ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2021 December: 27(4): 23~48 http://dx.doi.org/10.13088/jiis.2021.27.4.023

경제적, 산업구조적, 문화적 요인을 기반으로 한 주요 국가의 한국 품목별 수입액 예측 모형 개발: 한국의, 한국에 대한 문화적 요인을 중심으로*

전승표
한국과학기술정보연구원
글로벌R&D분석센터
(spjun@kisti.re.kr)

서봉군
한국생산기술연구원
국가산업융합지원센터
(bgseo@kitech.re.kr)

박도형
국민대학교 경영정보학부 /
비즈니스IT전문대학원
(dohyungpark@kookmin.ac.kr)

한국경제는 지난 수십년간 정부의 수출전략정책에 힘입어 지속적으로 경제 성장을 이룩해왔으며, 수출의 증가는 경제의 효율성 향상, 고용창출, 기술개발 촉진 등 우리나라의 경제 성장을 견인하는 주도적인 역할을 해왔다. 전통적으로 우리나라 수출에 영향을 미치는 주요 요인은 크게 경제적 요인과 산업구조적 요인이라는 두가지 관점에서 찾아볼 수 있다. 첫 번째, 경제적 요인은 환율과 글로벌 경기 변동과 관련된 것으로서, 환율이 우리나라 수출에 미치는 영향은 환율 수준 및 환율 변동성에 따른 영향으로 나누어 살펴볼 수 있으며, 글로벌 경기 변동은 세계 수입 수요에 영향을 미쳐 우리나라 수출을 좌우하는 절대적 요인으로 볼 수 있다. 두 번째, 산업구조적 요인은 국제 분업화의 둔화, 중국의 특정 수입품 자국내 대체 증가, 수출 주력 산업의 해외생산 형태 변화 등 산업이나 제품에 따라 발생한 고유한 특징이다. 가장 최근 글로벌 교류와 관련된 연구들을 살펴보면, 경제적 요인 및 산업구조적 요인과 더불어 문화적인 측면이 중요함을 여러 문헌에서 피력하고 있다. 이에 따라 본 연구에서는 각국의 한국 수입액 예측 모형에 문화적 요인을 함께 반영하여 예측 모형을 개발하고자 하였으며, 구체적으로 문화적 요인이 수입액에 미치는 영향을 PUSH-PULL 프레임워크 관점에서 반영해보고자 하였다. PUSH 관점은 한국이 자신의 브랜드를 개발하고 적극 홍보하는 관점으로 K-POP, K-FOOD, K-CULTURE 등으로 대표되는 한국의 브랜드에 대한 각국의 관심 정도로 정의할 수 있다. 또한, PULL 관점은 각 국가의 국민들의 문화적, 심리적 특징으로 해당 국가의 지배체계, 남성성, 위험 회피성, 시간에 대한 단기/장기 지향성 등으로 대표되는 각 국의 문화 코드로서 한류문화를 얼마나 수용할 성향을 띄고 있는지로 정의할 수 있다. 본 연구에서 제시한 최종 예측 모델의 고유한 특징은 Design Principle에 기반하여 설계한 것인데, 1) 신규로 추가한 데이터 소스를 통해 한국에 대한 관심 및 문화적 특성이 반영될 수 있는 모형으로 구축하였고, 2) 경제적 요인 등의 변화와 품목 및 국가 Code를 입력하면 예측값을 바로 불러올 수 있도록 실용적으로 편의성 있게 설계하였으며, 3) 이론적으로도 의미 있는 결과를 도출하기 위해서 입력과 목표 변수간의 관계를 해석 가능한 알고리즘을 중심으로 설계하였다는 점이다. 본 연구는 기술적 측면, 경제적 측면, 정책적 측면에서 의미 있는 시사점을 제시할 수 있으며, 수입액 예측 모형을 활용하여 중소•중견기업의 수출 지원 전략에 의미 있는 기여를 할 수 있을 것으로 기대된다.

주제어 : 무역, 수입액 예측, 디자인사고, 검색트래픽, 호프스테드 모형

논문접수일 : 2021년 8월 18일 논문수정일 : 2021년 9월 28일 게재확정일 : 2021년 10월 8일
원고유형 : 일반논문 교신저자 : 박도형

* 이 논문은 제9회 산업통상자원부 2021 공공데이터 활용 빅데이터분석 경진대회에 출품하여 우수상을 수상한 작품을 기반으로 작성되었음.
이 논문은 대한민국 교육부와 한국연구재단의 BK21 FOUR (Fostering Outstanding Universities for Research)의 지원을 받아 수행된 연구임.

1. Introduction

한국경제는 지난 수십년간 정부의 수출전략정책에 힘입어 지속적으로 경제 성장을 이룩해왔으며, 수출의 증가는 경제의 효율성 향상, 고용창출, 기술개발 촉진 등 우리나라의 경제 성장을 견인하는 주도적인 역할을 해왔다 (Yoon, 2020). 국가의 성장뿐만 아니라 기업의 관점에서도 수출은 내수시장의 의존성을 줄여줄 뿐만 아니라, 규모 및 범위의 경제 등에 있어 기업 성장의 디딤돌 역할을 하고 있다 (Park and Kang, 2020). 이러한 측면에서 수출에 영향을 미치는 요인들을 파악하고, 다년간의 시계열 데이터를 통해 수출액을 예측하는 연구들이 꾸준히 수행되어왔다 (Lee, Kim and Kim, 2017; Bae, Moon and Hwang, 2015; Son, 2011).

전통적으로 우리나라 수출에 영향을 미치는 주요 요인은 크게 경제적 요인과 산업구조적 요인이라는 두가지 관점에서 살펴볼 수 있다 (Financial Supervisory Service, 2019). 첫 번째, 경제적 요인은 환율과 글로벌 경기 변동과 관련된 것으로서, 환율이 우리나라 수출에 미치는 영향은 환율 수준 및 환율 변동성에 따른 영향으로 나누어 살펴볼 수 있다. 다만, 세계 시장에서의 경합도, 원자재 수입 비중 등에 따라 환율 변동의 영향은 수출 품목별로 상이한 것으로 확인되고 있다. 글로벌 경기 변동은 세계 수입 수요에 영향을 미쳐 우리나라 수출을 좌우하는 절대적 요인으로 볼 수 있으며, 환율이나 GDP(GDP 변화), 각 나라 간의 거리 관련 변수 등이 수입액 예측에 활용될 수 있다. 다만, 최근 몇 년은 세계 경제성장률 하락폭(↓0.8%p) 대비 세계 수입(↓2.7%p) 및 韓수출 증가율(↓7.9%p)이 크게 하락하여 글로벌 경기변동만으로 설명되지 않는 산업 구조적 요인에 의한 수출 둔화 가능성이 제기되고 있다. 이러한 측면은 산업이나 제품 고유의 요인들이 수입액 예측에 반영될 필요가 있는 것을 시사한다.

두 번째, 산업구조적 요인은 국제 분업화의 둔화, 중국의 특정 수입품 자국내 대체 증가, 수출 주력 산업의 해외생산 형태 변화 등 산업이나 제품에 따라 발생한 고유한 특징이다. 세계 교역량 증가에 기여하던 글로벌 가치사슬(GVC, Global Value Chain) 확대 추세가 글로벌 금융위기 이후 약화되고, 이에, GVC 참여도가 세계 4위 수준인 우리나라도 GVC 약화에 따라 수출에 부정적 영향을 받았을 것으로 예상된다. 대표적으로, 중국 정부는 제조업의 세계 시장 경쟁력 확보 등을 위해 수입 부품을 조립하는 등의 가공무역을 지양하는 정책과 수입을 통해 수요를 충당해 왔던 중간재의 자국내 조달이 증가함에 따라 관련 품목의 수입 비중이 감소하고 있다. 특히, 해외 생산은 직접적으로는 수출에 부정적 영향(수출 대체)을 미치기도 하지만 중간재 등 수출을 통한 긍정적 영향(수출 유발)도 큰 편이며, 글로벌 분업화 둔화, 중국의 수입 자국내 대체 진전 등 한국 수출(타국의 수입)에 영향을 미치는 구조적 요인들은 산업이나 제품별로 달라질 것이라 예상된다. 따라서, HS Code별 고유한 특징을 잡아내기 위해, HS Code에 따라 변화하는 해당 품목의 전체 글로벌 수입 변화, 특정 국가의 해당 제품 한국 수입액 변화 등을 고려해야 하고, 각국 정부의 방향성을 반영하는 관세 등을 산업구조적 요인으로 고려해야 할 필요가 있다.

글로벌 교류와 관련된 연구들을 살펴보면, 국내 수출 및 수입액에 영향을 미치는 요인으로서 경제적 및 산업구조적 요인과 더불어 문화적인 측면도 고려할 필요가 있다는 것을 여러 문헌에

서 피력하고 있다 (Choi, 2012; Jung and Kang, 2015; Kim and Ahn, 2012; Choi and Cho, 2018). 이에 따라 본 연구에서는 각국의 한국 수입액 예측 모형에 문화적 요인을 함께 반영하여 모형을 개발하고자 하였다. 특히, 문화적 요인이 수입액에 미치는 영향을 PUSH-PULL 프레임워크 관점에서 반영하고자 하였다. PUSH 관점은 한국이 자신의 브랜드를 개발하고 적극 홍보하는 관점으로 K-POP, K-FOOD, K-CULTURE 등으로 대표되는 한국의 브랜드에 대한 각국의 관심 정도로 정의할 수 있다. 또한, PULL 관점은 각 국가의 국민들의 문화적, 심리적 특징으로 해당 국가의 지배체계, 남성성, 위험 회피성, 시간에 대한 단기/장기 지향성 등으로 대표되는 각 국의 문화 코드로서 한류문화를 얼마나 수용할 성향을 띄고 있는지로 정의할 수 있다.

한국 문화의 PUSH 관점에서는 대표적으로 한류(Korean Wave)를 제시할 수 있다. Choi(2012)는 한국의 수출에 있어 한류는 양(+)의 효과를 가지는 것으로 분석되었는데, 한류의 진행 수준이 높을수록 더 큰 수출 효과가 있음을 확인하였으며, Kim and Ahn(2012) 는 한류와 한류 파생상품이 한국의 국가이미지와 한국산 소비재 선호도에 주는 영향을 분석하고, 한류가 한류파생상품보다 한국의 국가이미지에 주는 긍정적인 효과가 더욱 큰 것을 연구 결과로 제시하였다. Jung and Kang(2015)는 한류의 문화 콘텐츠 수출이 한국의 총수출에 미치는 영향을 117개국 전체 및 지역별로 구분하여 분석했는데, 실증분석 결과 아시아, 미주, 중동, 아프리카 지역에서 한류의 공간 효과가 유의미하게 나타났으며, 유럽과 대양주 지역에서는 비유의적으로 나타났다. 이처럼 한류는 우리나라의 교역량에 영향을 미치는 것으로 파악할 수 있으며, 본 연구에서는 한국 문화에 대한 관심과 주목 정도를 확인하기 위한 방안으로 검색트래픽을 활용하였다. 검색어는 K-POP은 물론 구체적인 컨텐츠로 2016~2019년 해외 활동에 두각을 나타낸 BTS, EXO, Wanna One의 국가별/연도별 검색량의 변화를 변수로서 추가하였다.

한국 문화에 대한 PULL 관점에서는 호프스테드 모형 기반 문화 코드(Culture Code)를 활용하였다. 호프스테드(Hofstede) 문화 모형은 비교문화 연구에서 활발히 활용되고 있으며, 국가문화 차원에 대한 비교와 이해를 제공한다고 평가받고 있다 (Choi, 2015). 또한, Hofstede(2011)에 따르면 한 국가의 문화란 정신의 집단적 프로그래밍으로 정의할 수 있으며, 각 나라의 문화를 정량적으로 모델링하여 국가별 Culture Code를 네 가지 차원(권력거리, 개인주의, 남성성, 불확실성 회피)로 구분하여 제시하고 있다. 이러한 네 가지 차원의 Culture Code가 국제무역에 중요한 영향을 미칠 수 있음을 여러 연구에서 제시하고 있는데, 예를 들어, Sung(2013)은 수출업체와 수입업체의 관계에 있어서 호프스테드의 조직문화 차원을 연구에 적용하였을 때, 불확실성 회피, 권력거리, 집단주의 등의 요인이 관계규범 차원에서 영향이 미침을 검증하였다. 또한, Park(2014)는 수출집약도 및 수출다양도에 호프스테드의 Culture Code를 연구에 반영하였고, 분석 결과 국가간 문화적 근접성이 높을수록 수출집약도가 증가하는 것을 확인하였다. 이와 유사하게 Park et al.(2019)는 호프스테드의 문화적 차이에 기반하여 교역상대국을 분류하고, 상대국의 국가 리스크(정치, 경제, 재무)가 우리나라와의 교역(수출/수입액)에 어떠한 영향을 미치는지 확인하였다. 구체적으로, 코드 차이가 적은 국가의 경우, 경제, 재무 리스크가 감소할수록 한국과의 교역

액이 증가하고, 코드 차이가 큰 국가의 경우, 정치, 경제 리스크가 감소할수록 수출액이 증가하였으며, 재무리스크가 증가할수록 수출액과 수입액이 증가함을 검증하였다. 이처럼 문화 코드가 우리나라의 교역액에 영향을 미치는 것을 확인할 수 있으며, 본 연구에서는 호프스테드 모형에서 정량화되어 있는 각국의 문화코드를 네 가지 차원을 기반으로 수집 정리하여 모형에 반영하고자 하였다.

최근에는 공공기관의 보유 데이터를 적극 활용하여 사회적 현안 해결에 대한 필요성이 대두되고 있다. KOTRA와 같은 국내 수출지원기관에서는 변화하는 무역시장을 정교하게 예측하기 위한 노력을 기울이고 있으며, 빅데이터 플랫폼을 고도화하기 위해 국가간 교역에 영향을 미치는 다양한 변인들에 대한 탐색을 시도하고 있다(KOTRA, 2021). 또한, 산업계에서도 특정 수출품목이나 개별 기업의 특성을 고려한 추천 서비스 확장에 대한 수요가 증가함으로써, 다양한 관점을 고려한 중소•중견기업의 맞춤형 추천 수출지원 서비스에 대한 실무적 필요성도 높아지고 있다.

이러한 측면에서 본 연구는 앞서 정리한 세가지 관점(경제적 요인, 산업구조적 요인, 문화적 요인)을 통합하여 주요 국가의 제품별 한국 수입액 예측 모형을 개발하고자 하며, 공공 빅데이터를 활용하여 기업이나 기관이 실무적으로 해결해야 할 혁신 모델을 제안해보고자 한다.

예측 모형의 개발은 '기존에 존재하지 않는 새로운 것을 창조하는 활동'으로 볼 수 있으며, 연구자, 데이터, 컨텍스트 등 여러가지에 따라 상이한 방향과 형태로 창조될 수 있기에 예측 모형 개발 방향성을 수립하는 것이 필요하다. 본 연구에서는 연구 모형 개발을 위한 대원칙 Design Principle을 연구 모형 개발 전에 정의하고 예측 모형을 개발하고자 한다. Design Principle은 디자인사고(Design Thinking)에서 제품, 서비스, 컨셉을 개발할 때, 의사결정의 지침이 되는 원칙을 정의하는 것으로 많이 활용되고 있다. 특히, 소비자의 니즈나 인사이트를 발굴한 후에 디자이너들이 창의성 있게 대상을 만들어 나갈 때, 디자이너들이 공통된 목표로 창의성을 발휘할 수 있도록 가이드 하는 방법론으로 알려져 있다. 본 연구 모형 개발의 세가지 Design Principle은 다음과 같다.

Drive for Holistic Tri-Dimensional Approaches

기존 연구에서 중요하게 다루어왔던 경제적 요인과 산업구조적 요인과 함께 문화적 요인을 예측모형에 고려하여, 완전하고 총체적인 예측이 될 수 있도록 하자.

Drive for Practical and System-friendly Architecture

예측 모형이 실무적으로 시스템화되어 개발될 수 있도록, 블랙박스가 아닌 모형의 구조와 변수간의 강도/방향성을 설명할 수 있고, 품목-국가 Code 입력을 통해 손쉽게 시스템적으로 수입액을 추출할 수 있으며, 주기적으로 각 변수의 데이터를 자동으로 획득할 수 있는 데이터 연계 체계를 구축하자.

Drive for Stable and Balanced Model Optimization

0부터 수억 이상의 넓은 범위를 갖는 연속형 변수의 예측을 위해, 모형 우수성 지표를 한 가지가 아닌 세 가지로 고려하여 특정 지표로 치중되지 않은 균형 있는 모형 채택 기준을 수립하자. 즉, 상관관계 지표를 통해 방향성을 잘 맞추면서, RMSE 지표를 통해 전체적으로 벗어난 예

측치를 줄이고, 마지막으로 MAPE 지표를 통해, 특정 데이터(수입액이 적거나 큰 국가)에게만 취약하지 않은 최적 모형을 선택하자.

2. Research Process, Design and Methodology

2.1. Research Process based on CRISP-DM

CRISP-DM (Cross Industry Standard Process for Data Mining)은 비즈니스 문제 인식 및 해결을 위한 대표적인 데이터마이닝(DM) 분석 방법론으로, 데이터마이닝 전문가들이 가장 널리 활용하고 있는 개방형 표준 프로세스 모델로 소개되고 있다 (Forbes, 2018). 기존 CRISP-DM은 6단계로 구성되어 있는데, 본 연구에서는 Design Principle 단계를 추가하여 총 7단계를 연구 프로세스에 적용하였고(<Figure 1 참고>), 각각의 과정은 다음과 같다:

1) Business Understanding: 과제의 목적과 요구사항, 도메인 지식을 파악하고 이해

2) Design Principle: 연구의 설계, 분석 방법, 결과 도출 등 연구의 방향성을 설계

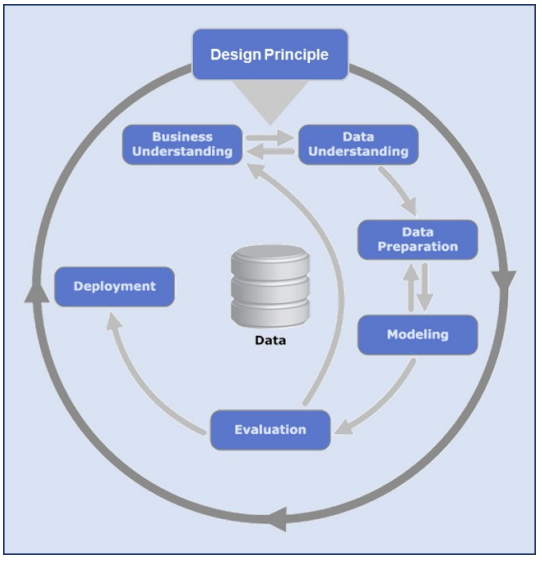
3) Data Understanding: 분석을 위한 데이터의 속성을 이해하고, 문제점 식별 및 인사이트 발견

4) Data Preparation: 데이터 정제, 통합, 새로운 데이터 생성 등 자료를 분석 가능 형태로 전환

5) Modeling: 다양한 모델링 기법과 알고리즘을 선택하고 파라미터를 최적화

6) Evaluation: 모형의 해석 가능 여부 및 결과가 목적에 부합하는지 평가

7) Deployment: 완성된 모델을 현업 상황에 맞게 수정 보안하여 적용



〈Figure 1〉 CRISP-DM Methodology with added Design Principle Step

Business Understanding 단계와 Design Principle 단계는 앞선 Introduction 섹션에서 설명되었고, 2장에서는 연구 설계 및 방법론과 관련된 Data Understanding과 Data Preparation 단계를, 3장에서는 연구 결과와 관련 있는 Modeling, Evaluation, Deployment 단계를 담을 것이다.

2.2. Data Understanding

본 연구에서는 KOTRA, World Bank, UN Comtrade 등에서 제공하는 데이터를 수집하여, 우리나라 수출에 영향을 미치는 경제적 및 산업 구조적 요인들을 구조화 하였으며, 문화적 요인들은 Google Trends, Hofstede Insights 등을 통해 얻어진 데이터를 활용하였다(Appendix의의 Table A 참조). 검색트래픽은 훈련(분석)용과 예측용에

Table 1. Tri-dimensional Variables Used in Predictive Model Development

	변수명	설명	단위;
경제적 요인	COUNTRYCD	ISO 국가 Code	숫자코드
	NY_GDP_MKTP_CD	GDP	US$
	NY_GDP_MKTP_CD_1Y	이전 년도 GDP	US$
	SP_POP_TOTL	인구 (연중 추정치)	명
	PA_NUS_FCRF	공식 환율 (미국 달러에 대한 현지 통화 단위, 월평균을 기준으로 한 연평균)	US$
	TRADE_COUNTRYCD	해당 연도 해당 국가의 전체 품목 수입금액	US$
	KMDIST	해당 국가와 한국과의 거리	km
	SeaDistance	해당 국가와 한국간의 선적 거리	nautical miles
	SNDIST	해당 국가와 수입 국가 간 평균 거리	km
산업 구조적 요인	HSCD	HS Code (품목 Code)	숫자코드
	TRADE_HSCD	해당 연도 해당 품목의 전세계 총 수입금액	US$
	TARIFF_AVG	해당 국가에서 해당 품목에 적용되는 평균 관세율	%
	TRADE_HSCD_COUNTRYCD	해당 연도 해당 국가의 해당 품목 수입금액	US$
	KR_TRADE_HSCD_COUNTRYCD	내년 해당 국가가 해당 품목을 한국으로부터 수입한 금액	US$
문화적 요인	BTS_2016-2018	2016 년부터 2018 년까지의 BTS 에 대한 해당 국가의 검색트래픽	점수 (0~)
	exo_2016-2018	2016 년부터 2018 년까지의 EXO 에 대한 해당 국가의 검색트래픽	점수 (0~)
	kpop_2016-2018	2016 년부터 2018 년까지의 KPOP 에 대한 해당 국가의 검색트래픽	점수 (0~)
	wanna_one_2016-2018	2016 년부터 2018 년까지의 Wanna One 에 대한 해당 국가의 검색트래픽	점수 (0~)
	IC_BUS_EASE_DFRN_DB	비즈니스 용이성 점수	점수 (0~100)
	PowerDistance	호프스테드 culture code - 권력거리	점수 (0~100)
	Individualism	호프스테드 culture code - 개인주의	점수 (0~100)
	Masculinity	호프스테드 culture code - 남성성	점수 (0~100)
	UncertaintyAvoidance	호프스테드 culture code - 불확실성 회피	점수 (0~100)

서 다른 값을 활용하게 되며, GDP와 같이 전년 대비 변화율이 반영되어 시간적 변화율을 고려하였다. <Table 1>에서 예측모형 개발에 활용된 변수들을 보여주고 있으며, <Table 2>에서는 변

Table 2. Descriptive Statistics of Tri-dimensional Variables Used in Predictive Model Development

	변수 명	N	평균	표준편차	왜도	첨도
경제적 요인	COUNTRYCD (불연속형변수)					
	NY_GDP_MKTP_CD	21,189	1,644,687,335,676	3,433,911,976,109	4.024	16.564
	NY_GDP_MKTP_CD_1Y	21,189	1,547,985,101,356	3,257,227,668,213	4.071	17.104
	SP_POP_TOTL	21,189	122,508,700	284,441,865	3.897	14.108
	PA_NUS_FCRF	21,189	1,666	6,049	4.078	16.112
	TRADE_COUNTRYCD	21,189	342,490,635,790	463,919,395,070	2.931	9.129
	KMDIST	21,189	7,952	4,054	0.602	0.502
	SeaDistance	21,189	6,893.90	3,596.61	-0.298	-1.257
	SNDIST	21,167	6,368	2,518	0.946	0.381
산업 구조적 요인	HSCD (불연속형변수)					
	TRADE_HSCD	21,189	15,169,280,110	39,156,772,230	6.102	42.526
	TARIFF_AVG	21,066	3	9	16.672	716.881
	TRADE_HSCD_COUNTRYCD	21,179	308,247,467	2,129,496,373	29.299	1172.481
	KR_TRADE_HSCD_COUNTRYCD	21,189	17,939,632	481,003,321	112.976	14,593.21
문화적 요인	BTS_2016	21,189	1,028.79	452.693	0.867	0.771
	BTS_2017	21,189	1,934.36	645.84	0.29	0.103
	BTS_2018	21,189	2,713.20	664.349	-0.352	-0.136
	exo_2016	21,189	888.66	529.894	1.086	0.421
	exo_2017	21,189	804.05	434.618	1.53	2.698
	exo_2018	21,189	633.96	307.918	1.812	4.078
	kpop_2016	21,189	44.95	42.352	3.278	13.606
	kpop_2017	21,189	47.21	39.031	2.601	9.069
	kpop_2018	21,189	53.77	38.786	2.421	8.515
	wanna_one_2016	21,189	23.43	33.892	4.197	20.733
	wanna_one_2017	21,189	131.18	139.897	2.335	5.179
	wanna_one_2018	21,189	185.92	221.664	2.498	5.454
	IC_BUS_EASE_DFRN_DB	21,189	71	11	-0.652	-0.258
	PowerDistance	21,189	62.47	22.695	-0.35	-0.713
	Individualism	21,189	47.01	24.076	0.258	-1.226
	Masculinity	21,189	51.84	18.841	0.051	0.852
	UncertaintyAvoidance	21,189	65.28	22.725	-0.438	-0.85

수들의 기술적 통계량이 제시되어 있다. 본 연구에서는 IBM SPSS 26 및 IBM Modeler 18를 개발 툴(Tool)로 활용하였다.

2.3. Data Preparing

Data Understanding의 모든 Case에서 단일 목표 변수 값으로 정의될 수 있는 두 가지 불연속 변수(국가 및 HS Code)의 조합을 확인할 수 있

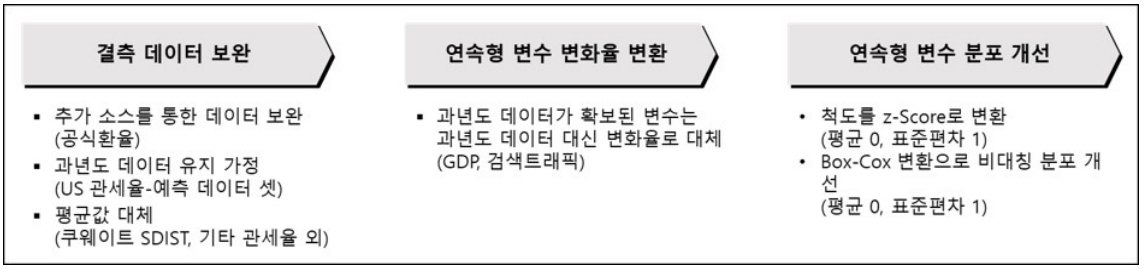


Figure 2. Data Preparing Process

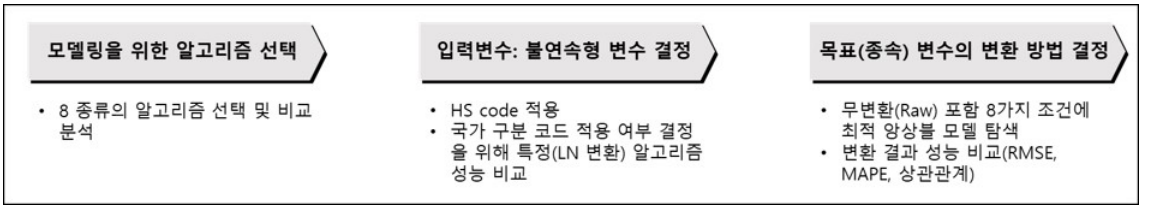
었으며, 일부 조합에서는 결측값이 확인 되었다. Data Understanding에서 나타난 연속형 변수들의 특징으로 첫째, 대부분의 Case에서 결측값이 없었지만 일부(0.7%) 결측값이 있는 것이 확인되었고, 연구의 목적상 결측값이 존재하는 경우도 적절히 예측 되어야 하였기 때문에 결측값의 보완이 필요하다 판단되었다. 둘째, 대부분의 변수는 당해 국가나 제품에 대한 특징을 설명할 수 있는 변수들이었지만, 일부 변수는 전년대비 변화도 설명할 수 있는 변수(GDP, 검색트래픽)라는 점에 착안하면 변화율이라는 변수를 파생시킬 수 있다는 것이 파악되었다. 셋째, 측정 단위뿐만 아니라 크기(Scale)와 분포가 매우 다양하기 때문에 데이터 학습에서 측정 단위나 분포에 따른 영향을 최소화할 필요가 있음이 확인되었다. 따라서, 결측값은 추가 정보원을 통해서 적절한 값을 추가 및 평균으로 대체하거나, 변환율이 측정 가능한 변수는 변화율 값을 파생시켜 활용하고, 변수간 크기 차이를 없애기 위해서 값을 표준화해서 활용하고자 하였다. 이상에서 설명한 세 단계의 데이터 준비(Data Preparing)과정을 정리하면 다음 <Figure 2>와 같다.

첫번째 단계인 결측 데이터 보완은 1) 추가 정보원(Source)에서 새로운 데이터 추가가 가능한지 확인하고, 2) 새로운 정보원에서 적절한 정보를 찾을 수 없을 때 과년도 자료가 제시된 경우 데이터 유지 가정(e.g., 예측용 데이터에서는 결측인데, 분석용 데이터에서는 존재하는 경우), 3) 상기 조건 모두에 맞지 않은 경우는 결측값을 평균으로 대체하고자 하였다. 예를 들어, PA_NUS_FCRF(공식 확률) 변수는 한국은행경제통계시스템(ECOS)에서 공식 환율(미국 달러에 대한 현지 통화 단위, 월평균을 기준으로 한 연평균)로 대체하였으며, 쿠웨이트 SDIST, 기타 제품 관세율 등은 평균 대체하였다. 두번째 단계인, 연속형 변수 중 과년도 데이터가 있는 경우는 변화율 변수를 파생시켰는데, 원래의 값이 국가별 차이를 설명할 수 있는 경우는 원래의 값도 분석에 같이 사용하였다. 예를 들어, GDP의 경우 전년도 기준으로 당해 년도 GDP 변화율을 새로운 변수로 파생시켰으며(GDP_growth), 당해 년도 GDP도 국가간 차이를 분석하는데 유의미할 수 있기 때문에 분석에는 활용하였다. 또한, 검색트래픽의 경우 전년도 기준으로 당해 년도 K-POP 관련 키워드 검색 변화율을 변수로 활용했는데(bts_t, exo_t, wanna_one_t, kpop_t), 검색트래픽은 해당국의 검색량의 변화를 상대적인 값으로 측정했기 때문에 국가간의 비교가 불가해 변화율만 분석에 활용하였다 (Jun et al., 2014). 마지막 세번째 단계에서는 측정 단위뿐만

〈Table 3〉 Improved Statistical Characteristics of Tri-dimensional Variables through Normalized Transformation

	변수 명	N	최소값	최대값	평균	표준편차	왜도	첨도
경제적 요인	COUNTRYCD							
	NY_GDP_MKTP_CD_trans	21,189	-0.471	5.109	0	1	3.976	15.895
	NY_GDP_MKTP_CD_1Y_trans	21,189	-0.476	5.205	0	1	4.024	16.564
	SP_POP_TOTL_trans	21,189	-0.420	4.430	0	1	3.897	14.101
	PA_NUS_FCRF_trans	21,189	-0.280	4.568	0	1	3.788	13.059
	TRADE_COUNTRYCD_trans	21,189	-0.726	4.309	0	1	2.884	8.655
	KMDIST_trans	21,189	-1.726	2.571	0	1	0.602	0.502
	SeaDistance_trans	21,189	-1.822	1.429	0	1	-0.298	-1.257
	SNDIST_trans	21,189	-2.086	3.458	0	1	0.901	0.194
산업 구조적 요인	HSCD	21,189	n/a	n/a	n/a	n/a	n/a	n/a
	TRADE_HSCD_trans	21,189	-0.370	10.249	0	1	6.572	50.857
	TARIFF_AVG_trans	21,189	-0.514	60.106	0	1	16.488	754.145
	TRADE_HSCD_COUNTRYCD_trans	21,189	-0.142	53.286	0	1	30.412	1265.400
	Sqrt_KR_TRADE_HSCD_COUNTRYCD	21,189	0.000	252379	1539	3945.948	20.363	920.190
문화적 요인	BTS_t_trans	21,189	-1.501	4.556	0	1	2.165	7.969
	exo_t_trans	21,189	-2.405	2.072	0	1	-0.418	-0.167
	kpop_t_trans	21,189	-2.195	4.789	0	1	2.112	9.186
	wanna_one_t_trans	21,189	-1.797	3.299	0	1	0.869	0.736
	IC_BUS_EASE_DFRN_DB_trans	21,189	-2.752	1.522	0	1	-0.744	-0.006
	PowerDistance_trans	21,189	-2.268	1.654	0	1	-0.350	-0.713
	Individualism_trans	21,189	-1.703	1.827	0	1	0.258	-1.226
	Masculinity_trans	21,189	-2.327	2.715	0	1	0.051	0.852
	UncertaintyAvoidance_trans	21,189	-2.521	1.440	0	1	-0.438	-0.850

아니라 크기와 분포에서 차이를 최소화하기 위해서 모든 연속형 변수는 표준화를 수행하였다. 즉, 연속형 변수들의 측정 단위에 의한 크기 차이와 분포를 개선하기 위해서 Z-score와 Box-Cox 변환을 적용하였다. 이상의 표준화 과정을 진행한 연속형 변수의 변환 결과는 아래의 표와 같은데, 결측값은 제거되어 모든 Case(21,189)를 활용할 수 있게 되었고, 평균은 0, 표준편차는 1로 조정되어 측정 Scale의 영향도 제거되었다. < Table 3 >에서 변수는 제곱근(SQRT)으로 변환된 사례를 제시한 것으로 왜도와 첨도가 개선된 것을 확인할 수 있다.



〈Figure 3〉 Modeling Process

3. Research Results

3.1. Modeling

Data Preparing에서 준비된 분석용 데이터를 바탕으로 예측을 위한 적절한 알고리즘을 선정하였는데, 본 연구에서는 복수의 알고리즘 조합을 활용하는 앙상블 모델링(Ensemble Modeling) 방법을 선정하였다. 예측용 앙상블 모델을 만들기 위해서 8가지 알고리즘(LSVM, CRT, CHAID, Exhaustive CHAID, Random Forest, GLM, LRA, NN)을 후보로 선택하였다. 선형회귀분석(Linear Regression Analysis, 이하 LRA)은 입력 변수와 목표 변수간 관계를 선형식으로 추정하는 모델이며, Generalized Linear Model(GLM)은 입력 변수들과 목표 변수들과의 관계가 정규분포에 따르지 않는 경우에도 설명이 가능한 대안 모델이다 (Kim and Hong, 2004). 기계학습 기법으로는 의사결정나무(Decision Tree), 신경망(Neural Network), 서포트 벡터 머신(Support Vector Machine)을 활용했는데, 의사결정나무 분석에서는 연속형 변수 예측이 가능한 CRT, CHAID 그리고 Exhaustive CHAID을 활용하였다. 특히, 의사결정 나무를 다수 만들어 예측력을 높이는 Random Forest 방법은 블랙박스 모델로 구분되기도 하지만 의사결정 나무 분석의 확장(중요도 분석 가능)이 가능하다는 측면에서 함께 활용하였다 (Choi et al., 2018). 기계학습 기법 중에서 서포트 벡터 머신은 주로 분류 문제에 활용되는데, 손실함수를 도입해서 회귀 분석에 활용할 수 있도록 확장한 Linear SVM(이하 LSVM)을 후보 알고리즘으로 활용하였다 (Choi et al., 2018). 마지막으로 기계학습 기법 중에서 신경망 기법은 뇌의 뉴런들의 상호작용을 모형화한 프로세스 알고리즘으로 비선형적이고 Noise가 많은 영역에서도 적합한 모형을 구축할 수 있다 (Kim and Hong, 2004). 다만, Gradient Boosting으로 구조화된 데이터 혹은 표 형식의 데이터에 적용하여 예측 모델 구축을 지원하는 강력한 Gradient Boost 의사결정 나무를 구현한 XGBoost 알고리즘도 있지만, 입력과 목표 변수의 사이를 설명할 수 없는 블랙박스 모형이라는 측면에서 본 연구의 Design Principle에 위배되어 활용을 지양하였다. 하지만, 추가적으로 모델의 평가 부분에서 본 연구에서 제안한 예측용 모델의 예측력을 설명하기 위해, XGBoost 알고리즘을 활용한 모델링 결과와 성능을 비교하였다.

모델링 과정은 1) 앙상블 모델링을 위한 후보 알고리즘 선택, 2) 불연속형 입력 변수의 조합을 경정하기 위한 성능 비교, 3) 목표 변수의 변환 방법을 결정하기 위한 성능 비교의 세 단계로 구성할 수 있다 (<Figure 3> 참조).

〈Table 4〉 Comparison of Algorithm Performance by Discrete Variable Utilization Method

HS Code 단독 활용				HS Code + 국가 Code 병행 활용			
알고리즘	상관관계	사용된 필드	상대오차	알고리즘	상관관계	사용된 필드	상대오차
Exhaustive CHAID	0.770	20	0.407	Exhaustive CHAID	0.781	19	0.390
Random forest	0.763	21	0.419	GLM	0.774	22	0.401
GLM	0.721	21	0.481	Random forest	0.743	22	0.454
NN	0.712	21	0.493	LSVM	0.736	22	0.494
CRT	0.708	20	0.499	NN	0.731	22	0.465
CHAID	0.707	11	0.501	CHAID	0.719	10	0.482
LSVM	0.658	21	0.615	CRT	0.711	21	0.494
LRA	0.560	21	0.687	LRA	0.568	22	0.677

첫번째 단계인 앙상블 모델링을 위한 후보 알고리즘 선택 과정에서는 입력과 목표 변수의 관계에서 선형 및 비선형 관계를 모두 고려하기 위해 앞서 설명된 바와 같이 8가지 후보 알고리즘을 선정하였다.

두번째 불연속형 변수를 선택하는 단계에서는 입력 변수 선택 과정에서는 명목형 변수를 'HS code' 단독으로 사용하거나 'HS Code + 국가 Code'를 병행하여 사용하는 두가지 접근법을 비교하였다 (<Table 4> 참조). 두가지 접근 방법 중 적절한 방법을 선택하기 위한 성능 지표는 상관관계와 평균 절대 오차(또는 상대 오차)가 개선되는 정도를 확인하였다. 여기서 목표(종속) 변수는 Ln으로 변환된 '차년도 HS Code/국가별 한국 수출액'이며, Ln(종속)변수 변환의 사례의 경우, HS Code + 국가 Code 병행의 경우가 HS Code만 단독으로 사용하는 것에 비해 상관관계와 상대 오차에서 개선의 효과를 보였다. 결과적으로, 입력 변수에서 불연속형 변수는 HS Code + 국가 Code를 병행하여 활용하는 것으로 결정하였다.

세번째, 목표 변수의 변환 방법을 결정하기 위한 성능 비교 단계에서는 변환없이 사용하는 경우(Raw-무변환)를 포함해서 여덟 가지 방법의 예측 성능을 비교하였다 (<Table 5> 참조). 목표(종속) 변수의 정규분포를 위해서라면 Box-Cox 변환을 활용할 수 있지만, 연속형 목표 변수의 예측 모형 개발을 위한 변환이기 때문에 여러가지 변환에 따른 예측 성능을 비교하는 방식으로 최적 변환 방법을 탐색하였다. 연속형 목표 변수의 예측 성능 측정에는 Mean Absolute Error(MAE), Mean Absolute Percentage Error(MAPE), Root Mean Square Error (RMSE)가 많이 활용되며, 이런 값이 가지는 한계를 보완하기 위해서 Rank Correlation과 같은 상관관계가 활용되기도 한다 (Mathaba et al., 2014). 본 연구에서는 예측 모델의 예측 성능을 비교하기 위해서 선행연구를 바탕으로 MAPE, RMSE 그리고 Pearson 상관관계를 활용하였다. 구체적인 목표 변수 변환 방법은 무변환(Raw), 로그 변환(Ln, Log), 지수 변환 3/4,

〈Table 5〉 Results of Ensemble Algorithm Configuration by Target Variable Conversion Method

알고리즘 / 변환	Raw (무변환)	Ln	Log	지수 (0.75)	지수 (0.67)	지수 (0.5)	지수 (0.33)	지수 (0.25)
Exhaustive CHAID		V	V				V	V
CHAID								
CRT								
Random forest		V	V				V	
NN								V
LSVM	V			V	V	V	V	
GLM	V	V	V	V	V	V	V	V
LRA	V			V	V			

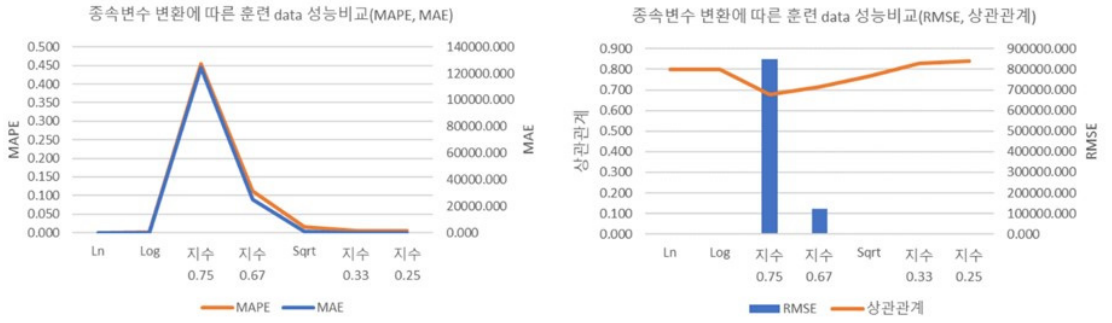
〈Table 6〉 Comparison of prediction performance according to target variable conversion method

변환 방법	Ln	Log	지수 0.75	지수 0.67	지수 0.50 (Sqrt)	지수 0.33	지수 0.25	Raw(none)
평균 오차	-0.002	-0.001	-3.218	-883.566	-69.800	-1.582	0.000	-165,127
MAE	1.413	0.613	124,327	24,808	1,035	39.188	8.992	39,673,742
MAPE	0.001	0.001	0.454	0.113	0.016	0.005	0.005	64.735
RMSE	2.006	0.87	847,650	123,872	2,529	66.988	13.436	398,687,427
상관관계	0.800	0.801	0.677	0.713	0.768	0.831	0.841	0.561
사용알고리즘	3	3	3	3	2	4	3	3

2/3, 1/2, 1/3, 1/4의 5가지를 고려하였는데, 이는 성능 지표들에서 Trade-off 관계가 관찰되었기 때문이다. 앙상블 모델 별 알고리즘 구성을 보면, 대체로 로그계열과 지수계열(무변환 포함)이 차이가 나며, 각각의 그룹에서 최적화된 알고리즘은 유사한 것을 확인할 수 있다.

목표 변수 변환 방법 별 앙상블 모델의 예측 성능을 비교한 결과는 <Table 6>에 제시되어 있는데, 지수 0.75, 0.67을 제외하면, 전반적으로 우수한 분석 성능이 나타나는 것을 확인할 수 있다. 또한, 지수 계열 변환은 지수가 낮아질수록 평균 오차 및 절대오차(MAE)가 낮아지고, 상관관계는 높아졌으며, MAPE는 대체적으로 성능이 우수해지는 경향을 보였다. 단, RMSE는 지수 0.5에서 가장 우수한 성능이 나타나 대조적인 결과가 도출되었다. 결과적으로, 변환된 종속 변수를 활용하는 것이 변환 없이(Raw) 종속 변수를 활용하는 것 보다 우수한 성능을 보여주고 있다.

목표(종속) 변수의 변환 방법에 따른 예측 성능을 그림으로 비교한 결과는 < Figure 4 >에 제



〈Figure 4〉 Comparison of prediction performance according to target (dependent) variable conversion

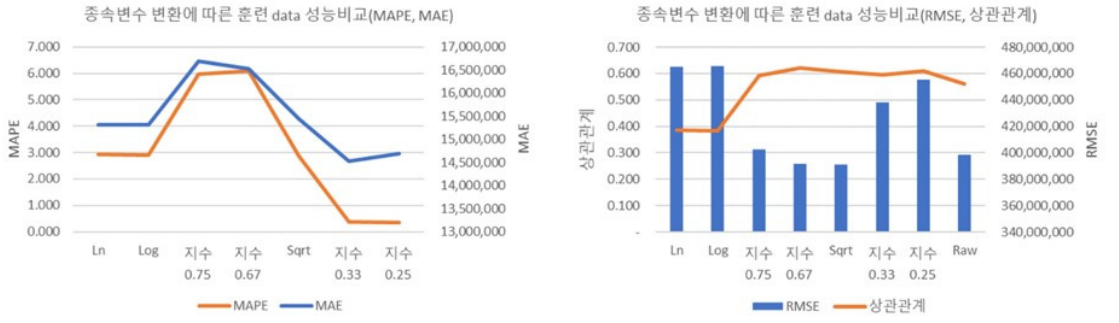
〈Table 7〉 Comparison of Prediction Performance According to Target Variable Conversion Method (After Target Variable Reduction)

변환 방법	Ln	Log	지수 0.75	지수 0.67	지수 0.50 (Sqrt)	지수 0.33	지수 0.25	Raw(none)
평균 오차	3,761,894	3,748,957	15,828,997	15,751,566	12,596,638	6,135,808	5,519,231	18,104,759 Real: 17.939,632
MAE	14,177,738	14,190,675	2,110,635	2,188,066	5,342,994	11,803,823	12,420,401	-165,127
MAPE	15,306,114	15,312,872	16,686,614	16,524,714	15,441,195	14,529,599	14,690,807	39,673,742
RMSE	2.922	2.894	5.977	6.073	2.862	0.376	0.348	64.735
상관관계	465,366,547	465,760,676	402,293,954	391,594,656	390,836,464	438,302,124	455,282,983	398,678,053

시되어 있으며, MAPE와 MAE, RMSE와 상관관계 결과가 각각 비교되어 있다. MAPE와 MAE가 기준일 때, 로그 계열(Ln, Log)과 제곱근(지수 0.5) 이하 지수 변환에서 우수한 성능을 보이는 것을 확인할 수 있으며, 여기서 두 값의 경향이 거의 차이가 없는 것은 변수 변환으로 이상값의 영향이 이미 상쇄되었기 때문으로 판단된다. 또한, RMSE와 상관관계가 기준일 때, 상관관계에서는 지수 0.75와 0.67에서 성능이 떨어졌으며, RMSE도 지수 0.75와 0.67에서 성능이 좋지 않은 것을 확인하였다. 상대적으로 무변환(Raw)은 성능이 크게 떨어지는 것으로 확인되었다. 최종적인 변환 방법은 이어지는 분석에서 수행 될 재변환(변환 변수의 환원 후) 성능 평가로 판단될 것이다.

3.2. Evaluation

앞서 Modeling에서 제시된 목표(종속) 변수 방법에 따른 예측 성능은 변환된 목표 변수 값에 대한 성능으로, 변환 방법에 따라 크기가 다른 원자료에 미치는 영향이 다르기 때문에 예측 결과를 환원한(원래 값으로 재변환) 결과에 대한 예측 성능을 비교해서 모델링 결과의 최종 예측 성능을 평가하였다. 목표 변수 변환 방법에 따른



〈Figure 5〉 Comparison of Predictive Performance According to Target (Dependent) Variable Conversion (Reduction)

환원 결과의 앙상블 모델 예측 성능을 비교한 결과는 다음 <Table 7>에 제시되어 있는데, 환원 전(변환된 목표 변수를 예측한) 성능(<Table 6> 참조)과 차이가 큰 것을 확인할 수 있다.

결과를 살펴보면, 로그 계열(Ln, Log) 변환은 환원 후 성능이 매우 크게 떨어지는 결과를 보이며, 스케일이 큰 변수에서 변환에 따른 왜곡 현상이 크게 나타났다. 지수 계열 변환의 경우, MAPE와 MAE는 지수가 작을수록 우수한 성능을 보였지만, RMSE는 오히려 지수가 클수록 우수한 성능을 보여서 Trade-off 현상이 나타났고, 상관관계에서는 일정한 경향이 나타나지 않았다. 로그 계열(Ln, Log) 변환과 낮은 지수 계열 변환에서는 종속 변수를 크게 과소 추정하는 현상이 나타났는데, 평균이 실제값에 크게 못 미치는 예측 결과를 보이며 모든 지표에서 우수한 결과는 나타나지 않았지만, 비교적 여러 지표에서 균형 잡힌 예측력을 유지한 목표 변수 변환 방법은 지수 0.5의 경우로 확인되었다.

목표(종속) 변수의 변환 방법에 따른 환원 후 예측 성능을 그림으로 비교한 결과가 다음 <Figure 5>에 제시되어 있는데, MAPE와 MAE, RMSE와 상관관계 각각의 결과가 비교되어 있다. MAPE와 MAE를 기준으로 보면, <Figure 4>와 비교할 때 여전히 지수 0.75와 0.67에서 성능이 떨어졌지만, 로그 계열(Ln, Log)도 제곱근(SQRT, 지수 0.5) 이하 지수 변환에 비해서 성능이 떨어지는 것을 확인하였다 (지수 0.5이하가 우수). 상관관계도 <Figure 4>의 환원 전 결과와 다른 경향이 나타나는데, 로그 계열(Ln, Log)은 무변환보다도 낮은 상관관계를 보여 성능이 불량한 것을 확인했으며, RMSE에서는 무변환보다 우수한 경우는 지수 0.67과 0.5로만 나타났다.

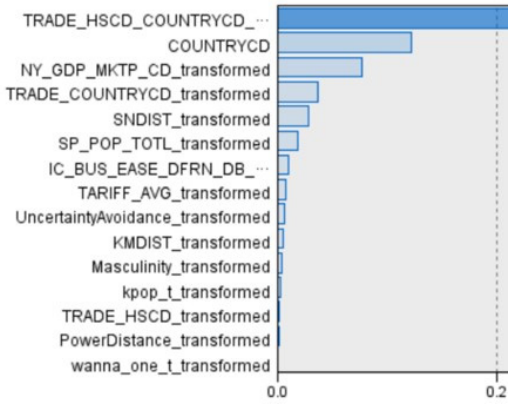
종속 변수 환원 후 성능 비교 결과 가장 균형 잡힌 즉 안정적 예측력을 유지한 지수 0.5(제곱근, Sqrt) 변환 모델을 최종 예측 모델로 선정하였다. 지수 0.5 변환 모델은 다른 변환 모델과 비교해서 안정성(Stable)이 돋보였는데, 특정 성능 지표로 치중되지 않는 균형 있는(Balanced) 성능을 보였다. 예측 결과의 크기에 대한 방향성을 의미하는 상관관계는 8개 목표 변수 변환 모델 중 3번째로 우수했고, 틀린 크기 값이 적어야 하는 RMSE는 8개 변환 모델 중에서 가장 우수했으며, 특정 데이터(수입액이 적거나 큰)만 틀려

〈Table 8〉 GLM Algorithm Modeling Result (Effect by Variable)

모형 효과 검정				
소스	제 III 유형			
	B	Wald 카이제곱	자유도	유의확률
(수정된 모형)	148.316	3402.689	1	0.01
HSCD	개별 적용	4704.504	498	0.01
COUNTRYCD	개별 적용	460.103	26	0.01
TARIFF_AVG_transformed	-77.415	12.797	1	0.01
SNDIST_transformed	360.315	20.015	1	0.01
TRADE_HSCD_COUNTRYCD_transformed	2292.818	14444.29	1	0.01

서는 안되는 MAPE에서는 8개 목표 변수 변환 모델 중 3번째로 우수했다.

최종 예측 (앙상블) 모델로 도출된 모델을 구체적으로 살펴보면, 세부 알고리즘에서는 GLM과 LSVM의 성능이 다른 알고리즘보다 성능이 뛰어났으며, 두 알고리즘이 제시하는 예측 값의 평균을 앙상블 모델의 예측 값으로 사용하였다. 앙상블 모델의 예측 변수 중요도는 다음 <Figure 6>과 같다.



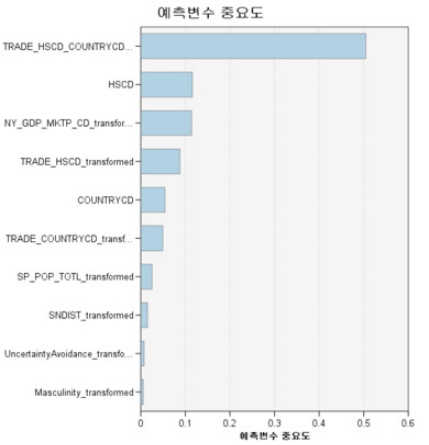
〈Figure 6〉 Predictor Importance of Predictive Ensemble Models (Top 15)

종속 변수를 환원(재 변환)한 예측 성능을 고려해서 입력 변수 중 명목형 변수는 HS/국가 Code를 모두 고려하고, 종속 변수는 제곱근(Sqrt, 지수 0.5)으로 변환한 후 앙상블 모델을 개발하였다. 앙상블 모델의 예측변수 중요도를 보면 국가별 HS Code 별 교역량(연속형)이 가장 높게 나타났으며, 다음으로 국가 Code(명목형)가 중요한 것으로 나타났다. 특히 추가로 고려한 호프스테드(국가 문화 특성) 지수와 한류 관련 검색 트래픽(한국에 대한 관심)도 예측에 유의미한 것으로 나타난 것이 주목할 만하다. 세부 알고리즘 중에서 GLM은 확률분포를 정규로서 연결함수는 항등으로 가정했으며, 결과는 Likelihood Ratio P <0.001로 나타났다. 구체적인 모형 효과 검정 결과는 다음 <Table 8>과 같으며, 관세가부(-)의 효과를 수입국 거리와 국가별 HS Code 별 수입량이 정(-)의 효과를 보였다. 또한, 여기서 HS Code와 국가 Code는 <Figure 9>와 같이 특정 계수 값을 가지게 된다.

〈Table 9〉 GLM Algorithm Modeling Results (Effect by Hs Code Dummy Variable)

모수 추정값							
모수	B	표준화 오차	95% Wald 신뢰구간		가설검정		
			하한	상한	Wald 카이제곱	자유도	유의확률
(수정된 모형)	148.315	446.1768	-726.176	1022.805	0.11	1	0.74
[HSCD=190219]	2937.587	570.3513	1819.719	4055.455	26.528	1	0.01
[HSCD=190230]	4280.404	573.0724	3157.202	5403.605	55.789	1	0.01
[HSCD=190590]	3047.024	569.4671	1930.889	4163.159	28.63	1	0.01
[HSCD=210390]	3190.36	569.5691	2074.025	4306.694	31.375	1	0.01
…	…	…	…	…	…	…	…

대상 필드	Sqrt(KR_TRADE_HSCD_COUNTRYCD)
모델 작성 방법	선형 SVM
입력된 예측변수 수	22
최종 모델의 예측변수 수	21
정규화 유형	L2
페널티 매개변수(람다)	0.1
회귀분석 전체자릿수(엡실론)	0.1
평균 제곱 오류	6,668,293



〈Figure 7〉 LSVM Training Conditions (Left) and Importance of Input Variables (Right)

마지막으로, 세부 알고리즘 중에서 LSVM의 조건 설정과 입력 변수의 중요도는 다음 <Figure 7>과 같다.

3.3. Deployment

본 연구에서는 Design Principle에 입각하여 실무적 활용도를 높이면서, 이론적으로 의미 있는 예측 프로세스를 추구하였기 때문에, XGBoost와 같이 예측력이 우수하지만 요인을 설명할 수 없는 블랙박스 모형의 사용을 지양하였다. 하지만, 향후 예측 모델 개선에 대한 연구 확장을 위해 앞서 제시한 예측 모델의 예측력을 블랙박스 모형(e.g., XGBoost 알고리즘)과 비교해보았으며, 예측의 정확도를 목표로 하였을 때는 XGBoost와 앙상블 모형을 병행하는 방법이 효과적인 것을 확인하였다. 다음 < Table 10 >에는 XGBoost를 활용한 예측 모형의 성능이 비교되어 있으며, XGBoost의 예측 성능 경쟁력을 확인할 수 있다.

동일한 데이터에 대해서 XGBoost는 높은 예

〈Table 10〉 Comparison of Prediction Performance According to Target Variable Conversion Method (Added XGBoost)

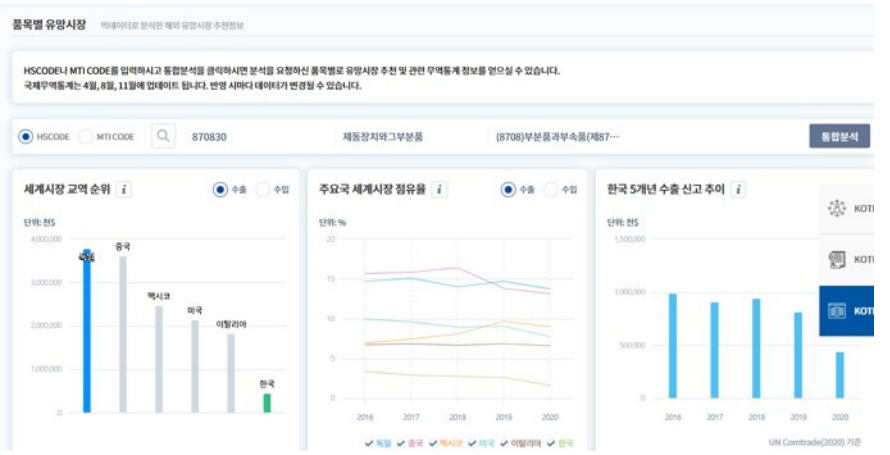
변환 방법	지수 0.75	지수 0.67	제곱근	제곱근	지수 0.33	지수 0.25	Raw(none)	Raw(none)
	앙상블	앙상블	앙상블	XGBoost	앙상블	앙상블	앙상블	XGBoost
평균	15,828,997	15,751,566	12,596,638	11,715,257	6,135,808	5,519,231	18,104,759 Real: 17.939,632	14,939,633
평균 오차	2,110,635	2,188,066	5,342,994	6,224,375	11,803,823	12,420,401	-165,127	-1.058
MAE	16,686,614	16,524,714	15,441,195	15,111,381	14,529,599	14,690,807	39,673,742	44,923,966
MAPE	5.977	6.073	2.862	2.710	0.376	0.348	64.735	65.571
RMSE	402,293,954	391,594,656	390,836,464	386,275,600	438,302,124	455,282,983	398,678,053	382,559,638
상관관계	0.593	0.621	0.606	0.624	0.595	0.61	0.561	0.607

측력을 보였으며, 종속변수의 변환 없이 수행된 분석에서도 변환된 변수를 활용한 앙상블 모델보다 RMSE와 상관관계에서 높은 경쟁력을 보였다. 다만, 평균 절대오차와 MAPE에서 성능이 취약한 것으로 나타났다. 본 연구에서 선택한 최종 모델(제곱근-앙상블)과 동일한 변환(제곱근)에 XGBoost를 적용한 경우 평균과 오차에서 다소 성능이 떨어졌지만, RMSE, MAPE, 상관관계에서 근소하게 우수한 성능을 확인할 수 있었다. 따라서 예측의 정확도에 초점을 맞추었을 때 향후 연구에서 XGBoost를 앙상블 모형과 병행하는 방법이 활용될 수 있을 것이다.

4. Conclusion

본 연구에서는 각 나라의 한국 수입액을 예측하기 위해 문화적 요인을 고려하였으며, 문화적 요인이 수입액에 미치는 영향을 PUSH-PULL 프레임워크 관점에서 정량화 하여 예측 모형 개발에 활용하였다. 특히, 본 연구에서 제시한 최종 예측 모델을 Design Principle에 기반하여 설계한 것을 특징으로 제시할 수 있는데, 1) 신규로 추가한 데이터 소스를 통해 한국에 대한 관심 및 문화적 특성이 반영될 수 있는 모형으로 구축하였고, 2) 경제적 요인 등의 변화와 품목 및 국가 Code를 입력하면 예측값을 즉시 불러올 수 있도록 실용적으로 편의성 있게 설계했으며, 3) 이론적으로도 의미 있는 결과를 도출하기 위해서 입력과 목표 변수간의 관계를 해석 가능한 알고리즘을 중심으로 설계하였다는 점이다.

분석 결과를 살펴보면, 개별 국가 및 HS Code에 따른 우리나라 수출액을 예측한 결과 국가별 HS Code 별 수입량이 압도적으로 높은 영향력을 가지고 있는 것으로 나타났으며, 해당국가의 수입액을 우리나라의 수출액이 넘어설 수 없다는 점을 감안하면, 분석 결과는 모델의 강건성을 간접적으로 보여준다고 할 수 있다. 앙상블 모델에서는 HS Code 별 수입량과 국가 Code가 중요한 예측 변수로 나타났으며, 국가별 문화 특성을 의미하는 호프스테드 지수와 한국에 대한 관심을 의미하는 검색트래픽 변수도 예측에 유의미한 영향을 미치는 것으로 확인되었다. 그 밖에 선행연구에 중요한 요인으로 설명된 경제여건(GDP)



〈Figure 8〉 Example of Big Data Service in KOTRA

이나 인구도 중요한 결정(입력) 요인으로 나타났으며, 해당 국가와 수입 국가간 평균 거리(SDIST)나 해당 구가와 한국과의 거리(KDIST)도 중요한 요인으로 나타났고, 비즈니스 용이성(IC_BUS)도 역시 중요한 요인으로 파악되었다. 특히, 최적 예측 모형에서 주목할 부분은 문화적 요인으로 호프스테드 지수에서 불확실성의 회피(Uncertainty Avoidance), 남성성(Masculinity), 권력의 격차(Power Distance)가 목표 변수 예측에 영향을 줄 수 있는 변수로 나타났으며, 한국에 대한 관심에서 K-POP이나 Wanna One에 대한 검색도 적지만 유의미한 영향이 존재할 수 있음이 확인되었다.

위와 같은 예측 모델 개발 결과는 기존 연구들에서는 수출에 영향을 미치는 주요 요인으로 경제적 및 산업구조적인 측면을 주로 다루어 왔지만, 최근 한국에 대한 관심의 증가, 한류 열풍 등의 추세를 고려하여 문화적인 측면을 함께 다룰 필요가 있다는 것을 시사할 수 있다. 특히, 한류와 같은 문화적 요인들은 일정 수준 통제가 가능하며, 이를 활용하여 수출 대상국가에 대해 국가적인 차원에서 특정 제품이나 서비스 등에 대한 광고 및 홍보활동을 수행할 수 있을 것이다. 또한, 수출 지원 기관(e.g., KOTRA)에서는 국내 기업이 해외에 특정 품목을 수출할 경우, 기업이 수출 대상 국가의 문화적인 특성을 이해할 수 있도록 안내해 주는 것이 필요할 것이다. 호프스테드 지수는 통제할 수 없는 요인이지만, 특정 국가의 문화적 특성을 나타내는 지표들을 기반으로 커뮤니케이션 코드와 같은 행동 모델을 만들어 낼 수 있을 것으로 예상된다.

본 연구는 기술적 측면, 경제적 측면, 정책적 측면에서 의미 있는 시사점을 제시할 수 있으며, 수입 예측 모형을 활용하여 중소•중견기업의 수출 지원 전략에 의미 있는 기여를 할 수 있을 것으로 기대된다. 첫번째, 기술적 측면에서는 KOTRA 등 수출 지원기관에서 제공하는 데이터 기반 수출 지원 서비스에 연계 및 활용될 수 있을 것으로 기대된다. 본 연구에서 개발한 예측 모델을 KOTRA 무역투자빅데이터의 품목별 유망시장 서비스에서 연계하여 (KOTRA, 2021), HS Code 및 국가별 시장 규모 예측 기반의 미래

지향적인 추천 서비스가 가능해 질 것이라 예상한다 <Figure 8 참조>. 현재 Random 효과(시계열)나 기술적 통계 서비스를 보완하기 위해 문화적 요인을 투입해본다면, 고정효과(Fixed Effect) 중심의 예측 모델 서비스가 추가될 수 있을 것이다.

두번째, 경제적 측면에서는 예측된 수출 시장 정보를 기반으로 수출 유망 제품이나 국가를 추천함으로써, 기업의 신속한 의사결정을 지원할 수 있고, 특히 문화적 요소까지 고려한 유망 제품이나 국가를 추천함으로써 수출확대에 기여할 수 있는 인프라 서비스를 개발할 수 있을 것이라 기대된다. 추가적으로, 이러한 서비스를 정교화하기 위해서는 HS Code 별 세부 예측 모델을 분화시켜 개발하는 것도 검토될 필요가 있다 (e.g., 500개 세부 예측 모델의 개발을 통해 KOTRA 빅데이터 플랫폼 서비스에 확대 적용 가능). 더불어, 향후 문화적 요인을 고려한 중소•중견기업의 적합성(기업이나 제품의 특성 비교) 모델을 개발하여 활용한다면, 유망 수출 국가나 제품 추천 모델의 서비스를 더욱 개선할 수 있을 것으로 기대된다. 세번째, 정책적 측면에서도 시사점이 존재하는데, 본 연구의 결과는 향후 수출과 관련한 정책의 방향성을 수립함에 있어서 객관적인 데이터에 기반하여 정책적인 선택이나 결정에 도움을 줄 수 있는 가이드라인이 될 수 있을 것이라 기대된다 (i.g., Evidence-based Policy). 본 연구에서는 예측 모델의 활용에서 개별 요인이 가지는 통계적 유의미성이나 중요도가 What-if 분석의 활용을 위한 의사결정에 중요하다고 생각했기 때문에 블랙박스 모델 활용은 지양하였는데, 이는 KOTRA와 같은 수출지원 기관이 정부나 공공기관에서 급격한 경제적 또는 산업구조적 변화에서 야기될 수 있는 수출의 영향을 예측하고, 대비하는 것에 도움을 줄 수 있을 것이다.

이와 같은 연구의 시사점에도 불구하고, 본 연구에서는 한 시점의 데이터만 가지고 모델링을 수행하여 시계열적인 특징을 반영하기 어려운 한계점이 존재한다. 향후 연구에서는 다년간의 시계열 데이터를 투입하여 분석한다면, 패널 자료의 고정과 임의 효과를 모두 고려한 정교한 예측 모형을 수립할 수 있을 것으로 예상한다. 또한 본 연구에서는 문화적 요인이 우리나라의 교역액에 미치는 영향을 확인하기 위해 호프스테드 모형에서 정량화 되어 있는 네 가지 차원의 문화 코드(권력거리, 개인주의, 남성성, 불확실성 회피)를 활용하였지만, 향후 연구에서는 경제적 의사결정에 영향을 주는 다양한 사회, 문화적 요인들을 고려하여 모형에 반영한다면 더욱 흥미로운 연구 결과를 도출할 수 있을 것으로 기대한다.

참고문헌(References)

Choi, C. H., J. S. Kim, D. H. Kim, J. H. Lee and D. H. Lee, "Development of Heavy Rain Damage Prediction Functions in the Seoul Capital Area Using Machine Learning Techniques", *Journal of The Korean Society of Hazard Mitigation,* Vol.18, No.7(2018), 435~447.

Choi, E. H. and W. Cho, ""The Impact of Korean Wave on Korean Cosmetics Exports to China"", *The Journal of Northeast Asian Economic Studies,* Vol.(30), No.3(2018), 21~43."

Choi, M. S., "The Effects of Korean Wave on Korea's Exportation", *International Commerce and Information Review,* Vol.(14), No.1(2012), 67~86.

Choi, S. W., "Testing the Validity of Hofstede's Cultural Dimensions", *Korean Public Administration Quarterly,* Vol.(27), No.4(2015), 1011~1032.

Financial Supervisory Service, *Key Factors Affecting Korea's Exports,* (2019), 1~14.

Hofstede, G., "Dimensionalizing cultures: The Hofstede model in context", *Online readings in psychology and culture,* Vol.(2), No.1(2011), 2307~0919.

Jun, S. P., Park, D. H. and Yeom, J, "The possibility of using search traffic information to explore consumer product attitudes and forecast consumer preference", *Technological Forecasting and Social Change,* Vol.86, (2014), 237-253.

Jung, M. S. and H. G. Kang, "The Effect of Korean Waves on Korean Exports using Spatial Effects Model", *International Business Review,* Vol.(19), No.2(2015), 1~22.

Kim, J. G. and S. Y. Ahn, "Hallyu's Effects on Exports of Korean Consumption Goods: A Comparative Study of Chinese and Vietnamese Consu", *Journal of International Trade and Industry Studies,* Vol.(17), No.3(2012), 193~217.

Kim, T. H. and H. K. Hong, "A Study on Apartment Price Models Using Regression Model and Neural Network Model Taehun Kim and Hankuk Hong", *The Korea Spatial Planning Review,* Vol.(43), (2004), 183~200.

KOTRA, Tri Big Trade investment big data service, http://www.kotra.or.kr/bigdata/guide, (2021), 15. Jun. (last accessed).

Lee, S. K., S. D. Kim and S. H. Kim "The Effect of Industrial • Trade Policies on the Export-Oriented Firms: The Comparison of Before-and-After FTA", *The Journal of the Korea Contents Association,* Vol.(17), No.11(2017), 135~142.

Mathaba, T., Xia, X., and J. Zhang, "Analyzing the economic benefit of electricity price forecast in industrial load scheduling", *Electric Power Systems Research,* Vol.(116), (2014), 158~165.

Moon, H. C., M. R. Bae and K. Y. Hwang, "The Impacts of the Service Quality of the Trade Promotion Agency on the Performance of Seoul Metropolitan and Local Exporting Firms", *International Commerce and Information Review,* Vol.(17), No.1(2015), 89~114.

Park, J. M., J. Hwang and J. S. Yu, "The Effects of Country Risk on the Cultural Contents Related Product Trades the Application of Gravity Function Model by Huntington's Civilizations and Hofstede's Cultural Dimensions", *Korean Journal of Journalism & Communication Studies,* Vol.(63), No.3(2019), 229~273."

Park, S. C., "The Impacts of Cultural Proximity on the Intensive and Extensive Margins of Exports", Korean Journal of EU Studies, Vol.(19), No.2(2014), 71~92.

Park, S. G. and S. M. Kang, "Analysis for Factors of Firms' Exporting in Industrial Organization Theory and Strategic Perspective", *The Journal of the Korea Contents Association,* Vol.(20), No.6(2020), 406~414.

Son, Y. J., "A Study on the Priority and Evaluation Criteria of the Korea Export Insurance System", *International Commerce and Information Review,* Vol.(13), No.1(2011), 179~196.

Sung, M., "The Effects of Organizational Culture Dimensions, Relational Norm Dimensions, and Participation on Conflict in Exchange

Relationship between Exporter and Importer", *Korean Journal of Marketing,* Vol.(28), No.3 (2013), 55~80.

Yoon, T. D., "Analysis on the Korea's Export to Brazil: Focused on the Causality with Korean GDP/total export", *Journal of International Area Studies,* Vol.(24), No.4(2020), 161~180.

Appendix A. Analysis data collection and development tool

예측 모델 개발을 위해서 활용한 변수들은 다음 Table A와 같으며, 각 변수들의 원 자료 원천(Original Source)도 같이 제시되어 있다.

Table A. Tri-dimensional variables and sources used in predictive model development

	변수 명	설명	Source
경제적 요인	COUNTRYCD	ISO 국가 Code	ISO 3166-1(Country Code) : 2020.03
	NY_GDP_MKTP_CD	GDP	World Bank (https://databank.worldbank.org/indicator/NY.GDP.MKTP.CD/1ff4a498/Popular-Indicators)
	NY_GDP_MKTP_CD_1Y	이전 년도 GDP	
	SP_POP_TOTL	인구 (연중 추정치)	World Bank (https://databank.worldbank.org/indicator/SP.POP.TOTL/1ff4a498/Popular-Indicators)
	PA_NUS_FCRF	공식 환율 (미국 달러에 대한 현지 통화 단위, 월평균을 기준으로 한 연평균)	World Bank (https://databank.worldbank.org/reports.aspx?source=2&series=PA.NUS.FCRF&country=)
	TRADE_COUNTRYCD	해당 연도 해당 국가의 전체 품목 수입금액	UN Comtrade (https://comtrade.un.org/data) $
	KMDIST	해당 국가와 한국과의 거리	CEPII (www.cepii.fr/distance/dist_cepii.dta, www.cepii.fr/distance/geo_cepii.dta)
	SeaDistance	해당 국가와 한국간의 선적 거리	https://sea-distances.org/
	SNDIST	해당 국가와 수입 국가 간 평균 거리	CEPII 데이터 활용하여 KOTRA 에서 가공
산업 구조적 요인	HSCD	HS Code (품목 Code)	UN Comtrade (https://comtrade.un.org/data)
	TRADE_HSCD	해당 연도 해당 품목의 전세계 총 수입금액	UN Comtrade (https://comtrade.un.org/data)
	TARIFF_AVG	해당 국가에서 해당 품목에 적용되는 평균 관세율	ITC (https://www.trademap.org/Index.aspx)
	TRADE_HSCD_COUNTRYCD	해당 연도 해당 국가의 해당 품목 수입금액	UN Comtrade (https://comtrade.un.org/data)
	KR_TRADE_HSCD_COUNTRYCD	내년 해당 국가가 해당 품목을 한국으로부터 수입한 금액	UN Comtrade (https://comtrade.un.org/data)
문화적 요인	BTS_2016-2018	2016 년부터 2018 년 까지의 BTS 에 대한 해당 국가의 검색트래픽	Google Trends (https://trends.google.com/trends/?geo=KR)

	변수 명	설명	Source
	exo_2016-2018	2016 년부터 2018 년까지의 EXO 에 대한 해당 국가의 검색트래픽	
	kpop_2016-2018	2016 년부터 2018 년까지의 KPOP 에 대한 해당 국가의 검색트래픽	
	wanna_one_2016-2018	2016 년부터 2018 년까지의 Wanna One 에 대한 해당 국가의 검색트래픽	
	IC_BUS_EASE_DFRN_DB	비즈니스 용이성 점수	World Bank (https://databank.worldbank.org/source/doing-business/Series/IC.BUS.EASE.DFRN.XQ.DB1719)
	PowerDistance	호프스테드 culture code - 권력거리	Hofstede insights(https://www.hofstede-insights.com/product/compare-countries/)
	Individualism	호프스테드 culture code - 개인주의	
	Masculinity	호프스테드 culture code - 남성성	
	UncertaintyAvoidance	호프스테드 culture code - 불확실성 회피	

Abstract

Development of the forecasting model for import volume by item of major countries based on economic, industrial structural and cultural factors: Focusing on the cultural factors of Korea*

Seung-pyo Jun** · Bong-Goon Seo*** · Do-Hyung Park****

The Korean economy has achieved continuous economic growth for the past several decades thanks to the government's export strategy policy. This increase in exports is playing a leading role in driving Korea's economic growth by improving economic efficiency, creating jobs, and promoting technology development. Traditionally, the main factors affecting Korea's exports can be found from two perspectives: economic factors and industrial structural factors. First, economic factors are related to exchange rates and global economic fluctuations. The impact of the exchange rate on Korea's exports depends on the exchange rate level and exchange rate volatility. Global economic fluctuations affect global import demand, which is an absolute factor influencing Korea's exports. Second, industrial structural factors are unique characteristics that occur depending on industries or products, such as slow international division of labor, increased domestic substitution of certain imported goods by China, and changes in overseas production patterns of major export industries. Looking at the most recent studies related to global exchanges, several literatures show the importance of cultural aspects as well as economic and industrial structural factors. Therefore, this study attempted to develop a forecasting model by considering cultural factors along with economic and industrial structural factors in calculating the import volume of each country from Korea. In particular, this study approaches the influence of cultural factors on imports of Korean products from the perspective of PUSH-PULL framework. The PUSH dimension is a perspective that Korea develops and actively promotes its own brand and can be defined as the degree of interest in each country for Korean

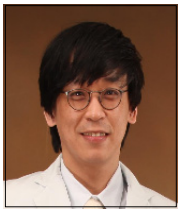
* This research was supported by the BK21 FOUR (Fostering Outstanding Universities for Research) funded by the Ministry of Education(MOE, Korea) and National Research Foundation of Korea(NRF).
** Global R&D Analysis Center, Korea Institute of Science and Technology Information
*** Korea National Industrial Convergence Center, Korea Institute of Industrial Technology
**** Corresponding author: Do-Hyung Park
School of Management Information Systems / Graduate School of Business IT, Kookmin University
77, Jeongneung-ro, Seongbuk-gu, Seoul, Republic of Korea
Tel: +82-2-910-5613, E-mail: dohyungpark@kookmin.ac.kr

brands represented by K-POP, K-FOOD, and K-CULTURE. In addition, the PULL dimension is a perspective centered on the cultural and psychological characteristics of the people of each country. This can be defined as how much they are inclined to accept Korean Flow as each country's cultural code represented by the country's governance system, masculinity, risk avoidance, and short-term/long-term orientation. The unique feature of this study is that the proposed final prediction model can be selected based on Design Principles. The design principles we presented are as follows. 1) A model was developed to reflect interest in Korea and cultural characteristics through newly added data sources. 2) It was designed in a practical and convenient way so that the forecast value can be immediately recalled by inputting changes in economic factors, item code and country code. 3) In order to derive theoretically meaningful results, an algorithm was selected that can interpret the relationship between the input and the target variable. This study can suggest meaningful implications from the technical, economic and policy aspects, and is expected to make a meaningful contribution to the export support strategies of small and medium-sized enterprises by using the import forecasting model.

Key Words : Trade; Forecasting Model for Import Volume; Design Thinking; Search Traffic; Hofstede Model

Received : Augus 18, 2021 Revised : September 28, 2021 Accepted : October 8, 2021
Corresponding Author : Do-Hyung Park

저 자 소 개



전 승 표

KAIST에서 경영학으로 석사학위를 취득하고, 고려대학교에서 과학관리학 전공으로 이학박사를 취득했다. 현재 한국과학기술정보연구원 글로벌R&D분석 센터에 책임연구원으로 재직 중이며, 과학기술연합대학원대학교(UST) 과학기술경영정책과 교수로 재직중이다. Technological forecasting and social change, Government Information Quarterly, Scientometrics, Energy policy, Internet research 등 해외학술지와 한국기술혁신학회지, 지능정보연구 등 국내학술지에 주저자로 다수의 논문을 게재했다. 주요 관심분야는 빅데이터를 활용한 수요 예측, 글로벌 R&D 동향 분석, 유망 기술 탐색, 기술가치평가, 중소기업 R&D 정책 등을 위한 지능형 정보 시스템 개발 연구이다.

서 봉 군

국민대학교에서 경영정보학(MIS) 석사학위를 취득하고, 동 대학원에서 공학 박사를 취득하였다. 현재 한국생산기술연구원(KITECH)에서 산업융합관련 규제 이슈 발굴, 신산업 규제개선 로드맵 수립 등의 업무를 수행하고 있다. 주요 관심 분야는 사용자/소비자 데이터 분석 및 인사이트 도출, 신산업부문의 과학기술정책 등이다.

박 도 형

KAIST 경영대학원에서 MIS 전공으로 석사/ 박사학위를 취득하였다. 현재 국민대학교 경영대학 경영정보학부/ 비즈니스 IT 전문대학원 부교수로 재직 중이며, 고객경험연구실(CXLab.)을 책임지고 있다(www.cxlab.co.kr). 한국과학기술정보연구원(KISTI)에서 유망아이템 발굴, 기술가치 평가 및 로드맵 수립, 빅데이터 분석 등을 수행하였고, LG전자에서 통계, 시선/뇌파 분석, 데이터 마이닝을 활용한 소비자 평가 모형 개발을 담당했었고, 스마트폰, 스마트TV, 스마트Car 등에 대한 Technology, Business, Market Insight 기반 컨셉 도출 프로젝트를 다수 수행하였다. 현재 주요 관심분야는 사회심리학 기반의 사용자/소비자의 행동 이론(User/Customer Behavior), 통계 및 인공지능 기법 기반의 사용자/소비자 애널리틱스(User/Customer Analytics), 디자인사고(Design Thinking) 기반의 사용자/소비자 경험 디자인(Experience Design)이다.