Journal of Korean Society of Industrial and Systems Engineering
https://doi.org/10.11627/jksie.2023.46.2.176

Vol. 46, No. 2 : 176-182, June 2023
ISSN : 2005-0461 / EISSN : 2287-7975

Machine Learning Model for Recommending Products and Estimating Sales Prices of Reverse Direct Purchase

Kyu Ik Kim · Berdibayev Yergali · Soo Hyung Kim · Jin Suk Kim†

Neoforce Co., Ltd.

역직구 상품 추천 및 판매가 추정을 위한 머신러닝 모델

김규익 · 베르드바에브 예르갈리 · 김수형 · 김진석†

주식회사 네오포스

With about 80% of the global economy expected to shift to the global market by 2030, exports of reverse direct purchase products, in which foreign consumers purchase products from online shopping malls in Korea, are growing 55% annually. As of 2021, sales of reverse direct purchases in South Korea increased 50.6% from the previous year, surpassing 40 million. In order for domestic SMEs(Small and medium sized enterprises) to enter overseas markets, it is important to come up with export strategies based on various market analysis information, but for domestic small and medium-sized sellers, entry barriers are high, such as lack of information on overseas markets and difficulty in selecting local preferred products and determining competitive sales prices. This study develops an AI-based product recommendation and sales price estimation model to collect and analyze global shopping malls and product trends to provide marketing information that presents promising and appropriate product sales prices to small and medium-sized sellers who have difficulty collecting global market information. The product recommendation model is based on the LTR (Learning To Rank) methodology. As a result of comparing performance with nDCG, the Pair-wise-based XGBoost-LambdaMART Model was measured to be excellent. The sales price estimation model uses a regression algorithm. According to the R-Squared value, the Light Gradient Boosting Machine performs best in this model.

Keywords : Learning To Rank, XGBoost, Recommendation System

1. 서 론

2030년까지 전세계 경제의 약 80%가 글로벌 시장으로 전환될 전망이고, 해외 소비자가 대한민국의 온라인 쇼핑몰에서 상품을 구매하는 역직구(reverse direct purchase) 상품 수출액이 연평균 55%씩 성장하고 있다. 2021년 기준 대한민국의 역직구 판매는 4,000만 건을 돌파하여 전년도 2,689만 건 보다 50.6% (1,361만 건)가 증가하였다. 역직구 시장의 급격한 증가는 코로나19 시대에 소비자의 비대면 상품 구매에 대한 선호도 증가와 국가간 경계가 점점 희석되어가는 글로벌 경제시대 온라인 소비패턴의 활성화에 기인한다. 특히, 대한민국의 역직구 시장은 한류 열풍이 시작된 2010년 이후 한류 콘텐츠가 글로벌 시장으로 뻗어 나가기 시작하면서 화장품, 패션, 액세서리 등 한류와 연관된 상품에도 관심이 집중되고 판매량의 증가로 이어졌다[5].

온라인 소비 트랜드의 가속화로 글로벌 대형유통업체들은 역직구 플랫폼을 정비하고 판매망을 넓히는 등 영토 확장에 나서고 있으며, 소규모 셀러도 참여가 늘고 있어 2020년 아마존 등 글로벌 전자상거래 플랫폼에 신규 등록한 대한민국의 온라인 판매 업체는 10만 8,724개로 전년(7

Received 31 May 2023; Finally Revised 22 June 2023;
Accepted 22 June 2023
† Corresponding Author : 0210kimjs@daum.net

만 9,033개) 대비 38% 증가하였다. 국내 온라인 유통업체가 해외시장에 진출하기 위해서는 다양한 시장 분석 정보를 기반으로 수출 전략을 마련하는 것이 중요하지만 국내 중소 셀러의 경우 해외시장 정보 부족, 현지 선호 상품 선택과 경쟁력 있는 판매가격 결정의 어려움 등 진입 장벽이 높은 실정이다[2, 3, 4, 10].

본 연구는 글로벌 시장 정보 수집과 활용 능력이 떨어지는 중소셀러에게 글로벌 쇼핑몰 판매정보와 상품 소비 트렌드를 수집·분석하여 유망한 수출 상품 및 적정 상품 판매가격을 제시하는 마케팅 정보를 제공하기 위하여 AI를 기반으로 하는 상품 추천 및 판매가 추정 모델을 제시한다[1].

2. 이론적 배경

2.1 상품추천 모델

본 연구에서 상품추천에 적용한 머신러닝 모델인 LTR (Learning to Rank) 기법은 최근에 검색 및 추천 분야에서 사용되고 있으며, 지도 학습 기반의 머신러닝 기법을 활용하여 정보 간의 순서를 정하는 방법론이다. 기존의 전통적인 지도학습이 단일 인스턴스에 대해 특정 분류 혹은 단일 값을 생성하는 방법이라고 하면, LTR은 개체들 사이의 연관성을 바탕으로 최적의 순서를 정하는 방법이라고 할 수 있다. 즉, 주어진 질의와 아이템 간의 연관성 점수를 손실함수를 통해 산출하여 아이템의 순서를 정하는 방법이다[6].

일반적으로 LTR에서 손실함수를 계산하는 접근법은 Point-wise, Pair-wise, List-wise로 나뉘며, 사용하는 접근법에 따라 서로 다른 방식으로 순위를 정한다. 각 손실함수를 사용하는 접근법의 개념은 다음과 같다.

Point-wise는 한 번에 한 개의 아이템만 고려하여 점수를 계산하고 목록을 정렬하는 방법이다. 이 방법은 가장 단순한 방법이나 목록의 전체 정보를 온전히 활용하지 못한다는 단점이 있다. 대표적으로 회귀 알고리즘이 있다.

Pair-wise는 한 번에 한 쌍의 아이템을 고려하여 둘 사이의 비교를 통해 최적의 순서를 도출하는 방법이다. 이 방법은 Point-wise보다 성능이 좋다는 장점이 있다. 대표적으로 RankNet, LambdaRank 등의 알고리즘이 있다.

List-wise는 목록의 전체 정보를 활용하여 최적의 순서를 도출하는 방법이다. 앞서 설명한 다른 방법에 비해 복잡하지만 좋은 성능을 기대할 수 있다. 대표적으로 SoftRank, ListNet, AdaRank, LambdaRank 등의 알고리즘이 있다.

LTR에서 사용하는 대표적인 머신러닝 알고리즘은 LambdaMART이며, Pair-wise, List-wise 접근법을 모두 사용할 수 있는 알고리즘이다. 이전부터 RankNet, LambdaRank 및 LambdaMART는 실제의 랭킹 문제를 해결하기 위한 알고리즘으로 성능이 입증되었다. 그 중에서도 LambdaMART의 경우 RankNet을 기반으로 하는 LambdaRank에 Gradient Boosted Decision Tree를 사용한 버전으로 기존의 LambdaRank보다 훨씬 뛰어난 성능을 보여준다[9].

XGBoost(Extreme Gradient Boostring)는 Gradient Boosted Decision Tree 알고리즘의 성능을 향상시키기 위해 과적합(Overfitting) 방지를 위한 파라미터가 추가된 알고리즘이다. 따라서 다른 Gradient Boost알고리즘보다 학습 속도가 빠른 것이 특징이다. 현재 XGBoost 라이브러리는 LTR에서 언급한 세 가지 접근법에 있어서의 랭킹 문제를 학습하기 위해 사용된다[7].

2.2 판매가격 추정 모델

국가별 현지 상품 최적 판매가 예측을 위한 모델로는 회귀모델(Regression)을 적용하였다.

그리고 현지 판매가격을 Luxury, Premium, Regular, Economy 4개 등급으로 나누어 추정하고 있으며 K-Means Clustering 모델을 사용하였다.

K-Means Clustering은 일반적으로 사용되는 분할 군집 기법으로 비지도학습의 머신러닝 알고리즘이다. 비슷한 특성의 데이터들을 k개의 군집을 만든 뒤, 각 군집 마다 거리 차이를 계산하고 이에 대한 분산 값이 최소화될 때까지 군집화를 진행하는 방식으로 동작한다.

3. 데이터 수집 및 전처리

3.1 데이터 수집

본 연구에서는 역직구 상품 추천 대상 국가로 미국, 중국, 베트남을 정하였다. 각 국가별 쇼핑몰 판매 실적 데이터 수집을 위해 미국의 Amazon, 중국의 Taobao, 베트남의 Shopee를 선정하고, 2021년 09월부터 2023년 04월까지의 판매 실적 데이터를 수집하였다. 수집 데이터는 판매상품, 판매량, 판매가격, 고객평점, 고객리뷰 등이며, 판매 실적 데이터는 월별로 집계하였다. 수집 기간 동안 수집된 데이터의 수는 16,574,283개이다. <Table 1>은 각 국가별 수집된 데이터의 수와 수집기간을 정리한 것이다.

<Table 1> Data Collection Status

Shopping Mall (Country)	Number of data	Collection period
Amazon (US)	1,453,647	2021/10 ~ 2023/04
Taobao (CN)	13,812,597	2021/09 ~ 2023/04
Shopee (VN)	1,308,039	2021/09 ~ 2023/04

3.2 데이터 통합(Integration)

수집된 원시 쇼핑몰 데이터는 각 국가별 쇼핑몰 데이터를 기반하기 때문에 상품을 분류하는 카테고리 체계가 상이한 문제를 가지고 있다. 본 연구에서는 대한상공회의소의 유통물류진흥원에서 정의한 KAN CODE 분류체계를 적용하여 모든 상품의 카테고리 분류를 재편성함으로써 이 문제를 해결한다.

또한 각 국가별 경제적인 상황을 고려하기 위해서 경제지표를 추가하였다. 추가된 경제 지표는 국내총생산(GDP, Gross Domestic Product) 지수, 소비자물가지수(CPI, Consumer Price Index), 환율 등이다.

상품의 선호도를 파악하기 위해서 상품별로 구매자들이 등록한 리뷰를 분석하여 감성사전을 구축하고, KAN CODE 분류 상품별로 긍정과 부정에 대한 각각의 점수 속성을 추가하였다

3.3 데이터 변환(Transformation)

각 쇼핑몰 데이터에서 상품의 가격은 각 국가별 화폐 단위로 표기가 되어 있다. 이에 대한 통일을 위해 미국의 화폐 단위인 달러($)를 기준으로 통일하였으며 데이터를 수집하는 시점의 월평균 기준 환율을 이용하여 상품 가격의 값을 변환하였다.

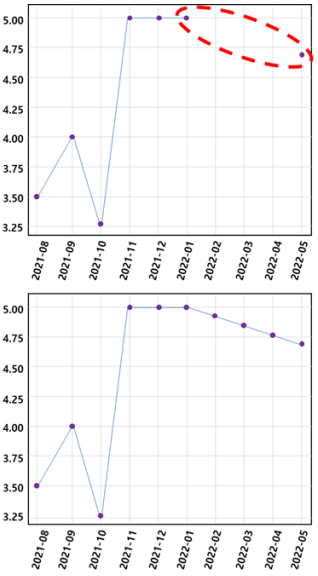
3.4 데이터 정제(Cleaning)

각 국가별 쇼핑몰 판매실적 데이터를 분석하여 학습데이터를 만들기 위해 원시 데이터를 정제하였다. 정제에는 중복 제거, 결손 보완, 상품분류코드(KAN, Korean Article Number) 매핑 등이 진행되었다.

각 쇼핑몰 데이터에 중복이 존재하는 경우 집계 시 편향된 데이터가 만들어질 수 있으므로 중복 정보를 제거한다.

특정 데이터의 속성값이 누락이 된 경우 선형 보간법을 통해 누락된 값을 보충한다. <Figure 1>은 선형 보간법을 이용하여 누락된 바디오일 상품의 평점 평균값을 도출한 결과이다.

특정 상품의 경우 카테고리 정보가 불명확하거나 여러 개의 KAN CODE에 분류되는 경우가 존재하고 있어 데이터의 모호함을 없애기 위하여 한 개의 상품은 한 개의 KAN CODE 로만 분류되게 하였다.

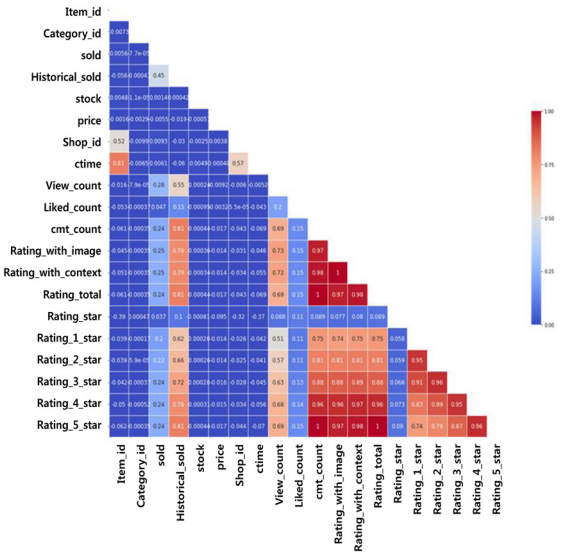


<Figure 1> Derivation of Missing GPA Values by Linear Interpolation

4. 학습데이터 구축

4.1 데이터 상관분석(Correlation Analysis)

수집된 판매실적, 경제지표 등의 데이터 상관분석을 통해 학습데이터 특징 추출에 사용할 유의미한 속성을 선택하였다.



<Figure 2> Correlation Analysis of Shopping Mall Data in Vietnam

베트남 쇼핑몰의 데이터의 경우, <Figure 2>와 같이 판매실적 데이터 중 cmt_count(댓글 개수), rating_total(총 평점 개수), rating_1_star(평점1점 개수), rating_2_star(평점2점 개수), rating_3_star(평점3점 개수), rating_4_star(평점4점 개수), rating_5_star(평점 5점 개수), rating_with_images(이미지가 포함된 평점 개수)는 강한 선형관계가 있는 것으로 분석 되었다.

데이터의 속성들 중 유사한 특징을 갖는 속성들은 모델 학습 시 연산 시간을 증가시킬 요인이 될 수 있고, 유사 속성을 제거하여도 분석 결과에 나쁜 영향을 주지 않으므로 제거하였다.

본 연구에서는 유의미한 속성을 선택하기 위해서 분산 팽창인수(VIF, Variance Inflation Factor)와 최소자승법(OLS, Ordinary Least Squares)을 이용하고 있으며, 그 절차는 아래의 설명과 같다.

	Variable	VIF-value
0	rating_5_star	inf
1	rating_total	inf
2	rating_4_star	inf
3	rating_3_star	inf
4	rating_2_star	inf
5	rating_1_star	inf
6	cmt_count	72550.61
7	rating_with_context	309.90
8	rating_with_image	208.43
9	ctime	3.32
10	item_id	3.10
11	view_count	2.36
12	shop_id	1.54
13	rating_star	1.24
14	sold	1.10
15	liked_count	1.06
16	price	1.01
17	category_id	1.00
18	stock	1.00

Remove index 0

	Variable	VIF-value
0	cmt_count	72550.61
1	rating_total	72376.29
2	rating_with_context	309.9
3	rating_with_image	208.43
4	rating_4_star	53.74
5	rating_3_star	33.34
6	rating_2_star	22.04
7	rating_1_star	11.13
8	ctime	3.32
9	item_id	3.10
10	view_count	2.36
11	shop_id	1.54
12	rating_star	1.24
13	sold	1.10
14	liked_count	1.06
15	price	1.01
16	category_id	1.00
17	stock	1.00

Remove index 0 .... Remove index 0

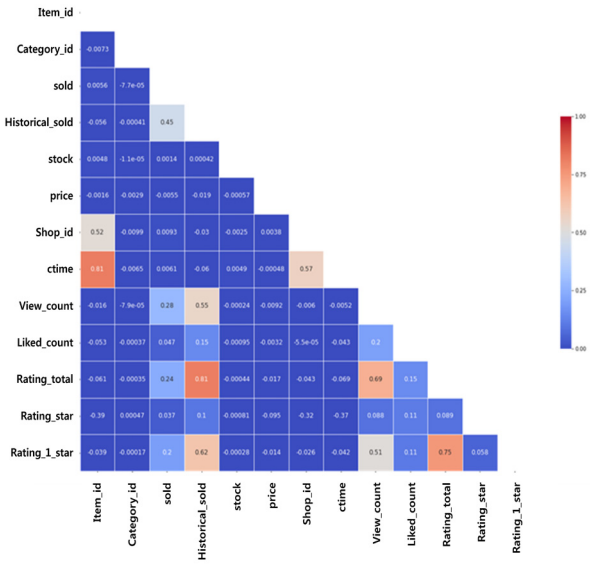
	Variable	VIF-value
0	ctime	3.29
1	rating_total	3.26
2	item_id	3.09
3	rating_1_star	2.32
4	view_count	2.01
5	shop_id	1.54
6	rating_star	1.24
7	sold	1.09
8	liked_count	1.06
9	price	1.01
10	category_id	1.00
11	stock	1.00

<Figure 3> Removing Properties with VIF

첫 번째로 VIF를 계산하여 값이 10미만으로 나오는 요소들을 제거한다. VIF값이 10보다 크면 독립변수 간에 다중공선성(multicollinearity)을 갖는다고 볼 수 있기 때문이다. <Figure 3>은 VIF 계산을 통해 상관관계가 높은 독립변수를 제거하는 과정이다.

두 번째로 VIF를 이용하여 찾아낸 속성들이 낮은 상관관계를 갖는지 추가 검증하기 위해 OLS를 적용하여 높은 R-squred Score와 낮은 P-Value를 갖고 있는지 확인한다. <Figure 4>는 OLS 분석 결과의 일부이다.

앞의 과정들을 통해 높은 상관관계가 있는 속성들을 제거한 결과 <Figure 5>와 같이 높은 상관관계가 있는 독립변수들이 제거되었다.



<Figure 5> Correlation Plot after Property Removal with VIF

rating_total

OLS Regression Results

Dep. Variable:	rating_total	R-squared:	0.822
Model:	OLS	Adj. R-squared:	0.822
Method:	Least Squares	F-statistic:	1.208e+05
Date:	Thu, 25 Nov 2021	Prob (F-statistic):	0.00
Time:	10:37:05	Log-Likelihood:	-2.0497e+06
No. Observations:	314320	AIC:	4.099e+06
Df Residuals:	314307	BIC:	4.099e+06
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2566.2221	5435.287	0.472	0.637	-8086.786	1.32e+04
item_id	9.674e-10	1.42e-10	6.816	0.000	6.89e-10	1.25e-09
category_id	-0.0200	0.054	-0.370	0.711	-0.126	0.086
sold	-0.2538	0.002	-162.790	0.000	-0.257	-0.251
historical_sold	0.1589	0.000	472.402	0.000	0.158	0.160
stock	-1.069e-08	3.24e-08	-0.330	0.742	-7.42e-08	5.28e-08
price	-1.169e-06	5.44e-07	-2.147	0.032	-2.24e-06	-1.02e-07
shop_id	-1.698e-08	2.21e-09	-7.670	0.000	-2.13e-08	-1.26e-08
ctime	-3.471e-07	1.56e-08	-22.258	0.000	-3.78e-07	-3.17e-07
view_count	0.0416	0.000	298.204	0.000	0.041	0.042
liked_count	-0.0021	0.000	-13.103	0.000	-0.002	-0.002
rating_star	-1.8112	0.133	-13.604	0.000	-2.072	-1.550
rating_1_star	42.3236	0.138	306.194	0.000	42.053	42.595

Omnibus:	255475.396	Durbin-Watson:	1.878
Prob(Omnibus):	0.000	Jarque-Bera (JB):	224470160587.815
Skew:	-1.648	Prob(JB):	0.00
Kurtosis:	4142.985	Cond. No.	1.47e+14

price

OLS Regression Results

Dep. Variable:	price	R-squared:	0.011
Model:	OLS	Adj. R-squared:	0.011
Method:	Least Squares	F-statistic:	289.7
Date:	Thu, 25 Nov 2021	Prob (F-statistic):	0.00
Time:	10:35:05	Log-Likelihood:	-4.5939e+06
No. Observations:	314320	AIC:	9.188e+06
Df Residuals:	314307	BIC:	9.188e+06
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.374e+07	1.78e+07	1.895	0.058	-1.16e+06	6.86e+07
item_id	-4.611e-06	4.65e-07	-9.915	0.000	-5.52e-06	-3.7e-06
category_id	-328.8937	176.967	-1.858	0.063	-675.745	17.957
sold	5.9339	5.318	1.116	0.265	-4.490	16.358
historical_sold	-3.3718	1.441	-2.340	0.019	-6.197	-0.547
stock	-2.765e-05	0.000	-0.260	0.795	-0.000	0.000
shop_id	-3.125e-05	7.25e-06	-4.309	0.000	-4.55e-05	-1.7e-05
ctime	-0.0002	5.11e-05	-3.873	0.000	-0.000	-9.78e-05
view_count	2.2372	0.518	4.320	0.000	1.222	3.252
liked_count	2.4110	0.533	4.520	0.000	1.366	3.456
rating_total	-12.5502	5.844	-2.147	0.032	-24.005	-1.096
rating_star	-2.509e+04	434.054	-57.798	0.000	-2.59e+04	-2.42e+04
rating_1_star	-900.0138	516.029	-1.744	0.081	-1911.415	111.388

Omnibus:	1014188.474	Durbin-Watson:	1.768
Prob(Omnibus):	0.000	Jarque-Bera (JB):	547454454646.966
Skew:	53.237	Prob(JB):	0.00
Kurtosis:	6457.497	Cond. No.	1.47e+14

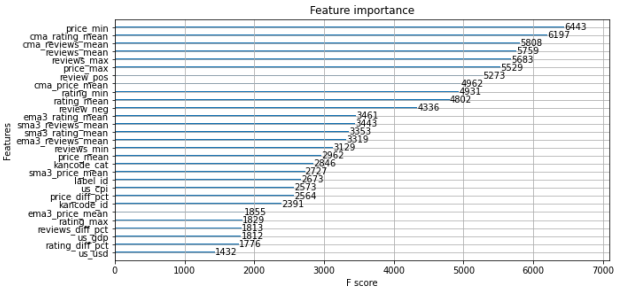
<Figure 4> Validate correlation with OLS

4.2 특징 추출(Feature Extraction)

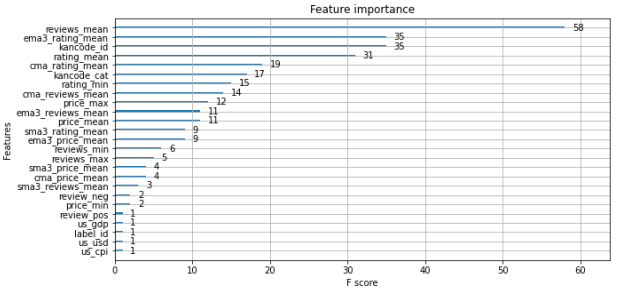
학습데이터를 구축하기 위해 상관도 분석을 통해 선택된 각 상품군의 가격, 평점, 판매량, 경제지표, 감성분석 데이터와 각각의 MIN, MAX, MEAN, MEDIAN을 산출하였다.

또한 상품 판매가격 추정을 위한 추세 데이터로 각 상품군의 가격, 평점, 판매량, 리뷰수에 대한 전월대비 증감률과 이동평균값을 산출하여 추가하였다. 이동평균값은 평균가격 속성을 활용하여 단순이동평균(SMA, Simple Moving Average), 지수이동평균(EMA, Exponential Moving Average), 누적이동평균(CMA, Cumulative Moving Average) 값을 사용한다. SMA는 특정 기간 동안의 값에 대한 증감을 수치화한 대표적인 지표이다. EMA는 시간이 흐름에 따라 가중치를 다르게 부여하는 방법으로 최근의 값이 더 높은 가중치를 갖는다. CMA는 기간에 대한 가중치 없이 데이터를 모두 평균을 내어 과거값과 최근값이 모두 동일한 가중치를 갖는다.

판매상품 특징 데이터들의 중요도 분석 결과는 다음과 같다.



<Figure 6> Importance by Feature in the Pair-wise Model



<Figure 7> Importance by Feature in the List-wise Model

상품 추천 모델의 경우 Pair-wise 접근법 기반의 모델에서는 최소판매가 속성의 중요도가 높게 나타났으며, List-wise 접근법 기반의 모델에서는 평균 리뷰수 속성의 중요도가 높게 나타났다.

5. 추천 모델 및 평가

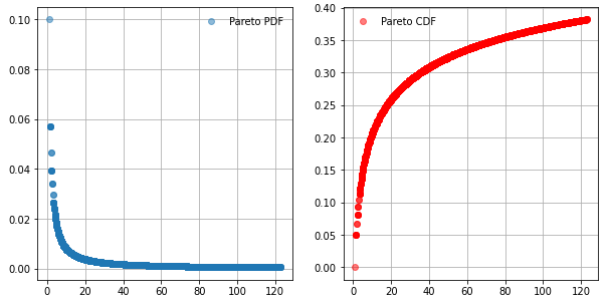
5.1 상품 추천 모델

AI 상품 추천 모델 학습을 위해서 Scikit-learn 라이브러리에서 제공하는 XGBoost를 사용하였으며, 그 중에서도 목적함수(Objective Function)를 기본 값이 아닌 Pair-wise 손실이 최소화되는 LambdaMART와 nDCG(Normalized Discounted Cumulative Gain)가 최대화되는 List-wise 접근법의 LambdaMART로 설정하여 모델을 구성하였다[8].

모델 파라미터로는 Tree에서의 레이어 수 10개, Gradient Bossted Tree 수를 KAN CODE의 수만큼 설정하였고, 목적함수의 계산식은 다음과 같다.

$$obj = \sum_i l(\hat{y_i},\ y_i) + \sum_k \Omega(f_k)$$

AI 상품 추천 모델의 상품 순위 예측에 사용된 연관성 점수는 다음과 같은 세부적인 과정을 통해 정의하였다. 첫째로 주요 특징을 통계적으로 조합하여 산정된 랭킹을 정렬하여 종합랭킹 변수를 생성하였다. 둘째로 파레토 분석을 통해 종합랭킹의 빈도를 확률밀도함수(PDF)와 누적분포함수(CDF)를 이용하여 분석하였으며 이 중에서 분포가 더 고르게 나타난 누적분포함수 기반의 파레토 분석 값을 선택하였다. 셋째로 누적분포함수 기반의 분석값을 토대로 최종적인 연관성 점수를 정의하였다.



<Figure 8> Pareto Analysis of Association Scores

본 연구는 학습된 모델을 기반으로 역직구 상품 추천을 위해 사용자가 입력한 KAN CODE, 수출하고자 하는 판매가격범위, 물량 등의 정보를 바탕으로 하여 예측된 국가별 연관성 점수를 비교하여 최적의 수출 국가와 상품군을 추천한다.

5.2 상품 추천 모델 평가

상품추천 모델 성능의 경우 <Table 2>와 같이 나타난다.

이때 사용한 성능 평가 지표는 MSE, RMSE, nDCG이며, 각각의 계산식은 다음과 같다.

$$MSE = \frac{\sum_{i}^{n}(y_i - \hat{y}_i)^2}{n}$$

$$RMSE = \sqrt{MSE}$$

$$nDCG_p = \frac{DCC_p}{IDCG_p}$$

그 중에서도 LTR과 같이 아이템 순위를 측정하는 것이 주요 목적인 경우, 이를 평가하기 위한 방법 중 가장 일반적인 방법이 nDCG이다.

<Table 2> Performance Comparison of Product Recommendation Model

Evaluation index	XGBoost-LambdaMART Model	
	Pair-wise	List-wise
MSE	34.38	47.33
RMSE	5.86	6.88
nDCG	**0.99**	0.96
MSE(min-max)	0.08	0.01
RMSE(min-max)	0.28	0.11
nDCG(min-max)	0.99	0.96

<Table 2>에서 보여주고 있는 것과 같이 Pair-wise 와 List-wise 접근법의 XGBoost-LambdaMART 모델의 nDCG 값은 모두 좋은 성능을 보이고 있으나 Pair-wise의 nDCG 값이 99%로 List-wise보다 더 좋은 성능을 나타낸다.

5.3 판매가격 추정 모델

AI 가격 추정 모델 학습을 위해서 pycaret 라이브러리에서 제공하는 회귀모델(Regression)의 Light Gradient Boosting Machine, Extreme Gradient Boosting 등의 알고리즘을 사용하여 학습데이터의 상품 평균가를 예측하여 가장 적합한 알고리즘을 선정한다.

5.4 판매가격 추정 모델 평가

Regression기반의 판매가 추정 모델의 경우, 성능은 <Table 3>과 같이 나타난다. 측정한 성능 지표는 MAE, MSE, RMSE, R-Squared이며 각각의 계산식은 위와 같다.

$$MAE = \frac{\sum_{i}^{n}|x_i - x|}{n}$$

$$MSE = \frac{\sum_{i}^{n}(y_i - \hat{y}_i)^2}{n}$$

$$RMSE = \sqrt{MSE}$$

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

성능 측정에 사용한 알고리즘 중에서 R-Squared 값이 0.9162을 나타낸 Light Gradient Boosting Machine 모델이 가장 좋은 성능을 보인다.

<Table 3> Performance Comparison of Price Estimation Model

Model	MAE	MSE	RMSE	R^2
Light Gradient Boosting Machine	42.24	90721.65	261.50	**0.9162**
Random Forest Regressor	37.71	125608.90	305.85	0.8793
Extra Trees Regressor	38.65	134051.94	313.95	0.8698
Gradient Boosting Regressor	42.59	150705.55	329.97	0.8523
Extreme Gradient Boosting	43.30	172593.44	355.48	0.8233
Decision Tree Regressor	48.90	211207.59	397.49	0.7812
AdaBoost Regressor	212.32	230009.48	430.48	0.7573

5. 결론 및 향후 계획

본 연구는 미국의 Amazon, 중국의 Taobao, 베트남의 Shopee 쇼핑몰을 대상으로 2021년 09월부터 2023년 04월까지의 상품 데이터를 수집하였으며, 수집된 데이터를 가공하여 AI 기반의 상품 추천 및 가격 추정 모델을 학습하였다. 상품 추천 모델의 경우 Learning To Rank 방법론 기반의 XGBoost-LambdaMART 알고리즘을 활용하였으며, 손실함수로 List-wise보다 Pair-wire를 사용하는 것이 더 높은 성능을 보였다. 판매가격 추정 모델의 경우는 Light Gradient Boosting Machine 기반의 Regressor를 사용하는 것이 가장 좋은 성능을 나타냈다.

향후 계획으로 본 연구를 바탕으로 역직구 사업을 추구하는 국내 중소 셀러에게 국가별 상품 추천 결과와 상품 판매가 추정 결과를 제공하는 서비스를 구축할 예정이다. 또한 현재 수집하고 있는 상품군이 의류, 전자제품, 화장품에 국한되어 있기 때문에 범용적인 서비스 제공을 위해서는 더 많은 상품군과 다양한 쇼핑몰 데이터를 확보하여

분석 정보를 확장할 필요가 있다.

Acknowledgement

This work was supported by the Technology Innovation Program (or Industrial Strategic Technology Development Program) (1415179185, Development of overseas market information analysis system for export small and medium sellers) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea)

References

[1] Jung, J.W., Kim, S.H., Yergali, B., Kim, K.I., Lee, J.S., Kim, J.S., A novel product recommendation system for global market, *Proceedings of the 10th International Conference on Big Data Applications and Services*, 2022, Jeju Island, Korea, pp. 45-50.

[2] Jung, M.J., Chung, J.E., and Yang, H.S., The Influences of Korean Wave and Product Image on Cross-border Shopping Intention for Korean Cosmetics in China, *Journal of Consumer Studies*, 2018, Vol. 29, No. 1, pp. 55-82.

[3] Kim, T.H. and Shim, W.J., An Empirical Study between Important Export Factors and Export Performance of Chinese B2C Reverse Direct Purchase Exporters, *Journal of International Trade and Insurance*, 2022, Vol. 23, No. 4, pp. 59-76.

[4] Kim, Y.D., A Study on the Improvement of Customer Satisfaction in Reverse Direct Purchases for the Promotion of Exports of Korean SMEs, *Journal of International Trade & Commerce*, 2021, Vol. 17, No. 2, pp. 313-330.

[5] Lee, Y.J., Endemik Sidae Jungsogieop Onlain Suchulisyu Bunseok Mich Jeongchaekjeok Sisajeom, *KIET Monthly Industrial Economics*, 2022, Vol. 284, pp. 28-37.

[6] Liu, T.Y., Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 2009, Vol. 3, No. 3, pp. 225-331.

[7] Shahbazi, Z. and Byun, Y.C., Product recommendation based on content-based filtering using XGBoost classifier, *International Journal of Advanced Science and Technology*, 2020, Vol. 29, pp. 6979-6988.

[8] Wang, Y., Wang, L., Li, Y., He, D., and Liu, T.Y., A Theoretical Analysis of NDCG Type Ranking Measures, *In Conference on Learning Theory*, 2013, pp. 25-54, PMLR.

[9] Xie, B., Tang, X., and Tang, F., Hybrid Recommendation Base on Learning to Rank, In *2015 9th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, 2015, July, pp. 53-57, IEEE.

[10] Yu, K.J., Jang, A.R., Sim, S.K., Oh, A.Y., and Lee, H.G., Construction of Emotional Dictionaries for Product Preference Analysis using Overseas Shopping Mall Reviews, *Autumn Annual Conference of IEIE*, 2022 Nov, pp. 755-756.

ORCID

Kyu Ik Kim | http://orcid.org/0009-0001-5251-2249
Yergali Berdibayev | http://orcid.org/0000-0001-7458-7009
Soo Hyung Kim | http://orcid.org/0009-0007-1485-0312
Jin Suk Kim | http://orcid.org/0009-0000-0600-0552