

[제4회 유통데이터 활용 경진대회 | 수요예측 부문]

## 2-Stage Prediction Framework: SCM 관점에서의 양상블 기반 매출·매입 예측

팀 명: move&move

팀 원: 진현준, 박천상, Tagir

### 1. 분석개요

본 분석의 목적은 도매공급망(SCM)에서 매출 및 매입 데이터를 동시에 예측하여 재고·발주 효율을 극대화하는 것이다. 중소유통 물류센터의 경우, 매출 변동에 따라 적절한 매입량을 결정하는 것이 핵심 과제이나, 기존의 단일 예측 모델은 매입과 매출 간의 연계 관계를 충분히 반영하지 못해 예측 오차가 크게 발생한다.

이를 해결하기 위해 본 연구에서는 **매출 예측 결과를 기반으로 매입량을 추정하는 2단계(2-Stage) 하이브리드 예측 프레임워크**를 제안하였다.

1단계에서는 XGBoost와 Random Forest 양상블을 활용하여 매출을 예측하고, 2단계에서는 예측된 매출 결과와 매입/매출 비율(Ratio) 변수, 시계열·환경 요인(온도, 강수량, 공휴일, 코로나 등)을 결합하여 매입량을 예측하였다.

이 과정을 통해 매출과 매입의 구조적 관계를 반영하면서도, 비정상 구간이나 변동성이 큰 시기에서도 안정적인 예측을 수행할 수 있었다. 최종적으로 본 분석은 **정확한 수요예측을 통해 공급망의 채찍효과(Bullwhip Effect)를 완화하고, 도매 물류센터의 재고관리 및 발주 의사결정 효율을 향상시키는** 것을 목표로 한다.

### 2. 분석 방법 및 절차

#### 2.1 데이터 전처리

본 연구에서 사용한 데이터는 **중소유통 물류센터 거래 데이터(매입·매출)** 와 **외부 보조 데이터**로 구성된다. 두 종류의 데이터를 각각 전처리한 후 병합하였으며, 특히 물류센터 거래 데

이터의 전처리 과정은 1차(기초 정제) 와 2차(고급 보정) 로 나누어 수행하였다.

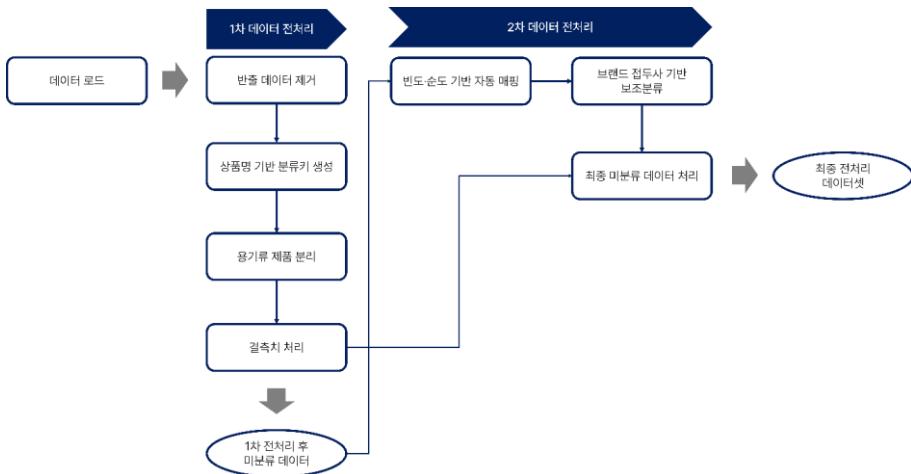


그림 1 중소유통 물류센터 거래 데이터 전처리 프로세스

### 2.1.1 중소유통 물류센터 거래 데이터

#### (1) 1단계 전처리: 기본 정제 및 분류키 생성

먼저 A물류센터와 B물류센터의 매입·매출 데이터를 불러온 후, 가장 먼저 '구분' 컬럼에서 반출행(예: 반출, 폐기, 샘플 등)을 제거하였다. 이는 반출 데이터가 실제 매입·매출 분석에 영향을 미치지 않기 때문이다.

다음으로, **상품명 기반 분류키(Classification Key)** 를 생성하였다. 이는 중분류 및 소분류 결측치를 보완하기 위한 핵심 절차로, 정규표현식을 이용하여 상품명 내 불필요한 정보를 제거하였다. 즉, 괄호 안 문자 제거 → 숫자 및 단위(ml, g, kg 등) 제거 → 특수문자 및 공백 제거 → 소문자 통일의 과정을 거친다. 이 과정을 통해 **상품명이 약간씩 달라도 동일한 분류키를 부여할 수 있다**. 예를 들어 “코카콜라 500ml PET”는 코카콜라라는 분류키로 표준화된다.

다음으로, **용기류(락앤락, 밀폐용기 등)** 상품을 별도로 구분하였다. 이는 식품류와 구분이 어려운 경우가 많기 때문에, 용기류 상품에는 분류키에 \_용기 접미어를 추가하여 별도 관리하였다.

상품명(원본)	분류기(전)	규칙 감지
락엔락 밀폐용기 2P	락엔락밀폐용기	용기류
보관통 3종 세트	보관통3종세트	용기류
젤리믹스 500g	젤리믹스	식품류
데미소다 애플 340ml	데미소다애플	식품류

상품명(원본)	분류기(후)
락엔락 밀폐용기 2P	락엔락밀폐용기_용기
보관통 3종 세트	보관통3종세트_용기
젤리믹스 500g	(그대로) 젤리믹스
데미소다 애플 340ml	(그대로) 데미소다애플

그림 2 용기류 제품 분리 예시

마지막으로, **분류 정보(대·중·소분류)** 와 **바코드(barcode)** 결측치를 보완하였다. 동일 분류 키 내에서 최빈값(mode)으로 결측치를 채우고, 대표 바코드를 동일 그룹 내 다른 행에서 가져와 보완하였다. 이로써 1단계 전처리에서는 결측치 보정이 가능한 항목을 모두 처리하고, 여전히 분류되지 않은 항목은 **2단계 전처리**로 넘긴다.

분류기	대분류(전)	중분류(전)	소분류(전)	바코드(전)
데미소다애플	음료	탄산음료	NaN	NaN
데미소다애플	음료	탄산음료	탄산음료	8801234...
락엔락밀폐용기_용기	NaN	주방잡화	보관용기	8807777...

분류기	대분류(후)	중분류(후)	소분류(후)	바코드(후)
데미소다애플	음료	탄산음료	탄산음료	8801234...
데미소다애플	(동일)	(동일)	(동일)	(동일)
락엔락밀폐용기_용기	생활/주방	주방잡화	보관용기	8807777...

그림 3 최빈값 결측치 처리 예시

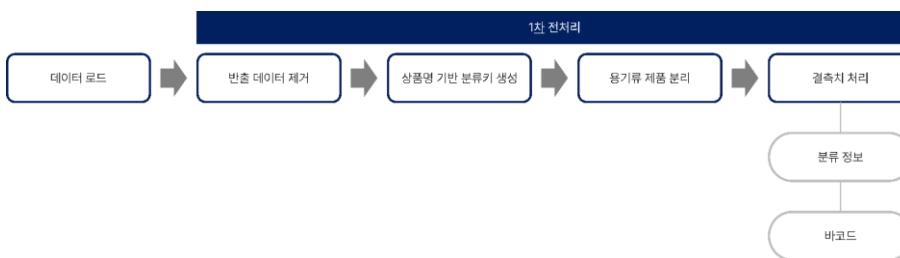


그림 4 1차 전처리 프로세스

## (2) 2차 전처리: 등장빈도 및 브랜드 기반 보정

2단계 전처리에서는 **등장빈도(frequency)** 와 **순도(purity)** 를 활용해 중분류와 소분류를 자동으로 매핑하였다. 분류기별로 라벨 등장 빈도를 계산한 후, 최빈 라벨의 비율이 0.9 이상인

경우(purity ≥ 0.9) 해당 라벨을 자동 할당하였다. 예를 들어, '코카콜라' 분류키에서 '음료' 90회, '식품' 10회 등장 시 순도는 0.9로 계산되며, '음료'가 자동으로 지정된다. 순도(purity)는 아래와 같이 구해진다.

$$purity = \frac{\text{최빈 라벨 빈도수}}{\text{전체 라벨 수}}$$

분류키	후보 라벨 분포	순도(purity)	최종 매핑
코카콜라	음료(90), 식품(10)	0.9	음료
삼다수	음료(100)	1	음료
락앤락	주방용품(85), 기타(15)	0.85	✗(자동 매핑 제외)

표 1 분류키별 순도(purity) 계산 예시

그러나 순도가 0.9 미만이거나 등장빈도(min\_support)가 기준에 미달하는 경우, **브랜드 접두사 기반 보정(Brand Prefix Refinement)** 을 적용하였다. 상품명에서 정규표현식을 이용해 브랜드명을 추출한 뒤, 브랜드-카테고리 매핑 표를 활용하여 분류를 보정한다.

브랜드	대표 분류
코카콜라	탄산음료
삼다수	생수
롯데	과자/음료
락앤락	보관용기

표 2 브랜드-대표 분류 매핑 예시

또한, 브랜드명 외에 상품명 내 단서 단어(keyword)도 함께 고려하였다.

- “콜라, 사이다, 제로, 캔” 등이 포함되면 → 음료 분류로 가중치 상승
- “용기, 보관, 밀폐” 등이 포함되면 → 주방용기 분류로 가중치 상승

이때 브랜드와 키워드가 일치하면 자동 분류가 이루어진다.

예를 들어,

- “코카 제로 355ml” → 브랜드 “코카” + 단어 “제로” → 탄산음료
- “락앤락 밀폐용기 2P” → 브랜드 “락앤락” + 단어 “용기” → 주방용기

이 과정을 통해 순수 데이터 기반으로는 분류가 어려운 상품명을 브랜드 단서로 보완하여 일관된 분류체계를 구축하였다.

상품명	브랜드	단서 단어	결과 분류
코카콜라 제로 355ml	코카	제로, 캔	탄산음료
삼다수 500ml	삼다수	물	생수
락앤락 밀폐용기	락앤락	용기	주방용기
롯데 스위트홈 비닐팩	롯데	비닐팩	생활잡화

표 3 상품명 내 단서 단어(keyword) 기반 자동 분류 결과

### (3) 미분류 데이터 처리

2단계 전처리 후에도 분류되지 않은 일부 데이터(미분류 행)는 별도로 분석 및 리포트용 데이터로 저장하였다. 이 데이터들은 추후 사전(dictionary) 확장을 위해 사용된다. 미분류 데이터는 다음 절차로 관리된다.

1. 미분류 항목을 집계하고, 순도·지지도·브랜드 검출 여부를 기록한다.
2. 반복적으로 등장하는 미분류 패턴을 CSV로 저장(unclassified\_summary.csv).
3. 브랜드명 누락 또는 오탈자 패턴을 추출하여 룩업 테이블 확장에 활용한다.
4. 노이즈 단어("기획", "증정", "세트" 등)는 별도 파일(noisy\_tokens.txt)로 관리한다.

이러한 미분류 분석 과정을 통해 전처리 파이프라인의 정확도를 지속적으로 개선할 수 있다.



그림 5 2차 전처리 프로세스

표 4 2차 전처리 프로세스

#### 2.1.2 외부 데이터

본 연구에서는 총 6개의 외부 데이터를 활용하였다.

(1) 온도, (2) 강수량, (3) 평일 여부, (4) 주말 여부, (5) 공휴일 여부, (6) 코로나19 확진자 수이다. 이러한 외부 변수들은 소비 패턴에 영향을 미치는 환경적 요인으로서, 예측 모델의 정확도를 향상시키기 위해 포함하였다.

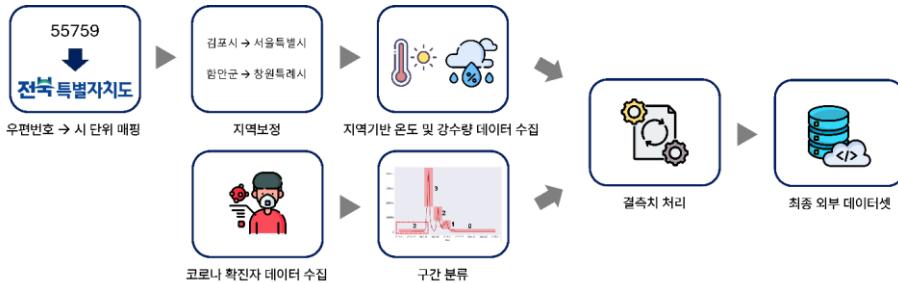


그림 6 외부 데이터 수집 및 전처리 프로세스

### (1) 온도 및 강수량 데이터

온도와 강수량 데이터는 **기상청 기상자료개방포털**에서 수집하였다.

먼저, 전처리가 완료된 중소유통 물류센터 거래 데이터에서 **우편번호를 기반으로 도로명 주소를 추출하고, 이를 시 단위로 매핑하였다.** 그러나 일부 지역의 경우 **기상관측소가 존재하지 않거나 자료가 불완전하였기 때문에, 동일 생활권 내 인접 지역의 기상 데이터를 대체값으로 활용하였다.**

예를 들어, 경기도 용인시는 수원시의 데이터, 함안군은 창원시의 데이터, 김포시와 성남시는 서울특별시의 데이터를 사용하였다.

이 밖에도 광명시는 서울특별시, 칠곡군과 경산시는 대구광역시, 논산시는 부여시, 고성군은 통영시, 창녕군은 창원시, 의령군과 사천시는 진주시, 익산시는 전주시, 평택시는 천안시, 음성군은 충주시의 데이터를 각각 적용하였다.

이러한 지역 보정은 **기상관측소 간의 거리와 지리적 유사성, 생활권 일치성을 고려하여 수행되었다.** 즉, 동일 기후대나 대권역(수도권·영남권·충청권·호남권 등) 내에서의 대체를 통해, 관측소 미보유 지역에서도 온도 및 강수량 데이터의 공간적 일관성과 연속성을 확보하였다. 또한 수집된 데이터 중 일부에는 결측치가 존재하였는데, 이는 **인접 지역의 유사한 온도 패턴 평균값을 이용한 보간(imputation)** 방식으로 보정하였다.

이 과정을 통해 기상 데이터의 결측 문제를 최소화하고, 모델 학습에 필요한 정확하고 안정적인 외부 변수를 구축할 수 있었다.

### (2) 평일·주말·공휴일 데이터

평일과 주말 여부는 파이썬 `datetime` 함수를 이용하여 자동 계산하였다. 공휴일 데이터는 국내 공휴일 정보를 정리한 GitHub 공개 자료를 활용하였으며, 연도별 공휴일을 일자 기준으로 매칭하여 각 거래일이 공휴일에 해당하는지를 표시하였다.

### (3) 코로나19 확진자 수 데이터

코로나19 확진자 데이터는 KDX(한국데이터거래소)에서 제공하는 "코로나 확진자 누적 데이터셋"을 사용하였다. 다만, 해당 데이터는 **2020년부터 2022년까지의** 누적 수치만 존재하며, 2023년 이후 데이터는 존재하지 않는다. 이에 단순히 0으로 처리할 경우, 시계열의 편차가 커지고 모델의 안정성을 해칠 우려가 있었다. 따라서 본 연구에서는 **확진자 수의 추세를 기반으로 4개의 그룹으로 구간화(categorization)** 하여 정성적 영향을 반영하였다. 특히, **2022년 2월 이전의 초기 확산기에는 실제 확진자 수가 적더라도 사회 전반의 활동 제한이 심했기 때문에, 영향도를 보정하기 위해 그룹 2(가중치 2)를 부여하였다.**

구분 기준	그룹 코드	비고
2022-02-01 이전	2	초기 팬데믹 시기, 사회적 영향이 극대화됨
17,000 명 초과	3	대규모 확산 단계
7,000 명 ~ 17,000 명	2	중간 확산 단계
2,500 명 ~ 7,000 명	1	경미한 확산 단계
기타(0~2,500 명 미만)	0	안정 단계

표 5 코로나19 확진자 수 구간화

특히, **2022년 2월 이전의 초기 확산기에는 실제 확진자 수가 적더라도 사회 전반의 활동 제한이 심했기 때문에, 영향도를 보정하기 위해 그룹 2(가중치 2)를 부여하였다.**

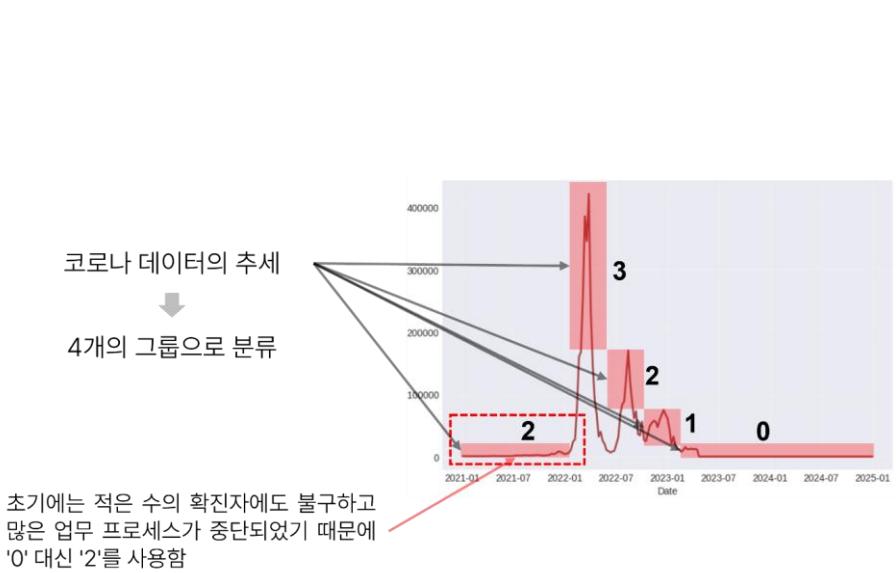


그림 7 코로나19 확진자 수 시각화

이렇게 완성된 두개의 데이터셋을 합쳐서 학습 데이터셋으로 만들었다.

### 2.1.3 일일 데이터를 주 단위 데이터로 변환

월별 매입·매출 데이터를 예측하기 위해, 기존의 일별 데이터를 주 단위로 변환하여 학습 데이터셋을 구성하였다. 일별 데이터는 일시적인 변동이나 결측이 많아 추세를 정확히 파악하기 어렵기 때문에, 이를 주 단위로 집계함으로써 데이터의 안정성과 대표성을 높일 수 있다. 또한, 주 단위 집계는 월 단위 예측을 위한 중간 수준의 시계열 단위로서, 지나치게 세밀한 일별 변동을 완화하면서도 월별 흐름을 충분히 반영할 수 있다는 장점이 있다. 따라서 이러한 이유로 주 단위 데이터로 변환하였다.

먼저, 공휴일 데이터를 활용하여 각 주차별로 해당 주에 포함된 공휴일 일수를 계산하였다. 이 값은 이후 분석 과정에서 해당 주의 영업 환경을 설명하는 변수로 사용되었다.

그다음, 매입 데이터와 매출 데이터를 각각 주 단위로 집계하였다. 매입 데이터는 한 주 동안의 총 매입 수량을 합산하였고, 외부 요인인 확진자 수는 주 평균으로 계산하였다. 매출 데이터 역시 주별 판매량을 합산하고, 온도·강수량·확진자 수는 주 평균값으로 변환하였다.

이 과정을 통해 일별 변동으로 인한 잡음을 줄이고, 주간 단위에서의 전반적인 패턴과 추세를 안정적으로 파악할 수 있도록 하였다.

다음으로, 일부 주차에서 데이터가 누락된 구간에 대해서는 결측 보정 작업을 수행하였다.

모든 데이터셋의 주차 범위를 최소~최대 구간으로 통일하고, 누락된 주차의 수량이나 매출 값은 0으로 대체하였다. 온도나 확진자 수와 같이 연속적인 변수의 경우에는 **인접 주차의 평균값을 활용한 보간(imputation)**을 적용하였다. 공휴일 정보가 없는 주는 0일로 처리하였다. 이와 같은 과정을 통해, 모든 주차가 연속적으로 포함된 완전한 형태의 주간 데이터셋이 구축되었다.

결과적으로, 일별로 산발적으로 흩어져 있던 데이터를 주 단위로 통합함으로써 **시계열적 연속성과 안정성을 확보하였으며**, 이후 예측 모델링에서 **보다 해석 가능하고 신뢰성 높은 결과를 얻을 수 있었다.**

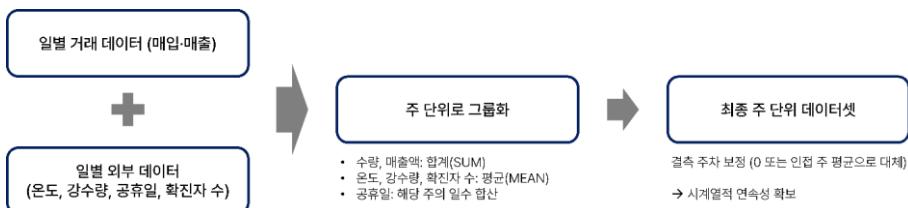


그림 8 일일 데이터를 주 단위로 변환 프로세스

#### 2.1.4 데이터 분리

모델 학습과 성능 평가의 객관성을 확보하기 위해 전체 데이터를 시간 순서(Time Series Order)에 따라 분리하였다. 2021년 1월 1일부터 2024년 5월 31일까지의 데이터를 학습 데이터(Training set)로 사용하였으며, 2024년 6월 1일부터 2024년 12월 31일까지의 데이터를 테스트 데이터(Test set)로 사용하였다.

#### 2.1.5 정확도 측정 지표

모델의 예측 성능을 평가하기 위해 **결정계수(R<sup>2</sup>)**, **평균절대백분율오차(MAPE)**, 평균제곱근 오차(RMSE)를 사용하였다. 이 세 지표를 선택한 이유는 서로 다른 관점에서 모델의 성능을 평가하면서도, 시계열 예측의 구조적 설명력·실무적 신뢰도·오차 안정성을 동시에 검증할 수 있기 때문이다.

**R<sup>2</sup> (Coefficient of Determination)**은 모델이 실제 관측값의 변동성을 얼마나 설명하는지를 나타내며, 즉 “모델의 전체 적합도(goodness of fit)”를 평가한다. 시계열 예측의 구조적 설명력(Explained Variance)을 확인하기에 적합하며, 값이 1에 가까울수록 모델의 예측이 실제 데이터와 유사함을 의미한다.

**MAPE (Mean Absolute Percentage Error)**는 예측값과 실제값의 상대적 오차를 백분율(%)로 표현하여, 스케일에 관계없이 직관적인 해석이 가능하다. 특히 물류 및 수요예측 분야에서는 정량적 오차를 실무적으로 비교하기 쉬운 지표로 널리 사용된다. 극단값(Outlier)에 다소 민감하지만, 정상 구간에서의 예측 정확도와 실무적 활용성(Practical Accuracy)을 함께 평가할 수 있다.

**RMSE (Root Mean Squared Error)**는 예측 오차의 제곱 평균의 제곱근으로, 큰 오차에 더 큰 가중치를 부여한다는 점에서 MAPE와 상호보완적이다. 이는 모델이 예측값을 얼마나 안정적으로 수렴시키는지를 나타내며, 오차의 절대 규모와 변동 폭(Volatility)을 평가하는 데 유용하다.

**MAE (Mean Absolute Error)**는 MAE는 예측값과 실제값의 절대적 차이의 평균을 나타내며, 오차의 평균적 크기(Average Magnitude of Error)를 직관적으로 보여준다. MAPE와 달리 백분율 변환을 거치지 않으므로, 데이터 단위(예: 수량, 매출액)\*\*를 그대로 유지한 상태에서 절대 오차 수준을 평가할 수 있다. 또한 RMSE에 비해 큰 오차의 영향을 덜 받기 때문에, **일반적 오차 분포에서의 대표적 지표**로 적합하다.

따라서, **R<sup>2</sup>**은 모델의 설명력(Explained Variance)을, **MAPE**는 예측의 실무적 신뢰도(Practical Accuracy)를, MAE는 예측의 평균 오차 규모를, **RMSE**는 예측의 안정성(Stability)을 각각 측정하는 보완적 지표로 활용하였다.

이 세 지표를 병행함으로써, 모델의 **통계적 정확도, 실무 적용 타당성, 절대적 신뢰 수준** 그리고 **예측 안정성을** 균형 있게 검증할 수 있도록 설계하였다.

## 2.2 EDA

모델의 정확도를 높이기 위해 학습 데이터셋을 사전에 탐색(EDA; Exploratory Data Analysis)하였다. 먼저, 일별 데이터를 주 단위로 집계한 후 **A물류센터와 B물류센터의 매입·매출 추세를 시각화하였다**. 그 결과, 매입 데이터는 비교적 일정한 추세를 보이는 반면, 매출 데이터는 두 센터 모두에서 일정한 패턴이 관찰되지 않았다. 즉, 매출의 변동성이 크고 시계열적으로 불규칙한 양상을 나타냈다. 특히, **A물류센터의 매출 데이터에서는 2023-01-13부터 2023-12-31까지의 기간 동안 매입 수량이 0으로 집계되었으며, B물류센터의 경우 2021-01-01부터 2023-04-30까지는 불규칙한 매입 추세를 보이고, 2023-08-01부터 2023-12-31까지는 매입 수량이 0으로 나타났다**. 이러한 구간은 실제 영업 활동이 중단되었거나 데이터가 누락된 구간으로 판단되어, 모델 학습에 포함할 경우 오차가 커질 가능성이 있다고 보았다. 따라서 본 연구에서는 **비정상 구간을 제외하고 정상적인 거래가 이루어진 구간**

만을 분석 대상으로 설정하였다. 이에 따라 아래와 같이 최종 분석 구간을 선정하였다,

- A물류센터는 2023-02-12 이전 구간과 2024-01-01 이후 구간,
- B물류센터는 2023-05-01~2023-08-01 구간과 2024-01-01 이후 구간

이와 같은 데이터 탐색 과정을 통해 비정상 시계열 구간을 사전에 제거함으로써, 모델이 정상적인 거래 패턴을 학습하고 예측의 신뢰도를 높일 수 있도록 하였다.



그림 9 A·B 물류센터 매입·데이터 추세

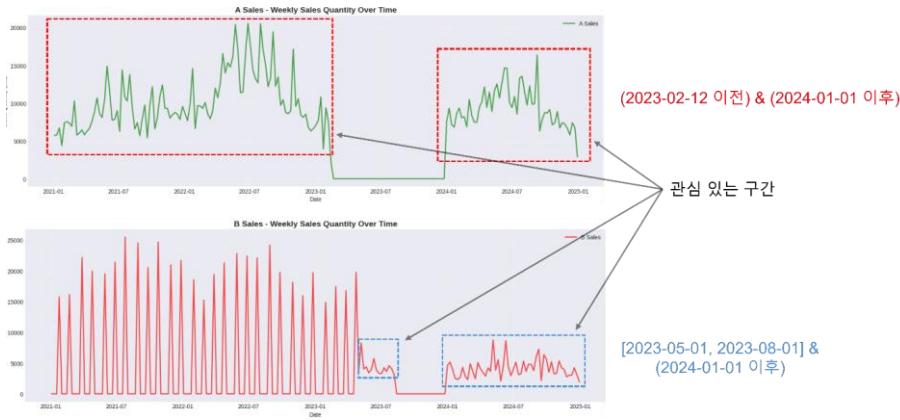


그림 10 A·B 물류센터 매출·데이터 추세

### 2.3 첫번째 예측 모델 실험 (XGBoost)

본 단계에서는 XGBoost 모델을 활용하여 매입·매출 데이터를 예측하였다. 예측의 정확도를 높이기 위해, 단순한 날짜·요일 정보 외에도 시계열적 특성(lag, rolling mean/std)과 환경적 요인(온도, 강수량, 코로나 확진자 수, 공휴일)을 포함한 다양한 파생 변수를 생성하였다.

변수명	설명	비고
온도	하루 평균 기온	일평균 섭씨 온도
강수량	하루 동안의 강수량(mm)	일별 강수량
확진자 수	해당 날짜의 코로나 19 확진자 수	일일 확진자 수
주말	토요일 또는 일요일이면 1, 평일이면 0	주말 여부
day_of_week	요일 (0 = 월요일, ..., 6 = 일요일)	숫자로 표현된 요일
day_of_month	한 달 중 날짜 (1~31)	일자 정보
month	월 (1~12)	월 번호
quarter	분기 (1~4)	연간 분기 구분
day_of_week_sin / day_of_week_cos	요일을 주기적으로 표현하여 주간 패턴을 부드럽게 반영	요일의 순환적 특성 인코딩
month_sin / month_cos	월을 주기적으로 표현하여 계절 변동을 반영	월의 순환적 특성 인코딩
수량_lag1 ~ 수량_lag5	이전 1~5 일의 판매 수량	단기 추세를 반영하는 지연 변수

<b>수량_rolling_mean_3 /</b>	최근 3 일 또는 5 일간의 평균 판매량	이동평균 변수
<b>수량_rolling_mean_5</b>	(현재일 제외)	
<b>수량_rolling_std_3</b>	최근 3 일간 판매량의 표준편차	단기 변동성 측정
<b>공휴일_lead1 ~ 공휴일_lead5</b>	향후 1~5 일 이내에 공휴일이 있는지 여부	수요 변화를 예측하기 위한 선행 변수

표 6 XGBoost 사용 변수

이러한 변수들은 시계열의 흐름과 환경적 요인을 동시에 반영하도록 설계되었다. 특히 **요일(day\_of\_week\_sin, cos)** 과 **월(month\_sin, cos)** 을 사인·코사인 함수로 순환 인코딩하여, 요일과 계절 주기의 연속성을 모델이 자연스럽게 학습할 수 있도록 하였다. 단순한 숫자형 인코딩은 '일요일(6)'과 '월요일(0)'이 멀리 떨어져 있는 것으로 인식되어 주기성이 끊기지만, 순환 인코딩을 적용하면 이러한 단절을 해소할 수 있다. 또한 **lag 변수(수량\_lag1~5)** 는 직전 며칠간의 판매 추세를 반영하여 시간의 흐름 속 수요 연속성을 학습하도록 하였고, **rolling mean·std 변수**는 일시적 급등·급락의 노이즈를 완화하고 단기적인 수요 수준과 변동성을 파악하는 데 활용되었다. 마지막으로 **공휴일\_lead 변수(lead1~5)** 는 향후 공휴일이 있는 시점을 미리 인식하게 함으로써, 휴일 전후의 매출 급변 패턴을 반영하였다.

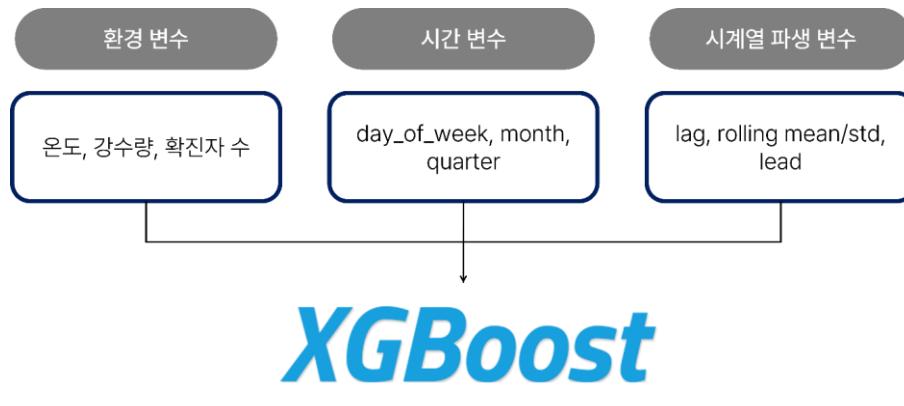


그림 11 XGBoost model 구조

### (1) 예측 결과

그림 12~19는 XGBoost 모델로 예측한 A/B 물류센터의 매입·매출 시계열 예측 결과를 나타

낸다. 매입 데이터(A/B)의 경우, 전반적인 평균 추세는 포착했으나, 급격한 변동 구간(피크 값)에서는 예측 오차가 크게 발생하였다. 이는 매입 데이터가 이벤트성 요인(프로모션, 단가 변동 등)에 크게 영향을 받기 때문으로 판단된다. 매출 데이터(A/B)는 비교적 실제값과 유사한 패턴을 보이며 예측 정확도가 높게 나타났다. 특히, 주기적 변동이나 단기 추세를 모델이 일정 부분 학습한 것으로 확인되었다.

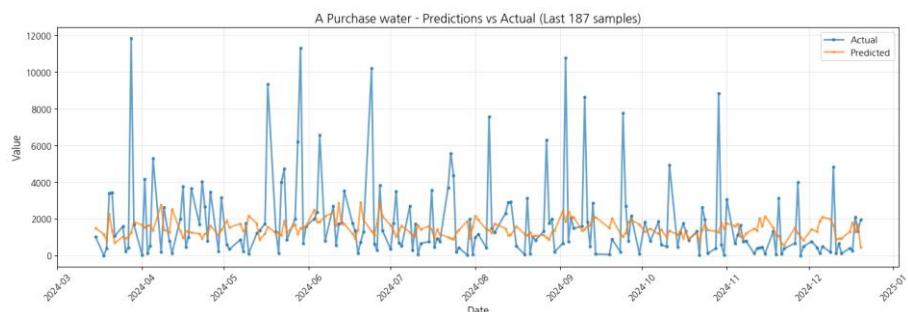


그림 12 A 물류센터 (중분류) 생수, 음료 건강 매입 예측 vs 실제

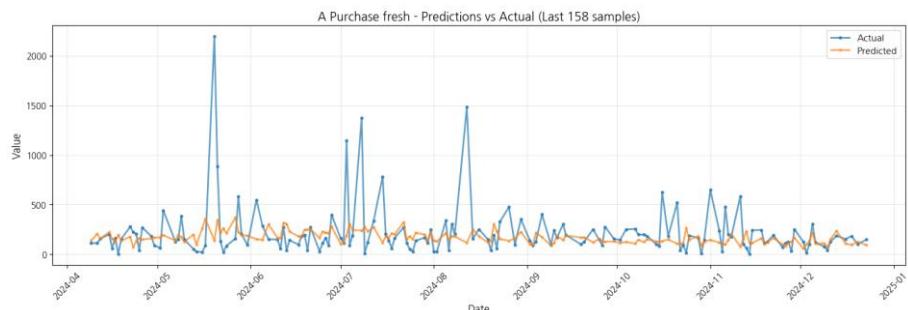


그림 13 A 물류센터 (중분류) 신선식품 매입 예측 vs 실제

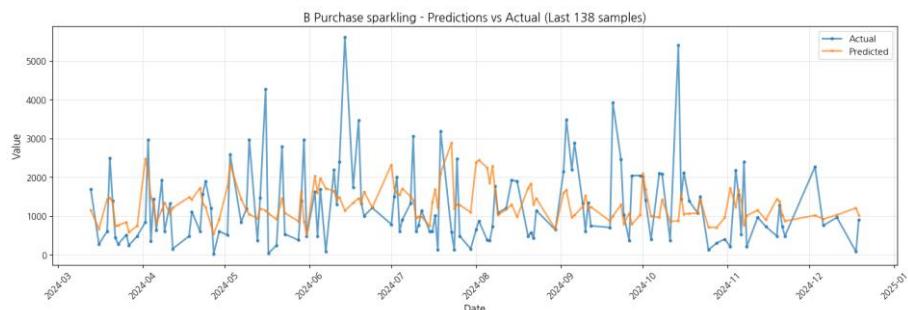


그림 14 B 물류센터 (소분류) 탄산음료 매입 예측 vs 실제

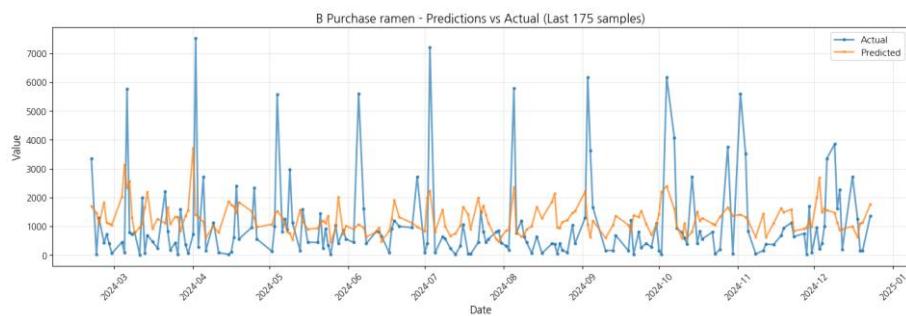


그림 15 B 물류센터 (소분류) 봉지라면 매입 예측 vs 실제

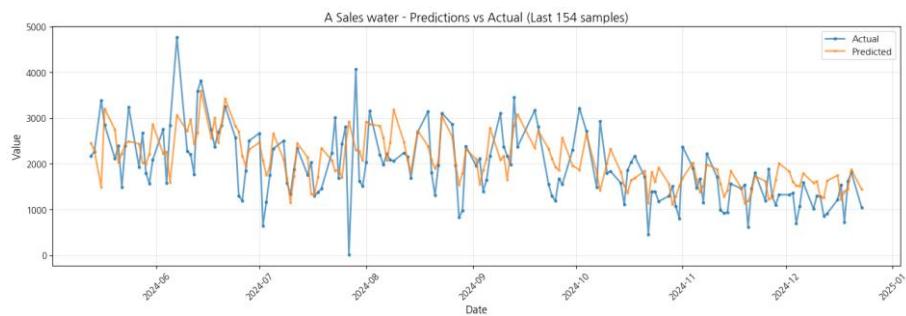


그림 16 A 물류센터 (중분류) 생수, 음료 건강 매출 예측 vs 실제

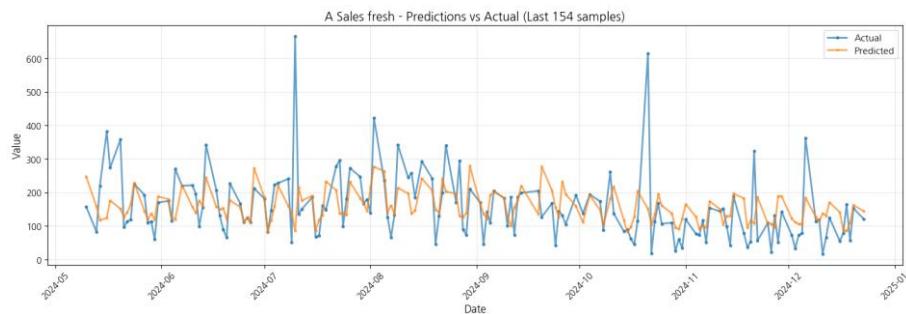


그림 17 A 물류센터 (중분류) 신선식품 매출 예측 vs 실제

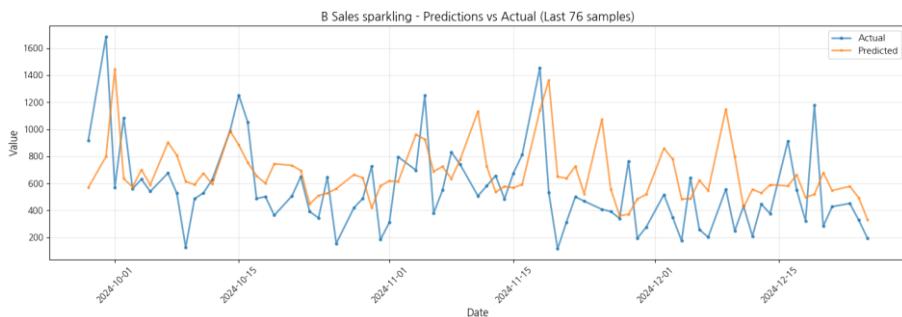


그림 18 B 물류센터 (소분류) 탄산음료 매출 예측 vs 실제

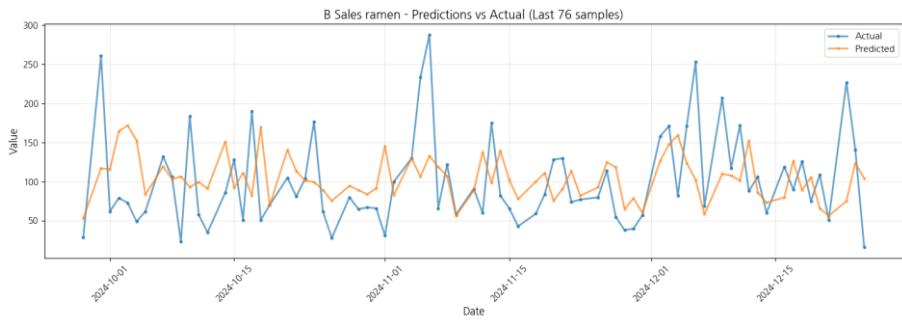


그림 19 B 물류센터 (소분류) 봉지라면 매출 예측 vs 실제

그림에서 보이다 싶이 매입 데이터의 경우 평균 추세에서 벗어나는 큰 값에서 대해서는 거의 예측하지 못하는 경향을 보인다. 반면 매출 데이터의 경우 비교적 정확하게 예측을 하지만 더 정확도를 높이기 위해 중요 변수만 추출하여 사용하고자 하였다. 그러기 위해서 첫번째 모델에서 예측에 있어 가장 중요한 역할을 하였던 상위 10개의 변수를 선택을 하였다.

## 2) 변수 중요도 분석

그림 20은 XGBoost 모델을 활용하여 A·B 물류센터의 주요 품목별 매입 및 매출 예측 시 변수 중요도(Feature Importance)를 분석한 결과를 보여준다. 각 그래프는 품목군(생수, 신선식품, 라면, 탄산음료)별로 상위 20개 변수를 제시하였으며, 모델이 어떤 요인을 중심으로 학습했는지를 시각적으로 확인할 수 있다.

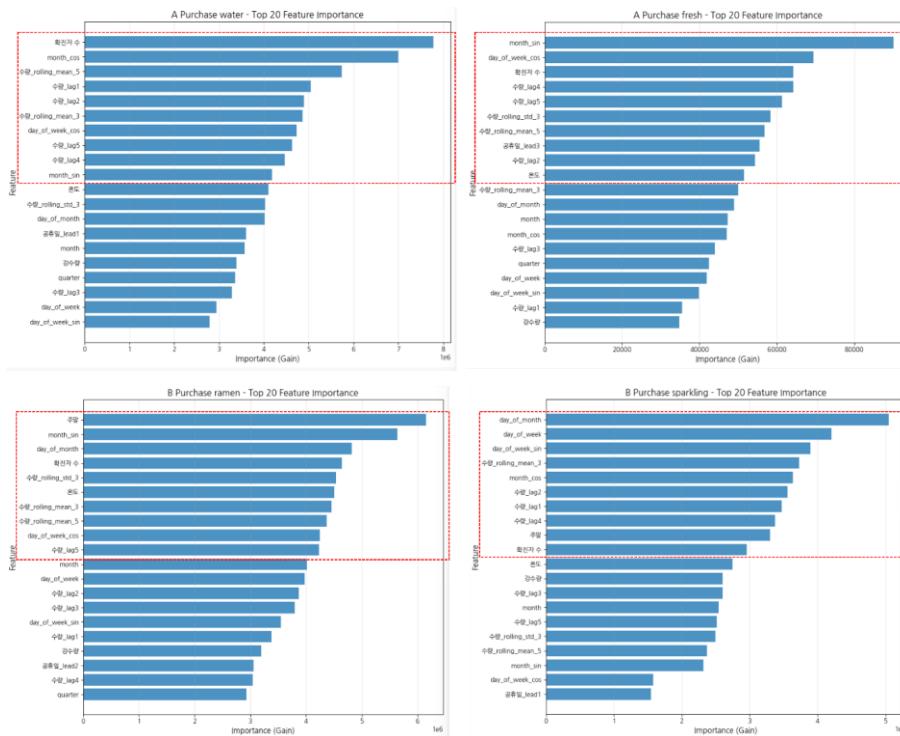
먼저 매입(Purchase) 예측 결과를 보면, A센터(상단 두 그래프)에서는 확진자 수, month\_cos, month\_sin, 수량\_lag1~5, 수량\_rolling\_mean\_3:5 등이 상위권에 위치하였다. 이는 매입이 코로나19 영향 및 월별·계절적 특성과 밀접하게 연관되어 있으며, 단기적인 시계열 패턴(lag·rolling 평균) 또한 주요 예측 요인으로 작용함을 의미한다. B센터의 매입 예측

에서도 확진자 수, month\_sin, day\_of\_week\_cos 등이 높은 중요도를 보여, **시계열 주기성과 외생 요인(공휴일, 계절 변화)** 이 매입량 변동을 설명하는 핵심 요인임을 확인하였다.

반면 **매출(Sales)** 예측에서는 전반적으로 day\_of\_week\_sin, day\_of\_week\_cos, 판매량\_rolling\_mean\_3, 판매량\_rolling\_mean\_5, month\_sin, month\_cos 등이 상위에 분포하였다. 이는 매출이 **요일·주기적 패턴과 단기 수요 흐름(이동평균)**에 강하게 영향을 받는다는 것을 보여준다. 즉, 소비자의 구매 주기와 계절성 요인이 매출 예측 정확도에 결정적인 역할을 하는 것으로 나타났다. 종합적으로 볼 때,

- **매입 예측:** 외생적 요인(코로나, 월별 특성, 공휴일 등)에 민감
- **매출 예측:** 내생적 요인(주기적 소비 패턴, 단기 추세)에 의존

이러한 결과를 바탕으로, 다음 단계에서는 각 품목별 분석에서 **상위 10개 중요 변수만을 선별하여 재학습을 수행하고, 불필요한 변수로 인한 모델 복잡도를 줄이는 방향으로 성능을 개선하고자 하였다.**



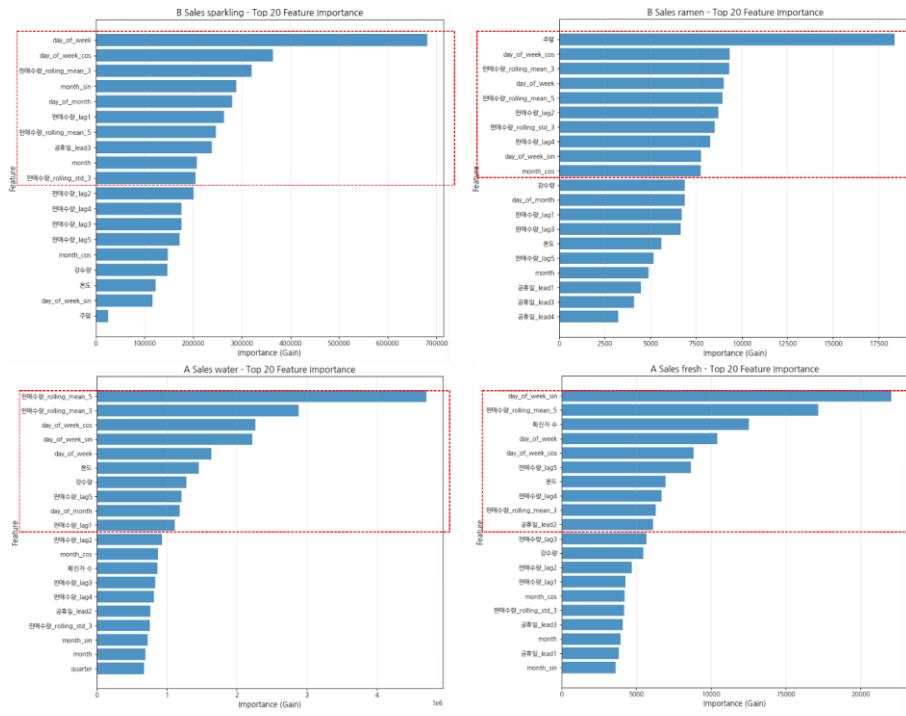


그림 20 A·B 물류센터 매입·매출의 변수 중요도 상위 10개

### (3) 한계 및 개선 방향

첫 번째 모델은 전반적인 추세를 잘 포착했지만, 매입 데이터의 급격한 변동을 제대로 설명하지 못하는 한계가 있었다. 특히, 평균 수준 근처에서는 안정적으로 예측하지만, 피크 구간에서 과소추정하는 경향이 뚜렷했다. 이에 따라 다음 단계에서는

- ① XGBoost의 단일 모델 한계를 보완하고,
- ② 데이터 내 비선형성과 이벤트 영향을 반영하기 위해  
랜덤 포레스트(Random Forest)를 결합한 양상블(Ensemble) 모델을 구성하였다

## 2.4 두 번째 예측 모델 실험 (Ensemble model ← XGBoost + Random Forest)

첫 번째 예측 모델(XGBoost 단일 모델)은 전체적인 추세를 잘 포착했지만, 매입 데이터의 변동 폭이 크고 이벤트성 요인의 영향이 커 급격한 피크값 예측에는 한계를 보였다. 이를 개선하기 위해 두 번째 단계에서는 **XGBoost와 Random Forest**를 결합한 양상블 모델을 적용하였다. 양상블 모델은 서로 다른 구조의 예측기를 병렬로 학습시켜, 각 모델의 약점을 보완하

고 예측 안정성을 높이는 방식으로 설계되었다. 특히, **XGBoost**의 **비선형 패턴 학습 능력**과 **Random Forest**의 **이상치 완화 효과**를 동시에 활용하여 매입·매출의 주간 변동성을 함께 반영하도록 하였다.

### (1) 모델 구성 및 학습 방식

두 번째 모델에서는 일 단위 예측 후 월 단위 집계 방식을 적용하였다. 이전 모델이 주 단위로 학습된 반면, 이번에는 더 세밀한 일별 단위 예측을 수행한 뒤 월별로 집계함으로써 월간 패턴의 누적 효과를 정밀하게 반영할 수 있었다. 또한 데이터의 수를 증강할 수 있어 충분한 학습을 할 수 있다고 생각하였다. 학습 과정에서 각 모델의 가중치를 최적화하여 예측 결과가 XGBoost 또는 Random Forest 중 한쪽에 치우치지 않도록 조정하였다. 이 방식은 랜덤성이 높은 매입 데이터의 노이즈를 완화하고, 매출 데이터의 주기적 흐름을 안정적으로 반영하는 데 기여하였다.

### (2) 예측 결과 분석

그림 21, 22은 양상블 모델을 이용한 A·B 물류센터의 월별 매입·매출 예측 결과를 보여준다.

메모 포함[준1]: 수정

**매입 데이터(A·B)**의 경우 두 센터 모두에서 실제값 대비 예측값의 차이가 여전히 크게 나타났다. 이는 매입이 특정 시점의 **프로모션**, **계약 변경**, **단가 조정** 등 **비정형 이벤트** 요인에 크게 의존하기 때문이다. 따라서 단순한 시계열·머신러닝 접근만으로는 충분한 예측 정확도를 확보하기 어려우며, 비즈니스 규칙 기반의 보정 로직(예: 발주 정책, 거래처 이벤트 캘린더 등)을 병행할 필요가 있다.

**매출 데이터(A·B)**의 경우 반면, 매출 데이터에서는 실제값과 예측값이 거의 일치하는 양상을 보였다. 특히 A물류센터의 경우 계절·월별 수요 변화에 대응하는 예측 정확도가 높았으며, 이는 양상블 모델이 단기적 트렌드와 주기성을 효과적으로 학습했음을 시사한다. 결과적으로 **매출 데이터에 대해서는 매우 우수한 예측 성능을 확인할 수 있었다.**

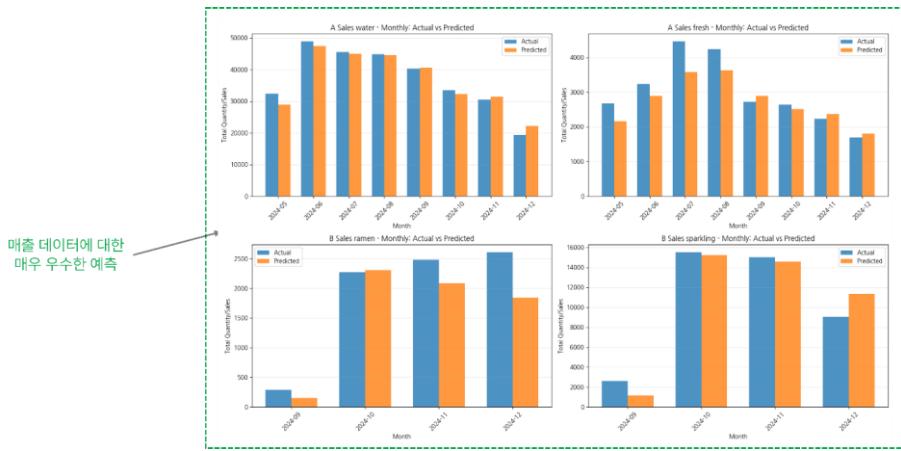


그림 21 A.B 물류센터 매입 월별 실제값과 예측값 비교(XGBoost + Random Forest → 양상블)

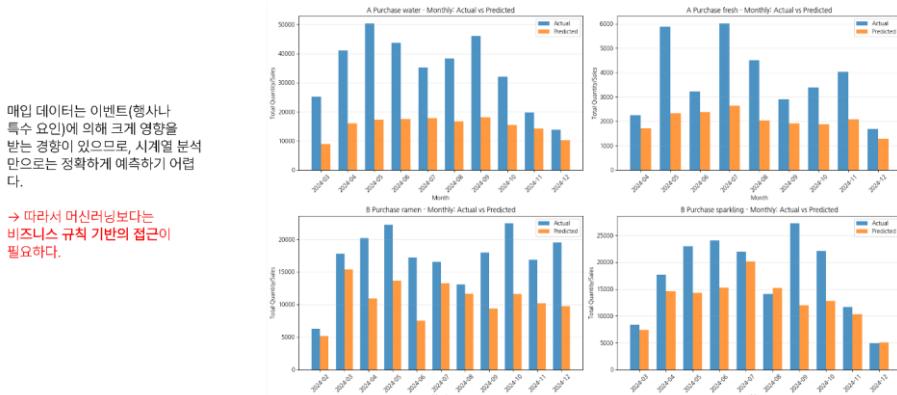


그림 22 A.B 물류센터 매출 월별 실제값과 예측값 비교(XGBoost + Random Forest → 양상블)

Dataset	Level	RMSE	MAE	R <sup>2</sup>	MAPE	%
A Purchase water	Daily	2475.69	1422.82	-0.2131	334.69	%
	Monthly	21305.12	19331.87	-2.5946	52.45	%
A Purchase fresh	Daily	287.94	136.48	-0.1547	143.70	%
	Monthly	2066.74	1738.37	-1.1335	41.41	%
B Purchase ramen	Daily	1520.91	847.09	-0.0893	349.86	%
	Monthly	7422.86	6516.52	-1.9499	35.33	%
B Purchase sparkling	Daily	1090.99	751.00	-0.1094	129.23	%
	Monthly	6995.78	5051.08	0.0140	23.13	%
A Sales water	Daily	615.79	453.27	0.3536	176.64	%
	Monthly	1785.03	1384.37	0.9621	4.72	%
A Sales fresh	Daily	90.52	56.62	0.1674	50.61	%
	Monthly	451.57	365.18	0.7442	11.07	%
B Sales ramen	Daily	59.12	42.43	0.0024	49.08	%
	Monthly	435.55	332.58	0.7875	23.60	%
B Sales sparkling	Daily	286.42	223.97	0.1238	51.44	%
	Monthly	1365.59	1104.41	0.9323	21.12	%

그림 23 성능 지표

### (3) 시사점 및 향후 방향

본 모델은 XGBoost의 비선형 학습 능력과 Random Forest의 평균 안정화 효과를 결합함으로써 예측의 전반적인 안정성과 해석력을 향상시켰다. 그러나 매입 데이터의 급격한 변동을 완전히 설명하기에는 한계가 남아 있으며, 이는 데이터 기반 접근만으로 설명할 수 없는 외생적 요인(행사, 계약 주기 등)에 기인한다.

따라서 향후에는 비즈니스 규칙 기반 예측 로직(Business Rule-based Forecasting)과 머신러닝 모델의 결합형 하이브리드 접근을 병행하는 것이 필요하다.

### 2.5 세번째 예측 모델 실험 (비율 기반 예측 모델)

앞선 두 모델에서는 매입 데이터를 직접 예측하는 방식을 적용하였으나, 매입 데이터는 외부 요인과 이벤트성 변동에 크게 의존하여 예측 정확도가 낮게 나타나는 한계가 있었다. 이

에 본 연구는 “매입은 일반적으로 미래 매출을 기반으로 결정된다”는 비즈니스적 관계에 주목하였다. 즉, **매출과 매입 간의 상관관계**를 정량적으로 분석하였다.

그림 19의 선 그래프에서 볼 수 있듯이, **빨간색은 매출, 파란색은 매입** 데이터를 나타내며, **A·B 물류센터 모두 유사한 추이**를 보인다. 이는 두 변수 간 일정 수준의 연관성이 존재함을 의미한다. 또한 산점도와 회귀선을 통해 매출이 증가할수록 매입도 증가하는 경향을 확인하였는데, **A센터의 결정계수(R<sup>2</sup>)는 0.05, B센터는 0.12**로 나타났다. 완벽한 선형관계는 아니지만, **매입이 매출에 영향을 받는 경향이 통계적으로 확인되었다**.

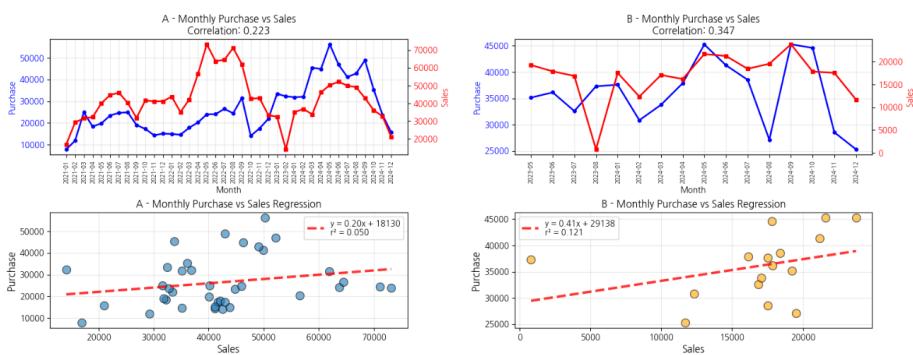


그림2424 매출과 매입 간의 관계

이러한 상관관계를 모델에 반영하기 위해 **비율(Ratio)** 개념을 도입하였다. 비율은 다음과 같이 정의된다.

$$Ratio(t) = \frac{Purchase(t)}{Sales(t)}$$

분석 결과, **A물류센터의 평균 비율은 0.678**, 즉 매출의 약 68% 수준으로 매입이 이루어진다는 것을 의미한다. 반면, **B물류센터의 평균 비율은 4.812**, 즉 매출의 약 4.8배 수준으로 매입하는 것으로 나타났다. 이러한 차이는 센터별 재고 정책 및 공급 주기의 구조적 특성을 반영한다. 또한 그림 20의 비율 추세를 보면, **A센터는 완만한 변동을 보이는 반면, B센터는 특정 월(8월)에 급등하는 패턴이 나타났다**. 즉, 매입·매출 비율에도 계절성(seasonality)이 존재하며, 이는 성수기·비성수기 등의 주기적 요인과 관련된 것으로 해석된다.

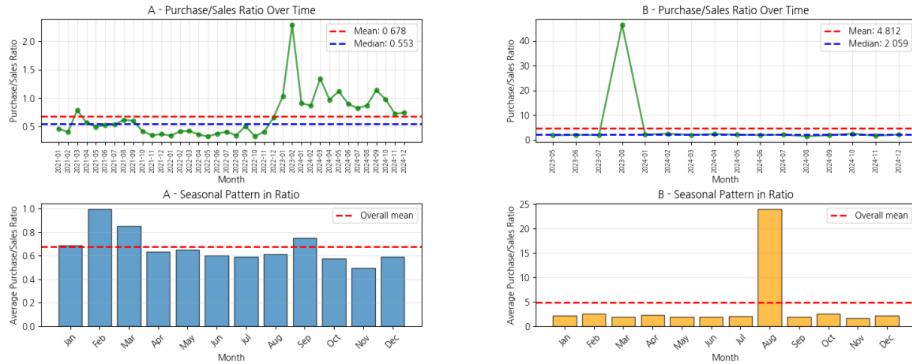


그림255 A·B 물류센터의 매입/매출 비율(Ratio) 추세 및 계절적 패턴

이러한 결과를 바탕으로 다음의 비율 기반 예측식을 제안한다

$$Purchase_{pred}(t) = Sales_{pred}(t) \times Ratio(t)$$

즉, 과거 매입 데이터만으로 예측하는 대신, 미래 매출의 예측값(Sales\_pred)\*\*과 \*\*과거로부터 학습된 매입/매출 비율(Ratio)을 곱하여 미래 매입을 간접적으로 추정한다. 이 접근법은 기존 머신러닝 모델이 설명하기 어려웠던 비정상적 매입 변동 구간에서도 안정적인 예측 성능을 보여줄 것으로 예상한다.

## 2.6 최종 Hybrid 모델 (Ratio + Machine Learning 결합형)

비율 기반 모델은 매출-매입 간 관계를 반영하는 데 효과적이었지만, 매출 급등·급락 구간에서는 비율 변화 자체가 불안정해지는 한계가 존재하였다. 이를 보완하기 위해 우리는 **2단계 (2-Stage) Hybrid 예측 모델**을 설계하였다.

### (1) 입력 변수 구성

1단계에서는 매출 예측 결과를 바탕으로 다양한 시계열 파생 변수를 생성하고,

2단계에서는 비율 특성(Ratio features)과 머신러닝 결과를 결합하여 매입을 예측하였다.

Hybrid 모델은 아래와 같이 **매출 예측 결과 기반 파생 변수와 비율 관련 변수**를 포함하여 설계되었다.

변수명	설명	비고
<code>sales_pred_current</code>	해당 날짜의 예측 판매량	당일 예측값
<code>sales_pred_lag1</code>	1 일 전의 예측 판매량	전일 기준 예측값
<code>sales_pred_lag3</code>	3 일 전의 예측 판매량	3 일 전 예측값
<code>sales_pred_lag7</code>	7 일 전의 예측 판매량	1 주일 전 예측값
<code>sales_pred_lag14</code>	14 일 전의 예측 판매량	2 주일 전 예측값

<b>sales_pred_lag21</b>	21 일 전의 예측 판매량	3 주일 전 예측값
<b>sales_pred_lag30</b>	30 일 전의 예측 판매량	1 개월 전 예측값
<b>sales_pred_ma7</b>	최근 7 일간 예측 판매량의 이동평균	단기 추세 반영
<b>sales_pred_ma14</b>	최근 14 일간 예측 판매량의 이동평균	중기 추세 반영
<b>sales_pred_ma30</b>	최근 30 일간 예측 판매량의 이동평균	장기 추세 반영
<b>sales_pred_trend7</b>	최근 7 일간 예측 판매량의 증감률(%)	단기 추세 변화율
<b>sales_pred_trend30</b>	최근 30 일간 예측 판매량의 증감률(%)	장기 추세 변화율
<b>hist_purchase_sales_ratio</b>	과거 매입 대비 판매 비율	재고 회전율 또는 수급 효율성 판단 지표
<b>hist_ratio_ma7</b>	최근 7 일간 매입/판매 비율의 이동평균	단기 비율 안정성 분석
<b>hist_ratio_ma30</b>	최근 30 일간 매입/판매 비율의 이동평균	장기 비율 추세 분석
<b>Month</b>	월 (1~12)	시계열 주기성 반영
<b>day_of_week</b>	요일 (0=월요일 ... 6=일요일)	주중·주말 패턴 분석용
<b>is_weekend</b>	주말 여부 (토·일요일이면 1)	주말 효과 반영
<b>month_sin / month_cos</b>	월(month)을 사인·코사인으로 변환한 값	계절적 패턴(주기성) 반영

표 7 입력 변수

모델 구조는 **XGBoost + Random Forest** 양상들을 기본으로 하고, 이 위에 비율 기반 보정 레이어(Ratio Adjustment Layer)를 결합하였다.

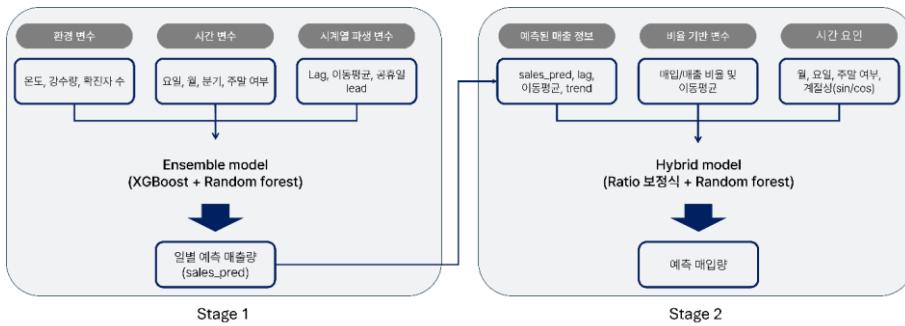


그림 2626 2-Stage Hybrid 모델 구조

## (2) 결과 및 성능 개선

그림 9는 Hybrid 모델의 예측 결과를 나타낸다. 단순 머신러닝 모델 대비 월별 예측 정확도

가 뚜렷하게 향상되었으며, 특히 B 물류센터는  $R^2 = 0.9669$ 로 매우 높은 설명력을 보였다.  
 A 물류센터 역시  $R^2 = 0.4023 \rightarrow 0.7949$ 로 개선되어, 모델의 안정성과 일반화 성능이 크게 향상되었음을 확인하였다.

이는 단순히 매입 데이터를 직접 예측하는 것보다, 매출 예측값과 매입/매출 비율을 함께 반영하는 것이 비즈니스 로직 측면에서 훨씬 현실적이고 효율적임을 시사한다

### (3) 시사점

결과적으로, 본 Hybrid 모델은

- ① 매출-매입 간 비즈니스 관계를 반영하고,
- ② 시계열적 패턴과 비율 안정성을 동시에 고려함으로써

기존 머신러닝 접근보다 훨씬 높은 예측 정확도와 해석력을 확보하였다.

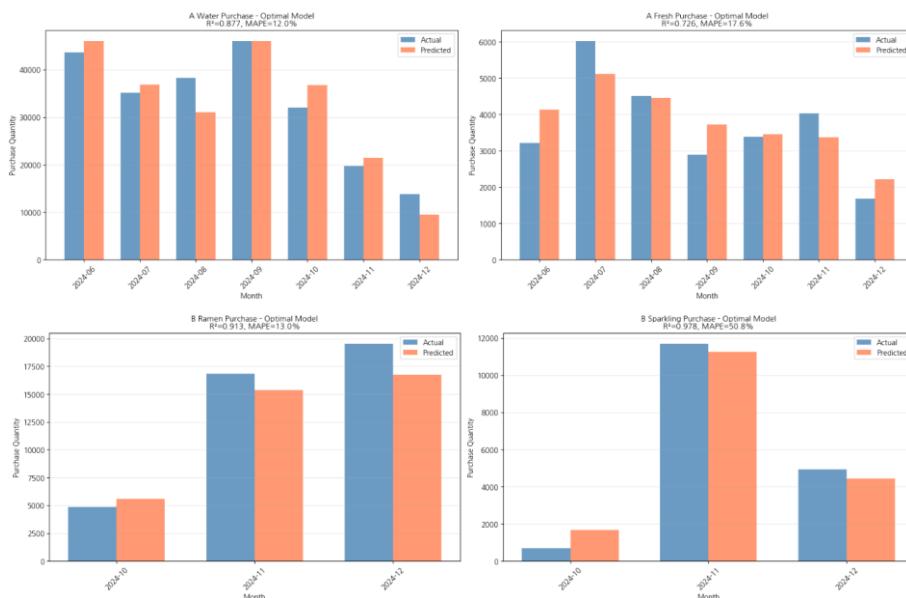


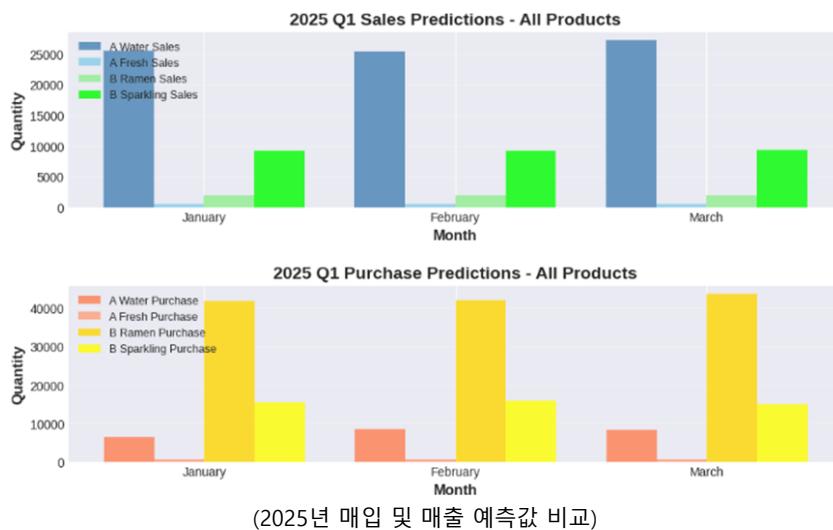
그림 2727 A-B 물류센터 Hybrid 모델의 실제값 vs 예측값 비교

### (4) 최종 모델 정확도

Model	RMSE	MAE	R2	MAPE
A물류 (생수) 매입	3876.70	3157.60	0.8775	11.98
A물류 (신선) 매입	662.15	565.98	0.7261	17.61
A물류 (생수) 매출	1785.03	1384.37	0.9621	4.72
A물류 (신선) 매출	451.57	365.18	0.7442	11.07
B물류 (탄산) 매입	675.81	631.29	0.9777	50.76
B물류 (라면) 매입	1887.26	1688.76	0.9127	13.04
B물류 (탄산) 매출	1365.59	1104.41	0.9323	21.12
B물류 (라면) 매출	435.55	332.58	0.7875	23.60

### 3. 분석결과

#### 3.1 월별 예측



A물류센터의 생수·음료·건강 분류는 매출이 1월 25,469에서 3월 27,240으로 완만하게 증가했고, 매입도 6,302 → 8,293으로 늘었다. B물류센터의 탄산음료는 매입이 1월 15,501에서 2월 15,852로 소폭 상승한 뒤 3월 14,897로 하락했다. 매출은 1월 9,147에서 3월 9,270으로 큰 변동이 없었다. 봉지라면은 매입이 1월 41,799에서 3월 43,713으로 꾸준히 증가했으며, 매출도 1,892 → 1,979로 소폭 상승했다.

#### 3.2 과거 대비 매입 및 구매 대비 판매 비율

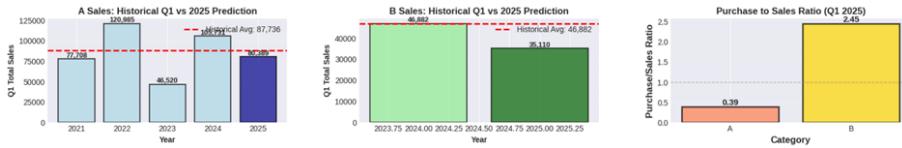


그림 28 2025년 예측값과 과거 데이터와의 비교 및 매입/매출 비율 비교

A물류센터의 매입은 과거 평균(87,736) 대비 8.4% 하락하였으며, B물류센터 또한 과거 평균(46,882) 대비 25.1% 하락하였음을 보여준다. 매입/매출 비율은 A물류센터는 0.39로 매입이 부족, B물류센터는 2.45로 매입이 과다하다. 이는 각각 품질 리스크, 재고 리스크를 불러올 수 있다. 공급망 전반에서 재고 부족과 재고 과잉의 양극화 현상이 확인되며 B물류센터는 매입 전략 재조정과 재고 회전율 관리가 필요하다.

### 3.3 일별 매출 패턴 및 월간 성장률



그림 29 시계열 패턴 및 2025년 1~3월 매입 및 매출 변화

A물류센터 일별 매출은 1,000~1,600 사이의 주별 반복 형태인 피크가 발생한다. 매입과 매출 모두 월간 성장률이 상승세를 보이며 안정적인 패턴을 유지한다. B물류센터에서는 200~800 사이의 불규칙한 스파이크가 존재한다. 이 기간의 일시적 매출 급증은 프로모션, 행사 등의 외부요인의 영향을 받은 것으로 추정된다. 월간 매출 성장률은 2~3월 구간에서 하락을 보였지만 반대로 매입 성장률은 상승세를 보여 매출 대비 매입이 증가하는 비효율적 구조를 보였다.

### 3.4 월별 매입-매출 비교 및 누적량



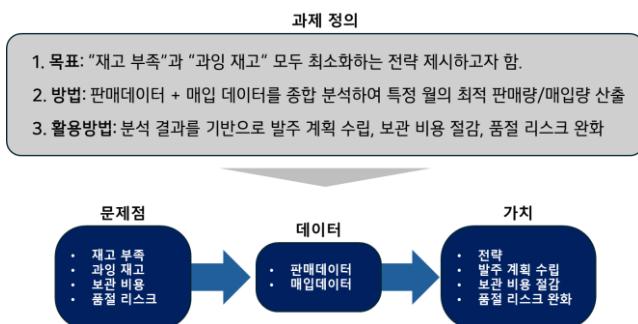
그림 30 2025년 월별 매입-매출 비교 및 누적량 그라프

각 물류센터별 매입 및 매출 변동은 있지만 전체적으로 A물류센터는 매출이 매입보다 큰 매입 부족을 보이며, B물류센터는 매출보다 매입이 큰 매입 과다를 보인다. 누적량을 보면 매입 및 매출 모두 상승하고 있지만, 각 물류센터 매입 및 매출의 폭이 증함을 보아 A물류센터는 **매입 부족**, B물류센터는 **매입 과다의 구조가** 지속적으로 강화되고 있음을 확인하였다.

#### 4. 활용방안

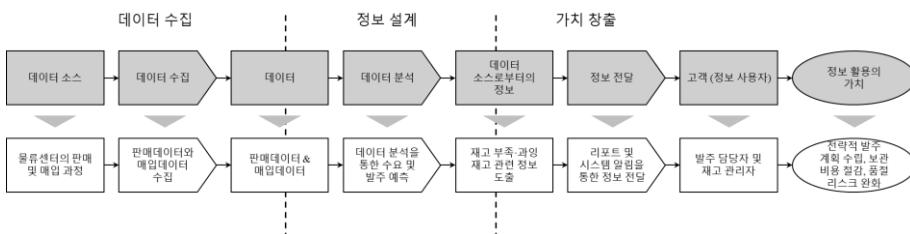
##### 4.1. 예측 기반 재고·조달 운영 활용방안

활용방안으로 서비스를 제안하고자 한다. 본 서비스는 과제의 핵심 요구("재고 부족"과 "과잉 재고"의 동시 최소화, 판매데이터·매입데이터의 종합 분석, 분석 결과의 실무적 활용)에 정합하도록 설계한다 (Fig. 1). 먼저 과제 문구를 해석하여 문제(재고 부족, 과잉 재고, 보관 비용, 품질 리스크)와 데이터(판매·매입), 그리고 창출해야 할 가치(전략적 발주 계획 수립, 보관 비용 절감, 품질 리스크 완화)를 명확히 규정한다. 이 규정 작업은 이후 전 과정의 설계 기준선으로 작동한다.



(과제 정의 및 분석 프레임워크)

이 기준선을 바탕으로 데이터 기반 가치 창출 프레임워크를 적용한다. 흐름은 Fig. 2와 같다.



(데이터 기반 가치 창출 활용방안 프레임워크)

구현 측면에서는 월·주·일 단위로 데이터를 정리하여 과거와 미래 구간을 시각적으로 구분하고, 미래 예측은 신뢰구간(불확실성 벤드)과 함께 표현한다. 정보 전달은 한 화면에서 경영·운영 의사결정을 바로 연결할 수 있도록 설계한다. 예컨대 발주 승인 흐름(권고→검토→확정), 위험 알림(임계치 초과 시 배지 또는 이메일), 주·월 단위 KPI 카드(절감액, 품절 경보 수, 대응 리드타임 등)를 일관된 스타일로 제공한다.

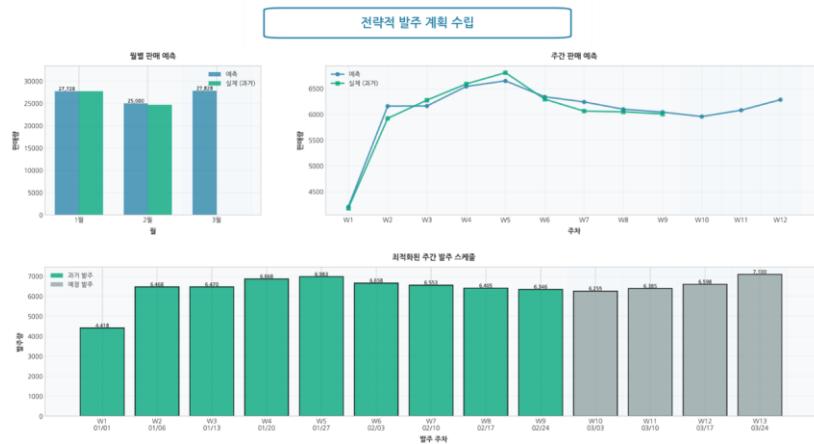
아래는 우리가 도출한 세 가지 핵심 가치(전략적 발주 계획 수립, 보관 비용 절감, 품절 리스크 완화)를 토대로 설계한 스마트 물류 서비스 콘셉트의 설명이다. 본 콘셉트는

Python(numpy, pandas, matplotlib)만으로 전면 구현하였으며, 전체 구성은 Fig.3에 제시한다. 각 서비스는 예측 결과를 단순 지표가 아닌 행동 가능한 의사결정으로 연결하는 것을 목표로 한다.



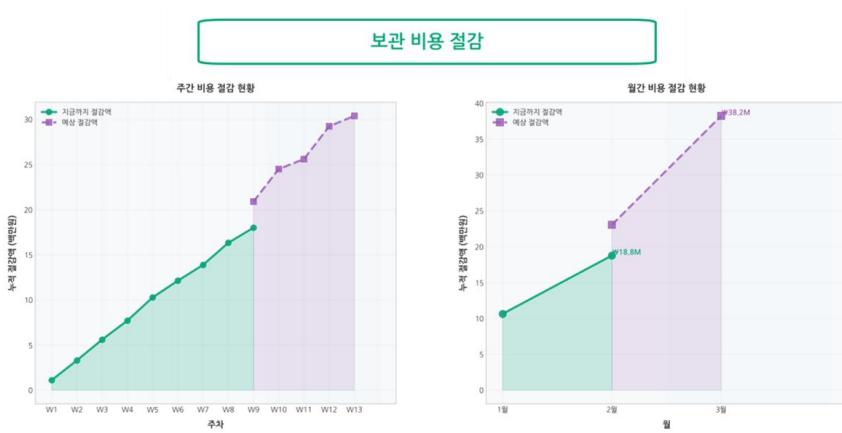
(스마트 물류 서비스 시연 예시)

(a) 전략적 발주 계획 수립 (Fig. 4): 목적은 향후 수요에 맞춰 발주 시점과 물량을 선제 정렬하는 데 있다. 서비스는 리드타임과 안전재고를 고려해 월·주 차원의 권장 발주 스케줄을 제공하고, 피크 전 선행 확보와 비성수기 완충 축소를 유도한다. 예측 모델은 미래 판매의 높낮 이를 보여 주어 언제·얼마나 발주해야 재고 부족과 과잉 재고를 동시에 줄일지 정량적으로 제시한다.



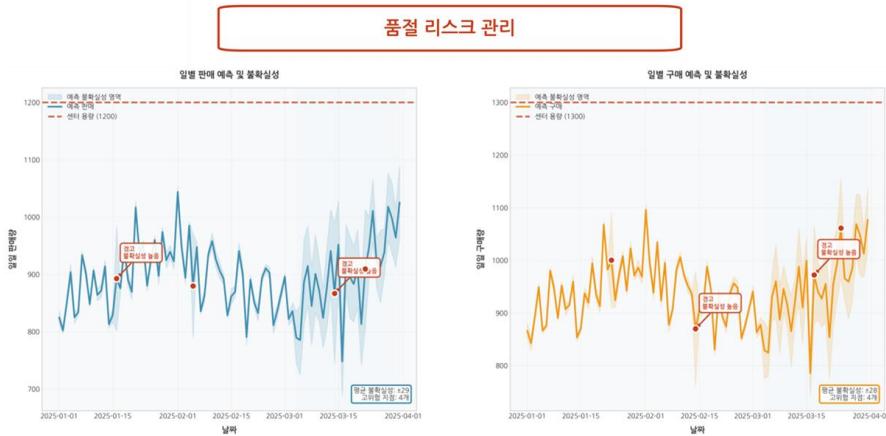
(기능 1. 전략적 발주 계획 수립 부분 시연 예시)

(b) 보관 비용 절감: 목적은 서비스 레벨을 유지하면서 평균 재고를 최적화해 보관 비용을 낮추는 데 있다. 서비스는 수요의 계절성과 변동 폭을 바탕으로 안전재고와 입고 타이밍을 재조정하고, 기준 정책 대비 누적 절감 경로를 시작화한다. 예측 모델은 필요 시점과 필요 물량을 드러내어 "많이 쌓아 안전하게"에서 필요할 때 필요한 만큼으로 운영 관성을 전환한다.



(기능 2. 보관 비용 절감 서비스 부분 시연 예시)

(c) 품질 리스크 완화: 목적은 품질 가능성의 조기 감지와 선제 대응이다. 서비스는 예측 평균과 불확실성 밴드를 현재 재고·입고 예정·센터 용량과 함께 비교하여 위험 구간을 경보한다. 예측 모델은 포인트 추정을 넘어 위험의 정도를 수치로 제공하며, 이 신호가 조기 발주·대체상품 전환·채널 재배분 같은 정책 스위치를 즉시 당기게 한다.



(기능 3. 품질 가능 리스크 관리 서비스 시연)

## 2. 한계점

- 본 모델은 월·주 단위 집계 성능은 견고하지만, 일 단위 급변 구간에서는 오차가 상대적으로 커진다. 이 한계는 발주 타이밍과 안전재고 산정의 정밀도에 직접적인 영향을 미친다. 특

히 프로모션 직전·직후, 비정상 이벤트(공휴일 편성, 이상 기상, 특정 채널 바이럴) 구간에서 예측 신뢰구간이 넓어지는 경향이 관찰된다. 따라서 일 단위 의사결정을 그대로 자동화하기보다, 불확실성 신호를 함께 제시하고 승인·검토 절차를 두는 편이 안전하다.

2) 또 하나의 구조적 한계는 Fulfillment 운영 규칙의 미반영이다. 매입은 순수 수요만으로 결정되지 않는다. 발주 주기, 최소발주수량(MOQ), 공급사 리드타임과 납기 편차, 도크·인력·설비의 처리 용량과 시간창, 입고 검수 속도, 센터 및 셀 용량 제약, 반품 역물류 간섭 등 다종 규칙이 현실 의사결정을 지배한다. 이 요소를 충분히 관찰·모델링하지 않으면 매입 측 예측에서 체계적 편향이 누적될 수 있다. 결과적으로 모델이 산출한 권장 발주량이 실제 실행 가능성이나 비용 최소화와 괴리되는 장면이 발생한다.

### 3. 향후 연구

1) 향후 연구는 불확실성 정량화 고도화와 운영 규칙의 체계적 내재화에 초점을 둔다. 먼저 불확실성 측면에서, 서로 다른 초기화·특징·구조를 갖는 양상을 모델을 구성하고 시점  $t$ 의 예측 집합  $\{\hat{y}_t\}$ 의 분산·사분위 범위(IQR)-예측구간(PI) 폭을 이용해 Uncertainty score  $U_t$ 를 정의한다. 직관적으로  $U_t$ 는 모델들이 서로 얼마나 일치하는지를 나타내며, 학습 분포에 충분히 노출된 패턴에서는 낮고, 낯선 패턴에서는 높아진다. 운영에서는  $U_t$ 가 임계치를 넘을 때 안전재고 가중(예:  $k\sigma$  반영), 선행 발주, 담당자 확인 요청을 자동 트리거한다. 시각화에서는 (c) 와 같이 밴드 폭의 변화 자체가 위험 신호로 기능하여, 현장 의사결정의 투명성과 대응 속도를 동시에 높인다.

2) 운영 규칙의 내재화를 위해서는 하이브리드 접근이 필요하다. 로그·리포트·현장 인터뷰를 통해 발주 주기, MOQ, 리드타임 분포, 입고 처리율, 피크 시간대, 도크·인력·셀 용량, 공급사 SLA를 체계적으로 수집한다. 수집된 규칙은 두 층위로 반영한다. 첫째, 규칙 기반 전처리/후 처리로 예측량을 실행 가능 영역으로 사상한다(예: MOQ 반올림, 도크 용량 초과 시 분할·스케줄 이동). 둘째, 수리 최적화/휴리스틱을 결합해 목적함수 (보관비 + 품절비 + 발주비) 최소화와 제약(MOQ, 리드타임, 일·주 입고 한도, 셀 용량)을 동시에 만족하는 현실적 발주 스케줄을 산출한다. 이때 시뮬레이션을 통해 대체 시나리오(분할 발주, 대체 상품, 우선순위 조정)를 빠르게 비교·평가하는 체계를 마련한다.

마지막으로 평가·데이터 측면에서, 일·주·월 다계층 지표와 비용 기반 메트릭(품질 비용·보관 비용·발주 비용의 가중 합)을 동시에 운영하여 모델 개선이 실제 비용 절감과 위험 완화로

이어지는지를 검증한다. 특징 확장(품목 계층, 채널, 공급사별 리드타임, 프로모션 플래그, 역물류 플로우 등)과 이상 이벤트 태깅을 강화하여 설명력과 재현력을 높인다.

## 5. 활용데이터 및 참고 문헌 출처 등

### 5.1 활용 데이터명

- 중소유통 물류센터 거래 데이터(매입·매출)

### 5.2 타기관 또는 민간 데이터 명

#### (1) 기상청 기상자료개방포털 기온분석자료

- 제공기관: 기상청 (<https://data.kma.go.kr/cmmn/main.do>)
- 데이터 url: <https://data.kma.go.kr/stcs/grnd/grndTaList.do?pgmNo=70>
- 주요내용: 지역별 기온
- 활용목적: 기온 요인이 매입 및 매출량에 미치는 영향 분석

#### (2) 기상청 기상자료개방포털 강수량분석자료

- 제공기관: 기상청 (<https://data.kma.go.kr/cmmn/main.do>)
- 데이터 url: <https://data.kma.go.kr/stcs/grnd/grndRnList.do?pgmNo=69>
- 활용목적: 강수량 요인이 매입 및 매출량에 미치는 영향 분석

#### (3) 월별 코로나 바이러스 감염증-19(COVID-19) 확진자 수 데이터

- 제공기관: KDX 데이터 거래소
- 데이터 url: <https://kdx.kr/data/view/25918>
- 활용목적: 확진자 수 요인이 매입 및 매출량에 미치는 영향 분석

#### (4) GitHub 개인저장소 대한민국 공휴일 가공 데이터

- 제공자: hyunbinseo
- 저장소 url: <https://github.com/hyunbinseo/holidays-kr>
- 활용목적: 월별 공휴일 수 요인이 업무일에 미치는 영향 분석

### 5.3 사용 라이브러리 및 개발 환경

- 개발 환경
  - o Google colab, Jupyter notebook, Visual studio code
- 라이브러리
  - o 데이터 처리
    - pandas, numpy, os, datetime, urllib.request, re

- 시각화
  - matplotlib, seaborn, tabulate
- 모델링 / ML
  - xgboost, lightgbm, sklearn, torch
- 통계 / 수학
  - scipy.stats, collections.Counter, itertools.combinations
- 보조 기능
  - Warnings
- 문서화 및 시각화 환경
  - Microsoft PowerPoint 2021 (결과 시각화 및 발표자료 제작)
  - Microsoft Word 2021 (보고서 본문 작성)
  - Flaticon (<https://www.flaticon.com/>)