한국빅데이터학회지
제8권 제2호, 2023, pp. 149-163 https://doi.org/10.36498/kbigdt.2023.8.2.149

기계학습을 이용한 풀필먼트센터의 실시간 박스 추천에 관한 연구

A Study on the Real-time Recommendation Box Recommendation of Fulfillment Center Using Machine Learning

차대욱 · 조희연 · 한지수 · 신광섭 · 민윤홍†

인천대학교 동북아물류대학원

요 약

지속적인 이커머스 시장의 성장으로 풀필먼트센터가 처리해야 하는 주문량은 증가하였고, 다양한 고객 요구사항은 주문 처리의 복잡성을 높이고 있다. 이러한 추세와 함께 최근 인건비 증가로 인해 풀필먼트센터의 운영 효율성이 기업 경영 관점에서 더욱 중요해지고 있다. 본 연구는 풀필먼트센터의 출고 프로세스 중 포장 작업 영역에 적용 가능한 박스 추천을 중심으로 연구를 수행하였다. 박스 추천을 하기 위해 과거 실적 데이터를 기계학습 모형의 학습 데이터로 사용하였다. 상품 정보, 주문 정보, 포장 정보, 배송 정보 4가지 종류의 데이터를 전처리, 변수 가공 과정을 거쳐 기계학습 모델에 적용하였다. 입력 벡터로는 상품 규격 정보에 해당하는 width, length, height 3가지 특성을 사용하였으며, 상품의 실수 정보를 구간별 정수 체계로 변환하는 변수 가공 과정을 통해 입력 벡터의 특성을 추출하였다. 기계학습 모형별 성능을 비교한 결과 GradientBoosting 모델을 적용하였을 경우 21개의 구간으로 상품 규격 정보를 정수로 변환하였을 때 95.2%로 가장 높은 정확도로 예측을 수행함을 확인하였다. 본 연구는 풀필먼트센터에서 잘못된 박스 선택으로 인해 발생하는 물류비용의 증가와 박스 포장 소요 시간의 비효율을 줄이기 위한 방안으로 기계학습 모형을 제시하며, 상품 규격 정보의 특성을 효과적으로 추출하기 위한 변수 가공 처리 방식을 제안한다.

■ 중심어 : 풀필먼트센터, 기계학습, 박스추천

Abstract

Due to the continuous growth of the E-commerce market, the volume of orders that fulfillment centers have to process has increased, and various customer requirements have increased the complexity of order processing. Along with this trend, the operational efficiency of fulfillment centers due to increased labor costs is becoming more important from a corporate management perspective. Using historical performance data as training data, this study focused on real-time box recommendations applicable to packaging areas during fulfillment center shipping. Four types of data, such as product information, order information, packaging information, and delivery information, were applied to the machine learning model through pre-processing and feature-engineering processes. As an input vector, three characteristics were used as product specification information: width, length, and height, the characteristics of the input vector were extracted through a feature engineering process that converts product information from real numbers to an integer system for each section. As a result of comparing the performance of each model, it was confirmed that when the Gradient Boosting model was applied, the prediction was performed with the highest accuracy at 95.2% when the product specification

2023년 11월 24일 접수; 2023년 12월 06일 게재 확정.
* 본 연구는 국토교통부/국토교통과학기술진흥원의 지원으로 수행되었음(과제번호: RS-2022-00156324)
† 교신저자 (yunhong.min@inu.ac.kr)

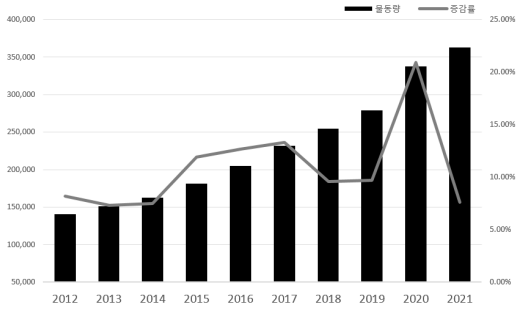
information was converted into integers in 21 sections. This study proposes a machine learning model as a way to reduce the increase in costs and inefficiency of box packaging time caused by incorrect box selection in the fulfillment center, and also proposes a feature engineering method to effectively extract the characteristics of product specification information.

■ Keyword : Fulfillment Center, Machine Learning, Box Recommendation

Ⅰ. 서론

국내외 이커머스 시장은 과거 10년간 지속적으로 상승하고 있으며, 최근 엔데믹 이후인 2022년에도 전년 대비 12%의 증가를 기록하며 이커머스 산업의 성장률은 증가하고 있다.[4] 이러한 이커머스 시장의 거래액 증가는 풀필먼트센터의 처리 물동량 증가로 연결된다.

(단위 : 만 박스)



〈그림 1〉 연간 택배 물동량 추이

(출처 - 한국통합물류협회(KILA) https://www.nlic.go.kr/nlic/parcelServiceLogistics.action)

이커머스 시장 성장과 함께 고객의 요구사항 또한 다양해지면서 이커머스 기업들은 고객에게 차별화된 구매 경험을 제공하기 위해 배송 서비스 고도화와 같은 전략을 취하고 있으며, 실시간 주문 처리를 통해 고객에게 빠른 배송과 같은 서비스를 제공하고 있다.

다양한 온라인 채널 및 해외 물품 거래 등의 증가로 인한 물동량 증가와 고객 서비스 다양화로 고객 주문을 빠르고 정확하게 처리할 수 있는 능력이 매우 중요하다.[2]

이러한 고객 요구사항에 대응하기 위한 이커머스 기업들의 전략은 주문을 처리해야 하는 풀필먼트센터의 주문 처리에 대한 난이도를 증가시키고 있으며, 풀필먼트센터의 운영비용의 부담 증가로 이어진다.

주문 처리에 대한 난이도 증가 외에도 지속적으로 상승하는 인건비로 인해 풀필먼트센터의 한정된 운영 시간 내 효율적인 작업에 대한 중요성이 높아지고 있다. 풀필먼트센터의 운영 시 발생하는 비용을 효율적으로 관리하지 못할 경우 재무 건전성의 악화로 이어져 기업 경영의 문제에 노출될 우려가 있다.

결과적으로 풀필먼트센터는 불특정 다수 고객의 서비스 만족도를 최대화하는 것을 목표로 물류 전반의 효율적인 운영을 통해 비효율을 최소화하여야 한다.

따라서 풀필먼트센터의 작업 비중이 높은 구간에서 개선 가능한 영역을 진단 및 개선이 필요하다. 풀필먼트센터 내 출고 작업은 전체 작업 중 약 75%의 비중을 차지하며, 이 중에서 피킹 후의 포장 단계에서 병목 현상이 가장 크게 발생한다.[2] 특히 풀필먼트센터의 출고 프로세스는 여러 가지 제약사항과 구성요소가 복잡하게 연관되어 문제가 발생해도 빠르게 식별하기 어렵고 현장 관리자의 주관적 판단에 의해 의사 결정이 이루어지고 있다.[3]

출고 프로세스의 비효율을 개선하기 위해 피킹 및 분류를 위한 다양한 장비와 자동화 설비를 도입하여 운영 효율을 높이려는 시도가 있지만, 출고 프로세스 중 포장 작업 영역은 자동화가 어렵고 인력 의존적인 작업이라는 한계가 있다.

또한, 합포장 작업 시 낱개 단위의 상품을 고려하여 적절한 포장 박스를 선택하는 것은 제품 수명 주기가 짧고, 다품종 소량의 주문이 주를 이루는 이커머스 산업의 특성상 선택의 어려움이 있다.

이러한 상황에서 포장 박스를 적절한 크기로 선택하지 못하면 과대 포장으로 인해 운송비용이 증가하고, 작은 박스를 선택하여 포장할 경우 재포장으로 인해 작업자가 제품을 포장하는 데 소요되는 시간이 증가하는 문제가 발생한다. 따라서 본 연구는 풀필먼트센터의 출고 프로세스 중에서도 포장 작업 분야에 활용 가능한 박스 추천을 통해 운영에 대한 효율을 증가시키기 위한 방안을 제시하고자 한다.

Ⅱ. 선행연구

2.1 박스 추천 관련 선행 연구

상품 체적과 무게를 고려하여 선정한 상품들로 최대 배낭 패킹 문제를 연구했다. 이 문제는 선택된 상품들을 배낭에 최대한 효과적으로 적재하여 적재율을 최소화하는 것을 목표로 한다. 주어진 배낭의 용량 내에서 상품을 선택하여 최대한 효율적으로 배치함으로써, 배낭의 적재율을 최소로 하는 문제를 다루었다.[6] 하지만 문제의 크기가 커질수록 계산 시간이 급격히 증가한다는 한계점을 언급하였다.

3차원 bin 크기 설계와 패킹 문제에 대한 해결을 위해, 그리고 해당 솔루션을 활용을 극대화하기 위해 혼합 생물 지리 기반 최적화(hybrid bio-geography-based optimization) 알고리즘을 제안하였다. 이는 공간 디자인에서 3차원 bin 크기를 효과적으로 계획하고 물체를 효율적으로 배치하는 방안으로 이를 통해 자원 활용을 극대화하고 설계된 공간의 효율성을 높이는 방안을 제시하였다.[8] 한계점으로는 하나는 최적의 bin(포장 상자) 개수를 고려하지 못하여, 적은 개수의 bin을 사용할 때 전체적인 효율성이 부족하다는 점이다. 이는 모델이 제한된 bin 개수로 인해 최적의 해결책을 찾지 못하는 한계를 언급하였다.

합포 주문을 효율적으로 포장하기 위하여, 주어진 화물을 포장할 수 있는 최소 포장 박스 크기의 도출을 목표로 하며 실제 물류 현장에 활용될 수 있도록 패킹 알고리즘의 수행시간을 고려하여 두 가지 발견적 기법을 제시하였다.[3] 한계점으로는 모의실험을 통해 데이터를 수집했기 때문에 실제 배송 과정에서 발생할 수 있는 다양한 요인들을 고려하지 못했다는 한계점을 언급하였다.

주어진 물품을 최소 수의 3차원 직사각형 bin에 포장하는 3차원 BPP 알고리즘 제안하였으며, 단일 bin, 3차원 bin 패킹 문제에 대한 알고리즘을 제안하여 최대 90개 항목의 인스턴스를 포함한 계산 결과로 많은 인스턴스가 합리적인 시간 제한 내에 해결될 수 있음을 검증하였다.[9] 한계점으로는 실제 데이터가 아닌 임의로 생성된 데이터를 사용했기 때문에 현실성이 떨어지는 점을 언급하였다.

최적화 알고리즘을 기반으로 하는 배낭 패킹 문제는 계산 소요 시간으로 실제 적용이 어렵다. 그러나 과거 데이터를 기계학습 모델의 학습 데이터로 활용할 경우 이러한 문제를 극복할 수 있다는 새로운 접근 방식을 제안하였다.[7] 연구에 활용한 입력 벡터로는 포장할 상품 수량, 포장할 상품의 총 무게, 근사 체적, 신선식품 포함 여부, 고객이 직접 픽업하는 경우 5개의 입력 벡터를 기계학습 학습 데이터로 사용하여 박스 사이즈를 예측하였다. 연구의 한계점으로는 체적 정보

를 알 수 없는 경우 중량에 기반한 근사 체적을 입력 벡터로 대체하는 방식으로 사용하였기 때문에 학습 데이터에 지나치게 과적합되어 있다는 점을 언급하였다.

2.2 선행 연구와의 차별성

기존 박스 추천 관련 선행 연구는 대부분 최적화 알고리즘 모형을 사용하였으며, 모형의 계산 소요 시간과 모의 데이터 사용 여부를 한계점으로 언급하였다. 본 연구는 일반적인 풀필먼트센터에서 수집 가능한 실적 데이터를 사용함으로써 모의 데이터의 한계를 극복하고자 하였다. 또한 모형의 계산 소요 시간에 대한 한계점을 극복하기 위해 기계학습을 활용하여 풀필먼트센터의 주문 처리 속도를 향상시킬 수 있는 방안으로 사용하는 휴리스틱 접근 방식이 아닌 기계학습을 활용한 박스 추천 관점으로 접근하고자 하였다.

또한 기계학습 모형에 기반한 선행 연구의 한계점으로 언급된 특정 화주사에 지나치게 과적합되는 문제를 해결하기 위해 수집 가능한 최소한의 변수인 상품의 가로, 세로, 높이 3개의 변수를 가공하는 방식을 적용하여 특성 추출 이후 기계학습 모형의 학습 데이터로 활용하였다.

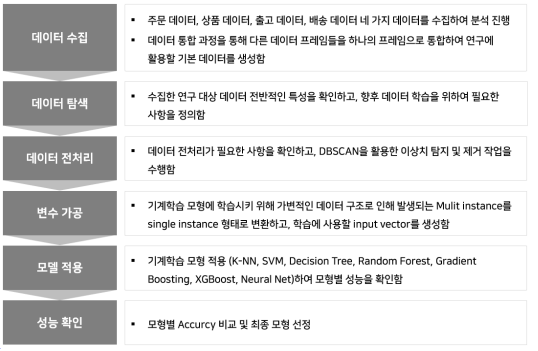
Ⅲ. 연구방법론

3.1 연구 절차

기계학습 모형에 학습 데이터로 사용하기 위해 전처리 과정을 거쳐 학습 가능한 데이터로 변환한다. 변환 과정에서 bin 개수를 활용한 변수 가공 방식을 적용하여 bin 개수가 1~30개일 경우 기계학습 분류 모델을 각각 성능을 비교한다.

학습 데이터 생성 이후 Boundary Based Model, Tree Based Model, Neural Net Based Model 3개의 하위 모델인 K-Nearest Neighbor, Support Vector Machine, Decision Tree, Random Forest, Gradient Boosting, XBGoost, Neural Net 7개의 모델을 통해 모델별 성능을 확인하고, 박스 크기를 예측하는 가능 성능이 좋은 최종 모델을 선정한다.

본 연구 <그림 2>와 같은 연구 절차와 내용으로 진행하였다.



〈그림 2〉 연구 절차

3.2 연구 대상 선정

풀필먼트센터에서 운영하는 셀러의 상품이 작은 상품 크기 또는 특정한 박스 사이즈에 편중되어 출고될 경우 이후 기계학습 학습 시 특정 데이터에 과적합 문제가 발생하게 된다. 학습된 데이터와 다른 유형의 데이터를 예측해야 하는 상황이 발생할 경우 모형의 정확도가 낮아지는 문제가 발생한다. 따라서 이후 연구 대상을 선정하기 위해 셀러의 취급 상품별 배송 현황을 확인한 결과는 <그림 1>과 같다.

종합 쇼핑몰의 경우 건강기능식품, 화장품, 완구류, 생필품을 취급하는 쇼핑몰로 상품 규격의 가로, 세로, 높이 분포를 확인할 경우 완구류, 화장품만을 취급하는 쇼핑몰에 비해 넓은 형태로 분포되어 있음을 확인할 수 있다. 상품별 규격 정보와 배송 건수를 확률 밀도 함수로 확인한 결과 완구류와 화장품에 비해 특정 영역에 밀집된 형태가 아닌 넓게 분포된 형태를 나타내는 것을 알 수 있다.

실제 완구류와 화장품만을 취급하는 쇼핑몰의 경우 극소 박스로 출고되는 비중에 98.7 %로 특정 박스 사이즈에 밀집되어 있다. 극소 비중이 높은 쇼핑몰을 대상으로 분류 모델의 성능을 확인하기에 부적합하여, 다양한 종류의 상품을 취급하는 종합 쇼핑몰을 대상으로 연구를 진행한다.

3.3 연구 모형 설계

연구 모형 설계는 3.1 연구 절차에 따라 진행한다.

3.3.1 데이터 수집

본 연구에 사용되는 데이터는 Online, Fulfillment Center, Delivery 세 개의 영역에서 생성되는 네 가지 종류의 데이터로 상품 정보, 구매 정보, 출고 정보, 배송 정보이며 데이터 수집 경로는 <그림 3>과 같다.

- 데이터 수집 기간 : 23.07~23.09 3개월 간 상품 정보 데이터, 주문 데이터, 출고 데이터, 배송 데이터를 수집
- 취급 상품 : 건강기능식품, 화장품, 완구, 애견용품

Online Store에 수집된 상품 정보와 구매 정보를 상품 코드 기준으로, Fulfillment Center 출고 정보와 Delivery 정보의 송장번호 기준으로 각각 통합하고 2개의 통합된 데이터는 주문 번호 기준으로 하나의 데이터 형태로 통합 작업을 수행한다.

이때 상품 정보의 width, length, height 정보와 order quantity를 input variable로 사용하고 배송 데이터의 box size column을 target variable로 설정하여 학습 데이터로 사용한다.

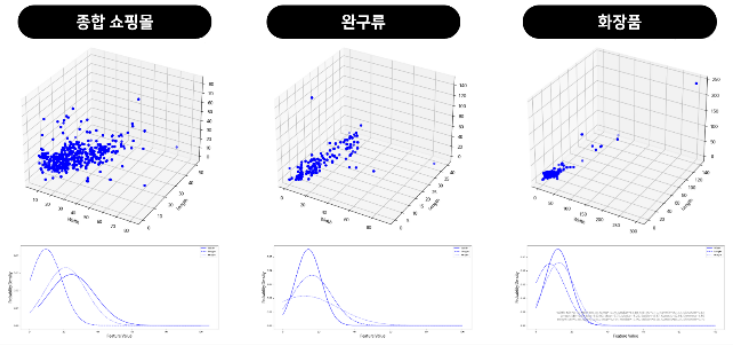
수집 이후 통합된 데이터 프레임 형태는 <표 1>과 같다.

주문번호 기준으로 통합된 데이터는 상품 품목을 기준, 품목별 주문번호 기준으로 구성되어 있으며, 고객의 구매 형태에 따라 가변적인 데이터 형태를 나타나게 된다. 가변적인 데이터 형태는 이후 기계학습 모델에 학습 데이터로 사용하기에 부적합하기 때문에 고정적인 크기를 갖는 데이터 형태로의 변환이 필요하다. 전처리 이후 변수 가공 과정을 통해 학습 가능한 데이터로 변환 작업을 수행한다.

3.3.2 데이터 탐색

데이터 탐색 과정을 통해 연구에 사용할 데이터 기본 현황을 분석한다. 연구의 Target variable로 사용할 배송 데이터 현황은 <표 2>와 같다.

target value는 총 19,403개로 다섯 개의 값으로 “극소”, “소”, “중”, “대형 #1”, “대형 #2”로 구성되어 있으며, 데이터 비중은 68.0%, 16.7%, 5.5%,



〈그림 3〉 쇼핑몰 상품 규격 및 배송 분포

〈표 1〉 통합 데이터 구조

order number	width	length	height	weight	product code	order quantity	invoice numver	box size
2387868462	9.5	9.5	12	0.17	B901324546043	1	510203501561	소
2023070196670211	26	3.5	24	0.58	B901324547517	1	510203501572	극소
2023070113517141	26	3.5	24	0.58	B901324547517	1	510203501583	극소
20230701-0000610	26	3.5	24	0.58	B901324547517	1	510203501594	극소
2023070112793481	26	3.5	24	0.58	B901324547517	1	510203501605	극소
3982718055	26	3.5	24	0.58	B901324547517	1	510203501616	극소
2023070194022131	26	3.5	24	0.58	B901324547517	1	510203501620	극소
20230702-0000418	5	3	15	0.5	B901324546063	1	510203501631	대1
2023070230992561	26	3.5	24	0.58	B901324547517	1	510203501642	극소
2143203509	26	3.5	24	0.58	B901324547517	1	510203501675	극소
20230630-0000177	8	5	24	0.25	B901324546969	1	510203501690	극소

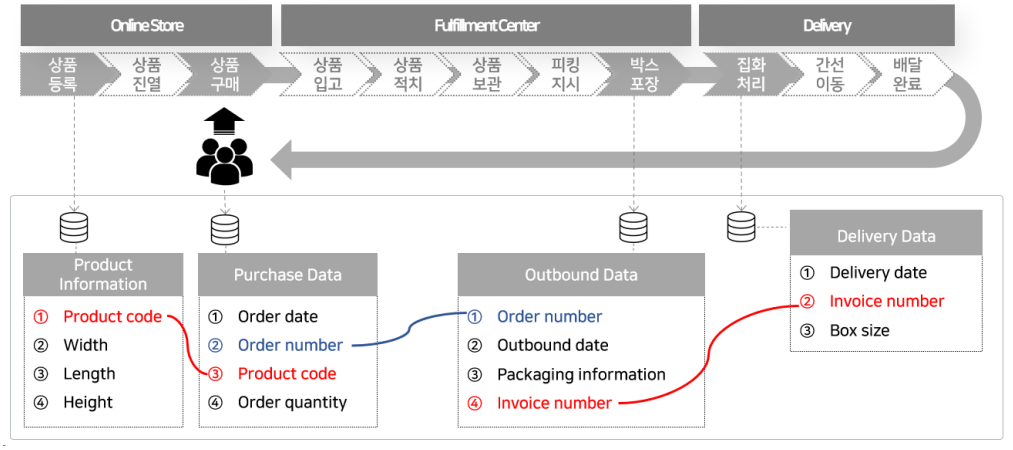
〈표 2〉 종합 쇼핑몰 배송 데이터 현황

box size	배송건수	비중	누계
극소	13,206	68.0 %	68.0 %
소	3,240	16.7 %	84.7 %
중	1,067	5.5 %	90.2 %
대형 #1	1,815	9.4 %	99.6 %
대형 #2	75	0.4 %	100.0 %
계	19,403	100 %	

9.4%, 0.4%로 “극소” 사이즈로 출고되는 비중이 68.0%로 가장 높은 비중을 차지하고 있으며, “대형 #2”로 출고되는 비중은 총 75건으로 0.4%로 가장 낮은 비중을 차지하는 것으로 나타난다.

“중” 사이즈로 배송되는 비중보다 “대형 #1” 사이즈로 출고되는 비중이 9.4%로 4% 높은 비중을 보이는 것으로 나타났는데, 이러한 배송 현황이 나타나는 이유를 확인하기 위해서는 주문 데이터 확인이 필요하며, 주문 데이터를 수량 기준으로 시각화한 결과 <그림 4>와 같다.

<그림 4>의 X축은 주문 단위 별 고객이 주문한 상품수량을 의미한다. X축 값이 1일 경우 고



〈그림 4〉 연구 데이터 수집 경로

객이 1개의 상품만 주문한 것을 의미하며, 총 7,321건 중 38%가 단포장으로 배송된 것을 알 수 있다. 2개 이상을 주문한 배송 건수는 12,082건으로 전체 배송 건수 대비 62%를 차지하는 것을 확인 할 수 있으며, 박스내 내품 수량이 9개 이상으로 구성되어 출고된 건수는 전체 대비 11.23%의 비중을 차지한다.

3.3.3 데이터 전처리

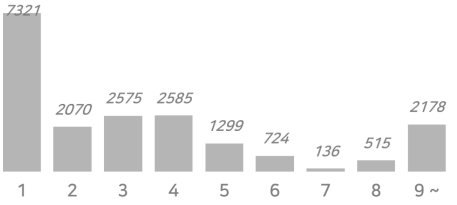
박스 크기를 예측하기 위해 본 연구에서 사용한 배송 데이터는 택배사에서 측정한 값으로 구성되어 있다. 택배사에서 측정한 박스 크기 값인 극소, 소, 중, 대형 #1, 대형 #2 값은 화주사와 정산 시 사용하기 위한 값으로 사용된다. 박스 크기 값은 실제 박스 규격과 중량의 최대값 기준으로 결정되며, 결정된 박스의 크기에 따라 박스 크기에 따라 배송 요금이 부과된다. 따라서 박스 중량 정보를 기준으로 박스 크기가 결정된 주문건을 제거하지 않은 상태로 박스 크기를 예측할 경우 배송 요금을 예측하는 모형이 될 수 있으므로, 중량을 기준으로 부과된 배송건의 주문의 전처리가 필요하다. 중량 기준으로 부과된 주문 건 외 박스 사이즈 결정 과정에서 발생되는 오류 또한 추가로 전처리가 필요하다.

중량 측정의 기준과 전산 처리 과정에서 발생하는 오류는 개별 확인 없이는 그 사실 여부를 파악하기 어렵다. 따라서 비지도 학습에 기반한 전처리 진행을 위해 DBSCAN 방식을 이용한 전처리 작업을 수행한다.

DBSCAN은 클러스터링에서 널리 사용되는 비지도 학습 알고리즘으로 데이터 셋에서 밀도가 높은 지역(군집)을 찾는다. 이후 데이터를 군집화하고 이상치를 감지하는데 활용한다.

본 연구에서는 가로, 세로, 높이 3차원 공간에 형성되는 밀도를 기반으로 이상치를 제거함으로써 중량을 기반으로 결정된 박스 크기와 전산의 오류로 인한 박스 크기 주문 건을 제거한다.

<그림 5>는 DBSCAN 방식을 적용하였을 경우 이상치 분포 현황이다.



〈그림 5〉 주문별 상품 수량

전체 주문데이터의 상품 규격(weight, length, height)을 대상으로 DBSCAN 적용 시 데이터 분포이다. 붉은 Data Point는 DBSCAN에 의하여 감지된 이상치를 의미하며, 전처리 단계에서는 밀도 기반의 군집으로부터 이탈된 독립된 Data Point는 weight 기반 측정된 값 또는 전산상 오류로 인한 값으로 가정하고, 해당 주문을 제거한다. 제거 전/후 현황은 <표 3>과 같다.

〈표 3〉 이상치 제거 전/후 배송건수 현황

박스크기	제거 전	이상치	주문건수	군집 수	제거 후
극소	13,206	665	5%	87	12,741
소	3,240	180	6%	22	2,660
중	1,067	52	5%	8	902
대형 #1	1,815	41	2%	9	1,774
대형 #2	75	5	7%	2	70
계	19,403	943	5%	128	18,460

인접한 Data Point 밀도 중심 기반으로 군집을 형성하는 비지도 학습 방법 중 하나인 DBSCAN 적용하였을 경우 총 128개의 군집이 형성되었으며, 이상치로 분류된 주문 건수는 총 943건으로 확인되었다. 전체 주문의 약 5%를 이상치로 제거하였으며 이후 18,460건을 대상으로 연구를 진행한다.

3.3.4 변수 가공 처리

3.3.3 데이터 전처리 과정을 통해 총 18,460개의 주문을 학습에 사용할 데이터 셋으로 선정하

였다. 그 중 네 개의 columns인 weight, length, height, order quantity를 input variable로 사용한다. 학습에 사용할 데이터의 구조는 <표 4>와 같다.

〈표 4〉 학습 데이터 구조

Order No.	Width	Length	Height	qty	Box size
Order #1	10	10	10	1	극소
Order #2	21.5	6.5	21.5	1	대형 #1
	21.5	6.5	21.5	1	대형 #1
	40.5	46	6.5	1	대형 #1
Order #3	40.5	46	6.5	1	극소
Order #4	38	15	23	1	소
	7.5	7.5	22	1	소
Order #5	7.5	7.5	22	1	극소
Order #6	17	5.5	12	1	대형 #1
	10	10	10	2	대형 #1
	10	10	10	6	대형 #1
	37	13.5	30.5	1	대형 #1

일반적인 온라인 주문 데이터는 품목별 주문번호를 기준으로 생성되며, 주문 단위로 묶음처리하였을 경우 <표 4>와 같은 데이터 형태로 이루어진다. width, length, height는 실수형 데이터로 상품에 대한 치수 정보를 담고 있으며, qty는 품목에 대한 주문 수량을 의미한다.

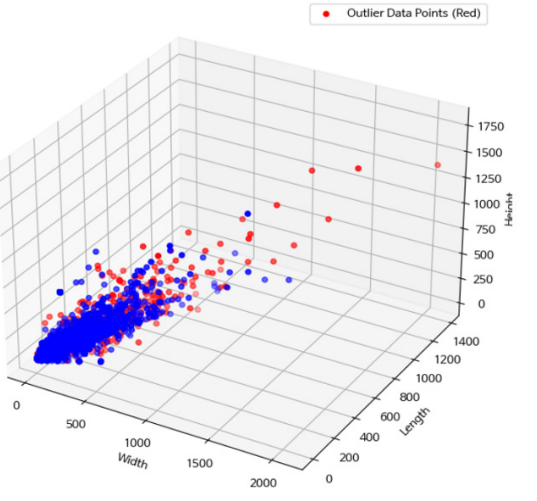
Order #6의 경우 고객이 총 4개의 품목을 주문하였으며, 각 품목별 1개, 2개, 6개, 1개 총 10개의 수량을 주문하였다. 이후 작업자의 상품 포장 작업을 거쳐 대형 #1 박스로 배송하여 최종 고객에게 배송되었음을 의미한다.

위와 같은 데이터 형태는 가변적인 속성 값을 갖고 있다. 주문을 수집할 때마다 주문의 형태는 다르기 때문에 실험 모형에 활용하기 위해서는 다중 인스턴스를 단일 인스턴스 형태로 고정된 속성값을 갖는 인스턴스 통합 과정이 필요하다.

본 연구에서는 인스턴스 통합 과정에서 Bin 개념을 활용하여 기존 실수 정보를 정수 형태로 변환하는 방식을 제안한다.

본 실험에서 제안하는 변수 가공 처리 방식은 기존 삼차원 공간 내 존재하는 상품 규격 정보를 설정한 bin의 개수에 의해 width, length, height 구간을 분할함으로써 삼차원 이상의 공간을 구성한다. 구간 분할 기준은 width, length, height 각각 상품 정보의 range(min, max)로 구간에 해당하는 경우 주문 수량을 더하는 방식이다.

변수 가공 개념도는 <그림 6>으로 order #2를 bin=3으로 설정하였을 때 처리 예시이다.



〈그림 6〉 DBSCAN 적용 시 이상치 현황

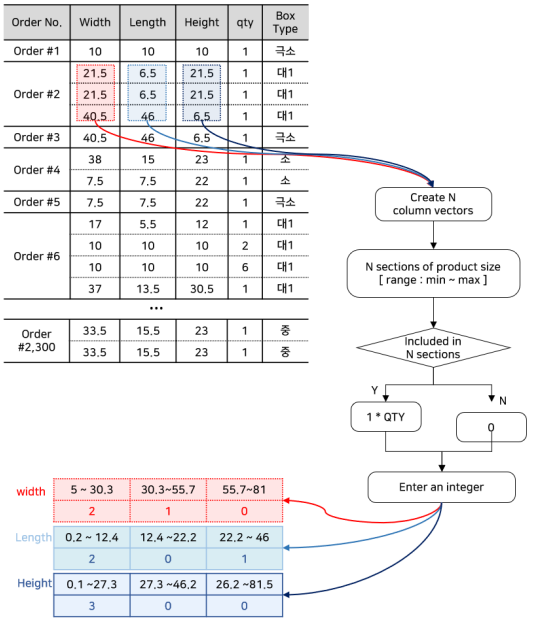
위 과정을 통해서 첫 번째로 다수의 품목으로 구성된 주문이 transpose됨으로써 단일 인스턴스 형태로 전환된다. 두 번째로 전체 상품 정보 범위를 bin의 개수로 분할하고, 분할된 컬럼에 해당할 경우 주문수량을 입력함으로써 기존 실수형 데이터가 정수 형태로 변환된다. 세 번째로 설정한 bin의 개수로 width, length, height 변수가 분할됨으로써 기존 삼차원 공간이 bin의 개수 * 3차원으로 차원이 확장됨으로써 고차원 데이터를 활용하여 모델의 예측 성능을 높일 수 있다.

Bin 개수에 따라 생성되는 학습 데이터는 <그림 7>과 같다. 본 연구에서는 bin의 개수를 임의

〈표 5〉 모형별 파라미터

Base model	model	parameter
Boundary Based Model	KNN	n=5
	SVM	kernel='linear', C=1.0
Tree Based Model	DT	max_depth=15 min_samples_split=7
	RF	n_estimators=500 max_depth=15 min_samples_split=7
	GRD	n_estimators=500 max_depth=15 min_samples_split=7
	XGBoost	n_estimators=500 max_depth=15 min_samples_split=7
Neural Net Based Model	NN	Hidden Layer 1: Neurons: 128, Activation Function: ReLU Hidden Layer 2: Neurons: 64, Activation Function: ReLU Output Layer: Activation Function: Softmax

로 1부터 30까지 증가시킴에 따라 모형별 성능을 확인한다. Bin이 3일 경우 총 9개의 입력변수가 생성되고, 4일 경우 총 12개의 입력 변수가 생성된다. 30일 경우에는 총 90개의 입력 변수가 생성된다.



〈그림 7〉 변수 가공 처리를 위한 구성도

변수 가공 처리 과정을 통해 생성된 단일 인스턴스 데이터셋을 이후 기계학습 모형에 적용한다.

3.3.5 모델 적용

연구 절차에 의해 생성한 학습 데이터는 비선형 구조의 데이터로 5가지의 카테고리를 예측하는 분류 문제로 정의할 수 있다. 따라서 bin의 개수가 증가함에 따라 가장 높은 성능을 보이는 모델을 선정하기 위해 Boundary Based Model, Tree Based Model, Neural Net Based Model 3개의 각기 다른 그룹을 기준으로 7개의 하위 모델을 선정하였으며, 모형별 특징은 다음과 같다.

3.3.5.1 적용 모형

Boundary Based Model은 Support Vector Machine (SVM)과 K-Nearest Neighbors (KNN)와 같이 결정 경계에 기반한 모델을 선정하였다. SVM은 데이터를 분리하는 최적의 결정 경계를 찾으며, KNN은 주변 이웃들의 경계에 따라 예측을 수행한다.

Tree Based Model은 Decision Tree, Random Forest, Gradient Boosting, 그리고 XGBoost와 같은 트리 기반 모델을 선정하였으며, 이 모델들은 트리 구조를 사용하여 데이터를 분할하고 예측을 수행한다. 세부 모형의 특징을 살펴보면 Decision Tree는 특징에 따라 데이터를 분할하여 특정 조건에 따라 데이터가 좌우로 분류된다. Random Forest는 여러 개의 의사결정 트리를 구성하고 각 트리의 예측을 평균하여 과적합을 줄이는 모델이다. Gradient Boosting은 약한 모델을 순차적으로 학습시켜 강력한 앙상블 모델을 만드는 방식으로 동작한다.

XGBoost는 Gradient Boosting을 기반으로 하며, 효율적인 부스팅 알고리즘과 특징 중요도 추정 기능을 제공하는 모델이다.

Neural Net Based Model로는 Neural Network (신경망) 모델을 선정하였으며, 신경망은 인공 뉴런과 그들의 연결로 구성된 모델로, 복잡한 비선형 관계를 학습할 수 있다.

비교 모형별 동일한 기준의 파라미터 값을 사용하였으며, 모형별 설정한 파라미터는 <표 5>와 같다.

모형별 파라미터 외 7개의 모형 모두 train, test set을 7:3으로 동일한 기준을 적용하였다.

3.3.6 실험 결과 및 분석

앞에서 설계한 모형에 대한 성능을 세 단계로 실험 결과를 확인한다. 첫 번째 bin의 개수가 증가함에 따라 모형별 성능 비교하고, 두 번째 DBSCAN 적용 전/후 모형별 성능 비교한다. 세 번째 최종 선정한 모형의 confusion matrix 확인을 통한 최종 모형의 성능을 확인한다. 첫 번째로 bin의 개수가 증가함에 따라 모형별 성능을 비교할 경우 <표 6>과 같다.

Boundary Based Model 기반 모델 중 K-Nearest Neighbors (K-NN)와 Support Vector Machine (SVM) 두 모델을 비교한 결과를 비교하였을 경우 K-NN 모델에서는 14개의 bin을 사용하였을 때 정확도가 92.7%이며, 반면 SVM 모델에서는 bin의 개수가 30일 때, 정확도가 92.3%로 측정되었다.

〈표 6〉 BIN 개수 증가에 따른 모형별 성능 비교

bin	Accuracy						
	KNN	SVM	DT	RF	GRD	XGB	NN
1	0.704	0.696	0.705	0.705	0.706	0.705	0.697
2	0.855	0.868	0.876	0.878	0.875	0.877	0.869
3	0.855	0.806	0.862	0.867	0.862	0.865	0.847
4	0.873	0.867	0.894	0.893	0.893	0.894	0.880
5	0.904	0.845	0.921	0.923	0.918	0.923	0.904
6	0.920	0.897	0.927	0.935	0.931	0.934	0.919
7	0.909	0.876	0.936	0.936	0.936	0.940	0.923
8	0.924	0.892	0.934	0.940	0.940	0.946	0.927
9	0.925	0.894	0.934	0.938	0.940	0.946	0.923
10	0.923	0.897	0.939	0.941	0.945	0.947	0.932
11	0.924	0.887	0.932	0.939	0.945	0.947	0.930
12	0.924	0.903	0.940	0.940	0.946	0.951	0.930
13	0.924	0.896	0.940	0.938	0.947	0.952	0.936
14	0.927	0.901	0.940	0.938	0.950	0.953	0.941
15	0.927	0.909	0.938	0.939	0.949	0.951	0.940
16	0.921	0.898	0.937	0.936	0.950	0.954	0.939
17	0.922	0.911	0.925	0.933	0.947	0.950	0.937
18	0.919	0.917	0.932	0.937	0.946	0.949	0.932
19	0.923	0.905	0.936	0.933	0.950	0.951	0.940
20	0.910	0.912	0.926	0.931	0.943	0.949	0.934
21	0.925	0.916	0.939	0.934	0.952	0.952	0.938
22	0.917	0.909	0.932	0.930	0.949	0.952	0.935
23	0.924	0.907	0.936	0.930	0.948	0.953	0.933
24	0.915	0.917	0.936	0.933	0.948	0.954	0.935
25	0.920	0.910	0.923	0.928	0.944	0.948	0.934
26	0.921	0.913	0.934	0.933	0.950	0.952	0.935
27	0.925	0.920	0.932	0.931	0.950	0.950	0.933
28	0.916	0.918	0.932	0.927	0.950	0.954	0.938
29	0.924	0.916	0.934	0.919	0.952	0.953	0.941
30	0.925	0.923	0.935	0.930	0.947	0.954	0.938

Tree Based Model인 DT는 bin의 개수가 12개일 경우 94.0%, RF의 경우 bin의 개수가 10개일 때 94.1%, GRD의 경우 bin의 개수 21개일 때 95.2%, XGBoost의 경우 bin 24개일 때 95.4%로 측정되었다.

Tree Based Model로서 Decision Tree (DT), Random Forest (RF), Gradient Boosting (GRD), 그리고 XGBoost 모델들의 정확도를 실험한 결과 Decision Tree에서는 12개의 bin일 경우 94.0%의 정확도를 나타냈으며, Random Forest는 10개의 bin에서 94.1%의 정확도를 확인하였다.

Gradient Boosting은 21개의 bin을 사용하여 95.2%의 정확도를 나타내었고, XGBoost는 24개의 bin에서도 동일한 95.2%의 정확도를 확인하였다. Neural Net Based Model 중 기본 모델인 Neural Net의 경우 bin의 개수가 29개일 때 94.1% 정확도를 나타내었다. 전체 모델을 비교하였을 경우 Tree 기반 모델 중 Gradient Boosting과 XGBoost 모델이 95.2%, 95.4%의 성능으로 가장 높은 정확도를 나타남을 확인하였다.

두 번째 각 모형별 가장 높은 성능을 보이는 bin 개수와 DBSCAN 적용을 통한 이상치 제거 전/후 모형별 성능을 비교할 경우 <표 7>과 같다.

〈표 7〉 이상치 제거 전/후 모형별 성능 비교

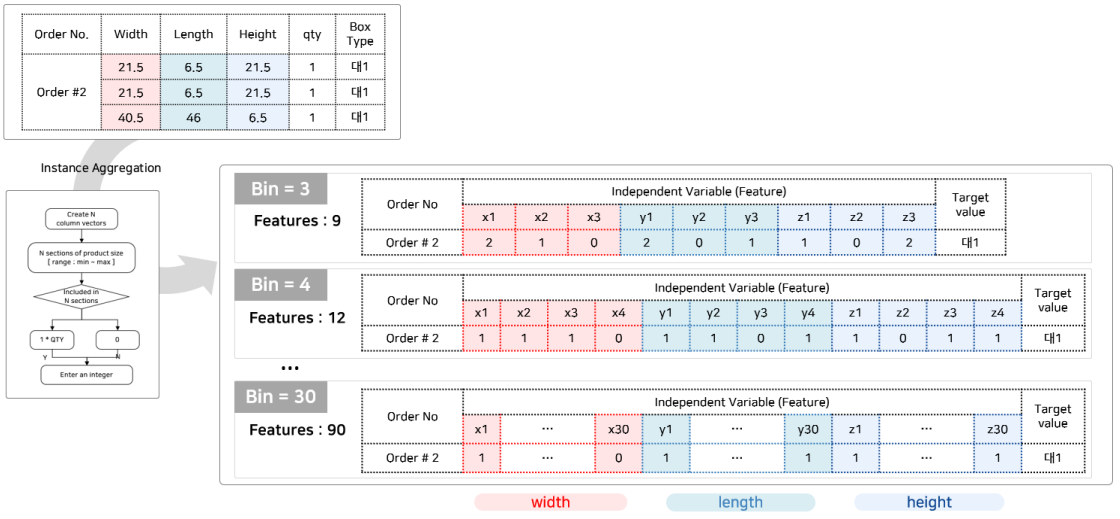
단일 인스턴스, 변수 가공 처리 이후 학습 데이터

model	이상치 제거 전		이상치 제거 후	
	BIN	accuracy	BIN	accuracy
KNN	15	0.901	15	0.927
SVM	28	0.909	30	0.923
DT	15	0.922	12	0.942
RF	16	0.934	10	0.941
GRD	14	0.94	21	0.952
XGBoost	16	0.94	24	0.954
NN	30	0.932	14	0.941

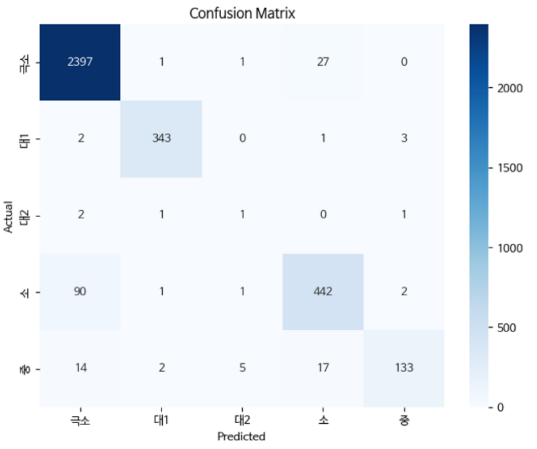
〈표 8〉 CLASS별 예측 성능 비교
(Gradient Boosting)

구분	극소	소	중	대형 #1	대형 #2
Recall	0.988	0.819	0.777	0.982	0.166
Precision	0.956	0.903	0.956	0.985	0.125

전체적으로 이상치를 제거하였을 경우 1~2% 정도 모형의 성능이 증가하는 것을 확인할 수 있다. Gradient Boosting과 XGB 모델 중 보다 적은 bin의 개수로 높은 정확도를 나타내는 Gradient Boosting모델을 최종 모형으로 선정한다. <그림 8>와 <표 8>는 최종 모형으로 선정한 Gradient



〈그림 8〉 변수 가공 처리 이후 학습 데이터 셋



〈그림 9〉 Confusion Matrix (Gradient Boosting)

Boosting 모델의 21개의 bin 구간에서의 confusion matrix와 박스 크기 별 recall, precision이다.

대형 #2 클래스의 경우 recall, precision을 확인한 결과 12~16%로 매우 낮은 것으로 확인하였다. 대형 #2의 클래스의 경우 출고 빈도가 낮은 문제로 타 클래스에 비해 관측 빈도가 낮다. 따라서 대형 #2의 경우 오버 샘플링을 통해 정확한 분류를 할 수 있도록 추가 학습이 필요하다.

재현율(recall)을 확인한 결과 출고 빈도가 낮은 대형#2의 경우를 제외하고, 극소와 대형 #1이 양극단 클래스로 가정할 경우 98.8%, 98.2%로 두 극단 클래스에서 양성 여부를 효과적으로 감지하고 있다. 소, 중 클래스로 넘어갈수록 성능이 떨어지는 이유는 상품의 길이에 의해 포장이 결정되는 경우로 인해 발생한 것으로 해석된다.

따라서 width, length, height 외에도 width + length, length + height, width + height 각 변수의 조합을 추가로 고려한 새로운 변수를 생성하여 분류 성능을 평가하는 추가 연구가 필요할 것으로 판단된다. 새로운 변수들은 기존 변수들의 조합을 통해 미처 고려되지 않은 특징을 도출하고, 이를 통해 양 극단 클래스(극소, 대형 #1) 외 중간 클래스 재현율을 향상시킬 수 있을 것이다.

정밀도(precision)를 확인한 결과 대형 #2, 극소를 제외하고 재현율에 비해 높은 성능을 보이는 것으로 확인된다. 하지만 정밀도의 예측 성능은 오분류로 인해 발생되는 비대칭 비용(Asymmetric Cost)과 직결된다. <표 9>는 박스 크기별 추가되는 운임 요금이고, <표 10>은 <그림 8> confusion matrix를 재구성한 것이다.

예를 들어, <표 10>의 파란색 하이라이트 좌하단 영역의 오분류는 과대 포장 되었을 경우로 <표 9>의 운임을 고려하였을 때 <표 11> 좌 하단 영역만큼의 오분류로 인한 추가 운임이 발생하게 된다.

<표 10>의 파란색 하이라이트 우 상단 영역의 오분류로 인해 발생하는 비용이 재포장 시간 15

〈표 9〉 박스 사이즈별 추가 운임

구분	극소	소	중	대형 #1	대형 #2
추가 운임(원)	0	400	1,100	3,000	1,0900

(출처 : 카페24 창업지원센터 공시자료)
https://soho.cafe24.com

〈표 10〉 confusion matrix 재구성

actual					(단위 : 주문건수)
극소	2397	27	0	1	1
소	90	442	2	1	1
중	14	17	133	2	5
대형 #1	2	1	3	343	0
대형 #2	2	0	1	1	1
	극소	소	중	대형 #1	대형 #2

predicted

〈표 11〉 오분류로 인한 추가 운임

actual					(단위 : 원)
극소		2,165	0	80	80
소	36,000		160	80	80
중	15,400	18,700		160	401
대형 #1	6,000	3,000	9,000		0
대형 #2	21,800	0	10,900	3,000	
	극소	소	중	대형 #1	대형 #2

predicted

초로 가정할 경우 최저 임금 9,920원(2023년 최저임금위원회 기준) 적용 시 <표 11>의 우상단 영역만큼의 비대칭 비용이 발생하게 된다.

“극소”의 경우 <표 10>의 confusion matrix 표를 통해 잘못 예측한 빈도를 확인해 보면 소(90) > 중(14) > 대형 #1, 대형 #2(2) 순으로 소를 다른 상자 크기로 잘못 분류하였음을 확인할 수 있다. 하지만 비용을 고려하였을 경우 극소 > 대형 #2 > 중 > 대형 #1 순서로 오분류로 인한 비대칭 비용 손실이 발생하였으며, 이러한 비용 손실을 고려하였을 때를 고려하였을 때 정확한 예측이 필요한 클래스의 우선순위가 변경됨을 확인할 수 있다.

Ⅳ. 결론

4.1 연구 결과

본 연구에서는 풀필먼트센터에서 상자 크기를 예측하기 위한 실험을 진행하였다. 네 가지 데이터 유형을 수집하여 최소한의 변수를 활용한 상자 크기 예측 성능을 7가지 모형별로 성능을 비교하였다.

연구 절차는 데이터 수집, 탐색 과정을 거쳐 DBSCAN 방식을 이용한 밀도 기반의 전처리를 진행하였고 그 결과 수집된 전체 주문의 5%를 이상치로 판단하여 제거하였다. 이후 변수 가공 처리 과정을 거쳐 만들어진 학습 데이터를 각 모형별로 학습을 진행하였고 이후 성능을 비교하였다.

성능 비교 결과, Gradient boosting 및 XGBoost 모델이 다른 모델에 비해 높은 95% 수준의 정확도를 보였다. 특히 Gradient boosting 모델은 더 적은 bin을 사용하면서도 유사한 성능을 나타내어 최종 모델로 선정하였다.

본 연구 결과를 통해 상품 규격 정보를 정수로 처리하고 해당 주문 수량을 나타내는 데이터로 변환함으로써 주문의 특성을 유지하면서, 이러한 기계학습 아키텍처를 통해 실시간으로 상자 추천을 처리할 수 있음을 확인하였다.

이는 포장을 위한 박스 추천 용도 외에도 풀필먼트센터에서 피킹 지시서 생성 시 박스 포장 작업을 고려한 주문의 분류, 전체 출하 물량의 크기를 예측을 통한 배차 용도로도 활용 가능함을 시사한다.

4.2 연구의 한계 및 향후 연구 방향

4.2.1 제품 특성에 대한 사항 미고려

최소한의 정보만을 이용하여 학습 데이터로 활용하는 과정에서 제품의 특성을 반영하지 못하였다는 한계점이 있다. 예를 들어, 파손되기 쉬운 상품의 경우 에어캡 포장과 같은 특별한 처리가 필요하지만 상품 규격 정보만을 이용하기 때문에 포장 시 고려가 필요한 물성을 반영하지 못하였다.

상품의 특성을 반영하기 위해서는 제품 특성 정보를 데이터화하는 작업이 필요하다. 하지만 제품의 수명주기가 짧은 이커머스 산업의 특성으로 인해 제품 특성에 관한 정보를 매번 수동으로 업데이트하기에 힘들다는 현실적인 문제가 있다. 이러한 문제를 해결하기 위해 상품 설명 또는 키워드 기반으로 제품 특성을 데이터화하기 위한 동적 태깅을 시스템에 관한 연구가 필요하며, 이를 통해 포장 시 고려해야 할 제품 특성에 관한 사항을 생성하고 추가적인 모델 설계를 통해 반영할 수 있을 것이다.

4.2.2 비대칭 비용(Asymmetric Cost) 미고려

최적 모형 선정 과정에서 정확도(accuracy) 지표를 이용하여 모델을 선정하였기 때문에, 클래스별 예측 성능을 높이기 위한 파라미터 튜닝 과정은 진행하지 못하였다. 이는 정확도만을 지표로 사용하였기 때문에 클래스 간에 발생하는 예측 오류에 대해 동일한 비용을 적용한 것으로도 해석 가능하다. 하지만 실제 오분류를 하게 되는 경우의 비용 차이를 고려하지 않으면, 모델이 특정 클래스에 대한 잘못된 예측 수행하였을 때 발생하는 손실을 충분히 반영하지 못하게 된다. 이러한 문제를 해결하기 위해 오분류 시 발생하는 비대칭 비용을 고려한 최적 모델의 선정이 필요하다. 모델의 하이퍼파라미터 튜닝 과정에서, 클래스 불균형이나 민감도를 조절할 수 있는 파라미터를 고려하여 모델이 학습할 때 특정 클래스의 중요성을 부여할 수 있을 것이다.

4.2.3 포장 효율성의 불확실성

실제 박스 규격 정보 기반이 아닌 다섯 가지 종류의 단순 상자 크기만 예측하였다.

과거 데이터를 활용하여 연구를 진행하였으며, 기존 처리된 실적 데이터가 최적의 선택이라는 전제를 가정을 두었다.

제품 조합별 상자 추천에 대한 고려가 이루어지지 못하였기 때문에 추천된 박스가 효율적인지는 확인이 어려운 점에서 한계가 있다. 이러한 효율성에 관한 부분을 확인하기 위해 DEA와 같은 방법론을 활용하여 효율성의 기준을 설계하고, 효율성 결과를 검토하면서, 상품 단위가 얼마나 효율적으로 포장되었는지 측정할 수 있다. 이러한 접근을 통해 추천된 박스의 효율성을 평가하고, 효율성을 고려한 예측 모델로 조정 가능할 것이다.

따라서 본 연구의 한계점으로 언급된 세 가지 사항에 대한 부분을 반영하여, 추가적으로 연구를 진행할 예정이다.

참 고 문 헌

[1] 권요한, “화물 적재 알고리즘에 관한 연구”, 석사학위논문, 2018.

[2] 김창현, “기계학습을 이용한 풀필먼트센터의 출고 운영의 위험 관리”, 박사학위논문, 2020.

[3] 류민지, “주문별 화물크기를 고려한 3D 화물 패킹 알고리즘 연구”, 석사학위논문, 2020.

[4] 문보영, “VFA 기법을 활용한 풀필먼트센터 운영 효율화 Case Study”, 박사학위논문, 2023.

[5] 이영종, “온라인 유통시장 성장에 따른 빠른 배송서비스 확산. 우정정보 정기간행물, 2019 여름여름(117), 23-37, 201.

[6] Fabio Furini, Ivana Ljubic, Markus Sinnl, “An effective dynamic programming algorithm for the minimum-cost mzximal knapsack packing prob-

lem", European Journal of Operational Research, 2017.

[7] Michael Heininger, Ronald Ortner, "Predicting packaging Sizes Using Machine Learning", Operations Research Forum, 2022.

[8] Mingzhou Chen, Jiazhen Huo, Yongrui Duan, "A hybrid biogeography-based optimization algorithm for three-dimensional bin size designing and packing problem", Computers & Industrial Engineering, 2023.

[9] Silvano Martello, David Pisinger, Daniele Vigo, "The Three-Dimensional Bin Packing Problem", Operations Research, 2000.

저 자 소 개

차 대 욱(Dae-Wook Cha)

· 2016년 2월: 한국교통대학교 도시교통공학과 (공학사)
· 2019년 2월: 인천대학교 동북아물류대학원 물류시스템학과 (공학석사)
· 2020년 3월~현재: 카페24 SCM팀
· 2022년 3월~현재: 인천대학교 동북아물류대학원 융합물류시스템학과 (박사과정)

<관심분야> 머신러닝, 딥러닝, 빅데이터 분석

조 희 연(Hui-Yeon Jo)

· 2022년 2월: 인천대학교 중어중국학, 물류학 (학사)
· 2022년 3월~현재: 인천대학교 동북아물류대학원 융합물류시스템학과 (석사과정)

<관심분야> 물류, SCM, 최적화

한 지 수(Ji-Soo Han)

· 2023년 2월: 안양대학교 글로벌경영학과 (학사)
· 2023년 3월~현재: 인천대학교 동북아물류대학원 융합물류시스템학과 (석사과정)

<관심분야> SCM, 최적화

신 광 섭(Kwang-Sup Shin)

· 2003년 2월: 서울대학교 산업공학과 (공학사)
· 2006년 2월: 서울대학교 산업공학과 (공학석사)
· 2012년 2월: 서울대학교 산업공학과 (공학박사)
· 2012년 2월~현재: 인천대학교 동북아물류대학원 교수

<관심분야> 빅데이터 활용, 솔루션

민 윤 홍(Yun-Hong Min)

· 2006년: 포항공과대학교 산업경영학과 (학사)
· 2012년: 서울대학교 산업공학과 (공학박사)
· 2012~2017년: 삼성종합기술원
· 2017년~현재: 인천대학교 동북아물류대학원 교수

<관심분야> 최적화, 인공지능