경북경제활동, 경북고용률, 광공업생산지수, 제조업가동률지수, 소비자물가지수, 생산자물가지수, 광공업-생산확산지수)*
- *KRX 정보데이터시스템(지수 및 보조자료, 구성지표 및 기타 경기지표, 경제활동 참가(%), 고용률(%))*
- *e커머스 풀필먼트 비즈니스를 위한 수요예측 모델 연구(논문)*

  김영남 1 , 모혜란 1 , 김현 2 1 숭실대학교 IT 정책경영학과 2 경희대학교 컴퓨터공학과

감사합니다