Journal of Channel and Retailing
Volume 31, Number 2(April), pp.25-48, 2026.

Korea Distribution Association(http://www.kodia.or.kr)
pISSN:1226-9263, eISSN:2383-9694
https://doi.org/10.17657/jcr.2026.4.30.2

중소 유통 공동 물류센터의 수요 예측을 위한 머신러닝 변수 선택 전략 연구*

안민정**, 한지영***, 정승환****, 노인준*****

본 연구는 중소 유통업체가 공동으로 이용하는 지역 기반 공동 물류센터의 운영 특성에 적합한 출하량 예측 모델을 제안하고자 한다. 공동 물류센터는 판매량이 적고 간헐적일 뿐 아니라, 개별 점포의 프로모션·행사와 같은 내부 운영 정보를 직접적으로 파악하기 어렵다는 구조적 제약을 지녀 대형 유통사의 방식과는 다른 접근이 요구된다. 이에 본 연구는 실제 한국의 공동 물류센터에서 수집된 2021~2024년 판매 데이터를 활용하여 랜덤 포레스트 기반 예측 모델을 구축하고, 두 가지 변수 선택 방법론인 Wrapper 방식과 Filter 방식을 물류센터의 데이터 특성과 운영 환경에 맞추어 비교·분석하였다.

Wrapper 방식은 변수 중요도 기반 Backward Stepwise Selection을 적용하였으며, Filter 방식은 피어슨 상관계수를 기준으로 상위 N개 변수를 선택하는 접근을 사용하였다. 분석 결과, 계산 효율성 측면에서 Filter 방식은 Wrapper 방식 대비 3~100배 빠른 처리 속도를 보였고, 예측 정확도(RMSE) 측면에서도 대부분의 N 구간에서 Wrapper 방식을 상회하는 성능을 나타냈다. 일반적으로 Wrapper 방식은 계산 시간이 오래 걸리지만 예측 정확도가 높은 것으로 알려져 있으나, 공동 물류센터의 출하량 예측과 같이 변수 수가 지나치게 많고 변수 간 계절성·동조성이 강한 도메인에서는 상관계수 기반 Filter 방식이 특히 효과적으로 작동하는 것으로 확인되었다. 특히 계절성을 띄고 간헐적인 수요 패턴을 보이는 소분류에서 Filter 방식의 예측력이 두드러졌으며, 이는 공동 물류센터의 데이터 구조와 높은 정합성을 갖는 결과이다. 본 연구는 중소 유통 물류센터의 풀필먼트 운영에 적합한 실무적 변수 선택 전략을 제시하였다는 점에서 의의가 있으며, 향후 지역 기반 물류 서비스 운영 효율을 제고하는 데 기여할 수 있다.

주제어 : 공동 물류센터, 수요 예측, 머신러닝, 변수 선택

Ⅰ. 서론

국내 유통 산업은 온라인 채널 확장, 다품종 소량 구매 행태 확산 등으로 빠르게 변화하고 있다(대한상공회의소, 2024). 대형 유통업체가 운영하는 기업형 슈퍼마켓은 지속적인 상승세를 보이고 있으며 다품종 상품 도입으로 경쟁력을 확보하고 있다. 그러나 소형 슈퍼마켓, 나들가게, 동네 마트와 같은 지역 기반 중소 유통업체는 대형 유통사의 온라인 물류체계에 편입되기 어렵기 때문에 지역 소비자의 수요를 감당하면서도 대형 유통사와 비교하여 경쟁력을 얻기 어려운 상황이다. 대형 유통사의 자체적 풀필먼트는 빠른 배송과 높은 SKU 다양성을 장점으로 가지지만, 이러한 서비스는 수도권과 대도시 중심으로 이루어져 지역 간 물류 서비스 편차 역시 심화되고 있다. 이에

* 이 논문은 2025년 한국유통학회와 한국전자정보통신산업진흥회의 학술데이터지원사업 지원을 받아 수행된 연구임.
이 논문은 2025년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2025S1A5C3A02006968).
** 연세대학교 경영학과 석사과정(amj5070@yonsei.ac.kr), 제1저자
*** 연세대학교 경영학과 석사과정(liz020928@yonsei.ac.kr), 제1저자
**** 연세대학교 경영학과 부교수(seunghwan.jung@yonsei.ac.kr), 교신저자
***** 고려대학교 경영학과 부교수(injoonnoh@korea.ac.kr), 교신저자

중소 유통업체는 주로 지역 기반의 공동도매물류센터에 의존하여 점포 운영에 필요한 상품을 조달하고 있다.

최근 한국산업진흥원은 공동도매물류센터의 지역 기반의 경쟁력을 강화하기 위해 디지털 유통물류센터 표준모델을 구축하고 있다(산업통상자원부, 2025). 기존의 풀필먼트 센터가 중소 유통사의 발주 및 출고가 이루어지는 창고의 역할을 했다면, 새로운 표준모델은 풀필먼트센터 기준의 생태계를 구축하여 중소 유통사, 온라인 쇼핑채널, 배달사 간의 상생적 생태계를 구축하고자 한다. 이러한 변화로 앞서 언급된 국내 유통 산업의 큰 특징인 온라인 채널 확장과 다품종 소량 구매에 대응하여 대기업과 비교한 경쟁력을 확보할 수 있다고 기대된다. 그러나 이러한 표준 모델이 가능하기 위해서는 풀필먼트 센터에 입주한 점포들에 대한 출고량 예측이 필수적이다. 특히 공동물류센터는 소매점의 작은 수요들을 통합하여 처리하는 특성상, 수요 예측의 정확성이 운영 효율성과 직결된다. 더 나아가 출고, 배송, 재고 보충, 인력 배치와 같은 운영 의사결정에 수요 예측을 실질적으로 활용하기 위해서는 정확성뿐 아니라 예측 결과를 신속하게 산출할 수 있는 계산 효율성이 필수적이다. 예측이 지연될 경우 현장에서 즉시 적용하기 어려워 실무적 효용이 크게 감소하기 때문이다. 이에 본 연구는 이러한 중소 유통물류 운영 환경을 반영하여, 높은 정확성과 신속성을 동시에 확보할 수 있는 출하량 예측 모델을 제안하고자 한다.

본 연구가 주목한 공동 물류센터 운영의 구조적 제약은 크게 두 가지로 정리될 수 있다. 첫째, 공동 물류센터는 다수의 소규모 점포를 대상으로 운영되기 때문에 판매량이 적고 변동성이 크며, 간헐적 판매가 매우 흔하게 발생한다. 실제 분석 대상 물류센터의 SKU 판매량 분포에서도 연간 판매량 50개 미만 SKU가 대다수를 차지하는 것으로 나타나, 개별 SKU 수준의 수요 패턴은 희소하고 불규칙적이다. 이러한 특성은 대형 유통사의 다품종·고빈도 판매 구조와 대비되며, 개별 품목의 출하량 예측을 어렵게 만드는 핵심 요인이다. 둘째, 공동 물류센터는 개별 점포의 프로모션, 행사, 진열 변경과 같은 내부 운영 정보를 직접적으로 확보하기 어렵다. 이는 대형 유통사가 활용하는 주요 예측 변수들을 동일하게 적용할 수 없음을 의미하며, 결과적으로 변수화가 가능한 정보가 제한된다는 문제가 발생한다.

본 연구는 이러한 두 가지 제약을 해결하기 위해 두 가지 방향에서 접근한다. 첫째, 개별 점포의 프로모션·행사와 같은 운영 정보를 직접적으로 확보하기 어렵다는 한계를 보완하기 위해, 상품군 간 판매가 동시적으로 움직이는 구조적 특성에 주목하여 관련 상품군의 수요 정보를 간접적 예측 변수로서 활용하고자 한다. 둘째, 판매량이 적고 간헐적이며 변동성이 큰 품목들로 구성된 데이터 구조에 적합한 예측 체계를 마련하기 위해 고차원·고변동 환경에서 효과적인 변수 선택 전략을 비교·검토한다. 특히 물류센터 도메인에서 시간 제약과 예측 정확도 간 균형을 확보할 수 있는 방안을 탐색하는 데 초점을 둔다.

한편 공동 물류센터의 수요를 예측하기 위해 본 연구는 매출(출고량) 기반 물동량을 수요의 지표로 활용한다. 이는 매입 데이터가 점포의 주문 의도나 발주 정책에 영향을 받는 반면, 매출은 실제로 물류센터에서 외부로 이동한 실질적인 수요를 반영하기 때문이다. 또한 출고량은 재고 소진, 차량 배차, 피킹 및 적재 인력 배치와 직접적으로 연결되어 물류센터 운영의 효율성을 달성하는 데 중요한 역할을 한다.

본 연구는 방법론의 측면에서 머신러닝의 변수 선택 방법 중 Filter 방식과 Wrapper 방식을 비교

한다. 공동 물류센터와 같은 고차원·고변동 데이터 환경에서 머신러닝을 적용하기 위해서 머신러닝의 성능을 높일 변수를 선정하는 것이 중요하다. 이에 본 연구는 예측 정확도가 높으나 시간이 오래 걸리는 Wrapper 방식과 예측 시간은 짧으나 정확도가 상대적으로 낮은 것으로 알려진 Filter 방식을 각각 공동 물류센터에 적합하게 모델링하여 공동 물류센터 환경에 보다 적합한 변수 선택 전략을 도출하고자 한다.

본 연구는 공동 물류센터와 같이 다수의 영세 점포가 참여하는 유통 센터에 특화된 머신러닝 기반 출하량 예측 방법을 실증적으로 제시한다는 점에서 의의가 있다. 기존 연구는 대부분 대형 유통사의 단일 물류체제를 전제로 하고 있으며, 공동 물류센터의 출하량 예측 모델 개발은 거의 시도되지 않았다. 따라서 본 연구는 이러한 공백을 보완하여 공동 물류센터 도메인의 특성을 반영한 상관관계 기반 머신러닝 모델을 제안하는 데에 연구적 의의가 있다. 머신러닝 기반의 정확성이 높고 신속한 출하량 예측 모델은 향후 지역 물류 서비스의 품질을 향상하고 중소 유통업체의 온라인 경쟁력 확보에 중요한 기초 자료로 활용될 수 있을 것이다.

논문의 구성은 다음과 같다. ‘II. 선행 연구’에서 물류센터의 문제를 다룬 연구와 머신러닝을 활용한 연구를 정리한 후, ‘III. 문제 정의’에서 공동 물류센터의 운영 환경과 데이터 구조를 분석하여 수요 예측에서 발생하는 핵심 문제를 정의한다. 이후 ‘IV. 모델’에서 이러한 문제를 해결하기 위한 예측 모델과 변수 선택 전략을 제시하고, Wrapper 방식과 Filter 방식의 설계 방식을 설명한다. ‘V. 분석’에서 실제 판매 데이터를 활용하여 두 방식의 계산 효율성과 예측 성능을 비교하고, 마지막으로 ‘VI. 결론’에서 결론과 시사점을 제시한다.

Ⅱ. 선행 연구

1. 물류센터 관련 연구

본 연구의 주요 목적은 머신러닝 기반의 수요 예측 방법론을 개발하여 물류센터 운영의 효율성을 높이는 데에 있다. 기존 연구에서 물류센터가 직면하는 핵심 문제로 작업 프로세스의 비효율성과 리드타임 지연이 주로 제기되어 왔다(Wan, 2022; Ye et al., 2025). 이러한 문제의 근본적인 원인은 불확실한 수요로, 이를 완화하기 위해 물류센터의 운영 효율화와 더불어 수요 예측 연구가 활발하게 이루어져 왔다(김영남, 2023; 민경창, 2022).

김영남(2023)은 전자상거래 풀필먼트 센터의 자원 관리를 위해 다양한 수요 예측 방법론을 이용하여 풀필먼트 센터 내 피킹과 패킹 작업 인력을 최적화한다. 해당 논문은 이동 평균 모델과 ML 기반의 모델 등 여러 알고리즘을 비교하며, 최종적으로 고객사 특성, 물성, 이벤트 등의 변수를 반영한 통합 모델을 제시한다. 방선호(2024)는 물류센터 출고 프로세스에 집중하여, 연관성 분석 알고리즘으로 상품의 피킹 우선순위를 정하고 프로세스 효율화를 달성한다. Ye et al.(2025)은 배송 시간 불확실성을 물류센터 운영의 가장 큰 문제점으로 인식하고, 배송 지연 분포를 예측하는 머신러닝 모델을 만들어 최적의 물류센터 배정 조합을 결정한다. Wan(2022)의 연구는 제조사가 단일 물류센터 대신 두 개의 물류센터나 공장 직출 구조를 활용할 때 리드타임 단축과 비용 절감이 이루어짐을 실증적으로 확인한다.

그러나 이러한 연구들은 대부분 개별 기업 단위의 물류센터를 분석 대상으로 삼고 있으며, 여러 업체가 공동으로 이용하는 공동 물류센터를 다룬

연구는 매우 제한적이다. 정지철(2015)의 연구는 공동 물류센터의 경쟁력 강화의 방법으로 입지·상품·가격·서비스 차원의 프레임워크를 제안하였고, Guo et al.(2024)은 중국의 RiRiShun Logistics 사례를 통해 공동 물류센터의 SKU 단위 주문 특성과 배송 시간 구조를 분석했다. 김정수(2025)는 공동도매물류센터 물동량 데이터를 활용하여 협업 필터링과 콘텐츠 기반 필터링을 통합하는 하이브리드 상품 추천 시스템을 제시했다. 하지만 이들 연구는 사례 기반 분석 또는 데이터 분석 과정에 그치거나 공동 물류센터에 특화된 수요 예측 모델을 개발하는 단계까지는 확장되지 못했다. 이에 본 연구는 기존 문헌의 한계를 보완하고자 소규모 유통업체가 공동으로 이용하는 중소 물류센터의 특성을 고려한 수요 예측 모델을 제안하여 중소 물류센터의 운영을 실질적으로 개선하고자 한다.

2. 머신러닝 기반 예측 기법

컴퓨팅 기술의 발전과 함께 머신러닝 기반의 수요 예측은 산업 현장과 학계에서 빠르게 확산되고 있다. 머신러닝 모델은 많은 입력 변수를 사용하기도 하는데, 고차원 데이터는 무관하거나 중복적인 특징이 포함되는 경우가 많아 학습 알고리즘의 성능을 저해하기도 한다(Yu & Liu, 2003). 따라서 예측 정확도 향상을 위해서 유용한 변수를 선별하는 변수 선택 과정(feature selection)이 중요하다.

머신러닝 기반 예측 연구에서 변수 선택 방법은 크게 Filter 방식과 Wrapper 방식으로 구분된다. 먼저 Filter 방식은 데이터의 통계적 특성을 이용하여 학습 알고리즘과 독립적으로 변수를 평가하고 중요 변수를 선별한다. 예를 들어 피어슨 상관계수, 카이제곱 통계량 등 단일 기준을 적용해 목표 변수와의 관련성이 낮은 변수를 사전에 제거한다. 이러한 방식은 계산 속도가 빠르고 확장성이 높아 고차원 데이터에서 유용하지만, 모델 구조나 변수 간 상호작용은 반영하기 어렵다는 한계가 있다. 반면 Wrapper 방식은 변수 선택을 모델 학습 과정에 통합하여 모델을 반복적으로 학습–평가하고, 그 결과를 기준으로 변수 조합을 조정해 나가는 탐색 기반 방식이다. 전진 선택(forward)이나 후진 제거(backward) 방식이 대표적이며, 변수 간 상호 작용을 고려해 예측력이 높은 변수 집합을 찾을 수 있다. 그러나 학습-평가-조정의 반복 과정으로 인해 계산 비용이 크며, 변수 수가 많을수록 탐색이 어렵다는 단점이 있다. 두 방식은 각각 예측 정확도와 계산 효율성의 측면에서 장단점이 있으며, 도메인 특성에 따라 적합한 기법이 달라질 수 있다. 이에 본 연구는 고차원·고변동성이 특징이며 동시에 예측의 신속성이 요구되는 공동 물류센터 데이터 환경을 고려하여, 두 변수 선택 방식 중 어떤 방법이 실질적 효용을 갖는지를 비교하고자 한다.

머신러닝 기반 예측 모델은 방법론적 특성에 따라 전통적 통계 기반 시계열 모형, 신경망 기반 모형, 그리고 트리 기반 앙상블 모형 등으로 구분될 수 있다. 전통적 시계열 모형으로는 SARIMA가 대표적이며, 신경망 기반 접근으로는 Neural Network 및 Transformer와 같은 딥러닝 계열 모델이 활용된다. 한편, 트리 기반 앙상블 모형에는 XGBoost와 Random Forest가 포함된다. 이 중 Random Forest(이하 랜덤 포레스트)는 다수의 의사결정나무를 부트스트랩 표본추출과 무작위 변수 선택을 통해 학습한 뒤 이를 평균화하는 대표적인 앙상블 학습 기법이다(Breiman, 2001). 랜덤 포레스트는 본 연구에서 중점적으로 사용된 머신러닝 모델이며, 변수 간 비선형 관계와 상호작용 효과를 사전 가정 없이 학습할 수 있고 개별 트리

의 분산을 줄여 과적합을 완화하고 높은 일반화 성능을 확보할 수 있다는 장점을 가진다. 특히 고차원 데이터 환경에서 변수 선택 과정과의 결합이 용이하며, 변수 중요도 지표를 통해 설명 변수의 상대적 기여도를 파악할 수 있다는 점에서 실증 분석에 적합한 방법론으로 평가된다.

최근 수요 예측 분야에도 머신러닝 기법을 이용한 연구가 다수 이루어지고 있다. Salari et al. (2022)은 온라인 주문의 실제 배송 시간 분포를 랜덤 포레스트 모델로 예측하여, 고객에게 가장 비용 효율적이고 정확한 약속 배송일을 실시간으로 결정한다. 민경창(2022)은 두 가지 머신러닝 모델, SARIMA와 랜덤 포레스트를 결합하고 상위 카테고리 예측을 하위 예측에 활용하여 단기 수요 예측 정확도를 높인다. 김혜선(2023)은 물류센터 출하 속도 향상을 위해 머신러닝 기반 연관분석으로 함께 구매되는 SKU 묶음을 추출하여 동일한 W/S에 배치하는 휴리스틱 알고리즘을 제시한다.

이러한 연구들은 머신러닝이 수요 예측에서 성과와 효율성을 향상시킨다는 것을 보여주지만, Wrapper 방식과 Filter 방식을 직접적으로 비교 분석한 연구는 드물다. 물류센터 데이터는 고차원적이고 변동성이 높은 것이 특징이므로, Wrapper 방식의 장점인 높은 정확도와 Filter 방식의 장점인 빠른 계산 속도가 실제 환경에서도 동일하게 나타나는지 검증할 필요가 있다. 따라서 본 연구는 Wrapper 방식 중 하나인 변수 중요도 기반 Backward 기법과 Filter 방식 중 하나인 상관관계 기반 변수 선택 모델을 비교한다. 두 모델을 물류센터 물동량 데이터에 적용하여 예측 시간과 성능의 측면에서 어떤 방식이 중소 유통 물류센터의 특성에 더 적합한지 실증적으로 분석하고자 한다.

Ⅲ. 문제 정의

1. 공동 물류센터의 수요 예측

본 연구는 기존의 머신러닝 기반 수요 예측 기법을 소규모·공동 물류센터의 운영 맥락에 적합하게 적용하고, 이러한 환경에서 발생하는 변수 선택 및 예측 문제를 분석하는 데 목적이 있다. 이에 따라 연구의 범위는 공동 물류센터 관점의 수요 예측으로 한정한다.

소규모 유통 점포는 대형 유통 기업과 달리 자체적인 물류센터를 구축하기 어려워 지역 기반의 공동 물류 허브를 이용해 재고와 판매를 관리한다. 이러한 공동 물류센터는 다수의 영세 점포를 대상으로 하여 비교적 작은 규모를 가지며, 개별 점포의 판매량이 작고 상품 회전율이 낮아 판매 구조가 대형 유통업과 본질적으로 상이하다. 특히 개별 SKU 단위의 판매는 빈도가 낮고 간헐적으로 발생하여 물류센터 차원의 수요 패턴이 불규칙적이고 어렵다. 이러한 구조는 물류센터의 안정적인 재고 운영과 배송 계획 수립을 어렵게 만들기 때문에, 공급 효율성을 확보하기 위한 정확한 수요 예측이 필수적이다.

공동 물류센터의 물동량 예측에는 크게 두 가지 어려움이 존재한다. 첫째는 변수 선택의 복잡성이다. 금융·마케팅 분야와 달리 물류 수요 예측에서는 외부 요인을 직접적으로 변수화하기 어렵다. 자체 물류센터를 운영하는 기업의 경우엔 그나마 프로모션 여부, 행사 일정, 신제품 출시 등 다양한 요인을 반영할 수 있지만, 공동 물류센터는 개별 점포의 운영 정보를 직접적으로 파악할 수 없어 이를 변수화하기가 구조적으로 어렵다. 이러한 제약을 보완하기 위해 본 연구는 다른 소분류 상품군의 판매 정보를 잠재 변수로 활용하는 방법을

고려한다. 특정 상품군의 판매량은 계절성, 보완재·대체재 관계를 반영하여 특정 상품의 프로모션 여부, 유행 정도 등 다양한 외부 요인을 간접적으로 반영할 수 있다. 이는 공동 물류센터 환경에서 직접 확보하기 어려운 정보를 보완할 수 있게 한다. 더불어, 7일, 30일 등의 물류 주기의 과거 판매량 정보는 요일 및 계절 정보를 간접적으로 제공할 수 있어 외부 정보를 별도로 입력하지 않아도 된다는 장점이 있다.

둘째는 이러한 방식을 차용할 경우 잠재 변수가 과다해짐으로 인해 발생하는 고차원성 문제와 이에 따른 시의성 저하다. 실제 유통·물류 환경에서는 상품군의 수가 수십에서 수십만 개까지 확대될 수 있으며, 시차항을 포함할 경우 변수 수는 상품군의 수에서 적게는 5배에서 많게는 수십 배까지 증가한다. 본 연구에서 사용한 데이터만 보더라도, 시차를 1, 7, 14, 21, 28, 30일로 설정했을 때 소분류 209개에 대해 총 1,254개의 잠재 변수가 생성된다. 이러한 잠재 변수들을 모두 예측 모델에 포함시킬 경우 모델의 복잡도 증가, 계산 비용 확대, 노이즈 유입, 과적합 등 다양한 성능 저하 요인이 발생한다. 따라서 다양한 상품군의 출하량 정보를 잠재 변수로 활용하기 위해서는 예측에 꼭 필요한 변수만을 선별하여 고차원성 문제를 완화하는 과정이 필수적이다.

그러나 잠재 변수의 수가 지나치게 많아질수록 변수 선택 과정 자체가 방대한 계산 시간을 요구하며, 이는 정확성과 시의성을 동시에 요구하는 물류 수요 예측의 특성에 부합하기 어렵다. 즉, 공동 물류센터 환경에서는 '빠르고 정확한' 변수 선택이 필수적이지만, 변수화가 가능한 외부 정보가 제한된 상황에서 잠재 변수 공간이 지나치게 커지기 때문에 이를 효율적으로 수행하는 것은 본질적으로 어려운 과제이다.

따라서 본 연구는 공동 도매물류센터의 운영 특성(고차원·고노이즈 구조, 변수화의 제약, 빠른 대응의 중요성)을 반영하여 이에 적합한 머신러닝 기반 변수 선택 및 수요 예측 방법론을 연구하고자 한다. 이를 통해 중소·영세 유통 점포가 경제 규모의 제약 속에서도 경쟁력을 확보할 수 있도록 공동 물류센터의 운영 효율화를 지원하는 데 기여하고자 한다.

2. 분석 데이터

본 연구의 분석에 활용된 데이터는 국내 K 물류센터의 2021~2024년 실제 판매 거래 데이터이다. 이 자료는 개별 판매 건별로 기록된 일별 거래 데이터로 구성되어 있으며, 각 상품이 어느 날짜에 얼마만큼 판매되었는지를 상세히 포함하고 있다. 분석 기간은 2021년 1월 2일부터 2024년 12월 31일까지로, 해당 물류센터의 연간 판매 활동을 포괄적으로 반영한 데이터셋이다.

데이터는 상품의 대·중·소분류, 바코드, 판매수량, 공급금액 등 다양한 변수로 구성되어 있으며, 주요 데이터 변수는 <표 1>에 제시되어 있다. 전체 상품 구조는 대분류는 17개, 중분류 77개, 소분류 209개로 구성되며, 이 중 소분류는 '당면', '일반국수' 등과 같이 실제 운영 단위에 가까운 세분화된 상품군 구성을 갖고 있어 정밀한 수요 패턴 분석이 가능하다. <표 2>는 면류·라면류 대분류의 중분류, 소분류, SKU의 예시를 제시한다. 이때, 소분류 정보가 누락된 상품은 분석 대상에서 제외하였으며, 나머지 데이터에 대해서만 분석을 수행하였다.

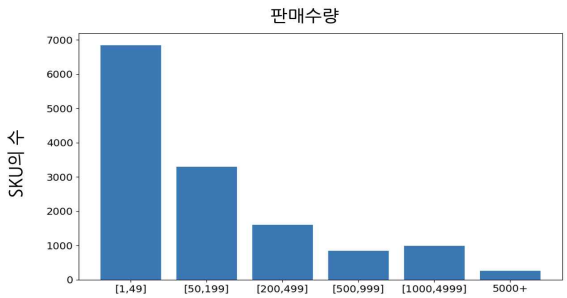
본 연구는 중소 물류센터의 운영적 특성을 반영하여 SKU 단위가 아닌 소분류 단위로 수요를 예측한다. <그림 1>에서 확인할 수 있듯이, 대부분의 SKU 판매량은 연간 50개 미만의 구간에 집중되어 있으며, 판매량이 증가할수록 SKU 개수가

<표 1> 데이터 개요

변수명	설명	예시
판매일	물류센터에서 해당 품목이 판매된 날짜	2024-01-02
매출처코드	개별 소매점을 식별하기 위한 고유 코드로 매출 발생 주체를 구분	00000
판매수량	거래에서 판매된 수량	12
규격	상품의 주요 단위 특성(중량, 크기, 개수 등)을 나타내는 변수	5kg, 100mL
상품 바코드	개별 상품의 바코드 기반 고유 SKU 코드	0000000000000
대/중/소분류	상품군을 계층적으로 분류	가공식품류, 밥·반찬류
공급금액	거래의 총 판매 금액	10,000

<표 2> 대, 중, 소분류 및 SKU 예시

대분류	중분류	소분류	SKU 예시
면류·라면류	기타라면	기타라면.면류.	(○○○)××국수중면 900g
	라면류	봉지라면	(○○○)×라면××맛멀티 120g×5
		용지라면	(○○○)××× 110g
	면류	기타 면류	(○○○)××쌀국수 92g
		당면	(○○○)××당면 300g
		일반국수	(○○○)××쌀국수××맛 92g
		파스타	(○○○)××××마카로니 500g
		칼국수, 수제비, 쫄면	(○○○)××××칼국수 900g



<그림 1> 연판매량에 따른 SKU 수

급격히 감소하는 긴꼬리 분포를 보인다. 이러한 구조에서는 개별 SKU 단위의 데이터가 희소하고 변동성이 커 수요 예측 모델의 안정적인 학습이 어렵다. 이에 반해 소분류 단위는 유사한 SKU의 판매가 자연스럽게 묶여 충분한 판매 규모를 확보할 수 있다. 특히 공동 물류센터를 이용하는 소규모 점포들은 동일한 지역 내에서 날씨, 요일, 이벤트, 계절성 등의 요인에 함께 영향을 받기 때문에, 특정 소분류 상품 간에 높은 상관 구조를 형성한다. 따라서 유사 소분류 간 상관성을 활용한 예측이 더욱 타당하며, 이는 대형 유통사보다 소규모 점포 집합체에서 더 강하게 나타나는 구조적 특징을 반영한 결과이다.

또한 공동 물류센터의 운영 의사결정은 피킹 인력, 차량 배차, 발주 주기와 같이 개별 SKU가 아닌 묶음 단위의 물동량을 기준으로 이루어지는 경우가 많다. 이와 같은 실제 운영적 의사결정의

특성을 고려할 때, 공동 물류센터 운영에 적합한 수요 예측 모델은 소분류 기준이라 판단하였다. 실제 중소 물류센터의 수요 예측 문제를 다룬 선행 연구에서도 SKU 단위가 아닌 상품 세분류 단위를 활용한 접근이 관찰된다(이상일, 2024).

이러한 이유로 본 연구는 SKU 단위의 원본 거래 데이터를 소분류 기준으로 일별 집계하여 분석용 시계열 데이터를 구축하였다. 전체 소분류 209개에 대해 소분류별 평균 일별 판매량을 집계한 요약 통계는 <표 3>과 같다. 값은 정수로 반올림하였으며, 1 이하의 경우 소수점 둘째자리로 반올림하여 표기하였다. 평균 일 판매량은 약 25개, 중앙값은 2개로 나타나 전반적인 분포가 오른쪽으로 치우친 형태임을 확인할 수 있다. 최소 0.00007개에서 최대 약 894개까지 범위가 매우 넓어 소분류 간 출하량의 편차가 크다는 점 또한 드러난다.

이와 같은 특징을 머신러닝 기반 수요 예측 모형에 효과적으로 반영하기 위해, 본 연구에서는 모든 소분류의 판매량을 로그 값으로 변환(log1p)하여 분석하였다. 이는 소분류 간 규모 차이에 따른 영향력을 완화하고, 모델 학습 과정에서 특정 대규모 소분류가 예측 결과를 과도하게 좌우하는 현상을 방지하기 위함이다. 이에 따라 이후의 변수 선택 및 예측 분석은 모두 로그 변환된 값을 기준으로 수행하였다.

또한, 본 연구에서는 변수로 사용하기 적합한 시차항을 탐색하기 위해 자기상관함수(ACF: auto-correlation function)를 분석하였다. ACF는 시계열 자료에서 현재 시점의 값과 과거 특정 시점(lag)의 값 간 상관관계를 측정하는 지표로, 시계열 내 반복적 패턴이나 주기성을 파악하는 데 활용된다.

<표 3> 소분류별 일별 평균 판매량 요약 통계

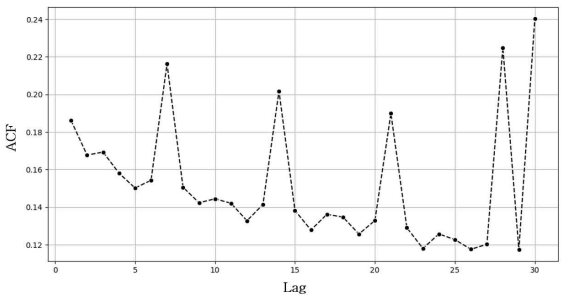
구분	최소	Q1	중앙	Q3	최대	평균
값	.00	.19	2	14	894	25

이에 따라, 물동량의 전체적인 주기적 특성을 확인하기 위해 전체 소분류의 수요를 합산한 총수요 시계열을 대상으로 ACF를 계산하였다. ACF 상위 10개 시차는 <표 4>에 제시하였으며, lag1부터 lag30까지의 ACF를 시각적으로 표현한 그래프는 <그림 2>에 제시하였다.

분석 결과, 7일 및 14일을 포함한 주 단위의 반복 패턴과 함께 30일 전후의 월 단위 시차에서도 비교적 높은 자기상관이 나타났다. 이는 물동량 데이터의 특성상 주간 단위의 운영 주기가 반영된 결과로 해석될 수 있다. 한편, lag 2~4와 같은 인접 시차는 상호 유사한 정보를 포함할 가능성이 있어, 단기 정보를 대표하는 lag 1을 포함하고

<표 4> ACF 기준 상위 10개 시차

순위	시차	ACF 값
1	0	1.000000
2	30	.240296
3	28	.224843
4	7	.216189
5	14	.201683
6	21	.189826
7	1	.186100
8	3	.169284
9	2	.167683
10	4	.158052



<그림 2> 시차항에 따른 ACF값

주간 및 중기 패턴을 반영할 수 있는 시차를 중심으로 변수 집합을 구성하였다. 최종적으로 1, 7, 14, 21, 28, 30일의 시차를 선정하였다.

Ⅳ. 모델

이 장에서는 공동 물류센터의 수요 예측에 적합한 변수 선택 방법을 탐색하기 위해 사용한 예측 모델을 소개한다. 특히 본 연구의 초점은 공동 물류센터의 데이터 환경에 적합한 변수 선택 기법을 규명하는 것이기에, 이를 분석하기 위해 여러 변수 선택 방법을 공통적으로 적용할 동일한 예측 알고리즘이 필요하다. 본 논문에서는 해당 알고리즘으로 시계열 예측 분야에서 널리 활용되는 랜덤 포레스트를 채택하였다. 랜덤 포레스트는 Breiman(2001)이 제안한 앙상블 학습 알고리즘으로, 개별 모델로는 결정 트리(decision tree)를 기반으로 한다.

랜덤 포레스트는 이러한 결정 트리를 부스트랩 샘플링과 무작위 변수 선택 전략을 통해 다수 생성한 뒤, 각 트리의 결과를 평균하거나 투표 방식으로 결합하여 최종 예측을 수행한다. 이는 개별 트리의 불안정성을 줄이고 예측력을 안정화하며, 복잡한 데이터 구조에서도 우수한 성능을 확보할 수 있게 해준다. 또한 변수 간 상호작용을 자동으로 포착하고 과적합을 효과적으로 방지하는 특성 때문에, 대규모·고차원 데이터가 빈번한 유통 수요 예측 환경에서도 적합한 모델로 평가된다(Breiman, 2001; Delgado et al., 2014). 최근 수요 예측 연구에서는 복잡한 비선형 구조와 고차원 변수 환경을 효과적으로 처리할 수 있다는 점에서 랜덤 포레스트가 다른 머신러닝 방법론보다 빈번하게 활용되고 있다(민경창, 2022; 이상일, 2024). 이에 따라 본 연구에서도 고차원성이 강한 공동 물류 데이터 환경을 고려하여 랜덤 포레스트를 기본 예측 모델로 채택하였다.

더불어, 랜덤 포레스트는 학습 과정에서 변수 중요도(feature importance)를 산출한다. 변수 중요도는 각 변수가 예측 성능에 기여한 정도를 나타내는 지표로, 예측에 중요한 변수를 식별하는 데 유용하다. 본 연구에서는 랜덤 포레스트의 불순도 감소 기반 변수 중요도(impurity-based importance)를 Wrapper 방식의 변수 제거 기준으로 활용하였다. 회귀 문제에서 불순도는 분산으로 정의되며, 각 변수의 중요도는 해당 변수가 트리 분할에 기여한 평균 분산 감소량을 기준으로 산출된다. 구체적으로, 랜덤 포레스트에서는 각 변수가 트리 분할에 사용될 때마다 예측 오차(분산)가 얼마나 줄어들었는지를 계산한다. 즉, 어떤 변수가 데이터를 더 잘 나누어 예측 성능을 개선했다면, 그 변수는 더 큰 분산 감소를 만들어낸다. 각 트리에서 이러한 분산 감소량을 모두 더한 뒤, 이를 전체 트리에 대해 평균하여 해당 변수의 중요도를 산출한다.

이 방식은 변수의 예측 기여도를 직관적으로 반영하기에 변수 선택의 기준으로 자주 사용되며, Wrapper 방식의 기준 지표로도 적합하다. 따라서 본 연구에서는 랜덤 포레스트를 공통 예측 모델로 채택하고, 변수 중요도를 Wrapper 방식의 평가 기준으로 활용하였다.

1. 벤치마크: 자기시차항(Self-Lagged Term)

먼저, 본 연구는 물류 데이터 환경에서 변수 선택 기법의 효과성을 검증하는 데 목적이 있으므로, 벤치마크 모델로는 변수 선택을 적용하지 않고 목표 시계열의 과거값만을 활용한 단순 모델을 설정하였다. 이러한 벤치마크는 연관 상품군의 변수를 도입하는 것이 실제 예측 성능 향상에 어

떠한 기여를 하는지를 판단하는 기준선으로 기능하며, Wrapper 방식과 Filter 방식이 각기 선택한 변수가 얼마나 추가적인 성능 개선을 제공하는지 비교하는 데에도 활용된다.

시계열 기반 머신러닝 예측에서 자기시차항을 포함하는 것은 가장 기본적이면서 안정적인 접근으로 널리 사용된다(Makridakis et al., 2018). 이러한 설정은 단순하지만 물류 수요의 핵심 패턴을 반영할 수 있어, 이후 제안하는 변수 선택 방법들과 비교하기 위한 기준선(baseline)으로 적합하다고 판단하였다.

2. Wrapper 방식: 변수 중요도 기반 Backward Stepwise Selection

Wrapper 방식은 변수 조합을 변경할 때마다 모델을 반복적으로 학습하고 그 성능을 비교하여 최적의 변수 집합을 찾는 방식이기 때문에 계산 비용이 매우 크다. 따라서 시간 제약이 큰 물류센터 의사결정 환경에서는 효율적인 탐색 전략의 선택이 필수적이다.

Wrapper 방식의 설계는 크게 다음과 같은 두 차원이 고려된다.

(1) 변수의 포함 여부를 판단하는 기준(criterion)

(2) 변수 조합을 탐색하는 전략(search strategy)

먼저 기준 측면에서는, 두 가지 접근이 가능하다. 하나는 모델이 제공하는 변수 중요도와 같은 사전 지표를 활용해 제거 대상을 결정하는 방식이고, 다른 하나는 각 변수를 하나씩 제거해 보면서 검증 성능을 비교하는 방식이다. 그러나 후자의 방식은 매 단계마다 모든 후보 변수를 개별적으로 제거해 재학습해야 하므로, 본 연구의 변수 규모에서는 계산 비용이 과도하게 증가한다. 이에 본 연구에서는 랜덤 포레스트의 변수 중요도를 제거 기준으로 활용하였다.

다음으로 탐색 전략 측면에서는 전수조사와 Stepwise 방식이 대표적이다. 전수조사는 모든 변수 조합을 평가하므로 이론적으로는 최적 조합을 보장할 수 있다. 그러나 본 연구의 경우 시차항을 포함한 잠재 변수가 약 1,200개 이상으로 확장되어, 전수 탐색은 현실적으로 계산이 불가능하였다. 실제로 변수 중요도 기반 전수 탐색을 제한적으로 시도하였으나, 너무 많은 학습 시간이 소요되어 실무 적용이 어렵다고 판단하였다.

이에 따라 현실적인 대안으로 Stepwise 전략을 채택하였다. Stepwise는 성능 개선 여부에 따라 탐색을 조기 종료할 수 있어 계산 부담을 크게 줄일 수 있다. Stepwise 전략은 Forward 방식과 Backward 방식으로 구분된다. Forward 방식은 소수의 변수에서 시작하여 성능이 개선되는 변수를 점진적으로 추가하는 방식이며, Backward 방식은 전체 변수를 포함한 상태에서 성능 기여도가 낮은 변수를 단계적으로 제거하는 방식이다. 두 방식을 모두 실험적으로 비교한 결과, 본 연구의 데이터 환경에서는 Backward 방식이 일관되게 더 좋은 예측 성과를 보였다. 이에 따라 최종 Wrapper 방식으로 변수 중요도 기반 Backward Stepwise Selection 방식을 채택하였다.

3. Filter 방식: 상관계수 기반 변수 선택

Filter 방식은 변수 선택을 위해 모델 학습을 반복적으로 수행하지 않고, 사전에 정의된 통계적 기준에 따라 변수를 한 번에 결정하는 방식이다. Wrapper 방식 대비 계산 비용이 매우 낮고 빠르다는 장점이 있으나, 그만큼 기준 선정의 타당성이 중요하다.

본 연구는 기준 선정을 위해 잠재 변수들이 “다른 상품군의 판매량”이라는 점에 주목하였다. 앞서 언급한 바와 같이 특정 상품군의 판매량은 계

절성, 보완재·대체재 관계, 프로모션 반응 등 다양한 외부 요인의 간접적 신호를 담고 있을 수 있다. 이러한 관계성이 존재한다면, 해당 상품군 간에는 양(+) 또는 음(−)의 상관관계가 나타날 가능성이 높다.

상관계수는 목표 소분류 판매량 시계열 y_t과 후보 변수 $x_{j,t}$ 간의 선형적 동조성(co-movement)을 정량적으로 포착하는 지표로, 피어슨 상관계수는 다음과 같이 계산된다.

$$\rho_j = \frac{\Sigma_{t=1}^{T}(x_{j,t}-\bar{x}_j)(y_t-\bar{y})}{\sqrt{\Sigma_{t=1}^{T}(x_{j,t}-\bar{x}_j)^2}\sqrt{\Sigma_{t=1}^{T}(y_t-\bar{y})^2}}$$

본 연구에서는 이 상관계수의 절댓값 $|\rho_j|$의 크기를 기준으로 상위 N개의 변수를 선정하여 예측 모델에 투입하였다.

이러한 상관계수 기반 필터 방식은 고차원 데이터에서 빠르고 효율적인 변수 선택 기법으로 연구되어 왔다. Yu and Liu(2003)는 상관 지표를 활용하여 관련성과 중복성을 동시에 고려하는 필터 알고리즘을 제안하였고, Haindl et al.(2006) 또한 변수 간 상호 상관 정보를 이용해 불필요한 변수를 제거하는 방식을 제시하였다. 두 연구 모두 Wrapper 방식 대비 상관 기반 필터 접근이 계산 효율성이 높고 고차원 데이터에 적합함을 실증적으로 보여주었다.

그러나 본 연구에서 제안하는 상관 기반 Top-N 필터 방식은 기존 연구들과 동일한 문제를 해결하려는 것이 아니라, 공동 물류센터의 수요 예측이라는 도메인 특수성에 맞추어 기존 상관 기반 필터 개념을 재구성한 방식이라는 점에서 차별성을 가진다. 기존 상관 기반 필터 기법들이 주로 일반적 고차원 데이터에서의 효율성 확보 또는 변수 간 중복성 제거(redundancy reduction)를 목표로 하는 반면, 본 연구는 공동 물류센터의 운영 환경에서 나타나는 도메인 고유의 패턴(계절성·동시판매·대체재/보완재 관계 등)을 반영해, 목표 시계열과의 상관성을 기반으로 예측에 실제로 도움이 되는 변수 집합을 실무적으로 빠르게 식별하는 데 목적을 둔다. 따라서 본 연구의 기여는 새로운 강력한 변수 선택 알고리즘을 제안하는 데 있는 것이 아니라, 물류 데이터의 구조적 특성과 시간 제약을 고려하여 기존 상관 기반 필터 접근을 도메인 적합적으로 변형·구체화한 실무 활용형 변수 선택 방식을 제시하는 데 있다.

이에 따라, 본 연구에서는 목표 소분류와의 상관관계를 기준으로 상위 N개 변수를 선정하여 모델에 투입하는 방식을 Filter 접근법으로 설정하였다.

이때, Filter 방식과 Wrapper 방식은 모두 동일하게 소분류별로 변수 선택 및 예측 모델을 독립적으로 구축하였다. 따라서, 두 방식 간의 차이는 변수 선택의 기준과 방법에만 있다.

이어지는 분석에서는 공동 물류센터 환경에서 이 두 가지 방식을 시간과 예측 성능의 측면에서 실증적으로 검토한다.

V. 분석

1. 실험 설계

앞서 논의되었듯, 세 모델(벤치마크, Wrapper, Filter) 모두 변수 입력 시 시차항은 ACF 값을 고려하여 1, 7, 14, 21, 28, 30일로 구성하여 동일하게 적용하였다. 전체 데이터는 80%의 학습용(Training Set)과 20%의 검증용(Validation Set)으로 분리하였다. Wrapper 방식의 경우 변수 선택 과정에서 추가적인 검증이 필요하므로, Training Set(80%)를 다시 60%의 학습 데이터와 20%의 테

스트 데이터로 재분리하여 변수 선택을 수행하였다. 이후 선택된 변수 집합을 기반으로 최종 모델을 Training Set 전체(80%)로 재학습시키고 Validation Set(20%)에서 성능을 평가하여 각 방식 간의 비교의 일관성을 확보하였다.

또한, 모델 성능 평가지표로는 RMSE(root mean squared error)를 사용하였다. RMSE는 오차의 제곱을 사용하기 때문에, 상대적으로 큰 예측 오차에 대해 더 큰 패널티를 부여하는 특성이 있다. 이러한 특성은 판매량과 같이 변동 폭이 큰 물류 수요 예측 문제에 적합하여, 기존 연구에서도 RMSE가 수요 예측 성능 지표로 널리 활용되어 왔다(Hyndman & Koehler, 2006).

더불어, 시간적 관점에서 실제 현업에서는 단순한 모델 학습 시간뿐 아니라 데이터가 입력된 시점부터 최종 예측 결과가 생성되기까지의 전체 처리 시간이 중요한 의사결정 기준이 된다. 따라서 본 연구에서는 현실적 타당성을 위해 각 모델별로 변수 선택에 소요된 시간, 모델 학습 시간, 그리고 결과 도출 시간을 모두 합산하여 비교하였다.

일반적으로 Filter 방식은 한 번의 계산으로 변수 선택이 완료되지만, 본 연구에서는 상위 N개의 변수를 기준으로 모델 성능이 어떻게 달라지는지를 분석하기 위해 다양한 N 값에 대해 반복적으로 모델을 학습하고 성능 변화를 관찰하였다. 이는 최적의 N을 적용했을 때의 성능을 Wrapper 방식 및 벤치마크 모델과 직접 비교하기 위함이다. 반면 N이라는 개념은 Filter 방식에서만 사용되며, Wrapper 방식과 벤치마크 모델에는 적용되지 않는다. 즉, 이 두 방법은 N 값에 따른 성능 변화(곡선)가 존재하지 않고 하나의 결과값만 나오게 된다. 따라서 Wrapper 방식과 벤치마크 모델은 각각의 단일 성능 값을 기준으로, Filter 방식에서 도출된 최적 N 지점의 성능과 비교하였다. <그림 3>은 본 연구에서 제안한 설계에 따라 구성된 각 모델의 절차를 요약한 Pseudo-code를 제시한다.

2. 결과

<그림 4>는 Wrapper 방식, Filter 방식, 그리고 벤치마크 모델 간의 시간적 효율성과 예측 성능을 비교한 결과를 나타낸다. 분석 결과는 물류 환경에서의 변수 선택 전략이 예측 모델의 실용성과 성능에 미치는 영향을 명확하게 보여준다.

먼저 시간적 측면에서 Filter 방식은 Wrapper 방식 대비 최소 약 3배에서 최대 100배까지 빠른 것으로 나타났다. 이는 물류센터와 같이 빠른 의사결정과 실시간·근실시간 분석이 요구되는 환경에서 중요한 결과이다. Wrapper 방식은 변수 조합을 반복적으로 학습·평가해야 하므로 근본적으로 계산 비용이 크며, 특히 수백 개의 상품군에 각기 여러 개의 시차항이 추가되는 물류 데이터에서는 그 부담이 더욱 증가한다. 이에 반해 Filter 방식은 단일 통계량을 기준으로 변수를 선별한 뒤 학습을 수행하기 때문에 시간 비용이 크게 절감된다.

놀랍게도 성능 측면에서도 Filter 방식은 예상보다 매우 우수한 결과를 보였다. N이 100~800인 구간에서 Filter의 예측력이 Wrapper에 비해 우수하였으며, N=300 구간에서 최적 성능이 확인되었다. 이는 기존의 머신러닝 분석 연구들이 보고한 패턴과는 다른 양상을 띤다(Yu & Liu, 2003). 이러한 결과는 물류센터 데이터의 도메인 특성이 반영된 것으로 해석된다. 즉, 물류 데이터에서는 상품군 간의 구조적 연관성 및 패턴의 동조성이 강하게 존재하므로, 상관관계 기반의 Filter 방식이 다른 데이터 영역에 비해 특히 효과적으로 작동하는 것으로 보인다. 그래프상에서 관찰된 성능 차이가 통계적으로 유의한지 확인하기 위해 단

Algorithm 1 Benchmark: 자기시차항

Require: Target series $\{y_t\}$, lag set $L = \{1, 7, 14, 21, 28, 30\}$, evaluation metric RMSE
Ensure: RMSE on validation set
(Lag construction)
Construct lagged features X_t:
  $X_t \leftarrow [\, y_{t-\ell} : \ell \in L \,]$
Split data into Train80 (first 80%) and Valid20 (last 20%)
Train Random Forest on Train80 using X_t
Compute RMSE on Valid20
return RMSE

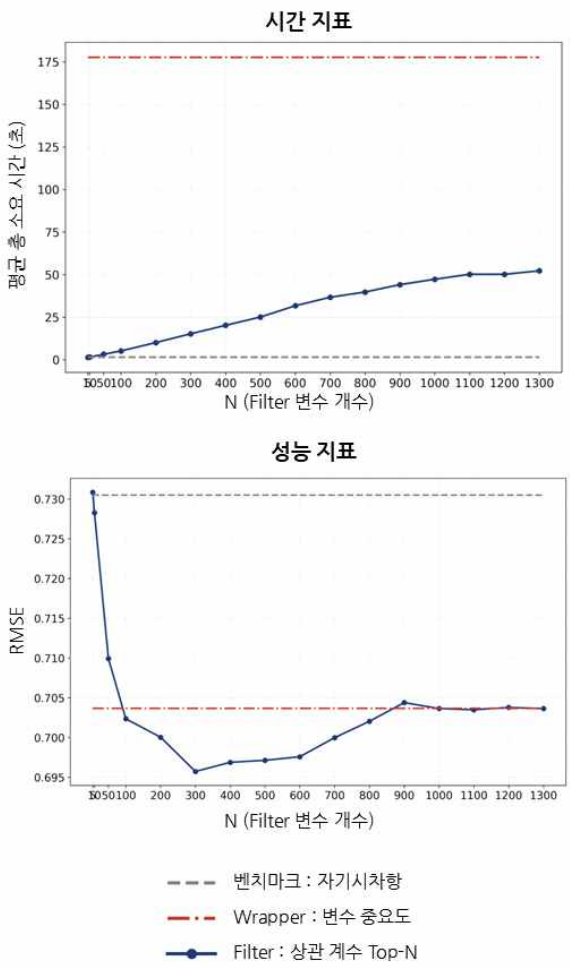
Algorithm 2 Wrapper: 변수 중요도 기반

Require: Target series $\{y_t\}$, candidate series $\{x_{j,t}\}$, lag set $L = \{1, 7, 14, 21, 28, 30\}$, evaluation metric RMSE
Ensure: Selected feature subset S, RMSE on validation set
(Lag construction)
Construct full feature set F:
  Include self-lags $y_{t-\ell}$ for $\ell \in L$
  Include cross-lags $x_{j,t-\ell}$ for all j and $\ell \in L$
Split data into Train80 (first 80%) and Valid20 (last 20%)
Within Train80, split into Train60 (first 60%) and Test20 (next 20%)
$S \leftarrow F$
function SCORE(S)
  Train Random Forest on Train60 using S
  return RMSE on Test20 using S
end function
curr_score $\leftarrow$ SCORE(S)
while $|S| > 1$ **do**
  Train Random Forest on Train60 using S
  Compute feature importance FI_j for all $j \in S$
  $v_{\min} \leftarrow \arg\min_{j \in S} FI_j$
  $S' \leftarrow S \setminus \{v_{\min}\}$
  new_score $\leftarrow$ SCORE(S')
  if *new_score* $<$ *curr_score* **then**
    $S \leftarrow S'$
    curr_score $\leftarrow$ *new_score*
  else
    break
  end if
end while
Train Random Forest on Train80 using final S
Compute RMSE on Valid20
return S, RMSE

Algorithm 3 Filter: 상관계수 기반

Require: Target series $\{y_t\}$, candidate series $\{x_{j,t}\}$, lag set $L = \{1, 7, 14, 21, 28, 30\}$, integer N, evaluation metric RMSE
Ensure: Selected feature subset S_N, RMSE on validation set
(Lag construction)
Construct full feature set F:
  Include self-lags $(y_{t-\ell})$ for $\ell \in L$
  Include cross-lag pairs $(x_{j,t-\ell})$ for all j and $\ell \in L$
Split data into Train80 (first 80%) and Valid20 (last 20%)
Compute Pearson correlation $\rho_{(j,\ell)}$ between $(x_{j,t-\ell})$ and y_t for all $(j, \ell) \in F$
Rank features in F by $|\rho_{(j,\ell)}|$ in descending order
$S_N \leftarrow$ Top-N features from F
Train Random Forest on Train80 using S_N
Compute RMSE on Valid20
return S_N, RMSE

<그림 3> 벤치마크, Wrapper, Filter 방식의 절차를 요약한 Pseudo-code



<그림 4> 벤치마크, Wrapper, Filter 기법의 소분류 기준 시간 및 성능지표

측 pairwise *t*-test를 수행한 결과, p 값이 .063으로 유의수준 10%하에서 Filter 방식이 Wrapper 방식보다 우수한 성능을 보이는 것으로 확인되었다. 또한, RMSE 차이의 분포가 정규성을 충분히 만족하지 않을 가능성을 고려하여 비모수 검정인 Wilcoxon signed-rank test를 추가적으로 수행하였다. 그 결과 p 값이 .005로 나타나, 보다 강한 수준에서 두 방법 간 성능 차이가 유의함을 확인할 수 있었다. 따라서 Filter와 Wrapper 기법 간의 성능 차이는 통계적으로도 일관되게 지지되는 것으로 판단된다.

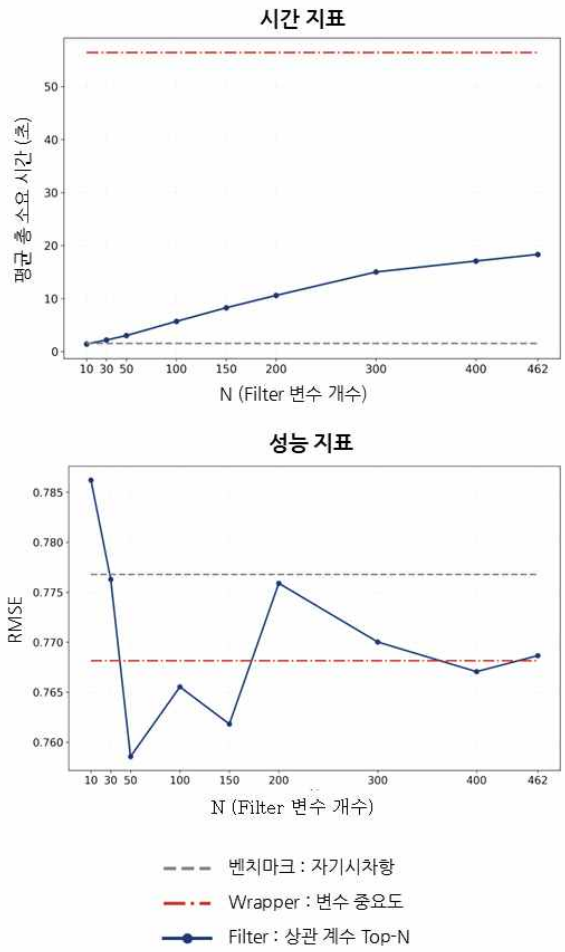
추가적으로, 두 방법에서 실제로 선택되는 변수의 수를 비교한 결과, Wrapper 방식은 평균적으로 약 1,250개의 변수를 포함하는 것으로 나타나 변수 축소 효과가 제한적으로 나타남을 확인하였다. 이는 고차원 입력 환경에서 Wrapper 방식이 변수 선택 과정에서 충분한 차원 축소를 달성하기 어려울 수 있음을 시사한다. 이러한 결과는 Wrapper 방식의 잠재력을 충분히 활용하기 위해서는 보다 광범위한 변수 탐색이 필요함을 의미하지만, 이는 공동물류센터와 같이 다수의 상품군과 시차항이 동시에 고려되는 환경에서는 계산 비용과 시간 측면에서 현실적인 제약이 크다.

더불어, 이러한 결과의 흐름이 중분류의 구조에서 또한 일관되는지를 확인하기 위해 중분류를 대상으로 동일한 분석을 진행하였다. 그 결과, 잠재 변수의 개수의 차이로 인해 소분류만큼 매끄럽지는 않더라도 대체적으로 동일한 양상을 보인다. <그림 5>는 중분류에서의 Wrapper 방식, Filter 방식, 그리고 벤치마크 모델 간의 시간적 효율성과 예측 성능을 비교한 결과를 나타낸다.

3. 논의

이 장에서는 앞선 결과를 보다 정교하게 이해하기 위해 몇 가지 해석을 제시한다. 먼저, Filter 방식이 Wrapper 방식에 비해 시간적 효율성과 예측 성능 측면에서 모두 우수한 결과가 나타난 이유는 크게 두 가지 측면에서 설명될 수 있다.

첫째, 공동 물류센터의 변수 구조 특성상 Wrapper 방식보다 Filter 방식이 더 유리하게 작동한다. 공동 물류센터의 수요 예측을 위해 다른 상품군의 판매량을 잠재 변수로 포함할 경우 변수의 수가 매우 많아지며, 그 과정에서 다수의 노이즈 변수가 함께 유입되는 고차원 구조가 형성된다. 특히 본 연구와 같이 시차항을 추가하는 경우 변수



<그림 5> 벤치마크, Wrapper, Filter 기법의 중분류 기준 시간 및 성능지표

수는 기하급수적으로 증가하여 탐색 공간이 폭발적으로 확대된다. 그럼에도 자기 시차항만을 변수로 포함한 벤치마크 모델에 비해 Wrapper 방식과 Filter 방식 모두 성능이 개선된 점을 고려하면, 다른 상품군 판매량을 변수로 도입하는 접근 자체는 분명 의미가 있다. 다만 이러한 고차원 환경에서는 계산 비용이 큰 Wrapper 방식의 구조적 한계가 명확하게 드러난다.

Stepwise 기반 Wrapper 방식은 탐욕적(greedy) 탐색 절차를 따르기 때문에 최적 변수 조합을 충분히 탐색하지 못할 가능성이 높으며, 이를 보완하기 위해 전수조사(exhaustive search)를 시도하더라도 변수 조합 수가 지나치게 커 현실적으로 불가능하다. 특히 물류 수요 예측은 신속성을 요구하는 의사결정 환경이므로 전수조사를 적용할 수 없으며, 결국 Stepwise 기법을 사용할 수밖에 없다. 그러나 Stepwise는 변수 수가 지나치게 많은 환경에서는 최적점을 제대로 찾지 못해 변수 선택 기능을 충분히 수행하지 못하는 한계가 있다.

둘째, 상관관계 기반 Filter 방식은 물류 데이터의 도메인 특성과 높은 정합성을 보인다. 물류 데이터는 시계열적 구조를 가지며, 상품군별 판매량이 유사한 패턴을 보이거나 반대로 움직이는 경우가 빈번하다. 이러한 패턴적 동조성(co-movement)은 계절성, 프로모션, 대체재·보완재 관계 등 실제 운영 요인의 영향을 함께 반영하는데, 상관계수는 이러한 판매 패턴의 연관성을 직접적으로 포착하는 지표로 기능한다. 따라서 단순한 상관 기반 필터링만으로도 의미 있는 정보 변수를 효과적으로 선정할 수 있고, 이는 본 연구에서 Filter 방식이 Wrapper 방식 대비 높은 예측 성능을 안정적으로 보인 이유로 해석할 수 있다.

4. 추가 분석

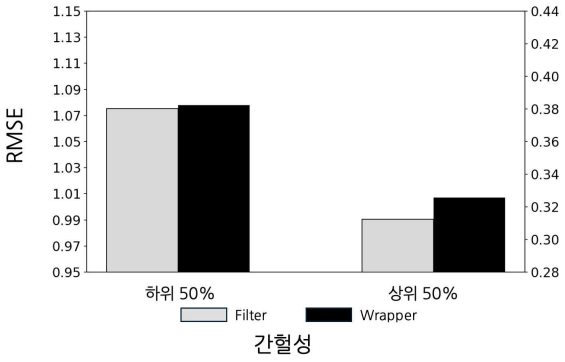
이 장에서는 결과에 대한 더 정교한 해석을 위해 상관관계 기반 Filter 방식이 Wrapper 방식에 비해 어떠한 수요 환경에서 강점을 나타내는지를 분석하였다. 구체적으로는 간헐성, 판매량, 변동성, 계절성의 네 측면을 기준으로 두 머신러닝 기법의 성능 차이를 분석하였다. 각 분석에서는 각 측면에 대해 각 품목을 중앙값을 기준으로 상위 50%, 하위 50%의 두 그룹으로 구분한 후 각 기법의 RMSE를 그래프로 도출하여 어느 그룹에서 Filter 방식이 더 우수한지를 시각적으로 확인하였다. 이후에는 통계적으로 유의한 분석을 위해

Filter 방식과 Wrapper 방식 간 RMSE 차이(Filter RMSE-Wrapper RMSE)가 두 그룹에서 유의미하게 차이 나는지를 이분산 t 검정을 통해 분석 하였다. 각 그래프는 상위와 하위 50% 그룹을 동일한 그림에 제시하기 위해 축의 스케일은 .02로 통일하였으며, 가독성을 위해 각 축의 기준점은 서로 다르게 설정하였다.

첫째, 간헐성의 경우 간헐성이 높은 상위 50% 품목군에서 Filter와 Wrapper 방식의 성능 차이가 하위 50% 품목군보다 유의미하게 크게 나타났다 (t=−2.091, p=.038). <그림 6>은 간헐성 수준에 따른 두 방법의 예측 성능 차이를 보여준다. 즉, 앞선 분석 결과와 동일하게 판매가 간헐적으로 발생하는 상품에서 Filter의 우수성을 다시 확인할 수 있다.

이때, 간헐성은 판매 발생의 연속성을 의미하며, ADI(average demand interval) 지표를 이용해 측정했다. ADI는 판매가 발생한 시점 사이의 평균 간격을 나타내는 지표로, 값이 클수록 수요가 간헐적으로 발생하는 품목임을 의미한다.

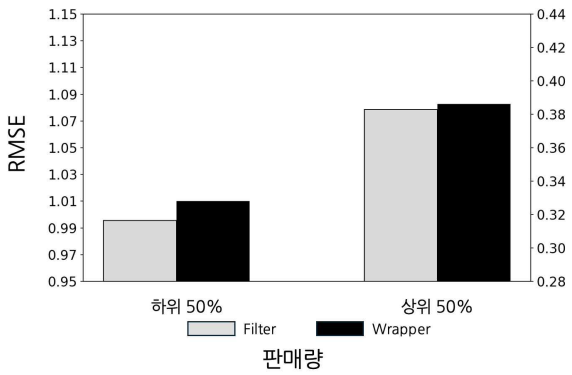
$$ADI = \frac{\sum_{i=1}^{N} t_i}{N}$$



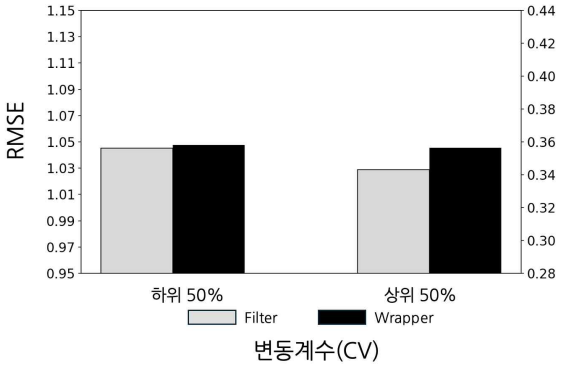
<그림 6> 간헐성에 따른 Filter와 Wrapper 성능 비교

둘째, 판매량 수준에 따른 동일한 분석을 수행한 결과(<그림 7>), 평균 판매량이 낮은 품목군에서 두 방식의 예측 오차의 차이가 전반적으로 Filter 방식에 유리한 방향으로 나타났다. 다만, 이 차이는 통계적으로 유의한 수준에서는 확인되지 않았다(t=1.498, p=.136). 이는 앞선 분석을 고려하면 매우 적은 저판매량 품목에서 Filter 방식의 우수성이 관찰되기는 하나, 판매량 단일 지표만으로는 두 방법 간 차이를 충분히 설명하기에는 한계가 있음을 의미한다.

셋째로, 판매량의 변동성에 관해서도 추가적으로 분석을 수행하였다(<그림 8>). 이때 변동성은 변동계수, 즉 표준편차를 평균으로 나눈 값으로 측정했다. 이때의 결과 역시 앞선 두 결과와 마찬



<그림 7> 판매량에 따른 Filter와 Wrapper 성능 비교

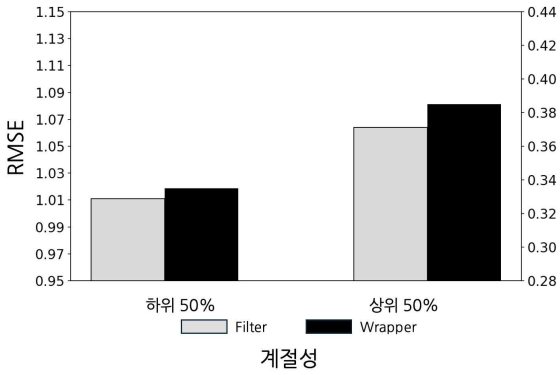


<그림 8> 변동성에 따른 Filter와 Wrapper 성능 비교

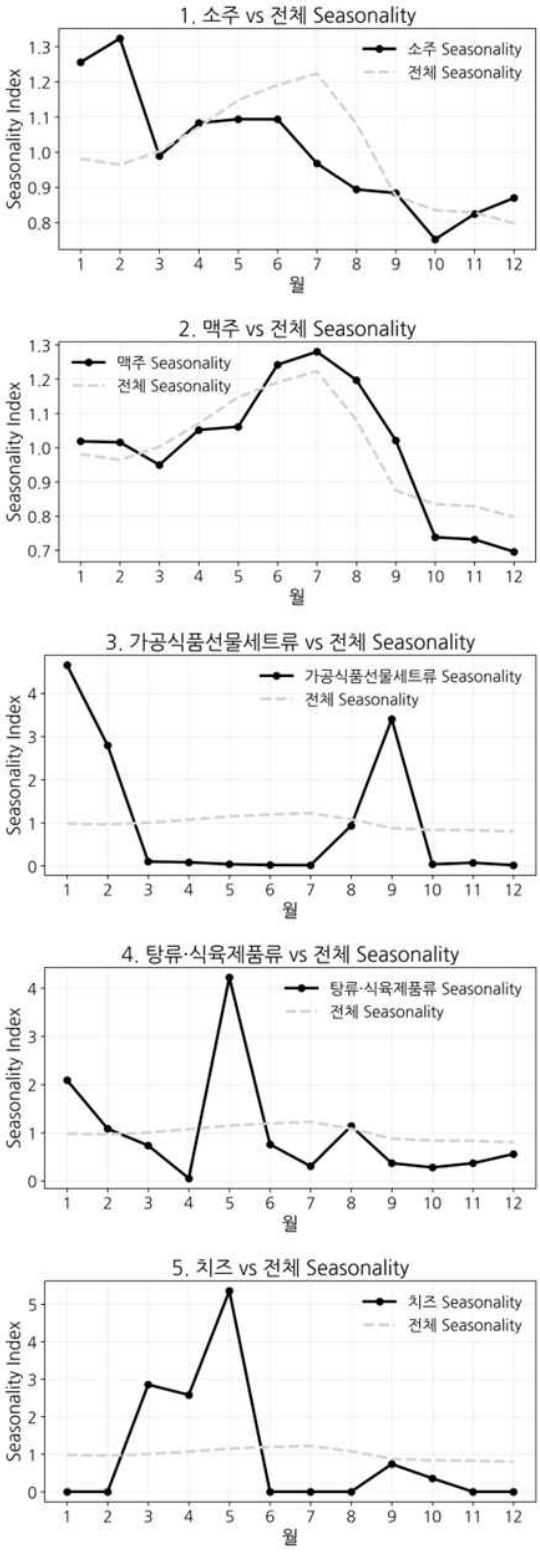
가지로 변동성이 높은 상품군에서 Filter의 상대적 성능이 뛰어났다($t=-2.114$, $p=.036$). 즉, 판매가 간헐적으로 발생하며, 판매량이 많지 않고, 주문의 변동성이 높은 상품에서 Filter의 성능이 일관적으로 높았다. 이 세 가지 속성은 모두 소규모 점포가 이용하는 공동 물류센터의 큰 특징이며, 이러한 환경에서 상관관계 기반의 Filter가 우수하다는 것을 확인할 수 있다.

마지막으로 계절성에 따른 분석을 진행하였다(<그림 9>). 계절성은 월별 변동계수로 측정하였다. 그 결과, 두 방식 성능 차이가 통계적으로 유의미한 것으로 밝혀졌다($t=-1.968$, $p=.051$). 즉, 월별 판매량이 뚜렷한 계절성이 나타나는 품목의 경우 예측 오차 차이가 Filter 방식에 유리한 방향으로 나타났다.

계절성의 경우 성능 차이가 많이 나는 품목들을 대상으로 월별 계절성을 계산하여 전체 품목들의 계절성과 함께 시각화하여 추가적으로 분석해 보았다. <그림 10>은 성능 차이가 가장 많이 나는 5개 품목의 예시이다. 그 결과, 성능 차이가 많이 나는 제품들의 경우 대체로 뚜렷한 계절성을 보이는 것으로 분석되었다. 이는 앞서 논의한 "상관계수가 계절성·동조성과 같은 구조적 시계열 패턴을 효과적으로 포착하기 때문에 좋은 성능을 나



<그림 9> 계절성에 따른 Filter와 Wrapper 성능 비교



<그림 10> 성능 차이 상위 5개 품목의 Seasonality

타낸다"는 해석을 실증적으로 확인해주는 결과이다. Filter 방식이 특히 우수했던 품목들이 실제로 높은 계절성·반복적 패턴을 보였다는 점은, Filter 방식의 우수성이 물류 데이터의 시계열적 구조와 정합적으로 나타난다는 강력한 근거가 된다.

더불어, 변수 선택이 실제로 어떠한 방식으로 이루어졌는지를 간략하게 확인하기 위해 명절 수요의 영향을 크게 받는 '가공식품선물세트류'를 대상으로 Filter 모델의 변수 중요도 상위 10개를 분석해보았다. <표 5>는 이를 보여준다. 그 결과, 과자선물세트류 및 기타 수산가공품과 같이 명절 수요에 따라 함께 변동하는 상품군이 주요 변수로 선택되었으며, 이는 Filter 방식이 계절성과 동조성을 직접적으로 포착하고 있음을 보여준다.

동시에, 탄산음료와 같이 계절적 소비 패턴을 공유하는 상품군뿐만 아니라, 당면이나 합성세제와 같이 직접적인 상품 연관성만으로는 설명하기 어려운 상품군 또한 중요한 변수로 선택되었다.

<표 5> 가공식품선물세트류의 변수 중요도

순위	이름	FI 값
1	선물세트 / 과자선물세트류 / 과자선물세트류_lag_1	.124
2	면류, 라면류 / 면류 / 당면_lag_28	.070
3	선물세트 / 가공식품선물세트류 / 가공식품선물세트류_lag_1	.036
4	음료, 차류 / 음료 / 탄산음료_lag_30	.033
5	조미료류 / 조미료 / 기타조미료_lag_7	.033
6	조미료류 / 식용유 / 참기름_lag_30	.032
7	일상용품 / 세제용품 / 세탁용합성세제_lag_30	.024
8	가공식품류 / 수산가공식품 / 기타 수산가공품_lag_7	.019
9	조미료류 / 장류 / 간장_lag_28	.018
10	일상용품 / 화장지류 / 물수건_lag_30	.018

이는 일부 변수 조합이 직관적인 상품 간 유사성이나 사전적 분류 기준만으로는 충분히 설명되지 않음을 보여준다. 이러한 결과는 머신러닝 기반 접근이 사전에 정의된 상품 간 관계를 넘어, 비선형적이고 잠재적인 수요 패턴을 포착할 수 있음을 시사한다. 더불어, 동일 상품군 내에서 유사한 시차 변수가 반복적으로 선택되기보다는, 서로 다른 상품군과 다양한 시차(lag)가 결합된 형태로 변수 선택이 이루어졌다. 이는 해당 방식이 정보 중복을 최소화하면서도 다중 시계열 구조를 효율적으로 반영하는 변수 집합을 구성함을 보여준다.

종합적으로 볼 때, 본 연구에서 관찰된 Filter 방식의 우수성은 단순한 알고리즘적 차이보다는 공동 물류센터 데이터의 구조적 특성에 기인한다. 즉, 수요가 간헐적이거나, 판매량이 적거나, 계절성과 같은 구조적 패턴이 강한 품목에서는 상관계수가 이러한 특성을 직접적으로 포착함으로써 변수 중요도 기반 Wrapper 방식보다 안정적이고 높은 예측 성능을 제공한다. 이러한 논의는 Filter 방식이 공동 물류센터와 같이 다수의 저판매량·고변동 품목이 혼재하는 환경에 특히 적합한 변수 선택 방법임을 시사한다.

VI. 결론

본 연구는 중소 유통업체가 공동으로 이용하는 지역 기반 공동 물류센터의 운영 환경에 적합한 출하량 예측 모델을 제안하기 위해, 머신러닝 기반 변수 선택 방법인 Wrapper 방식과 Filter 방식을 비교·분석하였다. 공동 물류센터의 데이터는 판매량이 적고 간헐적이며, 개별 점포의 프로모션·행사와 같은 내부 운영 정보를 직접적으로 파악하기 어렵다는 구조적 제약을 지녀 일반적인 대형 유통사의 SKU 기반 예측과는 다른 접근이

요구된다. 이에 본 연구는 소분류 단위로 데이터를 집계하고, 잠재 변수로 다른 소분류의 출하량을 선택하였으며 자기시차 기반 벤치마크 모델과 두 가지 변수 선택 방식의 성능을 동일 조건에서 평가했다.

분석 결과는 다음과 같이 요약된다. 첫째, 시간 효율성의 측면에서 Filter 방식이 Wrapper 방식보다 월등히 우수했다. Filter 방식은 Wrapper 방식 대비 최소 3배에서 100배까지 빠른 처리 속도를 보였으며, 이는 빠른 의사결정이 필요한 물류센터 환경에서 특히 중요한 의미를 갖는다.

둘째, 예측 성능의 측면에서도 상관 기반 Filter 방식이 일관되게 우수한 결과를 보였다. 5개의 소분류를 활용하는 수준에서도 RMSE가 벤치마크 대비 크게 개선되었고, N=300 구간에서 최적 성능이 나타났다. 반면 Wrapper 방식은 변수가 많아질수록 탐색 공간이 급격히 증가하여 최적 변수 조합을 탐색하지 못하는 한계가 드러났으며, 결과적으로 Filter 방식보다 낮은 성능을 보여주었다. 이는 기존 문헌들에서 Filter 방식의 한계로 지적되었던 낮은 정확도와 대비되는 발견이며, 고차원적 변수 공간과 간헐적 판매라는 공동 물류센터의 실제 운영 조건에는 Filter 방식이 정확도와 시간 측면에서 모두 유리하다는 것을 보여준다.

셋째, 성능 차이의 구조적인 원인을 추가적으로 분석한 결과, 판매량이 매우 적고 간헐적 패턴을 보이는 품목에서 Filter 방식의 우수성이 특히 두드러지는 것으로 나타났다. 전체 품목을 대상으로 실시한 *t*-test 기반 분석 결과 품목의 간헐성, 변동성, 계절성에 따른 Filter 방식과 Wrapper 방식의 유의미한 성능 차이를 검증할 수 있었다. 특히 성능 차이가 컸던 품목들은 월별 seasonality가 뚜렷한 경우가 많았는데, 이는 상관계수가 계절적 판매 패턴을 효과적으로 포착할 수 있음을 시사한다. 결과적으로 본 연구의 상관 기반 Filter 접근은 간헐적·고변동 품목이 다수 존재하며 잠재 변수로 다른 상품군의 출하량 정보를 활용해야 하는 공동 물류센터의 실제 운영 조건에 가장 적합한 변수 선택 방식임을 확인하였다.

본 연구의 학술적·실무적 시사점은 다음과 같다. 첫째, 공동 물류센터 도메인 특화형 변수 선택 기법을 제시했다. 외부 변수를 파악하기 어려운 공동 물류센터의 특성을 고려하여 다른 상품의 수요를 변수로 활용하는 예측 분석 모델을 제시하였으며, 이러한 예측 모델에서 드러나는 고차원적 변수 공간이라는 제약에서 가장 효율적인 방식이 Filter 방식임을 실질적으로 보여주었다. 둘째, 상관 기반 Filter 방식은 계산 효율성과 예측 성능을 동시에 확보할 수 있어, 향후 지역 기반 디지털 풀필먼트 모델 구축(예: 중소유통 디지털 물류센터 표준모델)에서 실질적인 분석 도구로 활용 가능함을 보였다. 중소 유통업체가 온라인 경쟁력을 확보하기 위해서는 출고량 예측 기반의 재고 운영과 배송 계획이 필수적이며, 본 연구는 그 기반이 될 수 있는 효율적인 모델을 제시했다는 점에서 의의가 크다.

그러나 본 연구는 몇 가지 한계를 가진다. 첫째, 분석 데이터는 특정 지역의 단일 물류센터에 한정되어 있어, 다양한 지역으로의 일반화 가능성을 검증하지는 못했다. 둘째, 본 연구는 판매량 데이터만을 활용하였으며, 명절 및 공휴일 정보·프로모션 정보·점포 속성 등 추가 변수의 결합 가능성을 충분히 고려하지 못했다. 셋째, 랜덤 포레스트 기반의 예측 모델에 한정하여 변수 선택 기법을 비교했기 때문에, 다른 머신러닝 모델(LightGBM, XGBoost)에서도 동일한 패턴이 나타나는지 추가 검증이 필요하다. 넷째, 본 연구는 운영적 입장을 고려하여 소분류 단위 대상으로 분석을 수행하여 SKU 단위의 물동량 예측에서도 동일한 결과가 나타나는지 검증하지 못했다.

종합하면, 본 연구는 공동 물류센터의 구조적 특성에 적합한 머신러닝 기반 출하량 예측 모델을 처음으로 제시하였으며, 상관 기반 Filter 방식이 계산 효율성과 성능 면에서 모두 유의미한 우위를 갖는다는 점을 확인하였다. Filter 방식의 우수한 성능으로 인한 정확한 물동량 예측은 공동 물류센터의 재고 운영 안정화, 피킹 및 배송 인력의 효율적인 배치로 이어질 수 있으며, 이는 중소 유통업체가 제한된 자원하에서도 물류 운영의 예측 가능성을 높이고 비용 변동성을 완화하는 데 기여한다. 이러한 관점에서 본 연구의 예측 프레임워크는 중소 유통업체의 물류 역량을 구조적으로 보완하고, 지역 기반 디지털 풀필먼트 생태계의 운영 효율성을 제고하는 데 실질적인 분석적 기반을 제공할 수 있을 것으로 기대된다.

논문접수일: 2025. 12. 02.
1차 수정본 접수일: 2026. 02. 13.
게재확정일: 2026. 03. 30.

이해 상충에 관한 보고

본 논문과 관련된 잠재적 이해 상충 관계가 없음을 보고함.

연구비 지원

이 논문은 2025년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2025S1A5C3A02006968).

감사의 글

이 논문은 2025년 한국유통학회와 한국전자정보통신산업진흥회의 학술데이터지원사업 지원을 받아 수행된 연구임.

연구 데이터 접근 가능성

본 연구에 사용된 데이터는 교신저자에게 합당한 요청 시 제공될 수 있음.

저자 기여 항목

연구개념화: 안민정, 한지영, 정승환, 노인준.
데이터 큐레이션/조사: 안민정, 한지영.
데이터 분석/검증: 안민정, 한지영.
방법론: 안민정, 한지영, 정승환, 노인준.
원고 초안 작성: 안민정, 한지영, 정승환, 노인준.
원고 검토 및 편집: 안민정, 한지영, 정승환, 노인준.
자금 조달/자원 확보: 정승환, 노인준.

윤리 심의 승인에 관한 보고

본 연구는 인간 및 동물 참여자가 없으므로 IRB/IACUC 심의가 필요하지 않음.

생성형 AI 사용에 관한 선언

본 논문은 생성형 AI의 사용과 무관함.

참고문헌

김영남, 류상천, 김현 (2023). 풀필먼트 센터 최적 운영을 위한 수요 예측 방법 연구. *전자공학회논문지*, *60*(4), 110-115.

김정수, 장명균, 김주영 (2025). B2B 유통 분야 상품 추천 시스템 연구. *유통연구*, *30*(3), 79-101.

김혜선, 김효은, 김기훈 (2023). 디지털 피킹 시스템을 위한 연관 분석 기반의 상품 배치. *한국생산관리학회지*, *34*(3), 413-434.

대한상공회의소 (2024). *2024 유통산업 백서*. 대한상공회의소.

민경창, 하헌구 (2022). 머신러닝과 시계열 기법 기반의 초단기 시간단위 수요 예측방법론 개발 연구. *로지스틱스연구*, *30*(3), 41-55.

방선호, 이강현, 류형하, 임지원, 황지혜, 신광섭 (2024). 연관성분석을 통한 물류센터 출고 프로세스 효율성 향상에 관한 연구. *한국빅데이터학회지*, *9*(2), 149-164.

산업통상자원부 (2025). *2025년도 디지털유통물류센터 표준모델 확산사업 시행계획 공고*, 산업통상자원부.

이상일, 유영웅, 나동길 (2024). 중소유통기업지원을 위한 상품 카테고리 재분류 기반의 수요 예측 및 상품추천 방법론 개발. *산업경영시스템학회지*, *47*(2), 155-167.

정지철, 박주영 (2015). 중소유통공동도매물류센터의 경쟁력 강화방안에 대한 사례연구. *경영교육연구*, *30*(1), 93-119.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research, 15*(1), 3133-3181.

Guo, X., Yu, Y., Allon, G., Wang, M., & Zhang, Z. (2024). RiRiShun logistics: Home appliance delivery data for the 2021 manufacturing & service operations management data-driven research challenge. *Manufacturing & Service Operations Management*, *26*(4), 1358-1371.

Haindl, M., Somol, P., Ververidis, D., & Kotropoulos, C. (2006). Feature selection based on mutual correlation. In *Iberoamerican Congress on Pattern Recognition* (pp. 569-577).

Hyndman, R. J. & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679-688.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLOS ONE*, *13*(3), e0194889.

Salari, N., Liu, S., & Sehn, Z. J. M. (2022). Real-time delivery time forecasting and promising in online retailing: When will your package arrive? *Manufacturing & Service Operations Management*, *24*(3), 1421-1436.

Wan, X. (2022). Omnichannel distribution to fulfill retail orders. *Manufacturing & Service Operations Management*, *24*(4), 2150-2165.

Ye, T., Cheng, S., Hijazi, A., & Van Hentenryck, P. (2025). Contextual stochastic optimization for omnichannel multicourier order fulfillment under delivery time uncertainty. *Manufacturing & Service Operations Management, Forthcoming*.

Yu, L. & Liu, H. (2003). Feature selection for

high-dimensional data: A fast correlation-based filter solution. *Proceedings of the 20th International Conference on Machine Learning (ICML-03)* (pp. 856-863).

Feature Selection Strategies for Machine Learning-Based Demand Forecasting in Joint Distribution Centers for SMEs*

Minjeong Ahn**, Jiyoung Han***, Seunghwan Jung****, Injoon Noh*****

ABSTRACT

Purpose: This study aims to develop a machine learning－based forecasting model tailored to the operational characteristics of regional joint distribution centers used primarily by small retail stores. Unlike large-scale fulfillment centers, these hubs face highly irregular and low-volume demand at the SKU level, making accurate forecasting particularly challenging. Furthermore, important predictors including promotions and store-level operational details are not accessible in this setting, creating structural constraints for effective variable selection. To address these challenges, this study proposes a forecasting framework that leverages demand information from related product groups as proxy variables and evaluates two feature selection strategies—Wrapper and Filter methods—to identify a scalable and accurate prediction approach.

Research design, data, and methodology: Using real transaction-level sales data from a Korean regional distribution center covering January 2021 to December 2024, daily sales were aggregated into 209 sub-categories. A Random Forest model served as the common predictive algorithm across three conditions: (1) a benchmark model using only self-lagged features, (2) a Wrapper approach employing backward stepwise selection based on feature importance, and (3) a correlation-based Filter approach selecting top-N subcategory features via Pearson correlation. Model performance was evaluated using RMSE on a hold-out validation set, and total computational time, which includes feature selection and model training process, was compared across methods.

Results: The Filter (correlation-based) method significantly outperformed the Wrapper method in both computational efficiency and predictive accuracy. The Filter approach required 3 to 100 times less compu-

* This study was supported by the 2025 Korea Distribution Association and Korea Electronics Association.
This study was supported by the 2025 Ministry of Education and National Research Foundation of Korea (NRF-2025S1A5C3A02006968).
** Master's Student, Department of Business Administration, Yonsei University, First Author
*** Master's Student, Department of Business Administration, Yonsei University, First Author
**** Assosiate Professor, Department of Business Administration, Yonsei University, Corresponding Author
***** Assosiate Professor, Department of Business Administration, Korea University, Co-Author

tation time than the Wrapper method and produced lower RMSE, with optimal performance observed when using approximately 300 top-correlated features. Performance gains were particularly pronounced for subcategories exhibiting low sales volume, intermittent demand, and strong seasonal or co-movement patterns. In contrast, the Wrapper method struggled with high-dimensional feature spaces and failed to identify optimal feature sets within reasonable time.

Conclusions: The findings indicate that in joint distribution center environments, correlation-based Filter selection is highly suitable, particularly in logistics settings characterized by sparse, volatile, and interdependent product demand. This approach simultaneously delivers accuracy and computational efficiency—both essential for operational decision-making in regional distribution hubs. Overall, the study contributes a domain-specific, data-driven methodology that enables more reliable shipment forecasting, improves fulfillment efficiency, and provides an analytical foundation for developing digital fulfillment models to enhance the competitiveness of small and medium-sized retailers.

Keywords: Joint Distribution Center, Demand Forecasting, Machine Learning, Feature Selection