ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2022 March: 28(1): 69~88 http://dx.doi.org/10.13088/jiis.2022.28.1.069

분리학습 모델을 이용한 수출액 예측 및 수출 유망국가 추천*

장영진
연세대학교 경영학과
(jiuoo3@naver.com)

원종관
부산대학교 경영학과
(jongkwan1@pusan.ac.kr)

이채록
부산대학교 경영학과
(li016010@naver.com)

최근 코로나19 팬데믹으로 인해 전 세계 경제와 외교 상황에 급격한 변화가 일어나고 있으며, 수출 의존도가 높은 한국은 이러한 변화에 큰 영향을 받고 있다. 본 연구에서는 기업의 수출전략 수립 및 의사결정 지원을 위해 차년도 수출액 예측 모델을 구축하고, 모델의 예측 결과를 바탕으로 수출 유망국가 추천 방식을 제안한다. 본 연구에서는 모델이 다양한 정보를 학습할 수 있도록 국가별, 품목별, 거시경제 변수 등 선행 연구에서 중요하게 사용된 변수를 다방면으로 수집하였다. 수집한 데이터를 분석한 결과, 국가와 품목에 따라서 수출액의 분포가 매우 비대칭적인 것을 확인할 수 있었다. 따라서, 모델의 예측 성능을 향상시키고 설명력을 확보하기 위해서 분리학습 방식을 사용하였다. 분리학습은 전체 데이터를 동질적인 하위 그룹으로 분리하고 개별 모델을 구축하는 방식으로, 본 연구에서는 수출액을 기준으로 5개 구간으로 데이터를 분리하였다. 모델 학습 과정에서 구간별 특성을 반영하여 구간1부터 구간4까지는 LightGBM을 사용하고, 구간5는 지수이동평균을 사용하였으며 이를 통해 모델의 예측 성능을 향상시킬 수 있었다. 모델의 설명력 확보를 위해서 추가로 구간별 모델의 SHAP-value를 계산하고 중요도가 높은 변수를 제시했다. 또한, 본 연구에서는 예측 모델을 기반으로 2단계 수출 유망국가 추천 방식을 제안했다. 효율적인 수출 전략 수립을 위해서 BCG 매트릭스와 국가별 점수 산출 방식을 사용하였고, 품목별 유망 국가 순위와 수출 관련 주요 정보들을 제공하였다. 본 연구는 다양한 정보를 학습한 머신러닝 모델로 여러 국가와 품목에 대한 예측을 실시하고, 이 과정에서 분리학습 방식으로 예측 성능을 향상시켰다는 점에서 의의가 있다. 또한, 현재 무역 관련 서비스들이 과거 데이터에 기반한 정보를 제공하고 있음을 고려할 때, 본 연구에서 제안한 예측 모델과 유망국가 추천 방식은 기업들의 미래 수출 전략 수립 및 동향 파악에 유용하게 사용될 수 있을 것으로 기대된다.

주제어 : 수출액 예측, 머신러닝, 분리학습, LightGBM, Shapley value

논문접수일 : 2021년 11월 25일 논문수정일 : 2021년 12월 21일 게재확정일 : 2021년 12월 23일
원고유형 : 급행논문 교신저자 : 원종관

1. 서론

한국 경제의 구조적 특징 중 하나는 수출의존도가 높다는 것이다. 많은 기업 활동이 글로벌 경제 및 외교 상황과 밀접하게 관련되어 있으며, 코로나 대유행과 같은 급격한 대내외 환경 변화는 기업 전반에 큰 타격을 입힐 수 있다. 특히 중소기업들은 이러한 경제적 불확실성이 지속되는 상황에 더 취약한데, 규모가 큰 기업에 비해 상대적으로 예측 불가능한 상황에 신속하고 적절

* 이 논문은 2021년도 '4단계 두뇌한국21사업(4단계 BK21 사업)'에 의하여 지원되었음.

한 대응을 하기 어렵기 때문이다(Morgan et al., 2020). 이러한 상황에도 불구하고 국가 경제에서 중소기업이 차지하는 비중은 점차 커지고 있다. 중소기업의 기술력과 부품경쟁력은 중소기업뿐만 아니라 대기업의 세계시장 경쟁력 확보에도 영향을 미치며 이는 국가 경쟁력과 직결된다(Lee et al., 2015). 따라서, 중소기업의 성장을 위해서 적절한 수출입 전략을 수립하는 것은 국가적 차원에서 중요한 과제이며, 최근에는 중소기업의 한정된 자원을 고려하여 향후 수출액을 예측하는 연구 사례가 증가하고 있다(Hong et al., 2017). 코로나 19 이후 수출증가율과 경제성장률의 관계가 밀접해지고 있기 때문에, 수출액 예측의 중요성은 더욱 높아지고 있다(Yun, 2021).

본 연구에서는 다양한 국가와 품목에 대한 수출액 예측 모델을 제시한다. 수집한 수출액 관련 데이터를 분석한 결과, 수출 품목과 대상 국가에 따라서 상이한 특징을 가지고 있다는 것을 확인할 수 있었다. 특히, 종속 변수인 수출액은 분산이 매우 크다는 특징을 가지고 있었다. 종속 변수의 분산이 큰 경우, 전체 데이터를 활용해 예측 모형을 구축할 경우 예측 성과가 떨어질 수 있다(Hong et al., 2010). 따라서 모델의 예측 성능 향상을 위해 수출액 구간에 따라서 그룹을 나누고 구간별로 모형을 구축하는 분리학습을 적용했다. 분석 알고리즘은 LightGBM을 사용했으며, 수출액 기준 최상위 구간의 데이터는 특수성을 고려하여 별도로 지수이동평균(EMA)을 이용하여 예측을 실시했다. 그 결과, 본 연구에서 제시한 분리학습 모델이 다른 비교 모델보다 뛰어난 예측성과를 달성했다. 또한, 모델의 예측 결과를 기반으로 2단계 품목별 유망 국가 추천 방안을 제시하였다. 예측 결과를 반영하여 각 품목에 대해서 국가들의 위치를 나타낸 BCG 매트릭스를 제시하고, Question Mark와 Star 그룹에 해당하는 국가들에 대해서는 점수를 산출하여 상위권부터 유망 국가로 추천했다. 추천한 국가들에 대해서는 해당 국가 및 품목에 대한 정보도 함께 제공하여 중소기업의 수출 전략 수립에 유의미한 정보를 제공하고자 하였다.

2. 이론적 배경 및 필요성

한국은 무역 의존도가 높은 편이므로 수출액 및 수입액 예측이 상대적으로 중요하다. 이와 관련하여 수출액을 예측하거나 중요 요인을 탐색하는 선행 연구들이 활발하게 진행되었다. 수출액 예측에는 다양한 변수가 사용되는데, 대표적으로 수요 요인, 가격요인, 환율요인, 정책 및 유가 요인이 있다. 이러한 변수들은 수출액 예측 모델의 성능을 향상시키고 모델의 설명력을 제공할 수 있다(Hong et al., 2017). 수출액 예측 시에는 대상 품목이나 국가에 따라서 중요 변수가 달라지기도 한다. 예를 들어, 농산물 수출액을 예측하는 경우, 거리를 나타내는 변수가 교역 패턴이나 추세 파악에서 중요하게 사용된다(Park, 2015). 일부 연구에서는 경기 지수를 이용한 수출 선행지수를 개발했다. 내수 경기를 나타내는 경기 선행지수는 수출실적에 선행하기 때문에, 이를 이용하여 수출국가의 경기를 통해서 수출액을 예측할 수 있다(Lee, 2006). 또한, 물동량을 이용하여 수출액 예측에 사용한 사례도 있다. 물동량 변동은 수출액과 밀접한 상관관계가 있으므로, 수출액 추세 변화에 대한 정보를 얻을 수 있다(Mo, 2015). 더 나아가, 물류 성과 지표는 국가 경쟁력을 파악에도 유의미한 지표로 사용될 수 있다(Kim, 2019; Son, 2020).

〈Table 1〉 Literature Reviews on Trade Exports Prediction

Authors	Methodology	Key variables	Purpose of Research
Hong, 2017	Building export function and predicting export amount	real GDP, real income, nominal exchange rate, FTA, FDI	Building export forecasting model
Park, 2015	Random effect model	GDP, GNI, population, distance	Forecasting agriculture trade volume
Mo et al., 2015	GARCH, EGARCH, GJR	Seaborne Trade	Predicting volatility of export volume in Gwangyang Port
Lee, 2006	Analysis of turning point, Correlation analysis	Composite leading indicators index, Export price index, Exchange rate	Construction of leading export index
Jeon, 2019	Multiple regression	International Logistics, Transport, Infrastructure	Competitiveness analysis using logistics performance index
Son et al., 2020	Fixed effect model, random effect model	Logistics performance index(LPI)	Verification of variables about domestic exports

2.1. 분리학습

분리학습은 전체 데이터를 동질적인 하위 그룹으로 분류하고 그룹마다 별도로 예측을 실시하는 방식으로, 예측 모델의 성능 향상시키고 설명력을 확보하기 위해서 사용된다. 분리학습이 사용된 사례는 다양한 분야에서 찾을 수 있다. 대표적으로 마케팅 분야에서 프로모션을 통한 구매액을 기반으로 고객을 등급화하고, 등급별로 고객의 구매액을 예측한 사례가 있다(Hong et al., 2010). 이동통신 분야에서도 고객 이탈을 예측하기 위해서 AdaBoost로 계산한 가중치를 기준으로 그룹을 분리하고 RandomF67orest로 분류 성과를 높인 연구 사례도 찾을 수 있다(Lu Ning et al., 2012). 시계열 데이터에도 분리학습을 적용한 사례가 있었는데, 공항 이용객 예측을 위해서 시계열 데이터를 추세와 변동 부분으로 나누고, 개별 학습 모델을 적용하여 예측 성능을 확보한 연구 사례를 확인할 수 있다(Joo et al., 2015). 부동산의 적정 가격을 예측하기 위해서 분리학습이 사용되기도 한다. 시장의 변동성에 따라 기간을 구분하고, 기간별로 적절한 예측 알고리즘을 사용하여 정확도를 향상시킬 수 있다(Bae et al., 2018). 아파트 거래가격을 예측하기 위해서 지역에 따라서 데이터를 분리하고 LIME 알고리즘을 통해서 예측 성능과 설명력을 확보한 사례도 있다(Cho et al, 2020). 이처럼 다양한 분야에서 분리학습을 이용하여 예측 성능을 높인 연구 사례를 확인할 수 있었다.

〈Table 2〉 Literature Reviews on Separate Learning

Authors	Method	Purpose of Research	Segmentation Technique
Hong et al., 2010	SVR	Customer purchase prediction	Separation based on customer purchase amount by promotion
Lu Ning et al., 2012	Random Forest	Churn management	Split customer group into two clusters based on the assigned weights obtained through Adaboost.
Joo et al., 2015	Seasonal ARIMA	Prediction of Passenger in Jeju Airport	Separation of airport passenger data into trend and volatility parts
Bae et al., 2018	ARIMA, VAR, SVM, RF, LSTM	Prediction of real estate price index	Separating stable market situation when the market situation changes rapidly and build model
Jo et al., 2020	ARIMA, RF, LSTM	Prediction of apartment price	Separating datasets by region and improving accuracy of prediction

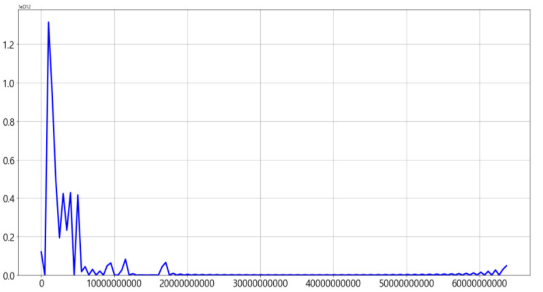
2.2. 연구 필요성

수출입 예측과 관련된 기존 연구에서는 특정 이론이나 변수의 영향력을 확인하거나 특정 상황, 국가, 경제권을 대상으로 연구가 진행된 사례가 많았다. 본 연구에서는 머신러닝 기법을 이용하여 다양한 국가와 품목에 대한 종합적인 수출액 예측을 시도하였다. 예측 모델에 다양한 정보를 학습시키기 위하여, 거시 경제 변수(경제성장률, 실질 GDP 등), 무역 관련 지수(LPI, RCA), 기업 관련 변수(기업 무역 활동성, 기업 생존율 등)와 같이 선행 연구에서 수출 예측에 유의하게 사용된 변수들을 다방면으로 수집하고 학습하였다. 이를 통해서 일반화 가능성이 높은 모델을 제시하고 보다 효율적이고 효과적으로 차년도 수출액을 예측하고자 하였다. 또한, 수출액 예측과 유망국가 추천 과정에서 설명력과 구체성을 확보하여, 중소기업들의 전략 수립에 유용한 정보를 제공하고자 하였다.

3. 분석내용 및 결과

3.1. 탐색적 데이터 분석

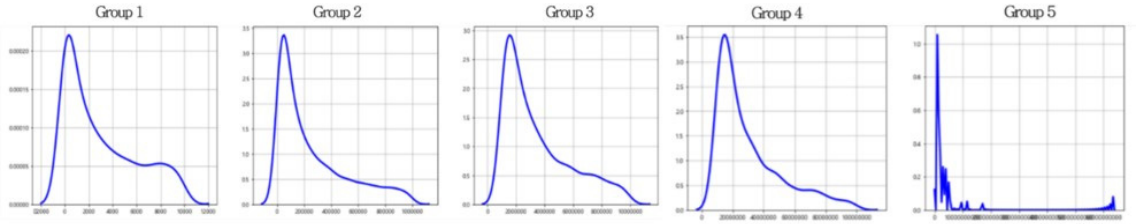
본 연구의 종속변수인 '국가별/품목별 수출액'의 기초통계량을 살펴보면, 평균이 17,939,632이고, 표준편차가 481,003,321로 국가와 품목에 따라서 수출액 차이가 크다는 것을 알 수 있다. 종속변수의 분포를 시각적으로 살펴보면 <Figure 1>과 같으며, 데이터가 비대칭적이며 이상치가 다수 존재하는 것을 확인할 수 있다.



〈Figure 1〉 Distribution of Exports

〈Table 3〉 Criteria of Separating Datasets (USD)

Group Name	Group Separation Criteria	Number of Rows
Group 1	Exports < 10,000	2,626
Group 2	10,000 < Exports < 1,000,000	11,093
Group 3	1,000,000 < Exports < 10,000,000	5,069
Group 4	10,000,000 < Exports < 100,000,000	1,995
Group 5	100,000,000 < Exports	406

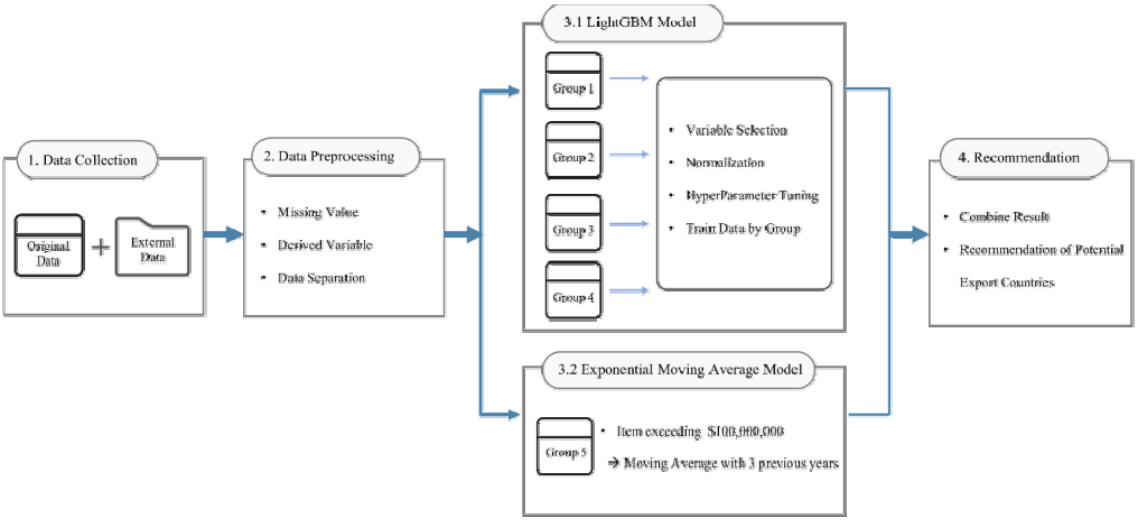


〈Figure 2〉 Distribution of Exports by Group

〈Table 4〉 Descriptive Statistics by Group(USD)

	Group 1	Group 2	Group 3	Group 4	Group 5	Whole Dataset
Mean	3,025.7	257,048	3,502,275	30,438,794	735,923,761	17,939,632
Standard Deviation	3,022	259,598	2,339,254	21,259,375	3,401,611,749	481,003,321
Max	9,996	999,996	9,998,596	99,796,985	63,695,330,669	63,695,330,669
Min	0	10,017	1,000,844	10,018,681	100,031,034	0

종속변수의 분산이 큰 경우에는 일반적인 지도학습 방식으로 데이터를 학습하면 예측 오차가 커질 가능성이 높다. 이러한 데이터 불균형에서 나타나는 한계점을 극복하기 위하여 본 연구에서는 분리학습 방식을 채택하였다. 전체 데이터에서 수출액 규모에 따라서 구간을 분리하고 별도로 학습을 실시했다. 기존 사례에서 수출액 규모를 분류한 기준을 살펴보면 다음과 같다. 한국무역통계 기준에 따르면 1만 달러 이하 / 1~10만 달러 / 10~30만 달러 / 30~100만 달러 / 100~500만 달러 / 500~1,000만 달러 / 1,000만 달러 이상으로 수출입액 규모를 분류할 수 있다. 선행연구에서는 국내 수출입 기업의 무역 규모를 10만 달러 미만 / 10~100만 달러 / 100~1,000만 달러 / 1,000만~1억 달러 / 1억 달러 이상으로 분류했다(유준수 외, 2019). 본 연구에서는 이러한 기준을 참고하여 <Table 3>과 같이 수출액 규모에 따라서 데이터를 5개의 구간으로 분류했다.



〈Figure 3〉 Research Framework

데이터를 종속변수에 따라서 구간별로 분류한 후, 구간에 따라 독립변수들의 값이 유의미한 차이가 있는지 확인하기 위하여 통계검정을 실시하였다. shapiro 검정과 levene 검정을 통해 구간별 변수들의 정규성과 등분산성 가정을 확인하였고 Kruskal-Wallis 검정으로 구간별 변수들의 차이를 확인하였다. 그 결과 유의수준 1%에서 주어진 독립변수들이 구간별로 유의미한 차이가 있음을 확인했다. 수출액을 기준으로 데이터를 5개 구간으로 분리한 후에 구간별 수출액 분포를 살펴보면 <Figure 2>와 같다.

<Table 4>에서 구간1~구간4의 경우 수출액의 표준편차가 감소하였지만, 구간 5의 경우에는 구간 분리 후에 오히려 표준편차가 7배가량 증가하였다. 따라서 구간 5에는 이상치가 다수 포함되어 있고 샘플의 수가 적다는 점을 고려하여 구간1~구간4와는 다른 방식으로 학습과 예측을 진행하였다.

또한, 각 구간의 주요 국가와 품목에도 차이가 있었다. 구간1의 경우 주로 중앙, 동남아시아 국가인 미얀마, 몽골, 카자흐스탄, 스리랑카의 비중이 높았으며, 구간5의 경우 중국과 미국이 큰 비중을 차지했다. 구간별로 주요 품목의 수출액 합계와 평균에서도 차이가 있었다. 구간1의 경우 유리제품, 인쇄물, 종이 품목의 비중이 높았으며, 구간5의 경우 석유, 자동차, 전자회로 관련 품목의 비중이 높았다.

3.2. 예측 모델

본 연구에서 제안하는 예측 및 추천 프레임워크는 <Figure 3>와 같고 각 단계별 세부 설명은 아래와 같다.

3.2.1. 사용 데이터

본 연구에서는 대한무역투자진흥공사(KOTRA)에서 제공한 2017년~2019년 국가 및 품목별 수출액 관련 데이터를 사용하였다. 또한, 추가로 총 16개의 외부 변수를 수집하였으며 해당 변수 목록은 <Table 5>와 같다.

〈Table 5〉 External Variable List

Variable name	Description	Variable name	Description
ratio_fuel	Ratio of exported fuel	eco_growth	Economic growth rate
ratio_ores_metal	Ratio of exported metal	trade_balance	Trade Balance to Korea
ratio_manufac	Ratio of exported manufacturing products	FTA	Whether South Korea has signed FTA in 2017
survival_rate_item (1~5 year)	Survival rate of each item	survival_rate_company (1~5 year)	Survival rate of company by country
marine_freight	World Seaborne Trade	air_freight	Air traffic trade
GDP_real GDP_Nominal GDP_Deflator	GDP deflator	Active company Entering company Existing company	The number of active/entering/exiting companies by country
CS_price	Consumer Price Index(CPI)	LPI	Logistic Performance index
Small_firm_export Mid_firm_export Big_firm_export	Exports and export growth rates of small/mid/big firm	Active money Entering money Existing money	The amount of active/entering/exiting money by country

〈Table 6〉 Derived Variable List

Variable name	Description	Variable name	Description
RCA	Revealed Comparative Advantage	hs_counrty_share	Exports of one item in one country / Whole exports of all items in one country
CAC	Comparative Advantages by Countries	range_distinct	Average distance – distance with South Korea(by item)
GDP_per_person	Gross Domestic Product per capita	SHARE_RATE	Market share
gravity	GDP / distance	GROWTH_RATE	Export increase / Decrease rate
country_share	Whole exports of one item in one country / Whole exports of one item in all countries	-	-

3.2.2. 데이터 전처리

일반적으로 데이터가 충분히 많은 경우에는 결측치 처리 과정에서 결측치가 존재하는 행을 삭제하는 것이 효과적일 수 있지만, 본 연구의 데이터는 하나의 행이 해당 국가와 품목의 정보를 나타내기 때문에 중요도가 크다. 따라서 정보의 손실을 줄이기 위해 UNcomtrade, WorldBank, ITC 등 외부 출처에서 결측치에 대한 정보를 탐색하여 추가하였다. 데이터에 대한 추가 정보를 얻을 수 없는 결측치는 평균으로 대체하거나 적절한 값을 선정하여 대체했다. 또한, 유의미한 변수를 모델에 포함하기 위하여 파생변수를 생성하였다. 파생변수는 기존 분석 및 연구 사례를 기반으로 해석 가능성을 중점을 두고 총 9개의 변수를 추가했다. 파생변수 목록은 <Table 6>과 같다.

3.2.3. 변수 선정

변수 선정 과정은 2단계로 진행하였다. 우선 범주형 변수의 경우 더미화를 실시하고, 각 더미 변수와 종속 변수의 등분산 검정 후 t-test를 실시해서 유의수준 5%에서 유의하지 않은 변수들을 제거했다. 이후 연속형 변수를 포함해서 다중선형회귀 모델에 대해서 양방향 Stepwise Selection을 실시하여, AIC가 감소하지 않을 때까지 변수를 하나씩 추가하고 제거하는 과정을 반복해서 최종 변수를 선정하였다. 본 연구에서는 구간별 특성을 반영하기 위해서 분리학습 모델을 채택하였기 때문에 모델 최적화를 위해서 구간마다 변수 선정을 실시하였다.

3.2.4. 분리 학습

LightGBM 모델

구간1~구간4는 LightGBM(Light Gradient Boosting Machine)을 이용하여 데이터를 학습했다. LightGBM은 여러 분야의 연구와 분석에서 사용되고 있는 분류 및 회귀 알고리즘이다. LightGBM은 기존 Gradient Boosting 알고리즘에 Gradient-based One-Slide Sampling과 Exclusive Feature Bundling 방식을 접목했다. 이를 통해서 Gradient Boosting Decision Tree보다 20배 빠른 속도로 학습하면서 높은 정확도를 확보할 수 있다(Ke, Guolin et al., 2017). 기존 부스팅 알고리즘은 전체 데이터를 학습하면서 수평으로 균형 잡힌 트리를 만들기 때문에 연산 시간이 오래 소요된다. 하지만 LightGBM은 leaf-splitting 통해서 수직 방향으로 트리를 우선적으로 확장하기 때문에, 빠른 속도로 데이터의 특성을 학습한다(Ma Xiaojun et al., 2018). 본 연구에서 사용된 데이터는 품목과 국가별 특징이 매우 다르기 때문에, 여러 특성을 효율적으로 학습하고 과적합을 방지하기 위하여 LightGBM을 예측 알고리즘으로 사용하였다. 변수선정 과정을 통해 선택된 입력변수에 대해서는 Min-Max scaling을 실시하였고, 종속변수는 로그변환을 실시했다. 주요 파라미터 8개에 대해서 Grid Search와 5-Fold Cross Validation 방식으로 최적 파라미터 조합을 탐색하였다. 이 과정을 구간마다 별도로 실시해서 구간별 데이터에 최적화된 파라미터 조합을 탐색하였다.

지수이동평균(Exponential Moving Average) 모델

데이터 탐색 단계에서 수출액 $100,000,000 이상에 해당하는 구간 5의 경우 데이터 분리 후에 오히려 표준편차가 증가한 것을 확인했다. 이 구간에는 이상치가 다수 포함되어 있고 데이터 샘플도 406개로 적다. 또한, 구간 5에 속한 품목들을 살펴보면 주로 석유, 승용차, 전자회로, 디램, 배터리 등 대내외 상황 변화에 상당히 민감하게 반응하는 품목임을 알 수 있다. 이러한 특성을 가진 데이터의 경우 일반적으로 과거의 데이터에서 패턴을 찾아내는 머신러닝 방식으로 학습하기가 어렵다.

따라서, 구간5에 해당하는 품목들은 LightGBM으로 학습하는 것이 아니라 전 3개 연도(2016년도~2018년)의 국가별, 품목별 한국으로부터의 수출액을 UNcomtrade에서 수집하여, 지수이동평균으로 예측값을 산출했다. 지수이동평균은 최근 값에 큰 가중치를 부여하고 과거의 값은 적은 가중치를 부여해 미래의 값을 예측하는 방식으로, 단기 변동성을 포착하거나 최근 추세를 반영하는 데 유리하다.

〈Table 7〉 Model Description

Model No.	Model Description
Model 1	Train whole dataset with LightGBM
Model 2	Group 1~Group 4 → LightGBM without separated learning method Group 5 → Exponential moving average
Model 3	Group 1~5 → LightGBM with separated learning method
Model 4 (Suggested Model)	Group 1~Group 4 → LightGBM with separated learning method Group 5 → Exponential moving average

〈Table 8〉 Prediction Accuracy by Model Structure (USD)

	Model 1		Model 2		Model 3		Model 4 (Suggested Model)	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Whole Dataset	8,404,585	202,572,091	4,303,968	46,595,205	6,799,060	136,949,211	4,097,177	46,533,099
Group 1	8,404,585	202,572,091	1,357,196	4,869,561	2,269	3,368	2,269	3,368
Group 2					108,084	167,801	108,084	167,801
Group 3					1,110,682	1,611,359	1,100,682	1,611,359
Group 4					8,552,013	13,574,801	8,552,013	13,574,801
Group 5			153,690,929	333,217,413	293,398,034	984,299,585	153,690,929	333,217,413

3.3. 예측 성능 평가

본 연구에서는 모델의 예측 성능을 두 가지 방식으로 비교 및 평가하였다. 첫 번째로 모델의 구조에 따른 예측 성능을 비교하였고, 두 번째로는 사용 알고리즘에 따른 예측 성능을 비교하였다. 평가 지표로는 모두 MAE와 RMSE를 사용했다.

3.3.1 모델 구조에 따른 예측 성능 비교

분리학습 모델의 유용성을 검증하기 위하여 모델 구조에 따른 예측 성능을 비교했다. 본 연구에서 제시한 분리학습 모델과 구조가 다른 모델들을 비교하였으며 사용한 예측 알고리즘은 LightGBM으로 통일했다. 총 4가지의 서로 다른 구조의 모델을 비교하였고, 각 모델에 대한 설명은 <Table 7>과 같다.

모델 1은 분리학습을 실시하지 않고 일반적인 방식으로 전체 데이터를 LightGBM으로 학습한 모델이다. 모델 2는 모델 1과 유사하지만, 구간5에 해당하는 데이터만 지수이동평균으로 예측한 모델이다. 모델 3은 전체 데이터를 구간1~구간5로 나눠서 분리학습을 하되, 지수이동평균을 사용하지 않고 모두 LightGBM을 사용한 모델이다. 마지막으로 모델 4는 본 연구에서 제안한 모델로, 구간1~구간4는 LightGBM으로 예측하고 구간5에 해당하는 데이터는 지수이동평균으로 예측한 모델이다.

〈Table 9〉 Prediction Accuracy by Algorithm (USD)

	Linear Regression		SVR		Random Forest		LightGBM	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Whole dataset	6,133,309	110,743,562	4,942,853	54,908,829	4,111,969	46,545,150	4,097,177	46,533,099
Group 1	2,325	3,413	2,216	3,024	2,194	3,239	2,269	3,368
Group 2	239,842	2,945,087	176,758	255,019	108,705	169,438	108,084	167,801
Group 3	8,308,857	205,334,174	3,645,118	59,003,138	1,100,400	1,620,096	1,100,682	1,611,359
Group 4	11,137,744	17,595,025	10,690,474	19,290,933	8,706,564	13,646,544	8,552,013	13,574,801
Group 5	153,690,929	333,217,413	153,690,929	333,217,413	153,690,929	333,217,413	153,690,929	333,217,413

모델 구조에 따른 성능을 비교한 결과는 <Table 8>과 같다. 전체 데이터셋(Whole Dataset)의 오차를 살펴보면 분리학습을 실시하지 않은 모델1의 오차가 가장 크게 나타났고, 본 연구에서 제안한 분리학습 모델인 모델4의 오차가 가장 작게 나타났다. 또한, 모델3과 모델4를 비교했을 때 구간 5에 대해서 LightGBM 보다 지수이동평균을 사용한 모델4의 오차가 작게 나타난 것을 확인할 수 있다.

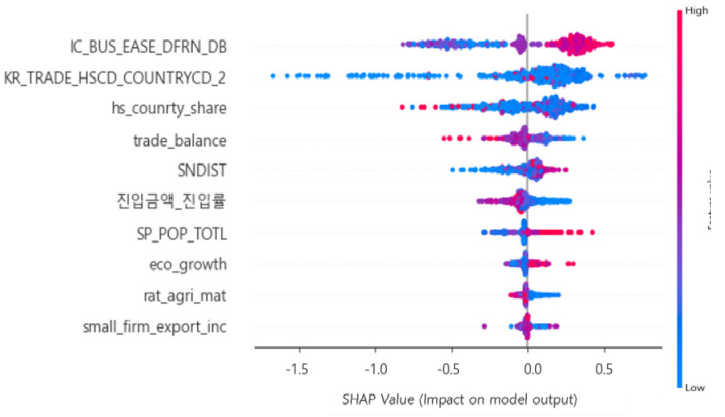
3.3.2. 사용 알고리즘에 따른 예측 성능 비교

사용 알고리즘에 따른 모델의 성능 비교를 위해서 본 연구에서 사용한 LightGBM과 다른 알고리즘의 예측 성능을 비교하였다. 알고리즘의 성능 비교를 위해서 모델의 구조는 앞서 언급한 모델4로 통일하였다. 비교 대상 알고리즘은 선행 연구 및 분석 사례를 참고하여 다중 선형회귀, Support Vector Regressor, Random Forest를 선정하였다. 각 모델의 예측 결과는 <Table 9>와 같다.

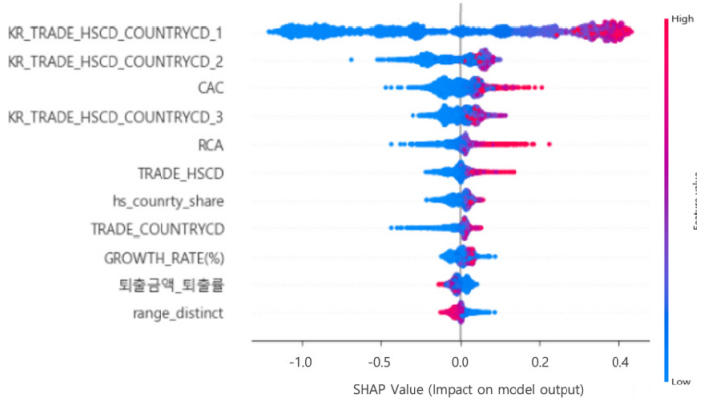
<Table 9>에서 구간5는 모두 지수이동평균을 사용하였기 때문에 MAE, RMSE가 동일하다. 전체 데이터셋(Whole Dataset)의 오차를 살펴보면, 선형회귀 모델의 MAE, RMSE가 가장 크게 나타난 것을 알 수 있다. 반면 LightGBM을 사용한 모델은 전체 데이터셋에서 오차가 가장 작게 나타났다. 구간별 데이터의 오차값에서도 전반적으로 LightGBM의 성능이 뛰어난 것을 확인할 수 있다.

3.4. 구간별 변수 중요도

본 연구에서 제안한 예측 모델의 설명력을 확보하기 위해서 LightGBM으로 학습한 구간1부터 구간4까지 모델의 변수 중요도를 Shap Value를 통해 추출하였다. Shap Value은 모든 가능한 변수 조합에 대해 하나의 특성의 기여도를 실제값과 예측값의 오차로 계산하는데, 주로 머신러닝 모델의 설명력을 제공하기 위한 목적으로 사용된다(Lundberg et al., 2017). 각 그래프에서 붉은색일수록 변수가 종속변수에 영향을 미치는 정도가 큰 것이고, X축을 기준으로 값이 커질수록 종속변수에 양의 영향을, 값이 작을수록 음의 영향을 미친다고 해석할 수 있다. 본 연구에서는 Shap Value 계산을 위해서 Python의 Shap 패키지를 이용하였으며 구간별 summary plot은 다음과 같다.



〈Figure 4〉 Group 1 - SHAP Value

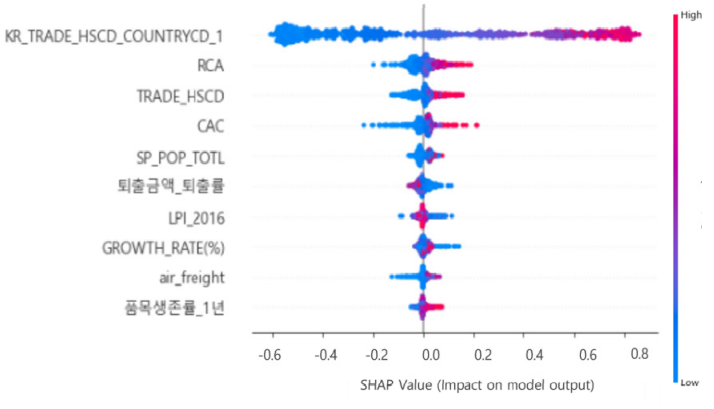


〈Figure 5〉 Group 2 - SHAP Value

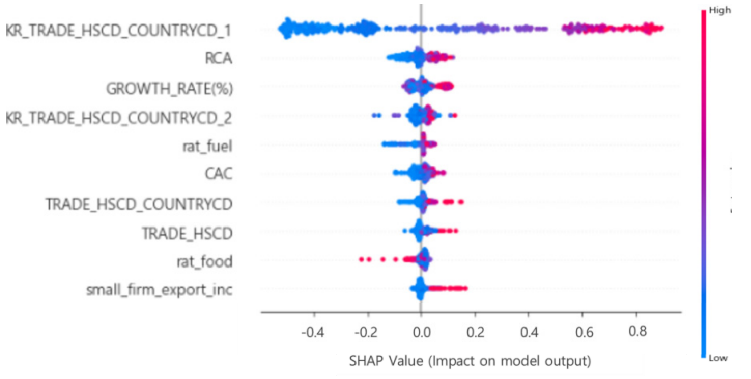
구간1 중요 변수들을 살펴보면, 비즈니스 용이성 점수(IC_BUS_EASE_DFRN_DB)의 중요도가 가장 높게 나타난 것을 확인할 수 있다. 이 경우 비즈니스 용이성 점수가 높을수록 예상 수출액이 높게 나타나고, 점수가 낮을수록 예상 수출액이 낮게 나타나는 경향이 있다고 해석할 수 있다.

구간2의 경우 1~3년 전 해당 항목의 수출액(KR_TRADE_HSCD_COUNTRYCD_1~3) 변수와 종속변수가 밀접하게 관련이 있음을 알 수 있다. 또한, 국가별비교우위지수(CAC)와 현시비교우위지수(RCA)도 중요 변수로 사용된 것을 확인할 수 있다.

구간3의 경우 전년도의 수출액(KR_TRADE_HSCD_COUNTRYCD_1) 변수가 가장 중요하게 나타났으며, 구간2와 동일하게 현시비교우위지수(RCA)와 국가별비교우위지수(CAC)도 유의하게 사용되었다.



〈Figure 6〉 Group 3 - SHAP Value



〈Figure 7〉 Group 4 - SHAP Value

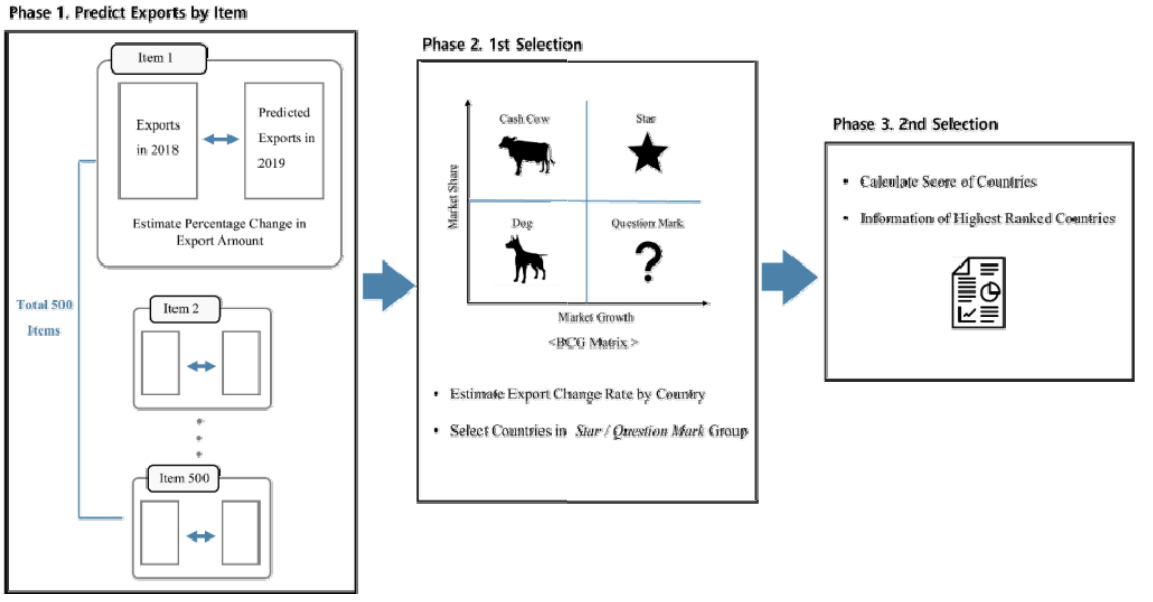
구간4의 경우 앞선 구간과 유사한 변수들이 중요하게 사용되었고, 수출증감률(GROWTH_RATE) 변수가 새롭게 추가된 것을 확인할 수 있다.

4. 품목별 유망국가 추천

본 연구에서는 국내 중소기업의 수출 의사결정 보조를 위해서 예측 모델의 결과를 기반으로 차년도 수출 유망 국가를 추천 방안을 제안한다. 추천과정은 <Figure 8>과 같으며 단계별 세부 설명은 다음과 같다.

4.1. 품목별 수출액 증감률 계산

먼저 LightGBM과 지수이동평균을 이용한 분리학습 모델로 각 품목별, 국가별 차년도 수출액 예측값을 계산한다. 그리고 각 품목마다 모든 국가의 당해 연도의 수출액과 차년도 수출 예측액을 비교해서 증감률을 계산한다.



〈Figure 8〉 Recommendation Process

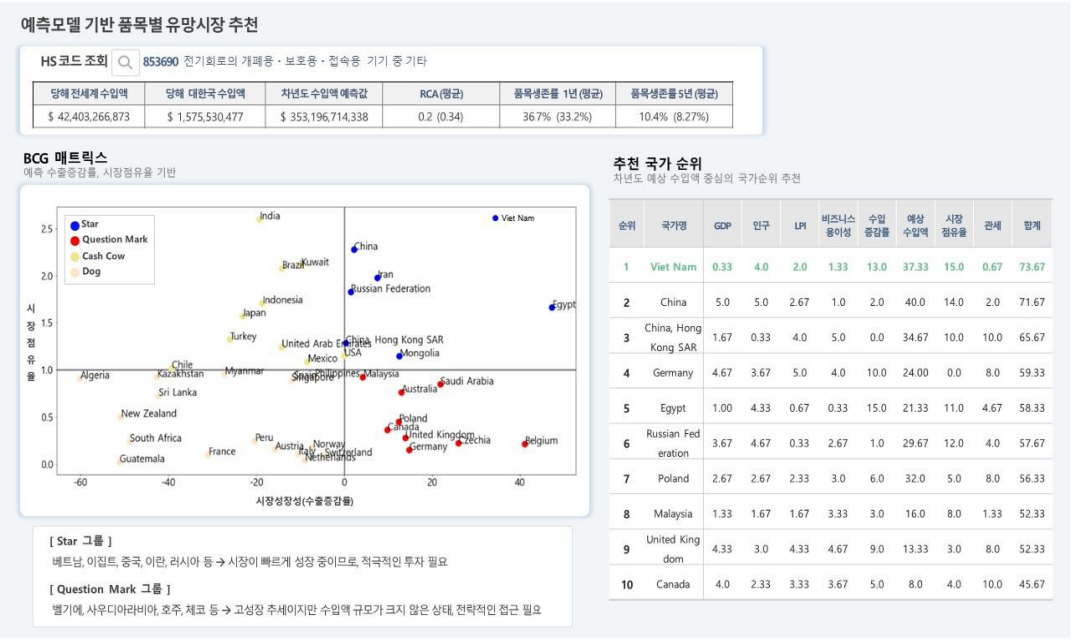
4.2. 1차 수출 유망국가 선정

수출액 증감률을 계산한 후에 BCG 매트릭스를 활용해서 1차 수출 유망국가를 선정한다. BCG 매트릭스는 x축을 수출 증감률, y축을 시장점유율을 기준으로 나타낸 것으로, 시장에서 기업 또는 상품의 상대적인 위치를 분석하기 위해서 사용된다. 수출 유망국가를 선정하는 과정에도 BCG 매트릭스가 사용되기도 한다(Kim et al., 2005). 본 연구에서는 개별 수출 품목을 하나의 시장으로 보고 여러 수출 대상 국가들의 상대적인 위치를 BCG 매트릭스로 제시하였다.

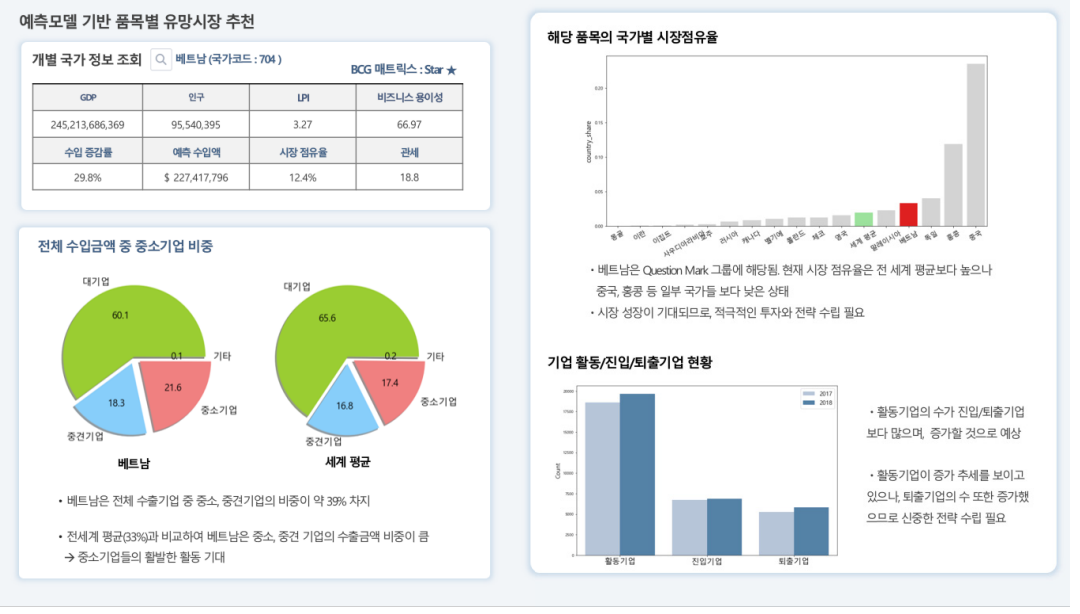
예측 모델의 결과를 수출 유망국가 추천 과정에 반영하기 위해서, x축의 값을 모델이 예측한 값을 이용하여 차년도 예상 수출 증감률을 사용하였다. y축인 시장점유율은 현시점의 데이터를 사용하였다. 그리고, BCG 매트릭스 상에서 수출 증감률이 높은 "Star"와 "Question Mark"에 해당하는 국가들을 일차적으로 수출 유망 국가로 선별하였다.

〈Table 10〉 Scoring Items and Weights

	Country index (20%)				Item index (80%)				Total
	GDP	Population	LPI	Ease of doing business	Change rate of income	Predicted Exports	Share	Tariff	
Weight	5%	5%	5%	5%	15%	40%	15%	10%	100%



〈Figure 9〉 Example of Recommendation Results by Item



〈Figure 10〉 Example of Recommendation Results by Country

4.3. 2차 수출 유망국가 선정

일차적으로 선정한 Star와 Question Mark 그룹의 수출 유망 국가들을 대상으로, <Table 10>의 8개 항목을 이용하여 종합점수를 산출하고 점수에 따라 추천 순위를 제시한다. 점수산출 방식은 한국무역협회의 수출 유망시장 점수산출 방식을 참고로 하되, 본 연구의 구조와 목적에 적합하게 항목과 비중을 수정했다.

<Figure 9>, <Figure 10>은 HS코드 591190 품목에 대한 유망국가 추천 예시 화면이다. <Figure 9>는 해당 품목을 조회한 예시 화면이다. 이용자가 특정 품목을 조회하면 해당 품목의 국가들에 대한 BCG 매트릭스와 추천 국가 순위를 계산하여 제공한다. 화면 중앙 좌측의 BCG 매트릭스를 통해서 해당 품목에 대해서 수출 대상 국가들의 상대적인 위치를 파악할 수 있다. 화면 우측에는 Star, Question Mark 그룹의 국가들을 대상으로 점수를 산정하여 순위대로 유망국가를 추천하였다. <Figure 9>에서는 점수가 가장 높은 베트남이 1순위로 추천되었다. <Figure 10>은 추천된 국가들에 대해서 국가별 세부 정보를 조회할 수 있는 화면이다. 화면 상단에 점수 산정에 사용한 항목들을 제시한다. 그리고 전체 수출금액 중 중소기업의 비중, 국가별 시장점유율, 활동/진입/퇴출기업의 현황에 대한 정보를 시각화 결과를 중심으로 제시한다.

5. 기대효과

5.1. 연구의의 및 기대효과

수출액 예측과 관련된 기존 연구에서는 주로 특정 국가나 상황에 대해서 분석을 진행하거나, 일부 변수의 영향력을 확인하는 방향으로 연구가 진행되었다. 본 연구에서는 머신러닝을 이용하여 여러 국가와 품목에 대한 종합적인 예측 모델을 제시했다는 점에서 의의가 있다. 다양한 정보를 학습한 모델을 구축하기 위해서 수출입과 관련이 있는 변수들을 다방면으로 수집하였으며, 추가로 파생변수를 생성하여 사용하였다. 이를 통해서 일반화 가능성 및 범용성이 높은 예측 모델을 제시하였다.

또한, 본 연구에서는 수출액 예측에 분리학습을 적용한 모델을 제안하였다. 기존 연구에서 수출액 예측에 분리학습을 활용한 사례는 많지 않았다. 수집한 데이터에서 수출액의 분포가 비대칭적인 점을 고려하여 전체 데이터를 수출액 규모에 따라서 분리하고, 구간별 특성을 반영하는 모델을 구축하였다. 이를 통해 수출액 예측 모델의 성능을 향상시키고 각 구간의 변수 중요도를 Shap value로 제시하여 구체적인 설명력을 확보할 수 있었다.

본 연구의 실무적 의의는 다음과 같다. 현재 대부분의 무역 정보 관련 서비스들은 과거부터 현시점까지의 수출입 데이터를 종합하고 분석한 정보를 제공하고 있다. 한편, 본 연구에서 제안한 모델은 차년도에 대한 수출액 예측 정보를 제공한다. 이를 통해서 미래 수출 동향, 모델의 구간별 변수 중요도, BCG 매트릭스, 추천 점수 등 다양한 정보를 활용할 수 있다. 본 연구에서 제안한 모델을 활용한다면 KOTRA와 같은 무역 정보 서비스 제공자 입장에서는 기존 서비스의 고도화를 통해서 서비스 효용성 및 만족도를 증대시킬 수 있을 것이다. 또한, 해당 서비스를 이용하는 기업 입장에서는 다양한 정보를 기업 상황에 맞게 적절하게 선택하여 활용할 수 있으므로, 효율적인 전략 수립이 가능할 것이다.

5.2. 연구 한계점

본 연구의 한계점 및 향후 연구 방향은 다음과 같다. 첫째, 본 연구에서 수집하여 사용한 독립변수는 GDP, 물가지수, 물동량 등 대부분 국가 단위의 범주형 변수이다. 이러한 변수는 국가별 특징을 학습하는 데에는 효과적이지만 품목별 정보를 학습하는 데에 한계가 있다. 본 연구에서 사용한 품목별 변수(관세, RCA 등) 외에 추가로 다양한 품목별 변수를 수집해서 사용할 수 있다면 예측 성능 향상과 더불어 다양한 해석을 제공할 수 있을 것이다.

둘째, 무역은 국제 관계, 정치, 외교와 밀접한 관련이 있다. 특히 코로나 19, 미·중 무역전쟁, 일본 수출규제와 같이 예측이 어려운 거시적인 변화가 발생하는 경우 교역 금액이 크게 변화할 수 있다(Ma et al., 2021). 따라서 현시점의 정형 데이터만으로는 미래에 대한 정보를 온전히 반영하여 학습하는 데에 한계가 있다. 머신러닝은 과거의 데이터에 패턴이 있음을 가정하고 이를 바탕으로 미래의 값을 예측하는 것이기 때문에, 이상치에 해당하는 품목들은 예측의 정확도가 떨어진다. 특히, 본 연구에서 구간 5에 속하는 품목들은 외부 환경 변화에 민감한 품목이라고 볼 수 있다. 따라서 이러한 품목들에 대해서는 추가 분석을 통해서 예측 결과를 보완할 필요가 있다.

참고문헌(References)

Bae, S. W. and J. S. Yoo, "Predicting the Real Estate Price Index Using Machine Learning Methods and Time Series Analysis Model," *Housing Studies Review*, 26.1 (2018), 107-133.

Cho, B. G., G. B. Park and S. H. Ha," Comparative Analysis for Real-Estate Price Index Prediction Models using Machine Learning Algorithms: LIME's Interpretability Evaluation," *The Journal of Information Systems,* 29.3 (2020), 119-144

Hong, S. W., et al., "A study on the export forecast model through the analysis of export determinants by country and economic region," *Korea Institute for Industrial Economics & Trade*, (2017), 1-142.

Hong, T. H. and E. M. Kim, "Predicting the Response of Segmented Customers for the Promotion Using Data Mining," *Information Systems Review,* 12(2), (2010), 75-88.

Jeon, H. J. and Y. M. Kim, "The Impacts of Logistics Performance on National Competitiveness," *KOREA INTERNATIONAL COMMERCIAL REVIEW*, 34 (2019), 99-116.

Joo, T. W. and S. B. Kim, "Time series forecasting based on wavelet filtering," *Expert Systems with Applications*, 42.8 (2015), 3868-3874.

Ke, Guolin, et al., "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, 30 (2017), 3146-3154.

Kim, J. R., "Korea's Export Competitiveness to EU through RCA-CAC Analysis," *Journal of European Union Studies*, 50 (2018), 73-105.

Kim, S. R., et al., "A Study on the Selection of the Export Promising Commodities and the Development of Supporting Policy in the area of Health Industry," *Korea Health Industry Development Institute*, (2005).

Kim, T. I., et al., "A Comparative Analysis on the Export Competitiveness between Korea and China: Focusing on RCA and TSI," *Asia-Pacific*

Journal of Business, 8 (2017), 57-73.

Lee, J. W., "Development of Export Leading Index and Analysis of Fitness," *KEXIM OVERSEAS ECONOMIC REVIEW*, (2006), 4-17.

Lee, M. S., S. B. Park and Y. S. Kwon, "A study on impacts of SMEs' recognition of the importance of cooperation and trust toward large customer companies upon enhanced inter-organization cooperations and the new," *Journal of the Korean Entrepreneurship Socieity,* 10 (2015), 71-94.

Lu Ning, et al., "A customer churn prediction model in telecom industry using boosting," *IEEE Transactions on Industrial Informatics*, 10.2 (2012), 1659-1665.

Lundberg, Scott M., and S. I. Lee, "A unified approach to interpreting model predictions," *Proceedings of the 31st international conference on neural information processing system*s. (2017), 4768-4777

Ma, H. S. and S. T. Kim, "A Study on the Changes in the Trade Environment of Korea and the Prospect of Trade between Korea and China after COVID-19," *The e-Business Studies*, 22.2 (2021), 89-103.

Ma, Xiaojun, et al., "Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning," *Electronic Commerce Research and Applications*, 31 (2018), 24-39.

Mo, S. W. and G. B. Kim, "Volatility of Export Volume and Export Value of Gwangyang Port Abstract," *Journal of Korea Port Economic Association*, 31.1 (2015), 1-14.

Morgan, T., Anokhin, S., Ofstein, L., and Friske, W., "SME response to major exogenous shocks: The bright and dark sides of business model pivoting," *International Small Business Journal*, 38(5) (2020), 369-379.

Oh, D. Y., "An Analysis of the Determinants of Korean SMEsnm,` Exports," *Journal of International Trade and Industry Studies,* 17.2 (2012), 135-159.

Son, K. W., H. S. Jo, and H. C. Moon, "The Determinants of Korea's Export Using Global Logistics Performance Index (LPI)," *Ocean Policy Research*, 35.2 (2020), 103-132.

Yoe, T. D. and S. D. Ki, "An Analysis of the Trade Relations between Korea and the Baltic Nations," *International Commerce and Information Review*, 20.4 (2018), 79-105.

Yoo, J. S. and E. H. Jang, "A Study on the Trade Items and Scale of Domestic Import and Export Companies on the Amount of Customs Refund," *THE INTERNATIONAL COMMERCE & LAW REVIEW*, 84 (2019), 189-207.

Yun, S. H., "COVID-19 Pandemic and Democracy, Economic Growth," *KIRI*, 517, (2021), 8-14.

Abstract

Export Prediction Using Separated Learning Method and Recommendation of Potential Export Countries*

Yeongjin Jang** · Jongkwan Won*** · Chaerok Lee****

One of the characteristics of South Korea's economic structure is that it is highly dependent on exports. Thus, many businesses are closely related to the global economy and diplomatic situation. In addition, small and medium-sized enterprises(SMEs) specialized in exporting are struggling due to the spread of COVID-19. Therefore, this study aimed to develop a model to forecast exports for next year to support SMEs' export strategy and decision making. Also, this study proposed a strategy to recommend promising export countries of each item based on the forecasting model.

We analyzed important variables used in previous studies such as country-specific, item-specific, and macro-economic variables and collected those variables to train our prediction model. Next, through the exploratory data analysis(EDA) it was found that exports, which is a target variable, have a highly skewed distribution. To deal with this issue and improve predictive performance, we suggest a separated learning method. In a separated learning method, the whole dataset is divided into homogeneous subgroups and a prediction algorithm is applied to each group. Thus, characteristics of each group can be more precisely trained using different input variables and algorithms. In this study, we divided the dataset into five subgroups based on the exports to decrease skewness of the target variable. After the separation, we found that each group has different characteristics in countries and goods. For example, In Group 1, most of the exporting countries are developing countries and the majority of exporting goods are low value products such as glass and prints. On the other hand, major exporting countries of South Korea such as China, USA, and Vietnam are included in Group 4 and Group 5 and most exporting goods in these groups are high value products.

* This work was supported by the 'BK21 FOUR (Fostering Outstanding Universities for Research)' in 2021
** School of Business, Yonsei University
*** Corresponding author: Jongkwan Won
School of Business, Pusan National University
2, Busandaehak-ro 63beon-gil, Geumjeong-gu, Busan, Korea
Tel: ***-****-**** E-mail: jongkwan1@pusan.ac.kr
**** School of Business, Pusan National University

Then we used LightGBM(LGBM) and Exponential Moving Average(EMA) for prediction. Considering the characteristics of each group, models were built using LGBM for Group 1 to 4 and EMA for Group 5. To evaluate the performance of the model, we compare different model structures and algorithms. As a result, it was found that the separated learning model had best performance compared to other models. After the model was built, we also provided variable importance of each group using SHAP-value to add explainability of our model.

Based on the prediction model, we proposed a second-stage recommendation strategy for potential export countries. In the first phase, BCG matrix was used to find Star and Question Mark markets that are expected to grow rapidly. In the second phase, we calculated scores for each country and recommendations were made according to ranking. Using this recommendation framework, potential export countries were selected and information about those countries for each item was presented.

There are several implications of this study. First of all, most of the preceding studies have conducted research on the specific situation or country. However, this study use various variables and develops a machine learning model for a wide range of countries and items. Second, as to our knowledge, it is the first attempt to adopt a separated learning method for exports prediction. By separating the dataset into 5 homogeneous subgroups, we could enhance the predictive performance of the model. Also, more detailed explanation of models by group is provided using SHAP values.

Lastly, this study has several practical implications. There are some platforms which serve trade information including KOTRA, but most of them are based on past data. Therefore, it is not easy for companies to predict future trends. By utilizing the model and recommendation strategy in this research, trade related services in each platform can be improved so that companies including SMEs can fully utilize the service when making strategies and decisions for exports.

Key Words : Exports prediction, Machine learning, Separated learning, LightGBM, SHapley value

Received : November 25, 2021 Revised : December 21, 2021 Accepted : December 23, 2021
Corresponding Author : Jongkwan Won

저 자 소 개

장 영 진
부산대학교 경영학과에서 학사학위를 취득하고, 현재 연세대학교 경영대학 정보시스템 전공 석사과정에 재학 중이다. 관심 연구분야는 추천시스템, 머신러닝, 빅데이터 분석이다.

원 종 관
부산대학교 경영학과에서 학사학위를 취득하였다. 현재 부산대학교 경영학과 경영정보 전공 석사과정에 재학 중이다. 주요 관심분야는 지능형 테크핀, 암호화폐 예측, 신용평가, 딥러닝, AI 등이다.

이 채 록
현재 부산대학교 경영학과 학사 과정에 재학 중이다. 주요 관심 분야는 주가 예측, 머신러닝, AI, 빅데이터 등이다.